UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2021-003
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 13/01/2021 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN INFORMATIQUE

by

Renato Manuel LEMOS BAPTISTA
Born on 3 December 1989 in Viseu (Portugal)

HUMAN MOTION ANALYSIS USING 3D SKELETON REPRESENTATION IN THE CONTEXT OF REAL-WORLD APPLICATIONS: FROM HOME-BASED REHABILITATION TO SENSING IN THE WILD

Dissertation defence committee

Dr. Djamila Aouada, Dissertation supervisor
Assistant professor/Senior research scientist, Université du Luxembourg

Dr. Björn Ottersten, Chairman
Professor, Université du Luxembourg

Dr. François Brémond, Vice-chairman
Research director, INRIA, Sophia Antipolis, Nice, France

Dr. Adeline Paiement
Associate professor, LIS, Université de Toulon, Toulon, France

Dr. Miguel Angel Olivares Mendez
Assistant professor/Senior research scientist, Université du Luxembourg

“A question that sometimes drives me hazy — am I or are the others crazy?”

Albert Einstein

Acknowledgements

I have met extraordinary people throughout my journey in the Interdisciplinary Centre for Security, Reliability, and Trust (SnT) at the University of Luxembourg. Firstly, I would like to thank my advisors Prof. Djamila Aouada and Prof. Björn Ottersten, for their unconditional support during my studies. I would like to emphasize my profound gratitude to Prof. Djamila Aouada for her seamless guidance and trust in me. Our conversations helped me to grow as a researcher, as a person, and to believe in my strengths.

Besides my advisors, I would like to thank all members of my thesis committee: Prof. Dr. François Brémond, Prof. Dr. Adeline Paiement, and Prof. Dr. Miguel Angel Olivares Mendez for accepting to review my thesis. Especially, to Prof. Dr. François Brémond for his insightful comments and discussions.

I would like to thank my immediate collaborators and friends, Dr. Michel Antunes, Dr. Enjie Ghorbel, Dr. Girum Demisse, Dr. Kassem Al Ismaeil, and Dr. Abd El Rahman Shabaeyk, for all the great discussions, either research-wise either to grow as a person. I would also like to thank my office mates Jevgenij Krivochiza, Konstantinos Papadopoulos, and Sumit Gautam, for all the moments we have shared and lived in that office. This will take a special place in my heart and will not be forgotten. Many thanks to all the Computer Vision, Imaging and Machine Intelligence (CVI^2) research group, including Anis Kacem, Alexandre Saint, Oyebade Oyedotun, Mohamed Adel, Inder Pal, Pavel Chernakov, Albert Sanchez, Kseniya Cherenkova, and Joe Lorentz. Thanks to Himadri, Gabriel, Liz, Luis, Anestis, Ahmed for this adventure.

I would like to thank my family, especially my parents and brother, mother and sister in

law, for all the support and great moments. Finally, my wife and best friend Viviana, thank you for all the support and love. This could not have been done if you were not with me. Thank you for giving me everything. Love you.

The work presented in this thesis was funded by the European Union's Horizon 2020 research and innovation project STARR under grant agreement No.689947. Many thanks to the EU project partners on STARR for the fruitful discussions. I would also like to thank HOPALE, France, and OSAKIDETZA, Spain, for the warm welcoming and for the productive discussions and journeys we have shared.

Index

List of Abbreviations

ADAM ..... Adaptive Moment Estimation

CCD ..... Charge-Coupled Device

CNN ..... Convolutional Neural Networks

CPN ..... Cascaded Pyramid Network

DNN ..... Deep Neural Networks

DTW ..... Dynamic Time Warping

GRU ..... Gated Recurrent Unit

HMM ..... Hidden Markov Model

IoT ..... Internet of Things

LARP ..... Lie Algebra Representation of body-Parts

LSTM ..... Long-Short Term Memory

KSC ..... Kinematic Spline Curves

MAS ..... Modified Ashworth Scale

MDS ..... Multi-Dimensional Scaling

MLDs ..... Moving Light Displays

MoCap ..... Motion Capture

MPJPE ..... Mean Per Joint Position Error

MSE ..... Mean Square Error

NPMSE ..... Normalized Pose Motion Signal Energy

PCA ..... Principal Component Analysis

RGB ..... Red Green Blue

RGB-D ..... Red Green Blue - Depth

RNN ..... Recurrent Neural Network

ROI ..... Region of Interest

SCI ..... Spinal Cord Injuries

ST-GCN ..... Spatial-Temporal Graph Convolutional Networks

TCN ..... Temporal Convolutional Networks

TSD ..... Trajectory Shape Descriptor

TVR ..... Time Variable Replacement

YOLO ..... “You Only Look Once”

List of Notations

M	Matrix M
$\mathbf{v}$	Vector $\mathbf{v}$
M^T	Transpose of matrix M
$\angle(\mathbf{a}, \mathbf{b})$	Angle between vectors $\mathbf{a}$ and $\mathbf{b}$
$\|\cdot\|$	L_2 norm
$\|\cdot\|_F$	Frobenius norm
$\mathsf{SE}(3)$	Special Euclidean group
$\mathsf{SO}(3)$	3D rotation group
$\arg\min(\cdot)$	The minimizing argument
$\arg\max(\cdot)$	The maximizing argument
$f \circ g$	Composition of functions $f(\cdot)$ and $g(\cdot)$
S_t	3D skeleton at time instant t
$\mathcal{B}$	Set of body-part matrices
$\mathcal{R}$	Set of rotation matrices
$\mathcal{G}(\cdot)$	Motion representation function

$\mathfrak{D}(\cdot,\cdot)$ ..... Distance between two curve representations

$\mathcal{Y}$ ..... Set of action labels

$\mathcal{V}(\cdot)$ ..... Function that maps an RGB sequence to the corresponding action label

List of Figures

List of Tables

Abstract

Human motion analysis using 3D skeleton representations has been a very active research area in the computer vision community. The popularity of this high-level representation mainly results from the large variety of possible real-world applications such as video surveillance, video conferencing, human-computer interaction, virtual reality, healthcare, and sports. Despite the effectiveness of recent 3D skeleton-based approaches, their suitability to real-world scenarios still needs to be assessed. Using these approaches in a real-world scenario can give new insights on how to improve them for reaching real-world standards. In this thesis, we propose new solutions to mitigate existing constraints for the deployment of 3D skeleton-based approaches in various real-world scenarios. For that purpose, we investigate two human motion analysis applications that are based on 3D skeletons, namely, home-based rehabilitation of functional activities and human motion analysis in the wild.

In the first part of this thesis, we propose a low-cost solution designed for supporting home-based rehabilitation of stroke survivors under the remote supervision of a therapist. To that end, we introduce the concept of color-based feedback proposals for guiding the patients in real-time while exercising. More specifically, color-based codes are visualized for informing the patient on the accuracy of the movement and on the adequacy of the posture. Feedback proposals are tailored to each patient's body anthropometry. An initial clinical validation shows an improvement of the posture and of the quality of motion when using the proposed feedback proposals.

In the second part of this thesis, we focus on human motion analysis in the wild in the context of cross-view action recognition. We propose and investigate different 3D human

pose estimation techniques from a single RGB camera in order to take advantage of 3D skeleton-based approaches. Indeed, given their 3D nature, 3D skeletons can overcome more easily the challenge of viewpoint variability in contrast to 2D-based approaches. To show the relevance of 3D pose estimation techniques in the context of human motion analysis, two different pipelines are proposed. The first pipeline makes use of a per-frame pose estimation approach. Per-frame pose estimation shows temporal inconsistency and small fluctuations in the skeleton joint locations over time. Considering this, the second framework is then based on a sequence-to-sequence pose estimation, providing, therefore, temporally consistent skeleton sequences that are more robust to sensing in the wild. These two pipelines show an improvement in recognition accuracy as compared to state-of-the-art approaches on two different well-known datasets. However, despite their relevance, 3D human pose estimation methods present some limitations. For example, their accuracy drops significantly in the presence of unseen environments or situations, *e.g.*, challenging camera locations, and outdoor conditions. For that reason, we introduce *3DBodyTex.Pose* dataset, an original dataset to address the challenges of camera locations and outdoor scenarios in the context of 3D human pose estimation. Moreover, 3DBodyTex.Pose offers to the research community new possibilities for the generalization of 3D human pose estimation from monocular in-the-wild images from arbitrary camera viewpoints.

Chapter 1

Introduction

Over the last two decades, analyzing human motion through visual information in an automatic manner has been widely investigated by the computer vision community. This has been mainly motivated by the significant number of possible application fields such as human-computer interaction [1], [2], surveillance [3], [4], healthcare [5], [6], sign language [7], [8], sports [9], [10].

Human motion analysis generally focuses on detecting, tracking, recognizing and understanding human movements from visual sequences. For analyzing human movements, a first step of feature extraction is usually required. In the literature, numerous approaches for encoding the spatio-temporal information of human motion have been introduced. These features should be sufficiently discriminative and relevant to the human motion in order to effectively make adequate decisions. The majority of surveys propose to classify human motion analysis methods according to the used representation [11]–[13]. For instance, in [13], Weinland *et al*. suggested to distinguish three classes of representations, namely, *image-based models*, *spatial statistics*, and *body models*. Image-based representations of actions do not require the detection and labeling of individual body parts. Generally, the features are computed densely on a regular grid inside the detected *Region of Interest* (ROI). On the other hand, spatial statistics, also termed as *spatial bags of features* representation, describe the sequence by using histograms of feature occurrence; thus without the need of modelling any geometrical structure between the feature locations. Finally, body models

Figure 1.1: Illustration of Johansson's experiments [14] to study the human body motion pattern of walking. The images extracted from [15].

are based on the detection of biologically meaningful joints; therefore, describing the human body as a kinematic tree connecting different body-parts. In computer vision, body models are also referred to as *poses* or *skeletons* that can be described in 2D or 3D [11].

The relevance of skeletons as a representation for human motion analysis has been initially demonstrated in the 1970s [14]. Johansson proposed to attach few *Moving Light Displays* (MLDs) to the human body and showed that humans are able to recognize human actions by only visualizing the MLDs. Figure 1.1 gives an example of Johansson's experiments, where the MDLs were placed and distributed on the human body joints against a uniform background. Furthermore, he also conducted experiments to analyze if humans are able to recognize actions directly from 2D patterns and if they intuitively reconstruct the 3D human body. In summary, Johansson stated that, *"The geometric structures of body motion patterns in man (...) are determined by the construction of their skeletons."*. Given the discriminative power of skeletons, this pioneer work in psychology has importantly motivated the use of skeletons.

Compared to other representations, 2D/3D poses allow a better discrimination of human motion, and are compact and easy to manipulate in real-time applications [16], [17]. Considering the 3D nature, 3D pose representations offer some advantages compared to 2D pose representations. 3D pose representations can easily handle viewpoint variability, background, and clothing variation [18], [19]. However, despite the multiple advantages of 3D pose representations, in the early 2000s they were moderately considered given the ne-

cessity of placing visual markers on the human body, limiting, therefore, their applicability in many realistic scenarios. As an example, one can cite the case of video surveillance, where it would not be possible to place markers on people in order to analyze their movements. With the recent advances in 3D imaging, *e.g.*, the release of RGB-D sensors, it became possible to estimate the 3D human pose in real-time without placing markers or landmarks. Thanks to that, a renewed interest for pose-based human motion analysis has been noted [16], [20], [21].

In this thesis, we investigate the use of 3D skeletons in real-world applications, and propose new approaches for human motion analysis under challenging constraints. The constraints can relate to a setup, to users, or to sensing conditions. We focus on two different applications of 3D skeletons in human motion analysis. First, we focus on introducing a low-cost healthcare solution for automatically analyzing and monitoring the movements and the posture of patients. Second, we propose ways to mitigate the constraints for deploying 3D pose estimation techniques from a monocular RGB camera to more real-world scenarios. In the following sections, we detail the motivation and the scope of this thesis, the different challenges addressed as well as the contributions of this work.

1.1 Motivation and Scope

As mentioned in the previous section, 3D skeleton-based representations have been widely investigated in the context of human motion analysis over the last years [18]. This is principally due to their potential range of applications. However, less works have attempted to concretely adapt these approaches to real-world applications such as home-based rehabilitation [6], [22], and abnormal behavior detection [23], [24]. While earlier approaches for human motion analysis have been demonstrated to be very effective and accurate [21], [25]–[27], their suitability to a real-world context still needs to be assessed. Furthermore, using these approaches in a real-world scenario can give new insights on how to improve them for reaching real-world standards. In this thesis, we focus on proposing ways to relax constraints for deployment of such approaches in more real-world scenarios. We present

two applications that use the 3D skeleton representation of the human body.

In the first part, we focus on introducing skeleton-based motion analysis methods in the context of home-based rehabilitation, specifically targeting stroke survivors. A large portion of the population of stroke survivors is affected by motor impairments, directly impacting their daily life activities. Exercising is crucial to recover and maintain some autonomy in functional activities [28]. The standard practice is to get an initial *on-site* rehabilitation with the physical presence of a healthcare professional, followed by home-based rehabilitation [29]. Then, patients are expected to regularly exercise in order to maintain and recover autonomy in daily life activities [30]. However, research has shown that stroke survivors do not exercise regularly for many reasons, such as fatigue, lack of motivation, confidence, and lack of exercising guidance [31]. Moreover, exercising alone at home may lead to musculoskeletal injuries if the exercises are not correctly performed. Hence, there is a need to monitor and guide patients in order to avoid such risky situations. For that reason, guidance feedback proposals need to be presented to the patient with the objective of improving the movement being performed. However, every patient has a different anthropometry and has a different way of moving. Consequently, guidance feedback proposals need to be adapted and tailored to that specific patient. We propose to explore and investigate solutions that are hand-free, real-time, and robust to camera placement variation. For that reason, we use the Microsoft Kinect sensor [32] to extract, in real-time, the skeleton information and perform home-based rehabilitation.

In the second part of this thesis, we focus on investigating and proposing means to alleviate the constraints for deploying 3D pose estimation techniques to more real-world scenarios in the context of human motion analysis. With the tremendous advances of deep learning, relatively accurate 3D pose estimation methods from a single RGB camera have emerged [33]–[38]. The effectiveness of these novel methods have been mainly tested using a *Mean Square Error* (MSE) criterion but very few works have started studying their suitability in a human motion analysis scenario. Therefore, there is a need for investigating solutions to relax constraints for the deployment of 3D human pose estimation to a larger variety of real-world scenarios. With that in mind, we propose to exploit 3D human pose estimation

techniques from a single RGB camera in the context of *cross-view action recognition*. Cross-view action recognition aims to correctly recognize actions observed from unseen views in the training phase. However, not all camera viewpoints are considered. Consequently, 3D skeletons are not well estimated for the cases where the subject is not fully visible or self-occluded, for these cases we denote them as *extreme viewpoints*, *e.g.*, camera is placed on top of the person [39]. Considering this, we propose a dataset to enhance robustness of 3D human pose estimation in the wild.

1.2 Challenges in Assisting Physical Activity Using 3D Skeletons

Human tracking and gesture therapy systems are used for monitoring and supporting the rehabilitation of stroke survivors at home [31], [40]–[44]. These home-based rehabilitation systems are advantageous not only because they are less costly as compared to *on-site* rehabilitation, but also because having such solutions at home and regularly available, the patients tend to exercise more often [31], [44]. Existing systems and research either:

1. combine exercises with video games as a mean to train people while keeping a high level of motivation [45], [46]; or
2. try to emulate a physical therapy session [31], [41].

Considering the first point, a common goal has been to provide visual feedback through a game-based platform. In more detail, the patients are guided to touch specific targets or lean the body towards a direction depending on the game development[6], [22]. In summary, the main challenge is to combine the two previous points into one system and adapt the guidance feedback proposals to the patient's body anthropometry for the purpose of home-based rehabilitation. In the following subsections, we present the challenges of home-based rehabilitation that are addressed in the first part of the thesis.

1.2.1 Computer-based Solutions for Home-based Rehabilitation of Stroke Survivors

Recent works tackle the problem of assessing how well people perform specific actions [9], [24], [31], [47], which can be used in rehabilitation, *e.g.*, to evaluate mobility and measure the risk of relapse. These works mainly focus on evaluating the performance of specific actions or motions that are being performed. However, not many approaches target providing feedback on how people can improve the motion being performed. Recently, [9], [31] presented methods to provide feedback proposals to improve the motion. Nevertheless, these feedback proposals are not optimal and involve a complex set of instructions for suggesting a particular body-part motion. Thus, there is a need to investigate solutions to ease how the feedback proposals are presented to the patients, especially if they exercise alone at home.

Home-based rehabilitation of stroke survivors is generally conducted in parallel with traditional *on-site* rehabilitation to ensure the effectiveness of the program, requiring the physical presence of the therapist. Moreover, the therapist is not able to control the home-based exercising because it is not possible to: 1) know if the patient completed the exercises; 2) guide and advise the patient while exercising; 3) detect misconducted exercises that can lead to musculoskeletal injuries, and 4) evaluate the exercise quality.

Recently, many rehabilitation-based systems have been designed [6], [22], [48]–[53]. However, they are mainly focusing on assisting the therapists in dedicated centers. They are hardly usable remotely (without the physical presence of the therapist) since they do not provide automatic real-time guidance feedback to support the patient while exercising, and to monitor postural defects. Consequently, there is a need to designing appropriate solutions that enable the home-based rehabilitation, emphasizing features that emulates the physical presence of the therapists.

1.2.2 Abnormal Motion Detection

In general, the quality of a given motion is estimated by measuring the deviation from what is normal. Abnormalities are usually detected by comparing a given motion to a normal

motion model [23], [24], [54]. In order to detect abnormalities, two main approaches exist: 1) assuming the abnormalities to be known *a priori*; and 2) not considering any prior information on what an abnormality is. Recent works used the 3D skeleton representation to detect abnormalities from a model of normal motions [23], [24]. However, there are two main challenges considering the 3D skeleton representation. First, the dimensionality of a 3D skeleton sequence is relatively high, *i.e.*, it directly depends on the number of joints and the number of frames of the skeleton sequence. Second, the speed or latency variation between the skeleton sequences. It is very unlikely that two motions are performed with the same speed, even if they are performed by the same subject.

1.3 Challenges in Cross-view Action Recognition

Understanding human actions is very challenging due to the variation in human motion, appearance, environmental settings, and camera view. In action recognition systems, tolerance to data variation resulting from different camera viewpoints has emerged as one of the main challenges in human action recognition in the wild [19], [26], [55]–[57], commonly referred to as *cross-view action recognition*. Recently, some methods have been introduced to tackle the case of cross-view action recognition directly from RGB images [55], [56], [58]–[60]. Such methods make use of 2D features as input, which are not view-invariant characteristics by definition and do not incorporate the radial motion information of the human body shape. To overcome these limitations, some methods have proposed to estimate 3D skeletons from RGB images [57], [61], [62]. It has been shown that adopting this strategy can enhance the accuracy of the recognition of human actions. Nevertheless, these works are based on per-frame pose estimation methods, leading to a temporal inconsistency in the 3D estimated skeleton sequence, negatively impacting action recognition accuracy.

1.4 Challenges in 3D Human Pose Estimation In The Wild

Thanks to the recent advances in *Deep Neural Networks* (DNN), the task of 2D human pose estimation has seen a significant improvement in results [63]–[65]. This has been mostly achieved thanks to the availability of large-scale datasets containing 2D annotations of humans in various conditions, *e.g.*, in the wild [66]. Similarly to 2D, 3D human pose estimation has shown a big progress. However, these advances are mainly considering controlled lab environments where the 3D information of the human is acquired using markers. Recently, many works focused on the challenging problem of 3D human pose estimation in the wild [34]–[36], [67]. These works differ significantly from each other but share an important aspect. They are usually evaluated on the same dataset used for training; thus, likely to have been over-optimized for specific datasets, leading to a lack of generalization. It becomes difficult to judge on the generalization, and more precisely for in-the-wild scenarios where variations coming from the background and camera viewpoints are always present. It is therefore necessary to investigate solutions to mitigate the challenging task of 3D human pose estimation in the wild.

1.5 Objectives and Contributions

In this thesis, we focus on assessing the suitability of 3D skeleton-based approaches to real-world scenarios. We investigate and propose solutions to alleviate the constraints for the deployment of these approaches in two real-world scenarios. Specifically, in a home-based rehabilitation scenario, and human motion analysis in the context of sensing in the wild. The main contributions are presented in the following subsections.

1.5.1 Visual and Human-Interpretable Feedback for Assisting Physical Activity

In order to provide feedback on how to improve the movement being performed, we propose to compute feedback proposals for body-parts, which are defined as configurations of

skeleton joints. In addition, the feedback instructions are presented in a friendly and easily understandable manner and do not require any pose constraints of joints configuration. Moreover, we study how to measure postural defects using 3D skeletons, and how to use these measurements for guiding the user in converging to a healthier and better posture. Two particular scenarios were analyzed. The first consists in examining the body posture while sitting on a chair, which is one of the main causes of health related issues in work environments [68], [69]. The second is related to sports and incorrect exercising in gyms, *e.g.*, weight lifting [70], [71], which causes many significant injuries. The experiments show that the provided feedback proposals are effective in guiding the users to improve the movement being performed as well as correcting the posture, converging to a healthy one.
This work has been published in [72], [73], [74], and [75].

1.5.2 Home Self-Training: Home-based Rehabilitation for Stroke Survivors & Clinical Evaluation

In the context of home-based rehabilitation, we adapt and merge our previous work [72]–[75] into one low-cost solution to support the rehabilitation of stroke survivors. In particular, we develop a system composed of two main applications: the therapist-side and the patient-side, strengthening the remote communication between the two sides. In the therapist-side application, the personalized exercise prescription is the key element. The therapist can tailor and update the exercises according to the condition of the patient. As for the patient-side application, the visual feedback proposals and the calibration phase are fundamental roles. Visual feedback proposals are relevant to guide and support the movement and monitor postural defects in a real-time manner. Since each patient has a different range of movement, the feedback proposals need to be adapted. Thus, feedback proposals are presented by considering the patient's body anthropometry. Hence, we propose to provide the visual feedback proposals by comparing them to the opposite limb movement, instead of using a predefined template. Furthermore, we perform a first evaluation with 10 healthy subject followed by an initial clinical evaluation on 10 stroke survivors. The conducted evaluation shows an improvement of the posture and the quality of the motion due to the provided

feedback proposals. According to therapists and patients, the application can be considered as reliable, simple to use, and positively impacting the psychology of the patients.
This work is the extended version of [73]–[75] and has been published in [76], and [77].

1.5.3 Abnormal Motion Detection using 3D Skeletons

In the context of abnormality detection, we propose to represent a 3D skeleton sequence as a 1-dimensional curve in order to reduce the dimensionality of the data by using a problem-specific knowledge. In this case, the gait pattern of subjects. Furthermore, we propose to represent such curve as a set of rigid-transformation matrices, resulting in a curve in the *Special Euclidean group* $\mathrm{SE}(3)$. Moreover, we use a deformation-based curve alignment function to analyze the time-variation in the motion; and therefore, proposing a motion-quality distance to emphasize and quantify variation in time. Experimental results show the superiority of the proposed approach as compared to existing works.
This work has been published in [78].

1.5.4 View-Invariant Action Recognition From RGB Data via 3D Pose Estimation

Among the most successful RGB-based approaches, knowledge transfer from 3D data has been commonly used to address the challenging problem of cross-view action recognition. In this thesis, we propose to address this challenge by relying on the recent effective *Convolutional Neural Network* (CNN)-based methods for estimating the 3D skeleton from a single RGB image [33], [79]. The proposed framework consists of two stages. First, we estimate the 3D skeletons from RGB images using two different methodologies, per-frame and sequence-to-sequence 3D skeleton estimation. Second, we propose to use an *Long-Short Term Memory* (LSTM)–based and a *Temporal Convolutional Network* (TCN)-based temporal models to effectively estimate the temporal dependency between skeletal pose estimates and classify them into action classes. Experimental results show that our approaches show an improvement in recognition accuracy in two different well-known datasets when com-

pared to existing methods.
This work has been published in [61] and [80].

1.5.5 Towards Generalization of 3D Human Pose Estimation In The Wild

3D human pose estimation approaches are usually evaluated on the same dataset used for training. Thus, these approaches have likely been over-optimized for specific datasets, leading to a lack of generalization to other domains. Considering this, we propose *3DBodyTex.Pose* dataset. It is an original dataset generated from high-resolution textured 3D body scans, similar in quality to the ones contained in the 3DBodyTex dataset introduced in [81]. 3DBodyTex.Pose is dedicated to the task of human pose estimation. Synthetic scenes are generated with ground-truth information from real 3D body scans, with a large variation in subjects, clothing, and poses. Realistic background is incorporated to the 3D environment using a cube mapping approach. 2D images are generated from different camera viewpoints, including challenging ones, by virtually changing the camera location and orientation. Experimental results show that retraining a state-of-the-art approach with 3DBodyTex.Pose significantly improves the performance of 3D human pose estimation in the wild.
This work has been published in [82], [83].

1.6 Publications

JOURNALS

1. **Baptista, R.**, Ghorbel, E., Shabayek, A.E.R., Moissenet, F., Aouada, D., Douchet, A., André, M., Pager, J., Bouilland, S.. "Home Self-Training: Visual feedback for Assisting Physical Activity for Stroke Survivors". Computer Methods and Programs in Biomedicine. 2019.

2. Ghorbel, E., **Baptista, R.**, Shabayek, A.E.R., Aouada, D., Oramaeche, M.G., Lago, J.O., Fernandez, L.O.. "Home-based Rehabilitation System for Stroke Survivors: A Clinical Evaluation". Journal of Medical Systems. 2020.

CONFERENCES

1. Antunes, M., **Baptista, R.**, Demisse, G., Aouada, D., Ottersten, B.. "Visual and Human-Interpretable Feedback for Assisting Physical Activity". In European Conference on Computer Vision (ECCV) Workshop. Amsterdam, The Netherlands, 2016.

2. **Baptista, R.**, Antunes, M., Aouada, D., Ottersten, B.. "Video-Based Feedback for Assisting Physical Activity". In 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Porto, Portugal, 2017.

3. **Baptista, R.**, Antunes, M., Shabayek, A.E.R., Aouada, D., Ottersten, B.. "Flexible Feedback System for Posture Monitoring and Correction". In IEEE International Conference on Image Information Processing (ICIIP). Shimla, India, 2017. ⋆ **Best paper award in computer-vision track.** ⋆

4. **Baptista, R.**, Ghorbel, E., Shabayek, A.E.R., Aouada, D., Ottersten, B.. "Key-Skeleton Based Feedback Tool for Assisting Physical Activity". In Zooming Innovation in Consumer Electronics International Conference (ZINC), Novi Sad, Serbia, 2018.

5. **Baptista, R.**, Demisse, G., Aouada, D., Ottersten, B.. "Deformation-Based Abnormal Motion Detection using 3D Skeletons". In IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA). Xi'An, China, 2018.

6. **Baptista, R.**, Ghorbel, E., Papadopoulos, K., Demisse, G., Aouada, D., Ottersten, B.. "View-Invariant Action Recognition From RGB Data via 3D Pose Estimation". In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.

7. Adel Musallam, M., **Baptista, R.**, Al Ismaeil, K., Aouada, D.. "Temporal 3D Human Pose Estimation for Action Recognition from Arbitrary and Challenging Viewpoints". In 6th Annual Conference on Computational Science & Computational Intelligence (CSCI), Las Vegas, USA, 2019.

8. **Baptista, R.**, Saint, A., Al Ismaeil, K., Aouada, D.. “Towards Generalization of 3D Human Pose Estimation In The Wild”. In IEEE International Conference Pattern Recognition (ICPR) Workshop. Milan, Italy, 2020.

PUBLICATIONS NOT INCLUDED IN THE THESIS

1. **Baptista, R.**, Antunes, M., Aouada, D., Ottersten, B.. “Anticipating Suspicious Actions using a Small Dataset of Action Templates”. In 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Madeira, Portugal, 2018.

2. Ghorbel, E., Papadopoulos, K., **Baptista, R.**, Pathak, H., Demisse, G., Aouada, D., Ottersten, B.. “A View-invariant Framework for Fast Skeleton-based Action Recognition Using a Single RGB Camera”. In 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Prague, Czech Republic, 2019.

3. Papadopoulos, K., Ghorbel, E., **Baptista, R.**, Aouada, D., Ottersten, B.. “Two-stage RGB-based Action Detection using Augmented 3D Poses”. In 18th International Conference on Computer Analysis of Images and Patterns (CAIP), Salerno, Italy, 2019.

1.7 Thesis Outline

This dissertation is divided into two parts and it is organized as follows:

I. Human Motion Analysis for Home-based Rehabilitation

- **Chapter 2**: This chapter presents the background regarding 3D skeleton representation and pre-processing. In addition, a brief introduction on 3D human pose estimation from RGB images is presented.

- **Chapter 3**: In this chapter, our contribution in providing feedback proposals to improve a movement being performed is introduced. Thus, visual feedback proposals are presented in a human-interpretable manner without specifying pose constraints of body joints configurations.
- **Chapter 4:** Exercising with an incorrect posture can lead to musculoskeletal injuries in a long-term. Hence, monitoring the postural defects in real-time is our next contribution and it is presented in this chapter.
- **Chapter 5:** In this chapter, we present a full system to support the home-based rehabilitation of stroke survivors. Furthermore, an initial clinical validation is conducted.
- **Chapter 6:** In this chapter, a curve-based representation of 3D skeletons to address the abnormality detection is introduced. Moreover, this representation is used to detect key-frames and quantify variation in time.

II. Human Motion Analysis In the Wild

- **Chapter 7:** In this chapter, our contribution to cross-view action recognition from RGB images is presented. A per-frame 3D human pose estimation technique is used to obtain the 3D skeleton. Furthermore, an LSTM-based approach is used to effectively model the temporal dependency between skeletal poses.
- **Chapter 8:** Similarly to the previous chapter, we use a different 3D human pose estimation technique, a sequence-to-sequence estimation. Moreover, a TCN-based model is used to learn the temporal dependencies of skeleton sequences and classify them into action labels.
- **Chapter 9:** In this chapter, 3DBodyTex.Pose dataset is introduced to address the challenge of 3D human pose estimation in-the-wild from arbitrary camera viewpoints, including challenging ones.
- **Chapter 10:** This thesis is concluded by summarizing the contributions and presenting various interesting future directions.

Chapter 2

Background

2.1 Introduction

Skeleton representation of the human body has been widely investigated in the research community since the 1970s and the work of Johansson [14]. Johansson started this study by attaching MLDs to the human body joints in a dark environment, refer to Figure 1.1. The objective of this study was to show that humans can recognize human actions by only observing 2D patterns. State-of-the-art approaches based on CNN architectures for 2D human pose estimation have shown great performance [65], [84], [85]. However, there are few significant challenges considering the 2D representation of the human body such as occlusions, and viewpoint variability. To handle these limitations, 3D skeletons have been recently considered as an alternative in the context of human motion analysis. In fact, thanks to the introduction of RGB-D sensors and the advances in deep learning, it became possible to estimate 3D poses without using markers attached to the human body. In this chapter, we start by describing the 3D human skeleton representation. Then, we present two different 3D skeleton estimation methodologies that are investigated and explored throughout this thesis.

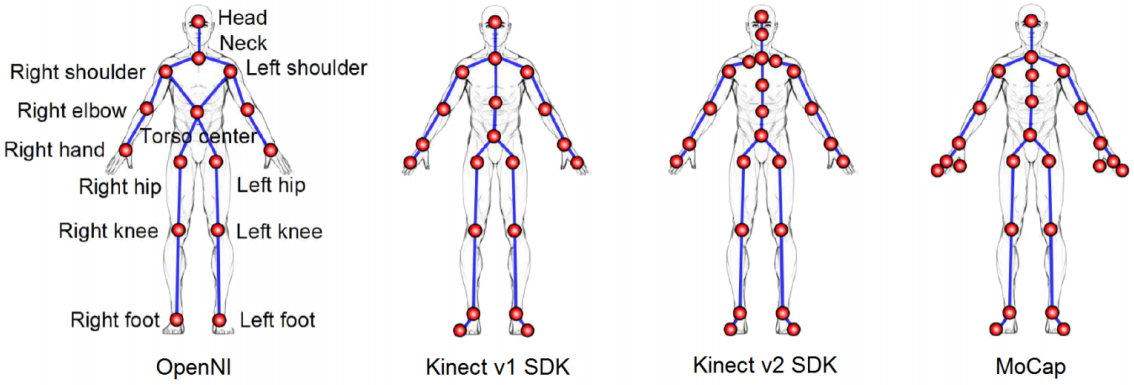


Figure 2.1: Examples of the skeleton structure of the human body obtained from different devices. Figure extracted from [18].

2.2 3D Skeleton Representation

In computer vision, a 3D skeleton refers to a set of human body joints that are semantically connected by segments following the human body structure. Figure 2.1 shows the skeleton structured from different devices. 3D skeleton-based representations are able to model the relationship of human joints and encode the whole body configuration. Considering this, a 3D skeleton S_t formed by N joints at a time instant t can be defined as,

$$\mathsf{S}_t = [\mathbf{j}_1, \cdots, \mathbf{j}_N], \tag{2.1}$$

where each $\mathbf{j}_i \in \mathbb{R}^3$ represents the 3D position of the joint i. Therefore, a sequence of 3D skeletons is denoted as $\mathsf{M} = \{\mathsf{S}_1, \cdots, \mathsf{S}_T\}$, where T is the number of skeleton frames. The core of the first part of this thesis is the analysis of the motion required for aligning body-parts with respect to a template skeleton pose $\hat{\mathsf{S}}_t$. Consequently, a template sequence is defined as $\hat{\mathsf{M}} = \{\hat{\mathsf{S}}_1, \cdots, \hat{\mathsf{S}}_T\}$. In order to compare two different 3D skeleton sequences, the sequences must be spatially and temporally aligned [21]. Thus, in the next subsections, we describe how the 3D skeletons are normalized, and how two 3D skeleton sequences are temporally aligned.

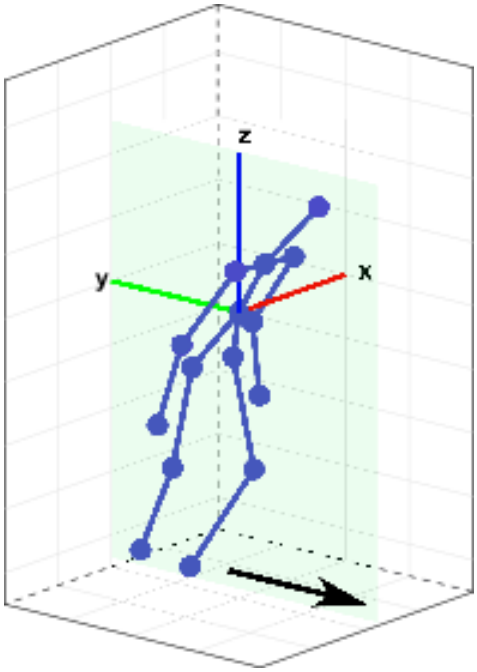


Figure 2.2: Skeleton normalization. The word coordinate system is placed at the hip center. The plane colored in green represents the y-z plane where the main direction of motion variation is happening. The arrow shows the main direction of the movement.

2.2.1 Skeleton Normalization

In order to compare two skeletons, the first requirement is that they need to be spatially registered. First, we assume that the joint corresponding to the hip joint to be the center of the coordinate system of the skeleton. Then, the skeleton is rotated such that the projection of the vector from the left hip to the right hip onto the x-y plane is parallel to the x-axis. Furthermore, each skeleton is normalized such that the body-parts lengths match the corresponding lengths of the reference skeleton. A skeleton is chosen to be a reference based on the application[1]. With such normalization, most of the motion variation is present in the y-z plane. Consequently, the direction of the x-axis is perpendicular to the main movement direction. Figure 2.2 shows an example of a normalized skeleton with the corresponding coordinate system and also the plane defining the main direction of the motion.

[1]For example, in the case of the home-based rehabilitation system presented in Chapter 5, the reference skeleton is a skeleton captured from the patient.

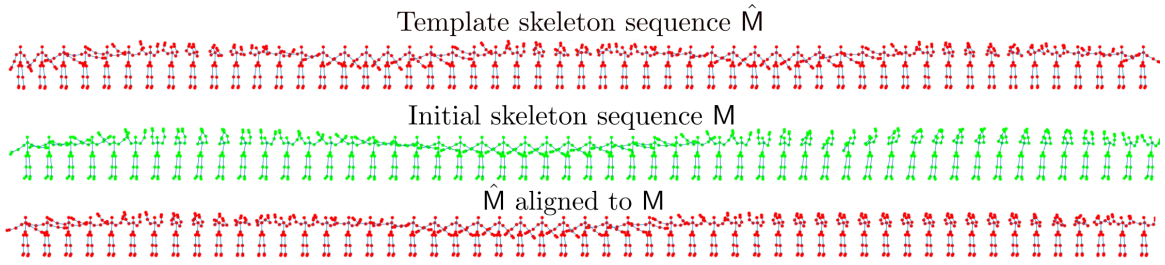


Figure 2.3: Temporal alignment of 3D skeleton sequences using DTW. The first row shows the template action $\hat{\mathsf{M}}$ (red), the second row shows the skeleton sequence M, and the third row shows $\hat{\mathsf{M}}$ aligned with respect to M via DTW.

2.2.2 Temporal Alignment

Different subjects, or the same subject at different times, perform a particular action or movement at different rates. In order to handle rate variations and mitigate the temporal misalignment of time series, *Dynamic Time Warping* (DTW) is usually employed [86], [87]. DTW is a widely known technique to find the optimal alignment between two temporal sequences which may vary in speed. A warping path $\phi = [\phi_1, \cdots, \phi_L]$ with length L, defines an alignment between the two sequences. The warping path instance $\phi_i = (m_i, \hat{m}_i)$ assigns the skeleton S_{m_i} of M to the skeleton $\hat{\mathsf{S}}_{\hat{m}_i}$ of $\hat{\mathsf{M}}$. The total cost C of the warping path ϕ between sequences M and $\hat{\mathsf{M}}$ is defined as

$$C_\phi(\mathsf{M}, \hat{\mathsf{M}}) = \sum_{i=1}^{L} d_c(\mathsf{S}_{m_i}, \hat{\mathsf{S}}_{\hat{m}_i}), \tag{2.2}$$

where d_c is a local cost measure. In this case, the local cost measure d_c is defined as the L_2 norm of the skeleton instances S_{m_i} and $\hat{\mathsf{S}}_{\hat{m}_i}$. Thus,

$$d_c(\mathsf{S}_{m_i}, \hat{\mathsf{S}}_{\hat{m}_i}) = \|\mathbf{s}_{m_i} - \hat{\mathbf{s}}_{\hat{m}_i}\|^2, \tag{2.3}$$

where $\mathbf{s}_{m_i}$ and $\hat{\mathbf{s}}_{\hat{m}_i}$ are the vector representation of the skeletons S_{m_i} and $\hat{\mathsf{S}}_{\hat{m}_i}$, respectively. Following equation (2.2), the alignment via DTW between the sequences M and $\hat{\mathsf{M}}$ is repre-

sented by DTW$(\mathsf{M}, \hat{\mathsf{M}})$ and is defined as

$$\mathsf{DTW}(\mathsf{M}, \hat{\mathsf{M}}) = \min\{C_\phi(\mathsf{M}, \hat{\mathsf{M}})\}. \tag{2.4}$$

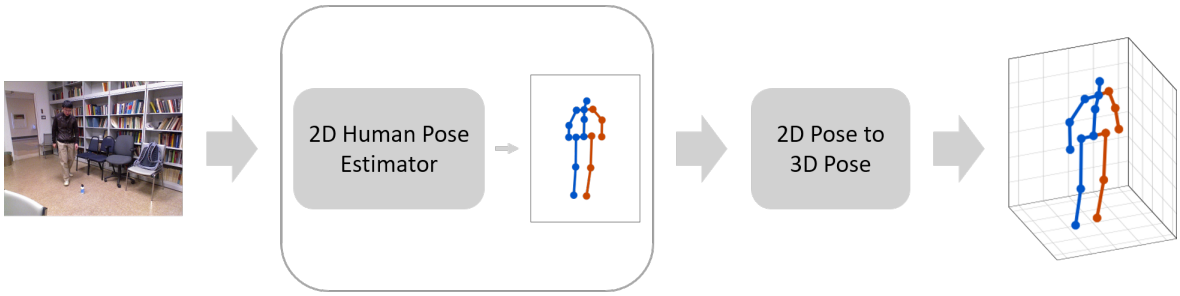
Figure 2.3 shows a temporal alignment example between two skeleton sequences.

2.3 3D Human Pose Estimation from RGB Cameras

3D human pose estimation aims to predict the locations of the human body joints in 3D space from RGB images. RGB-D sensors have made possible the real-time extraction of 3D human information using only depth maps [16]. However, these devices present some limitations when considering outdoor applications. They are very sensitive to lighting variation and their range is limited, restricting their applicability to real-world scenarios such as in video surveillance. Thus, there is a need to investigate new solutions to overcome these limitations.

Estimating 3D human pose from a single RGB image has been considered for a long time to be an ill-posed problem due to the loss of 3D information in RGB images. First attempts have been made to extract the 3D pose information from RGB images based on hand-crafted features and regression models [88]–[90]. Recently, and thanks to the advances in deep learning techniques, 3D human pose estimation has become more accessible [33]–[38], [91]–[93]. Generally, 3D human pose estimation approaches can be divided into two categories: 1) direct mapping of an RGB image to 3D skeleton, we refer to this as *RGB to 3D skeleton*; and 2) intermediate 2D pose followed by a depth estimation method to lift 2D pose to 3D skeleton, we refer to this as *2D to 3D skeleton*. Figure 2.4 illustrates the human pose estimation approaches of the two different categories. In the following subsections we describe these two categories and formalize the mapping to obtain the 3D skeleton from the respective input information.

(a) RGB to 3D skeleton.



(b) 2D to 3D skeleton.

Figure 2.4: Illustration of the two different categories of 3D human pose estimation approaches. Figure 2.4a illustrates the mapping from RGB directly to 3D skeleton. Figure 2.4b depicts the intermediate step of 2D pose estimation followed by the lifting to 3D skeleton.

2.3.1 RGB to 3D Skeleton

Given well labeled 3D data, 3D human pose estimation from RGB images can be formulated as a standard supervised learning problem. One of the pioneer works, Li and Chan [94] proposed to train a neural network to directly regress joint locations. Li *et al*. [95] presented a network structure that consists of a CNN for the image feature extraction followed by two sub-networks for transforming the image features and pose into a joint embedding. Tekin *et al*. [96] proposed to adopt an auto-encoder at the end of the networks. In contrast to regressing the coordinate of the joints, Pavlakos *et al*. [93] presented a coarse-to-fine learning approach, where a voxel representation for each joint is considered as the regression target. Pavlakos *et al*. [97] further trained the network with addition ordinal depth information of human joints as constraints. Iskakov *et al*. [37] presented a multi-view solution. Their work is based on learnable triangulation methods that combine 3D information from multiple views.

These works usually focus on a per-frame human pose estimation methodology, where

the goal is to learn a function $f(\cdot)$ that maps an RGB image I to a 3D skeleton estimate $\tilde{\mathsf{S}} \in \mathbb{R}^{3N}$ with N number of human body joints. Given an RGB image I, the corresponding 3D skeleton estimate is given by mapping I to $\tilde{\mathsf{S}}$ by using $f(\cdot)$, such that

$$\mathsf{I} \overset{f}{\mapsto} \tilde{\mathsf{S}} = [\tilde{\mathbf{j}}_1, \cdots, \tilde{\mathbf{j}}_N]. \tag{2.5}$$

2.3.2 2D to 3D Skeleton

Given 2D skeleton joint locations, inferring them to 3D skeleton joints is a very challenging task. Many works initially proposed to address this challenge by using pose constraints such as bone length [98], joint limits [99], etc. Relying on the recent advances in deep learning, a reasonable number of works have been emerging, proposing to learn the mapping between 2D joints and 3D joints. Some of the methods adopt off-the-shelf 2D human pose estimation modules to initially estimate the 2D skeleton, then extend the 2D skeleton to 3D skeleton. Martinez *et al*. [92] proposed a simple fully connected residual network to directly regress 3D coordinates from 2D coordinates. Moreno-Noguer [100] learned a pairwise distance matrix, which is invariant to image rotation, translation, and reflections, from 2D to 3D space. Tekin *et al*. [101] introduced a two-branch framework to predict 2D heatmaps and extract features from images, which are then fused to obtain the 3D skeleton. Mehta *et al*. [33] presented a fully convolutional pose formulation that regresses 2D and 3D joint positions jointly in real-time, denoted as *VNect*. Zhou *et al*. [34] proposed to couple together in-the-wild images with 2D annotations with indoor images with 3D annotations. In addition, a depth regression module is presented to predict 3D skeletons from 2D heatmaps with a proposed geometric constraint loss for 2D data. Yang *et al*. [67] adopted a generator presented in [34] and designed a multi-source discriminator with image, pairwise geometric structure, and joint location information. Pavllo *et al*. [36] presented a temporal convolutional network to predict 3D joints from 2D joint sequences. These methods have the advantage over RGB to 3D skeleton methods since they can easily utilize images from 2D human datasets. Mainly due to the availability of 2D annotated large-scale in-the-wild datasets [66], [102].

Given an RGB image I, the goal is to map the information related to the human present

in the image to the corresponding 2D locations of the human joints. Usually, the 2D poses are obtained either by using an off-the-shelf approach, or either using a CNN to regress the 2D joints together with a depth regression module. In general, we can obtain the 2D skeleton $\tilde{\mathsf{S}}_{2D} \in \mathbb{R}^{2N}$ with N joints by applying the mapping function $g(\cdot)$ such that $\tilde{\mathsf{S}}_{2D} = g(\mathsf{I})$. However, the interest is to lift the 2D skeleton to a 3D skeleton. To that end, a function $h(\cdot)$ is learned within the network to lift the 2D skeleton to 3D skeleton. Thus, given an image I, the 3D skeleton with N joints can be obtained by the composition of the function $h(\cdot)$ and $g(\cdot)$, $\tilde{\mathsf{S}} = (h \circ g)(\mathsf{I}) = h(g(\mathsf{I}))$. In other words,

$$\mathsf{I} \overset{g}{\mapsto} \tilde{\mathsf{S}}_{2D} \overset{h}{\mapsto} \tilde{\mathsf{S}} = [\tilde{\mathbf{j}}_1, \cdots, \tilde{\mathbf{j}}_N]. \tag{2.6}$$

In this category of 3D human pose estimation approaches, the function $h(\cdot)$ can be learned together with $g(\cdot)$ or separately where the input information for the network are 2D skeleton estimates.

Part I

Human Motion Analysis for Home-based Rehabilitation

Chapter 3

Visual and Human-Interpretable Feedback for Assisting Physical Activity

Physical activity is essential for stroke survivors for recovering some autonomy in daily life activities. Stroke survivors are initially subject to physical therapy under the supervision of a health professional, but due to economical aspects, home-based rehabilitation is eventually suggested. In order to support the physical activity of stroke survivors at home, this chapter presents a solution for guiding the user in how to properly perform certain actions and movements. This is achieved by presenting feedback in form of visual information and human-interpretable messages. The core of the proposed approach is the analysis of the motion required for aligning body-parts with respect to a template skeleton pose, and how this information can be presented to the user in the form of simple recommendations. Experimental results in three datasets show the potential of the proposed framework.

3.1 Introduction

Physical activity is vital for the general population for maintaining a healthy lifestyle. It is crucial for elderly people in the prevention of diseases, maintenance of independence and improvement of quality of life [103]. For stroke survivors it is critical and essential for recovering some autonomy in daily life activities [28]. Despite the benefits of physical activity, many stroke survivors do not exercise regularly due to many reasons, such as lack of motivation, confidence, and skill levels [31]. Traditionally, the stroke survivors are initially subject to physical therapy under the supervision of a health professional aimed at restoring and maintaining activities of daily living in rehabilitation centres [104]. The physiotherapist explains the movement to be performed to the patient, and continuously advises the patient how to improve the motion as well as interrupts the exercise in case of health related risk issues. Unfortunately, and due to the high economic burden [29], the *on-site* rehabilitation is usually of a short period of time and prescribed treatments and activities for home-based rehabilitation are usually suggested [30]. Sadly, stroke survivors, and more frequently older adults, do not appropriately adhere to the recommended treatments, because, among other factors, they do not always understand or remember well enough what and how they are supposed to do the physical treatment.

In order to support the rehabilitation of stroke survivors at home, human tracking and gesture therapy systems are being investigated for monitoring and assistance purposes [31], [40]–[44]. These home rehabilitation systems are advantageous not only because they are less costly for the patients and for the health care systems, but also because having it at home and regularly available, the users tend to do more exercise. A well accepted sensing technology for these purposes are RGB-D sensors (*e.g.*, Microsoft Kinect [32]) that are affordable and versatile, allowing to capture in real-time color and depth information [31], [44], [105]. Few works tackle the problem of assessing how well the people perform certain actions [9], [24], [31], [47], which can be used in rehabilitation, *e.g.*, to evaluate mobility and measure the risk of relapse. Pirsiavash *et al*. [9] propose a framework for assessing the quality of actions in videos. Spatio-temporal pose features are extracted and a regression

model is estimated that predicts scores of actions from annotated data. Tao *et al*. [24] also describe an approach for quality assessment of the human motion. The idea is to learn a manifold from normal motion, and then evaluate the deviation from it using specific measures. Wang *et al*. [47] tackle the problem of automated quantitative evaluation of musculoskeletal disorders using a 3D sensor. They introduce the *Representative Skeletal Action Unit* framework from which clinical measurements can be extracted. More recently, Ofli *et al*. [31] presented an interactive coaching system using the Kinect. The coaching system guides users through a set of exercises, and the quality of execution of these exercises is assessed based on manually defined pose measurements, such as keeping hands close to each other or maintaining the torso in an upright position.

In this chapter, we want to go one step further and not only evaluate, but also provide feedback in how people can improve the movement or action being performed. There are two main works that tackle this problem. In the computer vision community, the work of Pirsiavash *et al*. [9] is the most relevant. After assessing the quality of actions using supervised regression, feedback proposals are obtained by differentiating the scoring with respect to the joint locations, and then selecting the joint and the direction it should move to achieve the largest improvement in the score. In the medical community, Ofli *et al*. [31] provide assistive feedback during the performance of exercises. For each particular movement, they define constraints such as keeping hands close to each other or maintaining the torso in a upright position. These constraints are constantly measured during the exercise for assessing if the movement is performed correctly and in case predefined values for metrics on these constraints are violated, then corrective feedback is provided. While in [9] the corrective feedback is analysed per joint, which involves a complex set of instructions for suggesting a particular body-part motion (*e.g*., arm moving up), in [31] the motion constraints are action specific and manually defined.

3.2 Problem Definition

This section discusses the problem that we aim to solve. Given an action or movement as being a 3D skeleton sequence $\mathsf{M} = \{\mathsf{S}_1, \ldots, \mathsf{S}_t, \ldots, \mathsf{S}_T\}$ with T frames, the objective is to solve the following problem: given a template skeleton sequence $\hat{\mathsf{M}}$ and a subject performing a movement M, we want to provide, at each time instant, feedback proposals such that the movement can be iteratively improved to better match $\hat{\mathsf{M}}$. In our particular case, the goal is to align a given sequence M to a template sequence $\hat{\mathsf{M}}$. There are two possibilities, we either align M with respect to $\hat{\mathsf{M}}$, or vice-versa, $\hat{\mathsf{M}}$ with respect to M. Since we want to analyze each temporal instant of a given sequence M, it is reasonable to compute the temporal correspondences of $\hat{\mathsf{M}}$ with respect to M. As a first step, pre-processing on the input skeleton data is required. Existent approaches were previously introduced in the literature (*e.g.*, [21]), and are adapted to our specific problem, refer to Chapter 2.

3.3 Human-interpretable Feedback Proposals

After the spatial and temporal alignment processing described in Chapter 2, the skeleton instance $\hat{\mathsf{S}}_t$ in $\hat{\mathsf{M}}$ will be in correspondence with S_t in M. This section explains how to compute the body motion required to align corresponding body-parts of aligned skeletons $\hat{\mathsf{S}}$ and S, and proposes a method for extracting human-interpretable feedback from these transformations.

3.3.1 Body-part-based Representation

In line with recent research [106]–[109], we analyse the human motion using a body-part based representation. A skeleton S can be represented by a set of body-parts $\mathcal{B} = \{\mathsf{B}^1, \ldots, \mathsf{B}^k, \ldots, \mathsf{B}^K\}$. Each body-part B^k is composed by n^k joints $\mathsf{B}^k = \{\mathbf{j}_1^k, \ldots, \mathbf{j}_{n^k}^k\}$ and has a local reference system defined by the joint $\mathbf{j}_r^k$. Figure 3.1 shows the different body-parts defined for the dataset Weight&Balance.

Given the aligned skeletons $\hat{\mathsf{S}}$ and S, refer to Chapter 2, the objective is to compute the motion that each body-part of S needs to undergo to better match the template skeleton $\hat{\mathsf{S}}$.

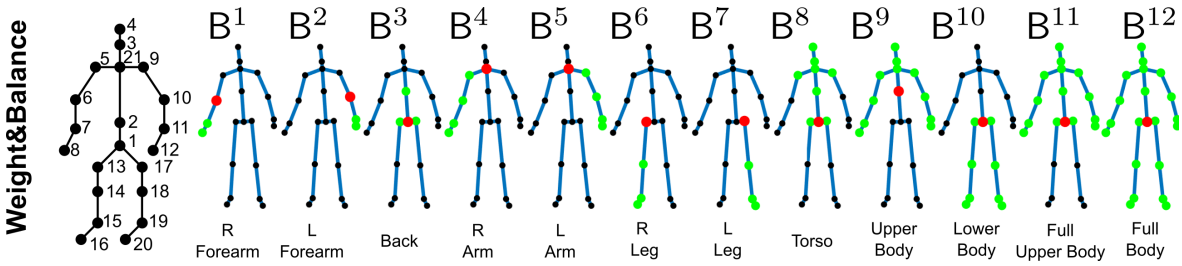


Figure 3.1: Proposed body-part representation. The skeleton of the dataset Weight&Balance is composed by 21 joints (left). 12 body-parts were defined. For each body-part, the composing joints are highlighted in green. The red joint corresponds to the local origin $\mathbf{j}_r^k$ of a each body-part (R=right, L=left).

This analysis is performed for each body-part using the corresponding local coordinate system. As a metric for measuring how similar is the pose of corresponding body-parts, we use the Euclidean distance as the scoring function. Following this, the error between B^k and $\hat{\mathsf{B}}^k$ is given by:

$$m^k = \sum_{j=1}^{n^k} ||\mathbf{j}_j^k - \hat{\mathbf{j}}_j^k||^2. \tag{3.1}$$

Remark that $||\mathbf{j}_r^k - \hat{\mathbf{j}}_r^k|| = 0$, because the previous computation is performed using the local coordinate systems that are assumed to be in correspondence.

3.3.2 Feedback Proposals

For providing feedback to the performer of skeleton S on how the movement can be improved to better match $\hat{\mathsf{S}}$, we compute the transformation that each body-part B^k needs to undergo for decreasing the scoring function m^k. We anchor the reference joints $\mathbf{j}_r^k$ and $\hat{\mathbf{j}}_r^k$ (refer to Figure 3.1) of the corresponding body-parts. The aim is then to compute the rotation $\mathsf{R}^k \in \mathrm{SO}(3)$ that minimizes the following error:

$$e^k(\mathsf{R}^k) = \sum_{j=1}^{n^k} ||\mathsf{R}^k \mathbf{j}_j^k - \hat{\mathbf{j}}_j^k||^2, \tag{3.2}$$

which can be computed in closed form. It is important to refer that since the human motion is articulated, depending on the movement being performed, a given body-part B^k may

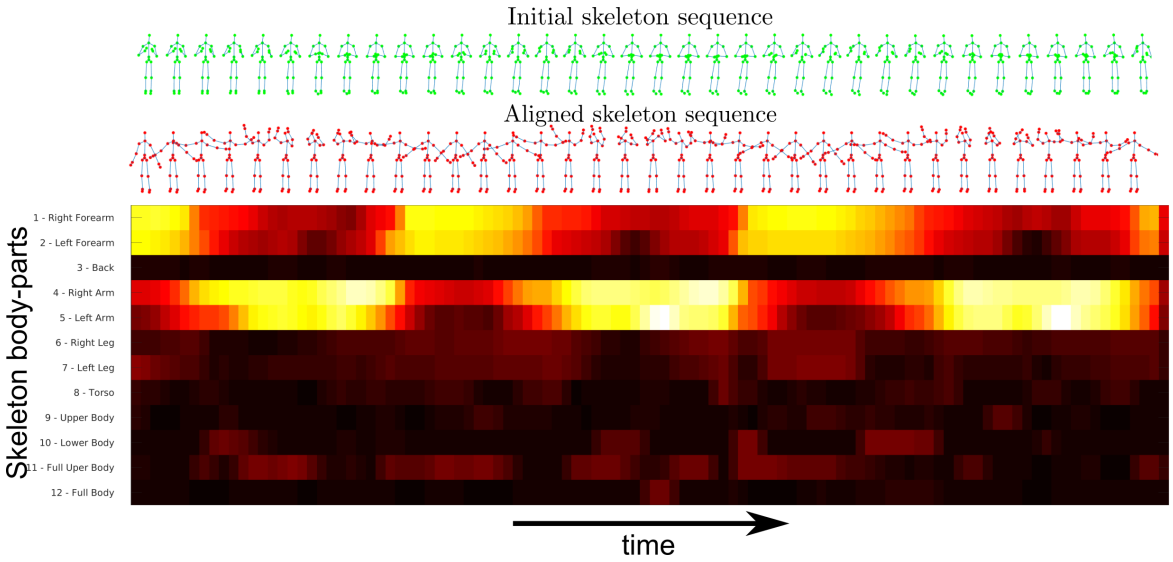


Figure 3.2: intensity of correcting motion feedback required for each body-part. (Top) Sequence M performing clapping, (middle) target sequence $\hat{\mathsf{M}}$ corresponding to the action waving using two hands after spatial and temporal alignment, and (bottom) the cost c_i^k (refer to Method 1) calculated for each temporal instant independently (the vertical axis corresponds to different body-parts, while the horizontal axis is the temporal dimension.

or may not move rigidly. This is not a critical issue because body-parts that do not moving rigidly have high joint matching error and will be considered not relevant by the method described next. Note that different body-parts B^k can contain subsets of the same joints, which implies that the transformation R^k will also have impact on the location of the other body-parts $\mathsf{B}^{l \neq k}$. Taking this into account, we want to compute a sequence of transformations $\mathcal{R} = \{\mathsf{R}_1, \ldots, \mathsf{R}_i, \ldots, \mathsf{R}_K\}$, one rotation $\mathsf{R}_i = \mathsf{R}^k$ for each body-part B^k, such that the first rotation R_1 has the highest decrease in the joint location error until R_K, which has the lowest impact in the human pose matching. This sorting is performed maximizing the following cost

$$c_i^k = m^k - e^k(\mathsf{R}^k), \tag{3.3}$$

where in iteration i, the body-parts B^k selected in the previous $i-1$ iterations are not taken into account. The pseudo-code of the overall scheme is shown in Method 1. Figure 3.2 show an example of the intensity of correcting motion pattern c_i^k for actions clapping and waving

Method 1: Computation of the sequence of body-part transformations that minimizes the skeleton matching error.

Input: S, $\hat{\mathsf{S}}$, B
Output: Sequence of rotations $\mathcal{R}$, list of body-part indexes $\mathcal{K}$
$\mathcal{L} := \mathcal{B}$, $\mathcal{K} = \{\}$, $\mathcal{R} = \{\}$, $i = 1$;
while $\mathcal{L} \neq \{\}$ **do**
  foreach $\mathsf{B}^k \in \mathcal{L}$ **do**
    compute R^k that minimizes $e^k(\mathsf{R}^k)$ (refer to Equation 3.2);
    $c_i^k = m^k - e^k(\mathsf{R}^k)$ (refer to Equation 3.3);
  $l := \underset{k}{\operatorname{argmax}} \left(c_i^k\right)$;
  $\mathsf{R}_i = \mathsf{R}^l$;
  $\mathcal{K} := \mathcal{K} \cup l$, $\mathcal{R} := \mathcal{R} \cup \mathsf{R}_i$;
  $\mathcal{L} := \mathcal{L} \setminus \mathsf{B}^l$;
  $i = i + 1$;

across time.

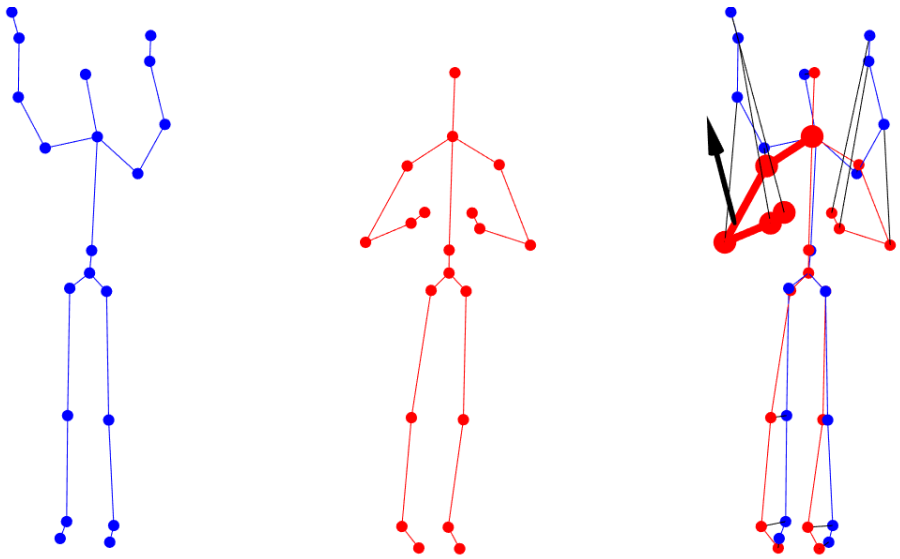
The rotations $\mathsf{R}_i = \mathsf{R}^k$ correspond to the motion required for the best alignment of B^k and $\hat{\mathsf{B}}^k$. However, it is difficult to present this rigid-body transformation as feedback proposals on, for example, a screen. For overcoming this, we compute feedback vectors for suggesting improvements on the motion. For each body-part, we pre-calculate the spatial centroid $\mathbf{c}^k$ (note that in case of single limbs, this point is located on the body-part itself). Then, the feedback vector anchored to $\mathbf{c}^k$ is defined as

$$\mathbf{f}^k = \mathsf{R}^k \mathbf{c}^k - \mathbf{c}^k. \tag{3.4}$$

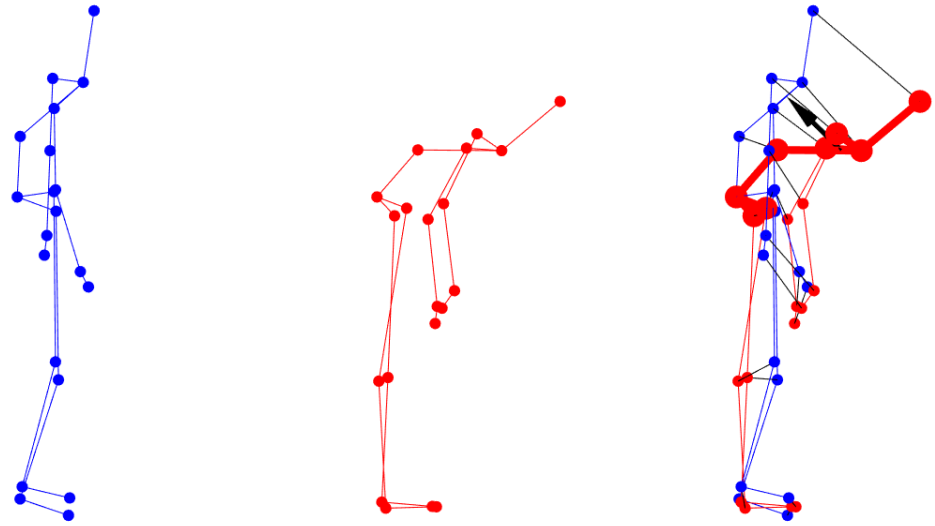
Figure 3.3 shows feedback vectors for two different pairs of actions being performed.

3.3.3 Feedback Messages

At this point, we have discussed how to compute the optimal rotation R^k for each body-part B^k, and how this transformation can be presented to a user in form of a feedback vector $\mathbf{f}^k$ anchored to the body-part centroid $\mathbf{c}^k$. Nevertheless, not all the persons have the same spatial awareness to realize how to perform the motion suggested by the feedback



(a) $\hat{\mathsf{S}} := \{waving\}$, $\mathsf{S} := \{clapping\}$



(b) $\hat{\mathsf{S}} := \{standing\}$, $\mathsf{S} := \{bending\}$

Figure 3.3: Two examples for feedback proposals. The target pose $\hat{\mathsf{S}}$ is shown in blue and the action being performed is shown in red. For each example, the third column shows superimposed the two skeletons, the matching joints (black lines) and the feedback vectors $\mathbf{f}_k$ (black arrows). Only the feedback proposal for R^1 is shown.

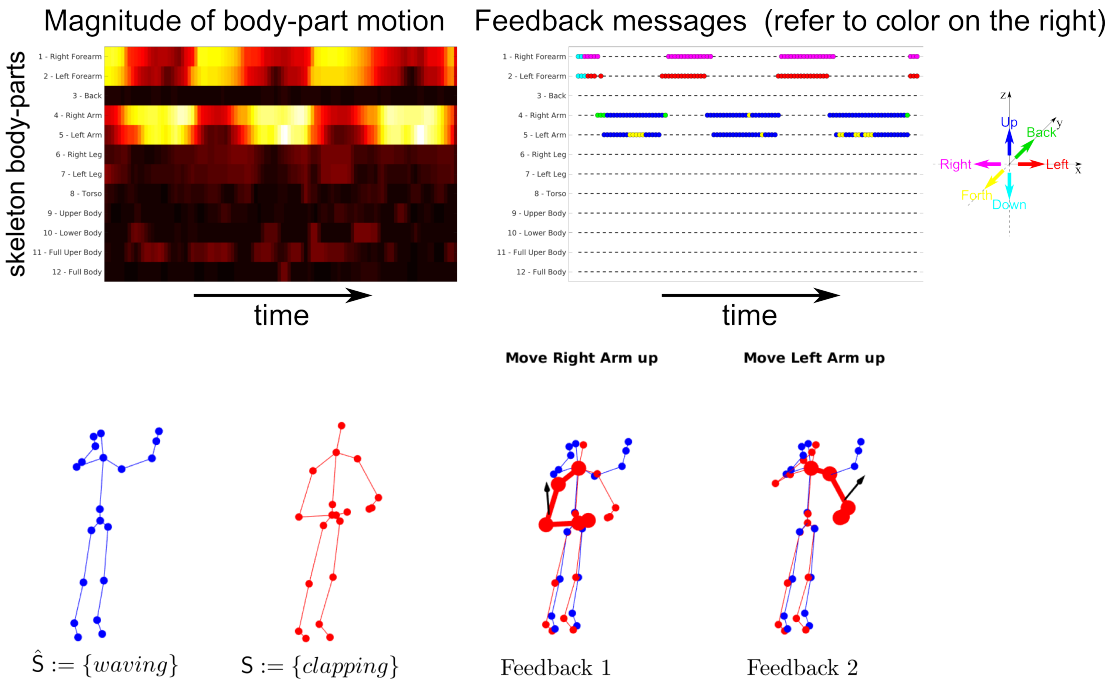


Figure 3.4: Feedback message proposals. The target action is waving using two hands and the movement being performed corresponds to clapping. (Top, left) The intensity c_i^k for each body-part B^k; (top, right) the feedback message proposals for the body-parts corresponding to R^1 and R^2. Each point corresponds to a particular message at a given time instant using the body-part name identified on the left and the color coding on the right, e.g., a blue point on the fourth dotted line corresponds to the message *Move Right Arm Up*. (Bottom) a particular instance of the template skeleton $\hat{S}$ (blue), an instance of the skeleton S (red), the feedback vectors for the body-parts corresponding to R^1 and R^2 (black arrows), and the corresponding feedback messages at the top.

vector $\mathbf{f}^k$ (refer to Figure 3.3). This difficulty is even more evident in cognitive impaired individuals [110]. In order to support the patient in improving their movements, we introduce in this section a system for presenting simple human-interpretable feedback messages that can be shown or/and spoken to the patient by the computer system.

Let us analyse the case of the body-part B^k that needs to undergo the largest motion $R_1 = R^k$. Initially, to each B^k was assigned a body-part name *BN*, *e.g.*, B^1 is the *Right Forearm* and B^8 is the *Torso* (refer to Figure 3.1). These labels are used directly for informing the user which body-parts should be moved. Then, the feedback vector $\mathbf{f}^k = [f_x^k, f_y^k, f_z^k]^T$

is discretized by selecting the dimension d with the highest magnitude $|f_d^k|$. The messages regarding the direction of the motion *BD* are then defined as:

- if $d = x$
  - if $f_x^k < 0$, then *BD* = *Right*
  - if $f_x^k > 0$, then *BD* = *Left*
- if $d = y$
  - if $f_y^k < 0$, then *BD* = *Forth*
  - if $f_y^k > 0$, then *BD* = *Back*
- if $d = z$
  - if $f_z^k < 0$, then *BD* = *Down*
  - if $f_z^k > 0$, then *BD* = *Up*

The feedback proposal messages are represented as the concatenation of strings:

$$\text{Feedback message} := \text{"Move"} + \text{BN} + \text{BD}. \tag{3.5}$$

Refer to Figure 3.4 for an example of feedback messages, where a color coding is used for identifying the directions *BD*.

3.4 Experiments

In this section, we experimentally evaluate the proposed system using three different sets of data. The first is called ModifyAction, and we use pairs of actions instances from the datasets *UTKinect* [111] and *MSR-Action3D* [112]. The objective is: given a person performing a particular action M, provide feedback proposals such that the person is able to perform a different action $\hat{\text{M}}$. The skeleton and body-parts used for this dataset are shown in Figure 3.5.

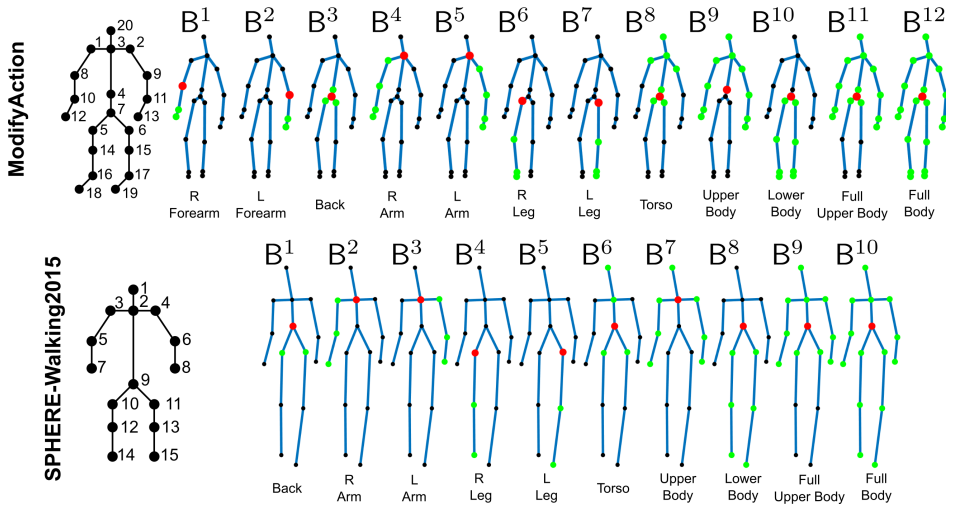


Figure 3.5: Proposed body-part representations. Each row shows the skeleton (black) and body-part configurations used for two different datasets. For each body-part, the composing joints were highlighted in green. The red joint corresponds to the local origin $\mathbf{b}_r$ of a each body-part.

The second dataset is SPHERE-Walking2015 that was introduced in [24]. The skeleton and body-parts used for this dataset are shown in Figure 3.5. It contains people walking on a flat surface, and it includes instances of normal walking and subjects simulating the walking of stroke survivors under the guidance of a physiotherapist. The objective in this regard is to analyse the difference in the walking pattern of normal subjects when compared to people with stroke.

Finally, the third dataset is new and is called Weight&Balance. This data was captured using the Kinect version 2. Refer to Figure 3.1 for a detailed description of the body-parts used. The idea is to simulate a person who suffered a stroke (refer to Figure 3.6): the *bad arm* issue due to the paralysis of an upper limb is simulated by lifting a kettle-bell using one of the arms, and the *balance* problem is replicated using a balance ball.

3.4.1 Experiments in ModifyAction

In this experiment, we studied the feedback messages to be show and the corresponding motion intensity over time considering 4 pairs of actions: 1) waving *vs.* clapping; 2) stand-

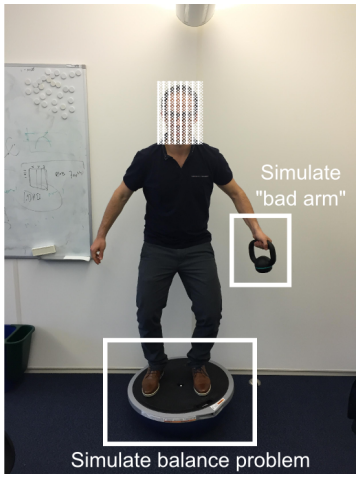


Figure 3.6: Weight&Balance dataset. We simulate the motion behaviour of a person who suffered a stroke: the *bad arm* issue due to the paralysis of an upper limb is simulated by lifting a kettle-bell using one of the arms, and the *balance* problem is replicated using a balance ball.

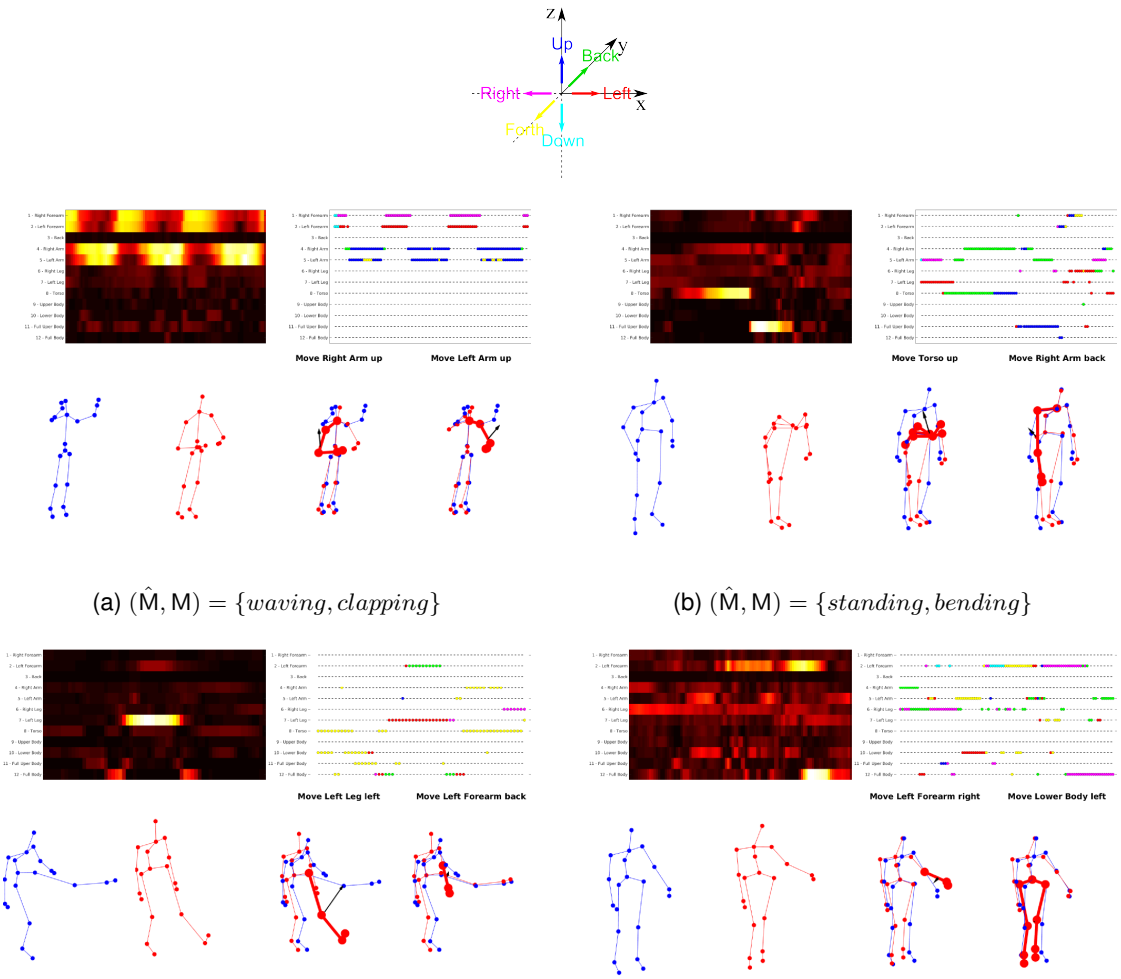
ing *vs.* bending; 3) side kick *vs.* forward kick; and 4) draw circle *vs.* draw X. Figure 3.7 shows experimental results of the proposed coaching system for the ModifyAction dataset.

3.4.2 Experiments in SPHERE-Walking2015

In the experiment of Figure 3.8, we compared the walking pattern of all the subjects with respect to the walking of healthy people (template action). It shows the intensity of correcting motion profile defined as the sum c_i^k across time for each subject. It is evident that stroke patients have a balance problem, because the body-part corresponding to the torso has high skeleton matching error, while also the stronger paralysis of one of the lower limbs can be identified. Figure 3.9 shows feedback proposals for normal people and stroke patients.

3.4.3 Experiments in Weight&Balance

The objective in this section is to simulate a simple physiotherapy session at home, and test if the feedback proposals are able to guide the user. We assume that a person needs to perform a template human pose $\hat{\mathrm{S}}$. The subject puts himself above the balance ball and lifts the kettle-bell. Giving only the guidance of the feedback vectors, body-part motion intensity



(a) $(\hat{\mathsf{M}}, \mathsf{M}) = \{waving, clapping\}$

(b) $(\hat{\mathsf{M}}, \mathsf{M}) = \{standing, bending\}$

(c) $(\hat{\mathsf{M}}, \mathsf{M}) = \{side\,kick, forward\,kick\}$

(d) $(\hat{\mathsf{M}}, \mathsf{M}) = \{draw\,circle, draw\,X\}$

Figure 3.7: Four experimental results for the ModifyAction dataset are shown. For each example, we show the intensity of correcting motion for each body-part (top,left); the feedback messages corresponding to R^1 and R^2 (top, right), refer to the color coding at the top; and the feedback vectors and messages for a particular temporal instant (bottom).

and feedback messages, the objective is to converge to the template pose without actually seeing it. The exercise lasts for 20 seconds and feedback proposals are shown at each time instant. The experimental results are shown in Figure 3.10 and Figure 3.11.

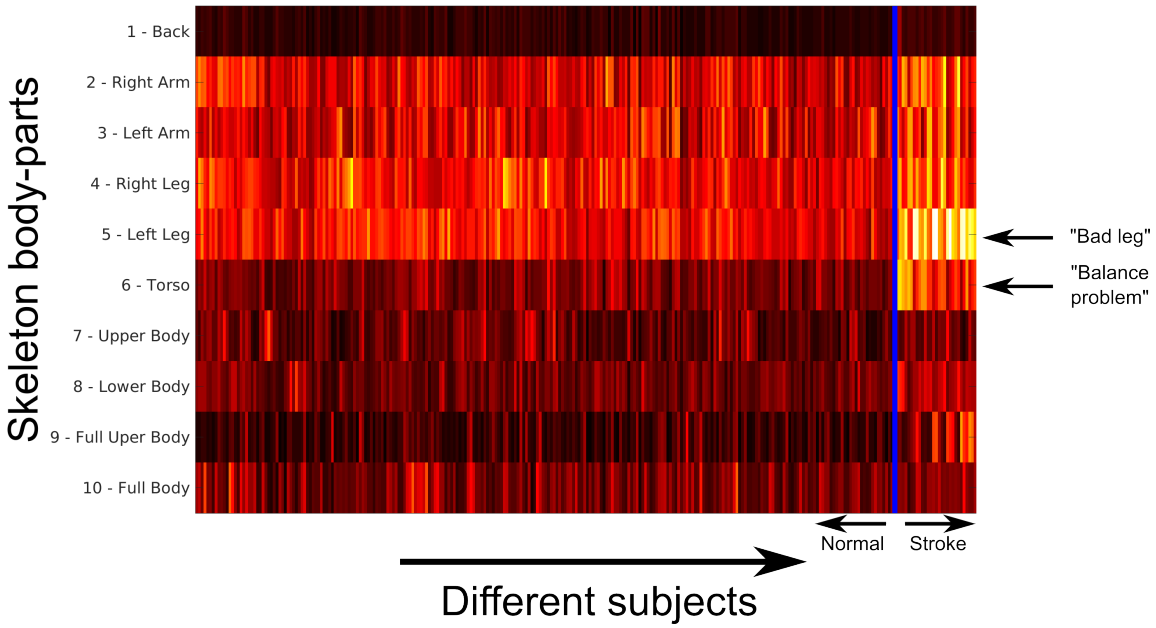
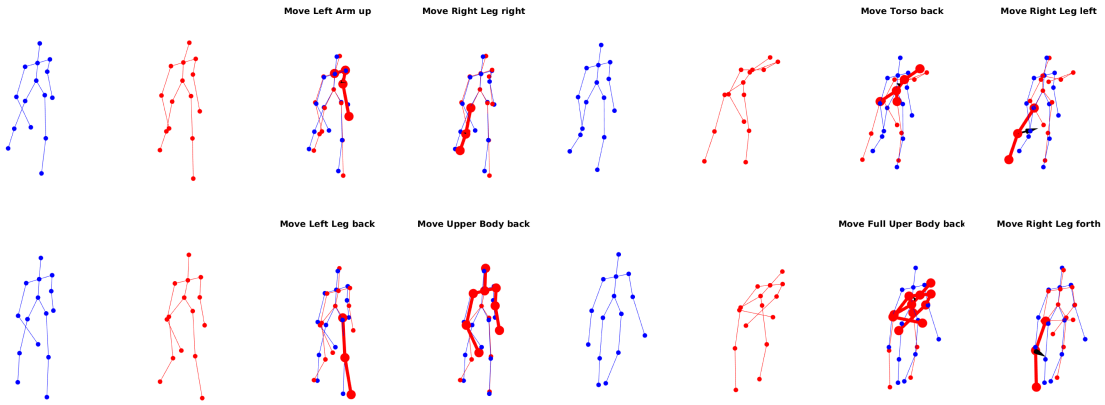
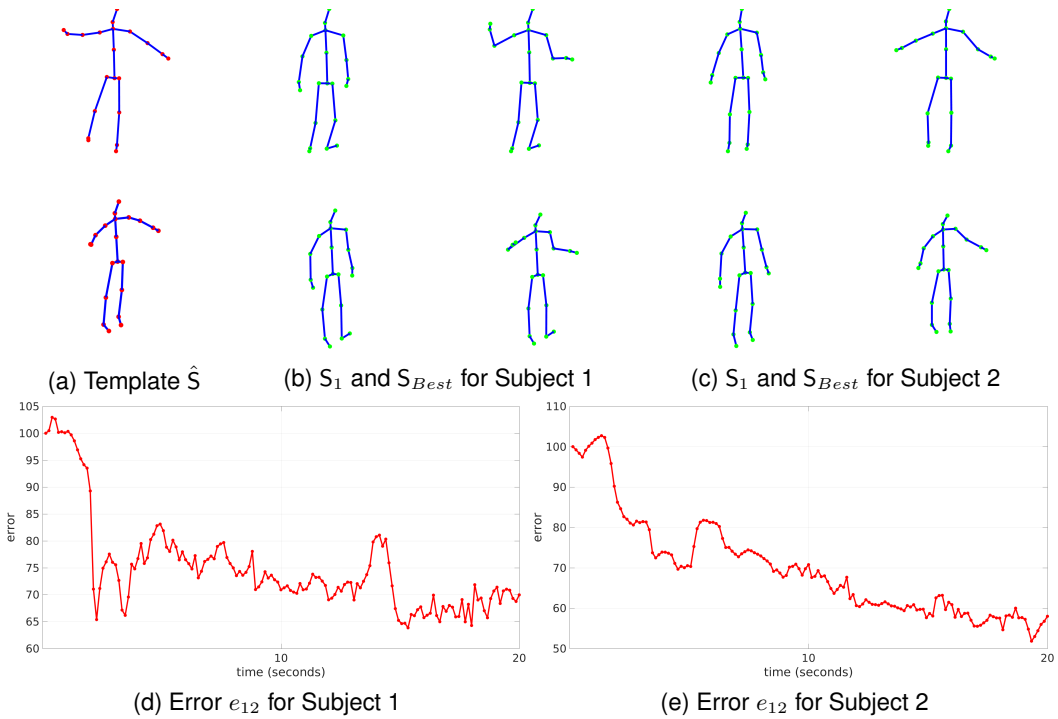


Figure 3.8: Intensity of correcting motion of different body-parts for different subjects. The subjects on the left of the blue line are healthy people, while the subjects on the right are the (simulated) stroke survivors.



(a) Normal (b) Stroke

Figure 3.9: Feedback proposals. The two subjects on the left are normal people, while the two subjects on the right are stroke survivors



(a) Template $\hat{S}$ (b) S_1 and S_{Best} for Subject 1 (c) S_1 and S_{Best} for Subject 2

(d) Error e_{12} for Subject 1 (e) Error e_{12} for Subject 2

Figure 3.10: Example 1 of Weight&Balance. (Top) two views of the template pose $\hat{S}$, and first pose S_1 and best pose S_{Best} for two subjects are shown. The best pose S_{Best} is the one that minimizes the error m^{12}. (Bottom) the relative error (difference between initial and current error divided by the initial error) in $\%$ for B^{12} is shown.

3.5 Conclusions

In this chapter, we have introduced a system for guiding a user in correctly performing an action or movement by presenting feedback proposals in form of visual information and human-interpretable feedback. Experiments show that the provided feedbacks are effective in guiding users towards given human poses. In the next chapters, we present methods to incorporate physiotherapy practices in the computation of feedback proposals, and validate the proposed framework in real-world conditions, focusing on home-based rehabilitation of stroke survivors.

(a) Template $\hat{S}$ (b) S_1 and S_{Best} for Subject 1 (c) S_1 and S_{Best} for Subject 2

(d) Error e_{12} for Subject 1 (e) Error e_{12} for Subject 2

Figure 3.11: Example 2 of Weight&Balance. (Top) two views of the template pose $\hat{S}$, and first pose S_1 and best pose S_{Best} for two subjects are shown. The best pose S_{Best} is the one that minimizes the error m^{12}. (Bottom) the relative error (difference between initial and current error divided by the initial error) in $\%$ for B^{12} is shown.

Chapter 4

Flexible Feedback System for Posture Monitoring and Correction

In the previous chapter, we presented a solution to provide understandable feedback messages to a user performing a movement or action. Therefore, in this chapter, we propose a framework for guiding users in how to correct their posture in real-time without requiring a physical or a direct intervention of a therapist or a sports specialist. In order to support posture monitoring and correction, this chapter presents a flexible system that continuously evaluates postural defects of the user. In case deviations from a correct posture are identified, then feedback information is provided in order to guide the user to converge to an appropriate and stable body condition. Experimental results in two scenarios (sitting and weight lifting) show the potential of the proposed framework.

4.1 Introduction

Posture assessment is important in health [113]–[115], sports [116] and in many work related tasks [117], [118]. Maintaining a correct posture throughout the day avoids injuries [119], and improves not only the physical condition but also self-esteem [120]. Posture analysis is usually performed by specialized therapists in health care centers [121] or specific

sports facilities, which usually involves high costs either for the patient and/or the insurance companies. Additionally, the analysis is performed at a moderate number of appointments throughout the year, and only the measurements of these appointments can be used for assessing the posture across time.

In order to support posture analysis, human tracking systems using RGB-D sensors (*e.g.*, Kinect) are being investigated and deployed for health-care and sports [122]–[124]. They can support the therapists for performing accurate physical measurements, and allow continuous visualization of posture metrics while performing specific exercises. In this chapter, we want to go one step further, and not only evaluate posture metrics (*e.g.*, [122]–[124]), but also provide real-time feedback to users in how to correct their posture automatically without requiring direct intervention of a therapist.

The proposed approach is inspired by the feedback proposed methodology presented in the previous chapter. In this chapter, we adapt the feedback messaged for specifically performing posture monitoring and correction. We do that by identifying the main body features of a correct posture (straight back and symmetric limbs), and providing real-time feedback for assisting users in correcting and maintaining a correct posture.

4.2 Proposed Approach

4.2.1 Definition of Correct Posture

As discussed in [119], posture is defined as the relative body joint dispositions at a given time, where every joint has an effect on the other joints. A correct posture is defined as a position in which minimum stress is applied to each joint. There are many features that define a correct posture, refer to [119] for a thorough analysis. We tackle two of them due to the fact that they can be analyzed using an affordable depth sensor (*e.g.*, Kinect) and also because they are simple to explain to the user. The first is related to having a straight back that is aligned with the gravity vector. The second is the balance between left and right limbs. The objective is that both legs and both arms should exercise the same force. This can be observed if the joints of arms and legs are symmetric with respect to a plane that

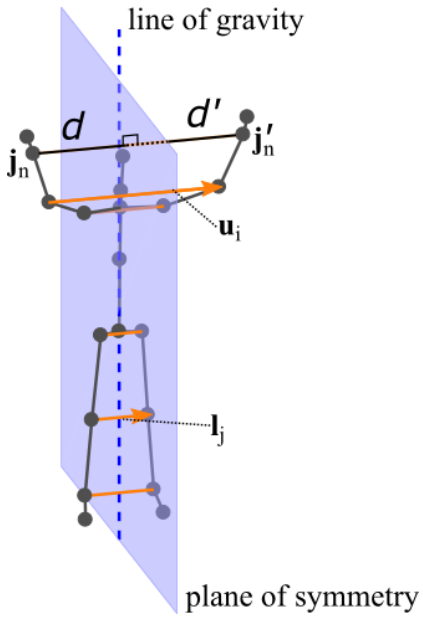


Figure 4.1: For a correct posture, the body joints of the limbs should be symmetric about the plane of symmetry that intersects the line of gravity [119]. The purple plane represents the plane of symmetry, which divides the skeleton into two parts. One part contains the left limbs (arm and leg, $\mathrm{B}_l^{(\uparrow,\downarrow)}$) and the other contains the right limbs (arm and leg, $\mathrm{B}_r^{(\uparrow,\downarrow)}$). The orange arrows connect corresponding joints on different parts, *e.g.*, the right elbow $\mathbf{j}_n$ is connected with the left elbow $\mathbf{j}'_n$. The vectors $\mathbf{u}_i$ and $\mathbf{l}_j$ identify the direction of the lines connecting corresponding joints.

intersects a straight line, called line of gravity in [119], and which divides the human body into two identical parts. Figure 4.1 illustrates the representation of a symmetric human body with respect to the plane of symmetry.

4.2.2 Metrics for Measuring Correct Posture

In contrast with the previous chapter, we further propose three measurements for evaluating postural defects. As shown in Figure 4.2, a skeleton S is divided into 5 body-parts, namely the back, the left and right arms, and the left and right legs. These body-parts were chosen because they can be used, as discussed in the next sections, to analyze general postural features discussed in the previous section in a simplified manner.

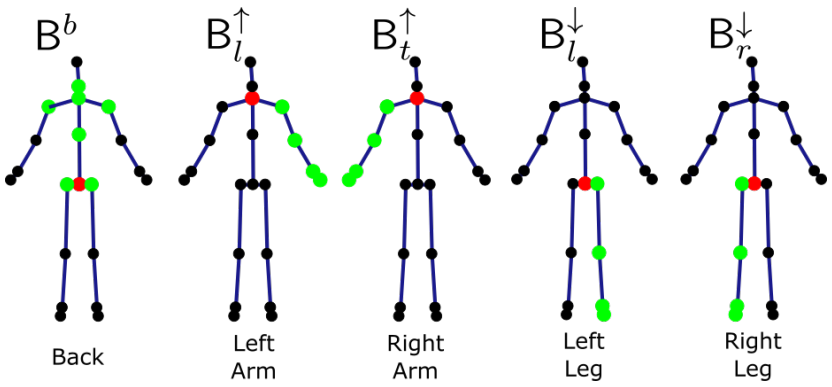


Figure 4.2: The human body is divided into 5 parts. The set of joints for each body part is highlighted in green and its local origin is the red colored joint.

Angle between Back and Gravity Vector

The objective of the first feature of correct posture is to have a straight back that is aligned with the gravity vector. Considering this, we propose to define the spine vector $\mathbf{w}$ as the vector that connects the hip joint, which is also the origin of the world coordinate system, with the neck joint. Since the skeletons are previously aligned so that the z-axis is aligned with the gravity vector, analyzing the deviation from a correct back posture is achieved by computing the angle θ between $\mathbf{w}$ and the direction $\mathbf{z}$ of the z-axis:

$$\theta = \angle(\mathbf{w}, \mathbf{z}). \tag{4.1}$$

The higher the angle θ, the worse is the back posture. The upper-part of the body (the first three body-parts, back, left and right arms) can be corrected by using the rotation $-\theta$ about the x-axis. Figure 4.3 shows the angle θ needed to rotate the upper part of the skeleton S such that it is aligned with the gravity vector.

Symmetry Between Upper and Lower Limbs

The second feature of a correct posture concerns the symmetry of the upper and lower limbs of the human body with respect to the plane of symmetry (refer to Figure 4.1). The plane of symmetry is defined as the plane that intersects the line of gravity and is aligned with

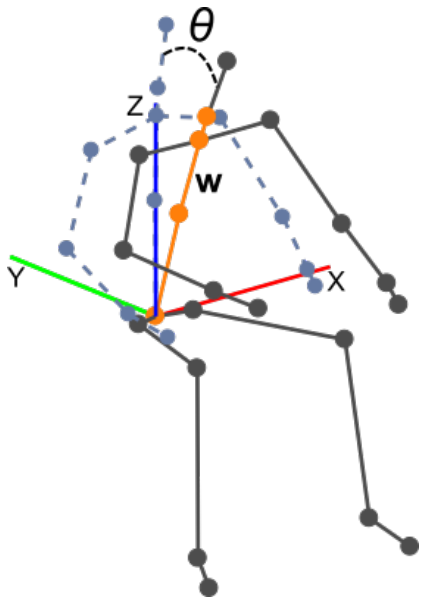


Figure 4.3: Angle θ between the spine vector $\mathbf{w}$ (orange color) and the gravity vector (blue color represents the z-axis).

the y-axis. As discussed in Chapter 3, since the skeleton is pre-normalized such that the world coordinate system is placed at the hip center, and rotated such that the projection of the vector from the left hip to the right hip onto the $x-y$ plane is parallel to the x-axis, the plane of symmetry correctly separates the left limbs ($\mathrm{B}_l^{(\uparrow,\downarrow)}$ in Figure 4.2) from the right limbs ($\mathrm{B}_r^{(\uparrow,\downarrow)}$ in Figure 4.2).

In order to achieve symmetry between the upper and lower body, the orthogonal distance between the joint $\mathbf{j}_n$ and the corresponding opposite joint $\mathbf{j}'_n$ with respect to the plane of symmetry should be equal. Let us define the distance d_n as the orthogonal distance between joint $\mathbf{j}_n$ and the plane of symmetry, and the same for the distance d'_n associated with joint $\mathbf{j}'_n$. Symmetry about the plane of symmetry is verified if $d = d'$, and, since the skeletons are normalized and centered with respect to the hip center, this is verified in case $\mathbf{j}'_n = [-j_{n_x}, j_{n_y}, j_{n_z}]$. Two corresponding body-parts $\mathrm{B}_l^{(\uparrow,\downarrow)}$ and $\mathrm{B}_r^{(\uparrow,\downarrow)}$ are symmetric if their joints are all symmetric about the plane of symmetry.

In order to simplify the analysis and visualization of the symmetry of joints about the plane of symmetry, we will also measure the angles of the lines connecting corresponding

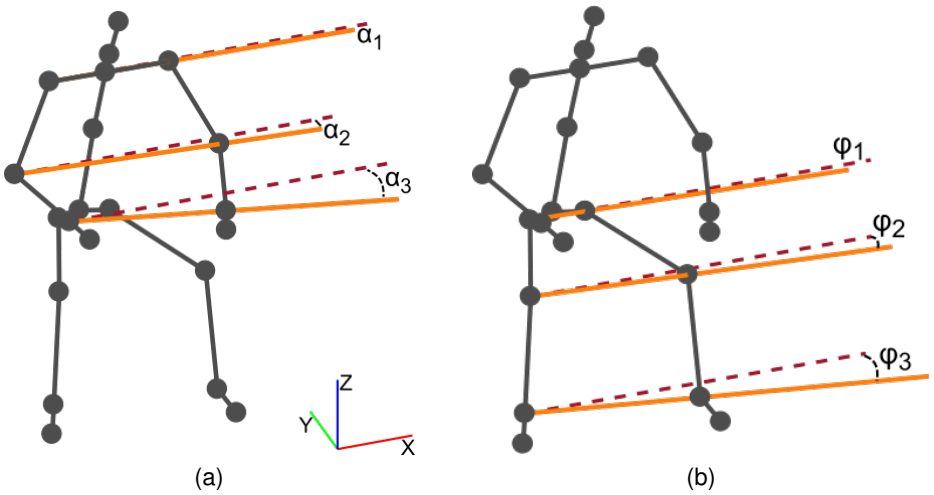


Figure 4.4: Representation of the critical angles between the lines connecting opposite joints (solid orange lines) and the lines parallel to the x-axis (red dashed lines).

joints, which will be called critical angles. Referring to Figure 4.4, let us define the vectors $\mathbf{u}_i$ and $\mathbf{l}_j$, with $i, j = 1, 2, 3$, as the vectors representing the directions of the lines that connect corresponding joints on different sides of the plane of symmetry. The vectors $\mathbf{u}_i$ concern the upper part of the body, $(i = 1)$ represents the shoulders, $(i = 2)$ the elbows $(i = 2)$ and $(i = 3)$ the wrists. Regarding the lower part of the body, $\mathbf{l}_j$ with $(j = 1)$ connects the hips , $(j = 2)$ the knees and $(j = 3)$ the ankles. Considering this, six critical angles are defined:

$$\alpha_i = \angle(\mathbf{u}_i, \mathbf{x}) \quad \text{with} \quad i = 1, 2, 3, \tag{4.2}$$

$$\varphi_j = \angle(\mathbf{l}_j, \mathbf{x}) \quad \text{with} \quad j = 1, 2, 3, \tag{4.3}$$

where, $\mathbf{x}$ is the direction of the x-axis. Figure 4.4 depicts the critical angles. The values of the critical angles for a correct posture should be as low as possible, ideally zero.

4.2.3 Posture Correction System

We explained in the previous sections how to compute and measure postural indicators, this is, the angle between the back and the gravity vector, and the distances and critical

angles between corresponding joints. This section explains how this information is used for assisting users in correcting their posture and converging to a more adequate physical state. The output of the proposed system is feedback suggestions, visual information and messages, in how to correct the back and have a symmetric body posture.

As depicted in Figure 4.5, the input to the system is a skeleton pose S of the user, which is acquired using the Kinect in this paper. The proposed algorithm is sensor independent and other technologies from which a skeleton can be estimated can also be used. The first step is to align and normalize S. Then, the angle θ between the skeleton back and the gravity vector is computed. This information is used as a first feedback indicator (feedback message 1 in Figure 4.5) and also used to correct virtually the current skeleton, obtaining S_c, for the next processing stages.

Given the corrected skeleton S_c obtained using a back rotation proportional to the angle θ, the next stage consists in identifying which lower and upper limbs of S_c should be moved so that the user's skeleton pose is symmetric about the plane of symmetry (refer to Figure 4.1). In order to achieve this, a database of correct skeleton poses for relevant postures and exercises is acquired using the supervision of an expert. For static postures like sitting, a discrete set of poses is sufficient, while for dynamic movements like lifting, a skeleton pose sequence is acquired. The corrected skeleton S_c is matched with one of the poses in the database (for dynamic movements, DTW is employed). Then, the Method 1 is used to identify the lower $\mathsf{B}^{(\downarrow)}$ and the upper $\mathsf{B}^{(\uparrow)}$ limbs that have the highest 3D error with respect to the template pose (highlighted in green in Figure 4.5). For an appropriate posture, these limbs should be a symmetric version of their counterparts about the plane of symmetry.

Let us define the operator s which reflects a body-part B about the plane of symmetry:

$$s(\mathsf{B}) = \bar{\mathsf{B}}, \text{ with } \bar{\mathbf{j}} = [-j_x, j_y, j_z]^\mathsf{T} \; \forall \, \bar{\mathbf{j}} \in \bar{\mathsf{B}}, \tag{4.4}$$

where $\bar{\mathbf{j}}$ is a general joint in $\bar{\mathsf{B}}$. Consider that the limb parts that had highest 3D error were $\mathsf{B}_l^{(\downarrow)}$ and $\mathsf{B}_l^{(\uparrow)}$ for the lower and upper parts, respectively (the same works for the right

limbs). Ideally, these body-parts should match the symmetric version of $\mathsf{B}_r^{(\downarrow)}$ and $B_r^{(\uparrow)}$, respectively, about the plane of symmetry. In order to guide the user to converge to a correct symmetrical posture, we compute the ideal symmetric versions:

$$\begin{aligned} \bar{\mathsf{B}}_r^{(\downarrow)} &= s(\mathsf{B}_r^{(\downarrow)}) \\ \bar{\mathsf{B}}_r^{(\uparrow)} &= s(\mathsf{B}_r^{(\uparrow)}). \end{aligned} \tag{4.5}$$

Finally, feedback proposals are obtained from the Method 1 that best align $\mathsf{B}_l^{(\downarrow)}$ with $\bar{\mathsf{B}}_r^{(\downarrow)}$ and $\mathsf{B}_l^{(\uparrow)}$ with $\bar{\mathsf{B}}_r^{(\uparrow)}$, respectively. From these matrices, feedback proposals are suggested to the user (feedback messages 2 and 3 in Figure 4.5).

4.3 Experimental Results

In this section, we evaluate the proposed posture assistance system using two different datasets, **sitting** and **lifting**, acquired using the Kinect v2 sensor.

4.3.1 Sitting

This dataset, acquired using the the Kinect v2 sensor, consists in different people sitting on a chair while writing or using a laptop. Generally, people tend to realize different body postures over time, many of which can cause serious physical injuries in the long-term. Usually, subjects start by having a correct posture, with a straight back aligned with the back of the chair and a symmetric posture of the upper limbs. As time goes by, the subjects start to feel tired of being in the same position and move the shoulders asymmetrically and bend the back towards the table. Having such an incorrect posture for a long period of time can cause serious injuries to the spine.

We tested our system using this dataset in order to assess if it could provide useful alerts to the user and support him in having a correct posture across time. In case the back angle or the critical angles are above a certain threshold, an alert is triggered and feedback proposals are spoken to the user by the system. Figure 4.6 shows an example where the

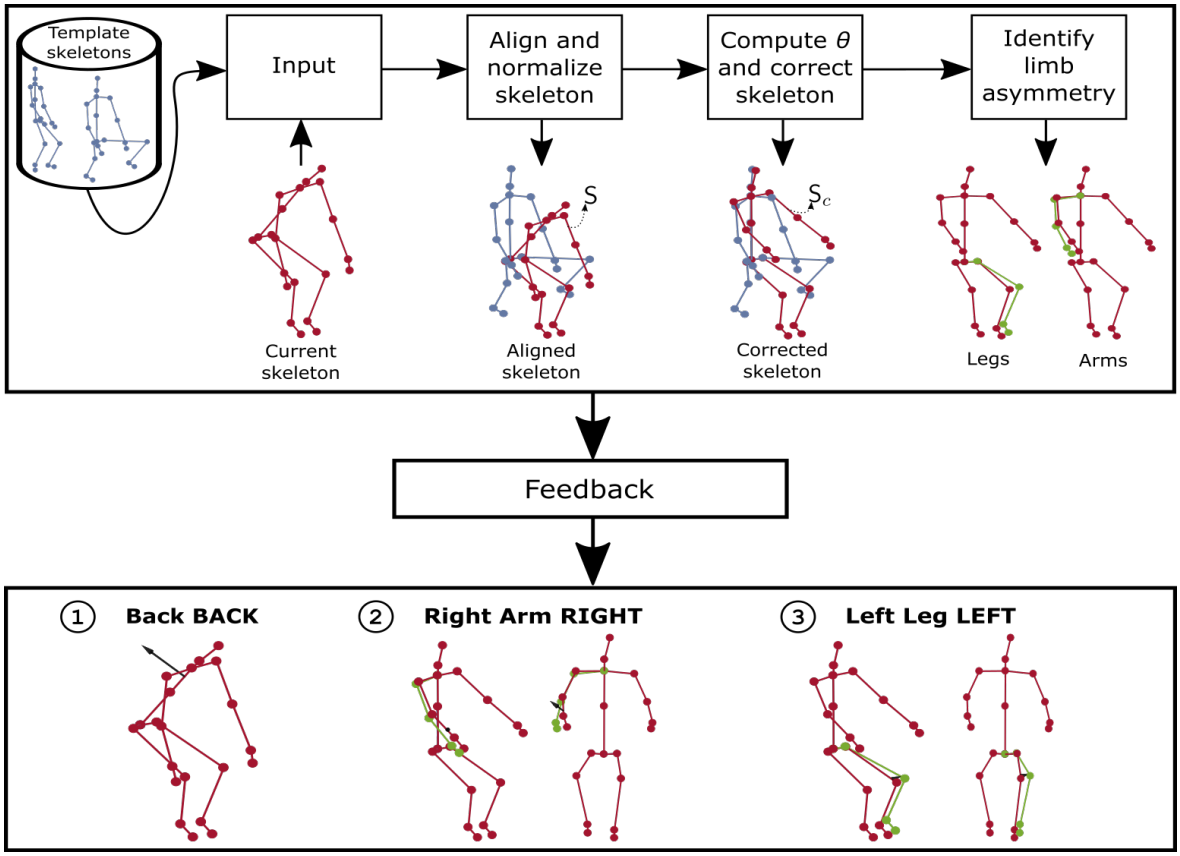


Figure 4.5: Overview of the different stages of the proposed approach. A database of template skeleton poses $\hat{S}$ (blue) is acquired using experts in the field relevant for the posture analysis application. The red skeleton S represents the current pose to which posture correction feedback should be provided. First, the skeletons are aligned and normalized and the angle for the back correction is computed. The corrected skeleton S_c is then generated by applying a rotation proportional to the angle for the back correction. Then, the lower and upper limbs to be moved are identified (green). Finally, feedback with information about the motion required to adjust the back, and the lower and upper limbs for converging to a correct posture is provided. The feedback is supplied in the form of visual information (black color arrows) and human interpretable messages.

feedback is proposed with the objective of correcting the back posture.

The objective is to study the posture of the subject while sitting on a chair during the working time. Considering this, we recorded a subject while sitting during 8 consecutive hours (regular working day time) with and without feedback proposals. The goal is to analyze

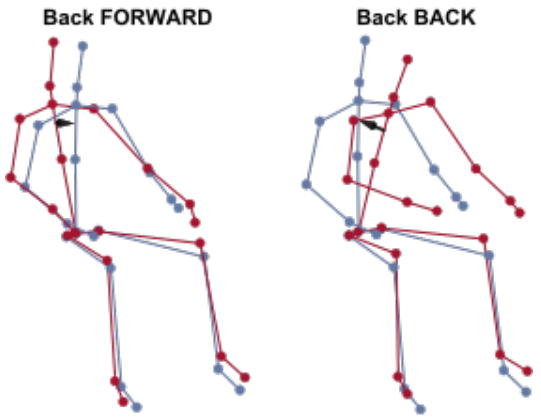
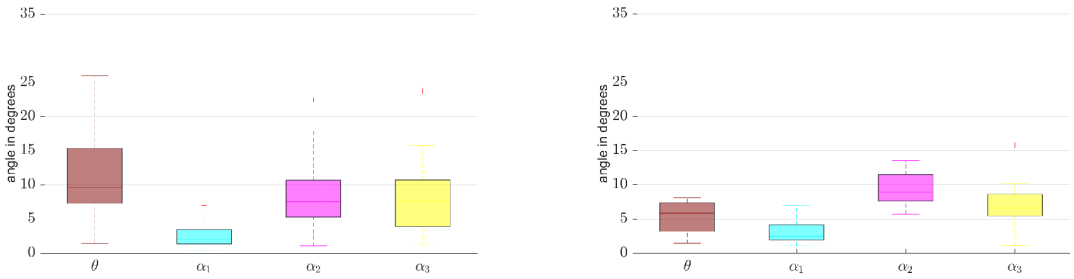


Figure 4.6: Feedback messages suggested by the proposed system to support the user in correcting the back posture.

the posture of the subject measuring the critical angles. Figure 4.7 shows the box plot over the critical angles with respect to a template correct posture for both experiments. Note that, in these experiments we do not evaluate the angles regarding the lower limbs of the subject due to the fact that while sitting, the lower limbs are not seen by the camera. Throughout the day, the subject has multiple postures while sitting due to the fatigue, these posture variations can be seen in Figure 4.7a where θ is the most affected angle. This angle θ concerns the angle of the back with respect to the line of gravity, concluding that the back of the subject is the most problematic body-part for this specific analysis. For the same experiment, we employed the feedback system to advise the subject and propose posture correction when predefined thresholds are reached. Figure 4.7b illustrates the critical angles for this experiment. Observing Figure 4.7, we conclude that the subject tends to correct his posture by following feedback proposals when an alert is provided, decreasing the values of the critical angles, specially the angle between the back and the line of gravity (θ).

4.3.2 Lifting

The objective of this experience is to analyze if the system is able to support and help a user in correctly lifting a weight. Most people incorrectly lift a weight by bending and executing most of the force using their back. Also, they tend to lose balance when lifting the weight



(a) No posture assistance.

(b) With posture assistance.

Figure 4.7: Box plot over the critical angles for two different experiments. In the first, no feedback was provided to the subject while sitting in a chair during the working time, Figure 4.7a. Figure 4.7b regards the second experiment, where the subject was informed to correct his posture following feedback proposals every time that his posture was considered as incorrect.

upwards, which causes an asymmetric body-posture and serious injuries. The ideal way of lifting a weight is to lower the upper body using a straight back and lifting the weight by exercising most of the force using the leg muscles.

The **lifting** dataset consists in multiple users lifting a metallic bar located on the floor, raise it over the head and then place it again on the floor. The experiment was performed by 100 different subjects following the same conditions (the same movement and the same bar). Figure 4.8a illustrates two examples (top rows) of the lifting exercise, the green skeleton sequence represents a correct posture for lifting and the red sequence represents a incorrect posture for lifting. Figure 4.8b depicts the back angle θ across time while lifting the bar, where θ_1 identifies a correct posture (top row), and θ_2 concerns the incorrect posture (second row). It is visible that θ_2 has a sudden increase when the user starts to bend to pick up the bar and also when leaving it on the floor. The reason of these high values of θ is that the user does not use the legs to apply the force to accomplish the lifting movement. Instead, the user bends the back to lift the bar and this is not the recommended posture to follow, causing severe injuries to the spine.

Each subject was asked to raise the bar two times. In the first, no instructions were provided. In the second, our system displays feedback alerts and messages on a screen

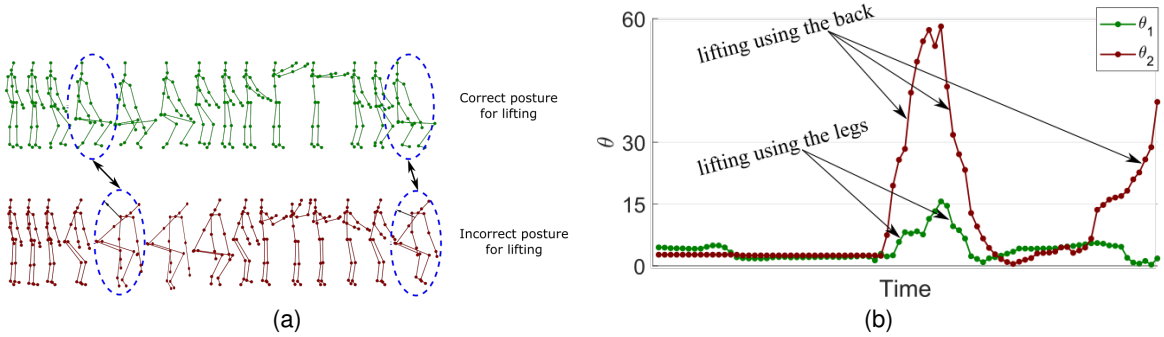
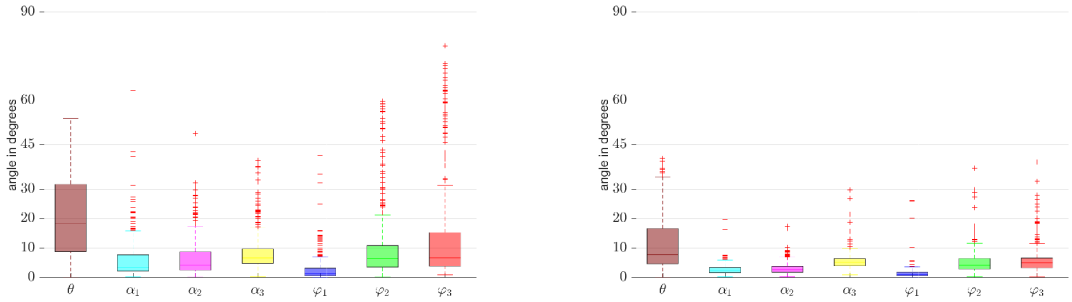


Figure 4.8: Figure4.8a shows an example of lifting sequence, where the exercise consists of picking a metallic bar, lift it over the head and leaving it on the floor. The green sequence (top row) illustrates a correct posture, and the red sequence (second row) shows an incorrect posture. The skeletons inside the blue dashed ellipse are examples where the back posture is particularly incorrect. Figure4.8b illustrates the angle θ over time, where θ_1 regards the correct posture (top row) and θ_2 the bad posture (second row).

in front of the user. Figure 4.9a shows a box plot over the angles for the first attempt (no feedback), and the angles for the second attempt are shown in Figure 4.9b. It is remarkable that, apart from the angle θ, the critical angles also had a significant decrease when compared with the first attempt. Resulting that, the user tends to correct the symmetry of the body following the feedback messages when applying force to lift the bar. Remark that with the proposed postural assessment and correction system, the user constantly has a more correct and healthier body posture, even for subjects without any experience in correct weight lifting.

4.4 Conclusions

In this chapter, we have proposed a system to guide users in how to correct their posture by providing real-time feedback without requiring a direct intervention of a therapist or a sports specialist. This is achieved by continuous monitoring of postural defects and using a database of correct skeleton poses for relevant postures and exercises acquired using the supervision of an expert as reference skeletons. Experimental results show that the provided



(a) No posture assistance.

(b) With posture assistance.

Figure 4.9: Box plot over the critical angles for two different attempts of the exercise. Figure 4.9a regards the first attempt, where no feedback was provided to the user while executing the lifting movement. Figure 4.9b concerns the second attempt, where feedback was provided to the user in order to correct the body posture.

feedback helps the user in converging to a healthy posture. The proposed system can be applied in different areas, such as rehabilitation at home (*e.g.*, for stroke survivors [31], [41], [44]), and in sports, *e.g.*, monitoring people in gyms and as soon as the posture of the user is not the most appropriated, the system generates an alarm with the objective of alerting the user to correct his posture.

In the next chapter, we present a low-cost solution designed for home-based rehabilitation of stroke survivors under the remote supervision of the therapist. Furthermore, we conduct an initial clinical validation in order to assess the usability of the home-based rehabilitation solution in real-world conditions.

Chapter 5

Home Self-Training: Visual feedback for Assisting Physical Activity for Stroke Survivors & Clinical Evaluation

With the increase of the number of stroke survivors [125], there is an urgent need for designing appropriate home-based rehabilitation tools to reduce health-care costs. The objective is to empower the rehabilitation of post-stroke patients at the comfort of their homes by supporting them while exercising without the physical presence of the therapist. In this chapter, a novel low-cost home-based training system is introduced. This system is designed as a composition of two linked applications: one for the therapist and another one for the patient. The proposed system is evaluated on 10 healthy participants and 10 stroke survivors without any previous contact with the system. Overall, the reported results suggest the relevance of the proposed system for home-base rehabilitation of stroke survivors.

5.1 Introduction

The number of stroke survivors in Europe is expected to increase by 25% in 2035 due to the growth of the aging population as reported in "*The Burden of Stroke In Europe*" report [125]. Hence, a great effort is being made to design solutions aiming at improving the quality of life of stroke survivors.

Thanks to the tremendous scientific and technological progress in computer science (*e.g.*, computer vision, *Internet of Things* (IoT), machine learning), a wide range of technology has been developed in this direction [126], [127]. More particularly, one can mention systems able to help clinical experts in the long-term rehabilitation process of stroke survivors [128]. However, as reported in [129], there exist very few medical institutions exploiting automatic computer-based tools. Thus, the design of a system supporting the rehabilitation of stroke survivors and more generally of disabled individuals is of major importance and can significantly impact medical development.

In a review of human motion tracking systems for rehabilitation, Zhou *et al*. [40] have highlighted the advantages of using markerless sensors since they present fewer restrictions, achieve good performance and are affordable. In [130], a virtual reality application has been introduced for brain injury rehabilitation. Lin *et al*. [131] proposed to use an eye-tracking device for rehabilitation of patients with dysfunctional eye movement. Also, in [132], [133], experimental results have proved that virtual reality applications favorably impact rehabilitation, while in [134], the *SonyPS2* potential for rehabilitation has been studied. To increase the reliability of such systems, some researchers have chosen to work with a multi-camera system. For instance, Lin *et al*. [135] employed a double *Charge-Coupled Device* (CCD) camera to capture the motion during rehabilitation sessions. In [136], the authors made use of a stereo-vision system in order to detect and assess the human motor reactivity under stimulation. On the other hand, some works have reinforced vision-based methods by combining markerless acquisition systems with other kinds of sensors or technologies. For example, Mirelman *et al*. [137] proposed a robotic-virtual reality integrated system to train post-stroke patients. Similarly, in [138], the authors associated Virtual Reality to *Wii*

gaming technology for stroke rehabilitation.

Recently, the availability of RGB-D cameras (*e.g.*, Kinect) has considerably boosted computer vision-based rehabilitation systems. In addition to RGB images, these cameras are able to capture in real-time depth images and 3D human skeleton sequences [16]. In fact, the 3D human skeleton, considered as a high-level representation, allows better discrimination of motion and is easy to manipulate [16], [17]. Following this trend, researchers used the Kinect for the rehabilitation of individuals with motor disabilities [5], [6], [22], [48]–[52].

In [22], a game-based rehabilitation application has been developed and the two acquisition systems OptiTrack[1] and Kinect have been compared in this context. Moreover, the authors investigated the capability of the Kinect as a robust tool for *Spinal Cord Injuries* (SCI) rehabilitation. The results have shown that the performance of both sensors is comparable. Clark et al. [139] also compared marker-based (VICON[2]) and markerless-based (Kinect) systems to assess the lateral trunk lean angle in healthy participants. Using an individualized calibration, the authors were able to obtain a small mean difference of 0.8°+/- 0.8°. In [48], the authors presented a virtual rehabilitation system for stroke survivors composed of a Kinect and a haptic glove for tactile feedback. Bao *et al*. [49] introduced a Kinect-based virtual reality training for the upper limbs after subacute stroke. Also, Zannatha *et al*. [50] proposed a rehabilitation system for stroke survivors by combining Kinect, a humanoid robot and ergonometric signals. Lozano-Quilis *et al*. [51] have proposed a system using virtual reality and natural user interfaces for the rehabilitation of patients with multiple sclerosis. In [6], a virtual reality-based exergame for post-stroke rehabilitation has been introduced and called Motion Rehab AVE 3D. Recently, Spasojevic *et al*. [53] presented a Kinect-based application to provide support to medical doctors during the clinical evaluation phase. In [52], a system allowing the therapist to tailor an exercise according to the patient has been proposed.

The aforementioned rehabilitation based systems have been mainly designed to assist clinicians in dedicated centers and are hardly usable remotely (without the physical presence of the therapist) since they do not provide real-time automatic feedback. In fact, to the best

[1] `https://optitrack.com/`
[2] `https://www.vicon.com/`

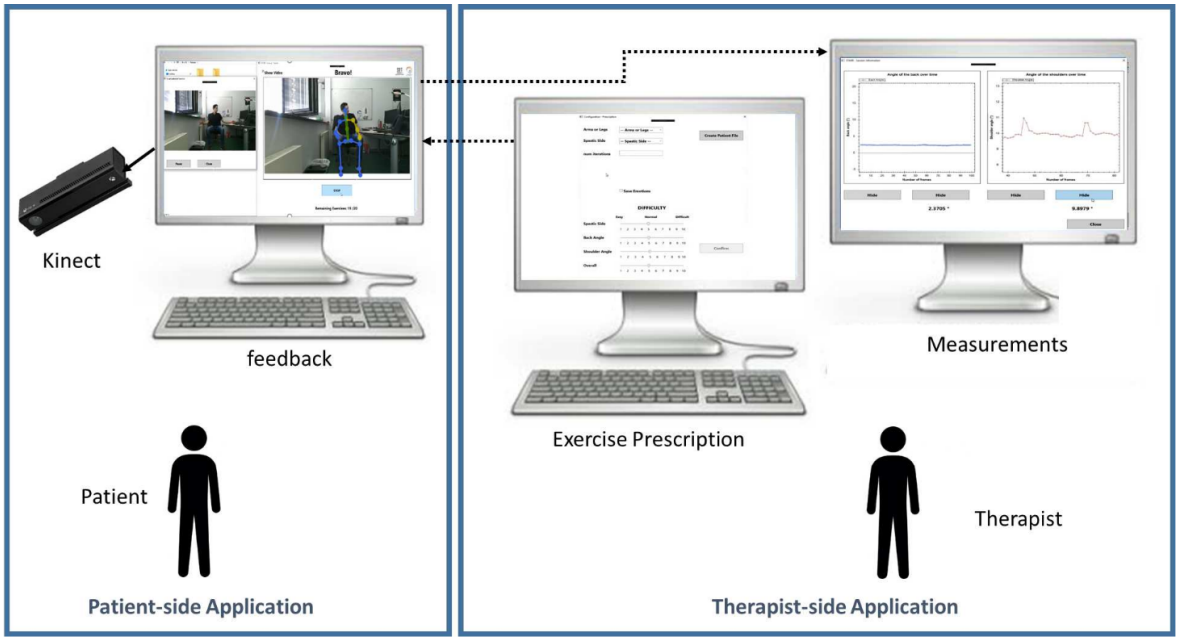


Figure 5.1: Overview of the proposed system dedicated to stroke survivors. The system consists of the combination of two end-user applications called: 1) the therapist side application; and 2) the patient side application; shown respectively on the left and right sides of the figure.

of our knowledge, very few works have proposed a computer-based system for home-based rehabilitation.

In this chapter, a markerless and affordable home-based rehabilitation system is designed for stroke survivors to orient them during the training is proposed. Figure 5.1 illustrates an overview of the proposed system. Our system is designed as follows:

1. Kinect is chosen for the acquisition as it is a low-cost markerless RGB-D sensor with acceptable and reliable measurement accuracy. In principle, any RGB-D sensor can be used.

2. To ensure the continuous communication between the therapist and the patient, our system is composed of two connected applications; one for each.

   (a) The therapist-side application enables remote personalized prescription of exercises and visualization of the measurements per exercising session.

(b) The patient-side application enables real-time and automatic visual feedback [72]–[75], [140] (clinically validated) as well as reporting on how well the exercise is performed [74], [141].

3. The measurements of each exercising session performed by the patient are automatically communicated to the therapist.

The proposed system exploits our previous work presented in Chapters 3 and 4, where we proposed two different visual feedback techniques to guide stroke survivors in self-rehabilitation scenarios. In summary, we present a system where a number of innovative computer methods were proposed, as listed below:

1. The development of a relevant and instantaneous computer-vision based feedback to guide and support the movement of the patient using a single RGB-D sensor;
2. The development of a relevant and instantaneous computer-vision based feedback to monitor the postural defects of the patient using a single RGB-D sensor;
3. The development of a relevant abstract measurement for evaluating the correctness of the global motion of the patient;
4. The optimization of the previously mentioned components in order to have a real-time interaction with the patients;
5. The full home-based training system that is composed of two applications which gathers all the components mentioned above.

As for the clinical point of view, this work has the following contributions:

1. A complete system architecture composed of two applications: the therapist side and the patient side.
2. A personalized prescription is done by the therapist where (s)he can tailor exercises according to the patient's profile and also update them given the exercising results.
3. A new feedback measurement that quantifies the global quality of the exercise.

4. An evaluation of spastic limb movement by comparing it to the equivalent healthy limb movement, instead of using a given template.

5. A first validation of the proposed system by conducting experiments on 10 healthy participants.

5.2 Methodology

5.2.1 Clinical Motivation

The primary goal of the proposed system is to support the rehabilitation of stroke survivors at home. Exercising is crucial for them to recover some autonomy in their daily life activities [28]. Unfortunately, many stroke survivors do not exercise regularly due to multiple reasons, such as fatigue, lack of motivation, confidence and skill levels [31].

Traditionally, stroke survivors are initially subject to physical therapy under the supervision of a health professional with the objective of restoring and maintaining activities of daily living known as functional activities in rehabilitation centers [104]. Consequently, they are continuously advised by experts on how to improve their movements and monitored in order to avoid health risks [29]. Unfortunately, due to the high economic burden [29], *on-site* rehabilitation is generally prescribed for a short period of time and recommended treatments and activities for home-based rehabilitation are suggested [30]. As an alternative, home-based rehabilitation, or self-rehabilitation programs are usually proposed since they are not expensive and do not involve the presence of a therapist [142], [143]. To that aim, the therapist usually explains and demonstrates the exercises to be performed by the patient. In addition, the therapist also provides the patient with a booklet containing an illustrated description of the prescribed exercises [143].

Having this in mind, the proposed system aims to monitor and guide stroke survivors while exercising without the physical presence of the therapist. To ensure adaptive clinical monitoring, continuous communication between the patient and the therapist is maintained allowing the therapist to follow the patient evolution and to adapt the exercises to its specific

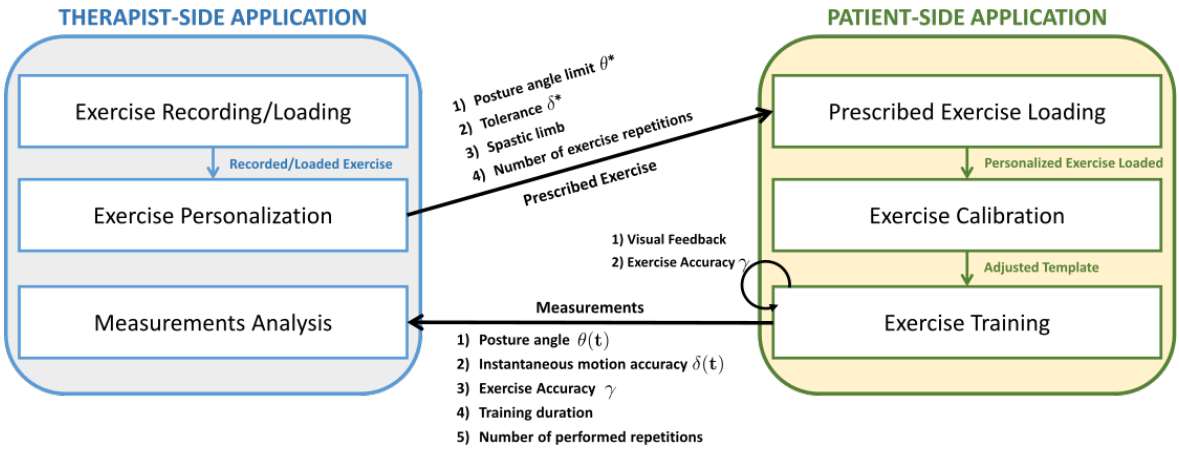


Figure 5.2: The architecture of the proposed system. Internal communication and functionalities of the proposed system composed of the therapist side application and the patient side application.

conditions. Some systems including this communication feature have already been designed as in [52], [144].

However, compared to earlier systems, we propose to include real-time, visual and easily interpretable motion feedback, presented in Chapters 3 and 4, tailored to the therapist prescription. This is different from what exists in the market, as for example, in [52], where feedback in a virtual reality game is proposed in the form of binary feedback, *i.e.*, correct or incorrect action. It is given after each repetition, independently of the therapist prescription. In [9], more sophisticated feedback proposals are introduced. Nevertheless, the corrective feedback is analyzed per joint involving a complex set of instructions for suggesting a particular body-part motion which is hard to interpret by the patient. Thus, the two simple novel color-coded feedback proposals are designed to instantly guide the patient on how to correctly perform the exercise and to warn him to avoid damaging compensatory movements. Compensatory movements are defined as the appearance of new motor patterns resulting from the adaptation of remaining motor elements or substitution [145]. For example, motor compensations can relate to the movement patterns that incorporate trunk displacement and rotation, scapular elevation, shoulder abduction, and internal rotation [146], [147].

Moreover, in order to evaluate the global quality of the exercise, we propose a novel ab-

stract measurement to be computed after each repetition. There are different attempts in the literature aiming at quantifying the exercise quality such as [9], [47]. These existent metrics rely on the comparison of the performed exercise with manually predefined templates, which do not take into account the patient's specificities. To overcome this issue, we propose an original way of defining the template: the patient is asked to do the exercise with the equivalent healthy limb; then, the exercise of quality is computed by comparing the spastic limb movement to the equivalent healthy limb movement. Thanks to this process, the anthropometry, as well as the movement particularity of the patient, are taken into consideration allowing a more subtle analysis.

5.2.2 System Overview

The proposed system is designed for home-based rehabilitation of stroke survivors under remote clinical control. For this reason, the proposed system is composed of two linked end-user applications, one for the therapist and one for the patient, as shown in Figure 5.1. For more clarity, in the rest of this chapter, we will refer to these two applications as the *therapist side* and the *patient side* applications.

Using the therapist side application, the therapist transfers a prescription containing personalized exercise(s) to the patient and receives relevant measurement describing each home-based training performed by the patient. For this application, only a basic camera and a computer are needed. The patient side application allows the interpretation of the therapist prescription, the presentation of visual feedback to the patient while training and transferring the training measurement data to the therapist. Figure 5.2 depicts in more detail the global architecture of the system.

For a deeper understanding of the system, we propose to describe a typical usage scenario of the system constituted of 6 steps: 1) the therapist starts by recording a video regarding a specific rehabilitation exercise; 2) the prescription or in other words, the exercise personalization, is created based on the patient's profile. Then, the prescription is automatically sent to the patient; 3) starting from the moment that the patient receives the prescription, the patient loads the exercise personalization that was prescribed by the therapist; 4) the patient

is first asked to perform a calibration task. Such a calibration task corresponds to the exact same exercise that was prescribed but with the healthy limb. This calibration process allows to adapt the exercise constraints and parameters to the anthropometry of the patient; 5) the patient starts training his spastic limb. To improve and correct his movement, the patient is constantly oriented with visual and understandable feedback proposals while training; and 6) at the end of the training session, a report containing relevant measurements describing the quality of the spastic limb motion is communicated to the therapist.

5.2.3 Therapist Side Application

As illustrated in Figure 5.2, the therapist side application is formed by three main components: 1) record/load an exercise; 2) prescribe a personalized exercise, and 3) analyze measurements.

Exercise Recording/Loading

The therapist loads or records rehabilitation exercises that can be prescribed to a particular patient. The recorded/loaded exercise constitutes a reference video which shows an exercise performed by a therapist.

Exercise Personalization

After choosing the exercise to be prescribed, the therapist adapts it according to patient-specific conditions by adjusting the following parameters:

- Maximum Posture angle (θ^*). It represents the maximum back angle with respect to the vertical plane for which the posture of the patient is considered as acceptable;
- Spastic limb. It specifies the spastic body-part to be trained;
- The number of exercise repetitions prescribed to the patient (n_1);
- Tolerance (δ^*). It defines the overall tolerated error. The lower the tolerance is, the stricter the feedback proposals are.

Measurements Analysis

This functionality allows the therapist to follow and analyze the home-based training sessions without requiring the physical presence of the patient. Each time that the patient performs a training session, a detailed report is sent to the therapist through dedicated communication service, as shown in Figure 5.2. This report contains the following measurements which are detailed in the upcoming section:

- Posture angle θ. It represents the back angle with respect to the vertical plane during the whole training session;
- Instantaneous exercise accuracy $\delta(t)$ at each instant t of the spastic limb;
- Duration of the training session;
- Number of repetitions (n_2) of the exercise performed by the patient during the training session;
- Exercise accuracy γ. It quantifies the quality of the patient motion based on the temporal alignment proposed in [141].

Such measurements allow the therapist to assess and evaluate the progress of the patient while using the proposed home-based rehabilitation system. This would also allow the therapist to have a clear understanding of the patient progress and to adapt the parameters to better fit the rehabilitation of the patient.

5.2.4 Patient Side Application

The patient side application enables the patient to load the prescribed exercise, calibrates the exercise with respect to the healthy limb, and train the spastic limb, as depicted in Figure 5.2.

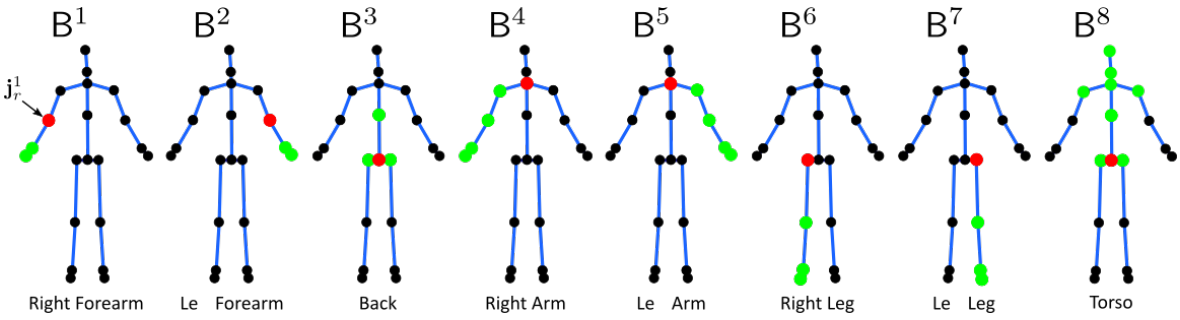


Figure 5.3: Body-part representation. The set of joints for each body-part is highlighted in green and its corresponding local coordinate system in red color.

Prescribed Exercise Loading

The patient should load the prescribed exercise that was created and shared by the therapist. This allows the system to set up all the movement constraints that are related to the patient. Such constraints directly define how strict should be the feedback proposals to be provided and also the number of repetitions asked by the therapist.

Exercise Calibration

This step is necessary to create the reference motion allowing the evaluation of the movement the spastic limb. The patient is therefore asked to calibrate the system with the equivalent healthy limb. For example, let us assume that a given patient has a spastic right arm and that the therapist has consequently prescribed some exercises involving it. Then, during the calibration phase, the patient is asked by the system to do exactly the same exercise with the left arm.

Since we make use of an RGB-D sensor, the 3D skeleton information extracted from depth images [16] is used to capture the movement during the calibration and later during the training session. Not very accurate skeletons encode the human pose (in contrast to RGB or depth images). Thus, the human body motion is represented using the spatial position of the skeleton joints, *e.g.*, left and right shoulders, elbows, wrists, etc. At each instant t, let $\mathrm{S}(t)$ be the captured skeleton composed of N joints. For optimal use of the skeleton, it is important to overcome skeleton viewpoint variation, camera positioning variability, and

anthropometry variation. Consequently, the same pre-processing of skeleton normalization and spatial alignment presented in previous chapters is applied to the skeleton sequence.

To differentiate it from the rest of the skeleton, we denote the body-part of interest by $\mathsf{B}^{limb}(t) = [\mathbf{j}_1^{limb}(t), \cdots, \mathbf{j}_{n_{limb}}^{limb}(t)]$, which is composed of n_{limb} joints. Figure 5.3 shows the body-part representation that is used.

Hence, in the calibration step, the joint positions of the side-opposed limb are recovered. To make them comparable to the spastic limb, we apply an axial symmetry. The obtained trajectories, varying over time t, representing the reference motion are denoted by $\hat{\mathsf{B}}^{limb}(t)$.

Exercise Training

While training, the patient is guided in how to correctly perform the proposed exercise and to avoid movement compensation. Consequently, the feedback proposals can be divided into three distinct parts: 1) the instantaneous motion feedback; 2) the posture monitoring, and 3) the measure of the exercise accuracy. While the two first feedback proposals are visually provided to the patient in real-time, the last one is reported to the patient as a percentage score after finishing each training iteration.

Instantaneous Motion Feedback

The objective of motion feedback is to support the patient while performing the prescribed exercise. To that end, the feedback proposals are provided at each time instant in order to iteratively help the patient improving the movement of the body-part of interest.

First, to measure the similarity between the reference body-part movement $\hat{\mathsf{B}}^{limb}(t)$ and the spastic body-part movement $\mathsf{B}^{limb}(t)$ at each instant t, we use the equation (3.1) to obtain the score function $m(t)$. Hence, the joints of the spastic and the reference body-parts respectively denoted as $\mathbf{j}_r^{limb}$ and $\hat{\mathbf{j}}_r^{limb}$ (colored in red in Figure 5.3) are anchored in the same local coordinate system. Consequently, the aim is to compute the rotation $\mathsf{R} \in \mathrm{SO}(3)$ that minimizes the error in equation (3.2). Thus, at each instant t, the intensity of correcting motion cost $\delta(t)$ defined in equation (3.3) relates to the difference between the

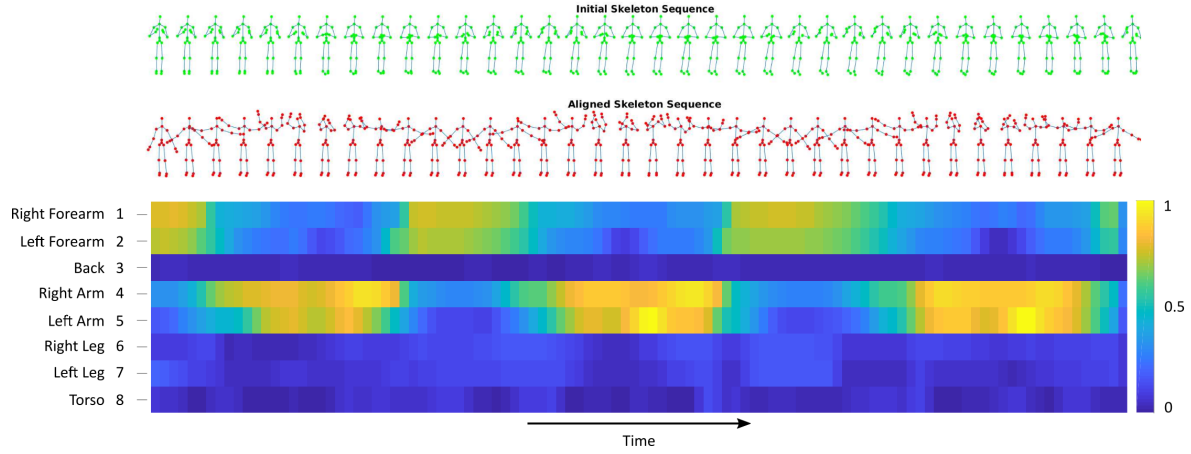


Figure 5.4: Intensity of correcting motion cost $\delta(t)^k$ of the feedback required for each body-part k. (Top) Skeleton sequence performing the clapping movement; (middle) skeleton sequence corresponding to the waving movement using two hands after spatial and temporal alignment; and (bottom) the intensity of correcting motion cost $\delta(t)^k$ calculated for each temporal instant independently (the vertical axis corresponds to different body-parts, while the horizontal axis regards the temporal dimension).

score function $m(t)$ and the error $e_{\mathsf{R}}(t)$.

For a given body-part k, Figure 5.4 shows an example of the intensity of correcting motion cost $\delta(t)^k$ for the movements of clapping and waving hands, where the cost is computed for each instant t. For the alignment of both skeleton sequences, DTW as presented in Chapter 2.

Compared to Chapter 3, only the body-part of interest is taken into account. Thus, the feedback is presented by highlighting the body-part of interest (in this case, the spastic limb) using $\mathsf{Color}_{\mathsf{B}^{limb}}(t)$, which is computed based on the intensity of correcting motion cost $\delta(t)$ as in [75]. It is defined as

$$\mathsf{Color}_{\mathsf{B}^{limb}}(t) = \begin{cases} green & , \text{ if } \delta(t) \leq \delta^*, \\ red & , \text{ otherwise} \end{cases}, \tag{5.1}$$

where the threshold δ^* represents the tolerance fixed by the therapist during the exercise personalization phase. Note that, δ^* plays an important role during the exercise personal-

ization. This parameter is highly correlated with the feedback that is provided to the patient while exercising. Considering that the feedback proposals are presented based on the reference motion, there is a need to sample the reference motion with the objective of guiding the patient to iteratively perform the proposed motion. As for the sampling function, we assumed a uniform distribution of the reference motion. Consequently, the parameter δ^* defines the threshold that the patient needs to reach in order to sequentially advance on the uniform distribution of the reference motion. In fact, if δ^* is relatively high, the provided feedback might be irrelevant. This motivates the interaction between the therapist and the patient. The more the therapist knows the limitations of the patient (by fixing the threshold δ^* accordingly), the more suitable the system is. The color transition from green to red (or vice-versa) is done gradually based on the current value of $\delta(t)$, as shown in Figure 5.5c. The green color feedback expresses the correctness, while the red one indicates the inverse.

Posture Feedback

As mentioned in Chapter 4, when stroke survivors are asked to perform specific movements, they tend to use the trunk to help in performing the movement. It is undesired since it can induce musculoskeletal injuries. Thus, the posture feedback is given using a feature reported in Chapter 4, which checks if the patient keeps a straight back aligned with the gravity vector. Figure 4.1 illustrates the patient's body representation divided by the plane of symmetry. The spine vector $\mathbf{w}$ is defined as the vector that connects the hip joint, which is also the origin of the world coordinate system, to the neck joint, as illustrated in Figure 4.3. Considering that the skeletons are previously normalized and aligned so that the z-axis is aligned with the gravity vector, analyzing the deviation from a correct back posture is achieved by computing the angle θ presented in equation (4.1).

Taking into consideration the clinical practices, the error of the angle that is introduced by the Kinect is of high importance. In [139] the authors evaluated the lateral trunk lean angle during the gait training using two different setups: 1) marker-based system (VICON); and 2) marker-based system (Kinect). In this study, the mean error of the Kinect sensor was of 3.2° +/- 2.2° when compared to the VICON system. However, with an individualized calibra-

tion, the authors were able to reduce the mean error to 0.8° +/- 0.8°. According to [148], such a range of error is acceptable for clinical gait analysis. Furthermore, they suggest that errors that are not larger than 5° may not have an impact on clinical interpretation. While these results can not be directly applied to the postural angle measured in the present study, it provides the first threshold of clinical acceptability.

Considering the feature θ, the posture monitoring feedback is also presented in a real-time color-based way. In this case, the highlighted body-part is always the back denoted by $\mathrm{B}^{back}(t)$ and B^3 in Figure 5.3. The color-based feedback $\mathrm{Color}_{\mathrm{B}^{back}}(t)$ at each instant t is defined as

$$\mathrm{Color}_{\mathrm{B}^{back}}(t) = \begin{cases} green & , \text{ if } \theta \leq \theta^* \\ red & , \text{ otherwise} \end{cases}, \tag{5.2}$$

where θ^* is the threshold fixed by the therapist during the exercise personalization step. This threshold indicates the maximum allowable angle to consider the patient posture as correct. The transition from one color to other changes gradually with the value of the current angle θ as presented in Figure 5.5c.

Exercise Accuracy

In contrast to the two previous types of feedback, this one is reported to the patient after performing the prescribed exercise. While the previous motion feedback and posture feedback instantly give an indication about the correctness of the local movement, the exercise accuracy provides a global evaluation. Since the skeleton joints of each body-part are provided, the movement of each body-part can be seen as a set of joints spatially varying over time. Thus, they can be considered geometrically as a set of joint trajectories. Thus, to evaluate the quality of the spastic body-part movement B^{norm}_{limb}, we propose to compare it to the movement of the healthy equivalent body-part $\hat{\mathrm{B}}^{norm}_{limb}$ assimilated to joint trajectories as well. This is done by computing the similarity between each pair of equivalent spastic and healthy joint trajectories. This is an original approach as compared to the common one where a template

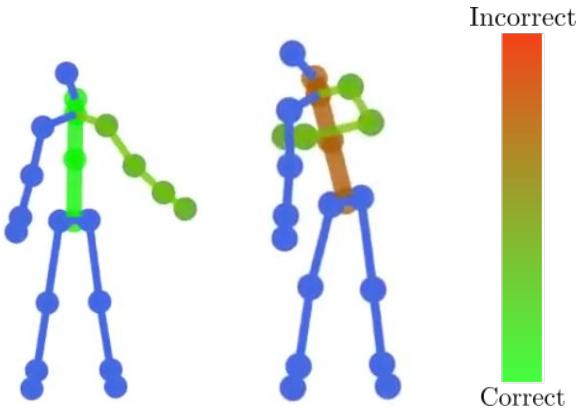


Figure 5.5: Feedback proposals in terms of the color code. Figure 5.5a illustrates the correct position of the body-parts of interest and a good posture. On the other hand, Figure 5.5b shows an example in which the patient uses the back to compensate the movement, resulting in red color feedback. Figure 5.5c depicts the color range used in the application to model correct and incorrect body position and bad posture.

is used as a reference. This further contributes to the system capability for personalization as the proposed accuracy is relative to each patient and not to an absolute measurement. Indeed, the patient specificities such as anthropometry or the way that the patient performs movements are considered.

Despite this, the execution rate variability resulting from different ways of performing a given movement can bias this comparison. For this reason, we propose to employ *Time Variable Replacement* (TVR) method of [141], which reduces this rate variability impact. This method reparametrizes the numerical joint trajectories by changing the time variable by a rate-invariant variable. We illustrate the concept of the temporal normalization TVR in Figure 5.6. Indeed, before applying the normalization, the movement of the reference body-part (blue) and the movement of the spastic one (green) are expressed in different temporal ranges, making them difficult to compare (Figure 5.6a). After normalization, the two movements are reported to the same range with a similar distribution of movement over time (Figure 5.6b). In this work, we used the *Normalized Pose Motion Signal Energy* (NPMSE) proposed by [141]. The obtained normalized trajectories corresponding to the joint j of the reference and the spastic body-parts are respectively denoted as $\hat{\mathbf{j}}_i^{limb,norm}$ and $\mathbf{j}_i^{limb,norm}$.

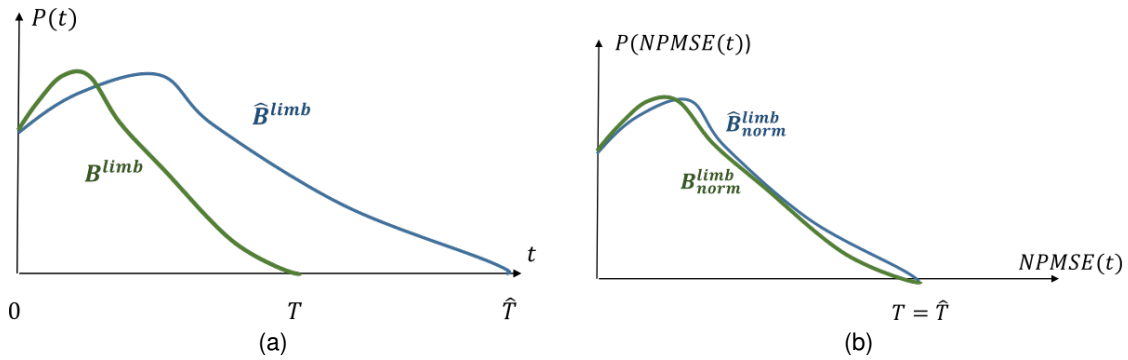


Figure 5.6: Example of the TVR method effect on two similar trajectories with different execution rate. While the green trajectory represents the motion of the spastic body-part, the blue one symbolizes the motion of the reference body-part. Figure 5.6a, the trajectories are plotted as functions of time. We can observe that the support of the two functions is different ($\hat{T} \neq T$). Figure 5.6b, after the change of the variable time by NPMSE, it can be noted that the trajectories vary in the same range $[0, T]$ and the two trajectories encoding similar movements look more similar.

To compute the similarity between the two body-parts, the Euclidean distance D between each couple of equivalent joints belonging to the spastic and the reference body-parts is then computed as

$$D = \sum_{i=1}^{n_{limb}} \|\hat{\mathbf{j}}_i^{limb,norm} - \mathbf{j}_i^{limb,norm}\|_2. \tag{5.3}$$

Thus, the smaller the distance is, the more similar the reference and spastic movements are and the better the movement quality of the spastic body-part is.

For an easier interpretation of this distance by the patient, the exercise accuracy measurement is defined as a percentage γ computed as

$$\gamma = \begin{cases} 100 \times \frac{\delta^*}{D} & \text{if} \quad \delta^* \leq D \\ 100 & \text{otherwise} \end{cases}. \tag{5.4}$$

We recall that δ^* represents the tolerated error adjusted by the therapist while prescribing the exercise.

(a) (b) (c) (d)

Figure 5.7: Illustration of the exercise used for the experiments: (a)-(d) are organized in the chronological order.

5.3 Evaluation on Healthy Participants

In this section, the preliminary results realized on 10 healthy subjects in order to evaluate the color-based feedback are reported. In the following, the implementation details, the experimental protocol, as well as the results and discussion are detailed.

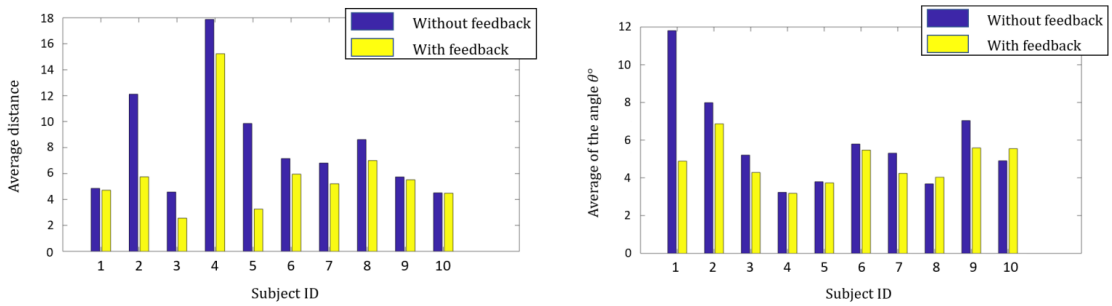
5.3.1 Implementation Details

The proposed system has been implemented in Visual C# within the Windows Presentation Foundation (WPF) framework where XAML, an XML-based language, was used to define and link the interface elements. To run any of the applications, a standard PC with Windows 10 as an operating system is required.

Description of setup: Microsoft Kinect v2 sensor connected to the laptop. The system runs at an average frame rate of 25 frames per second (fps).

5.3.2 Experimental Protocol

To analyze the impact of the color-based proposed feedback, we carry out two different experimental sessions (without and with feedback). During both sessions, the participants are asked to perform a unique exercise described in Figure 5.7. Advised by clinicians, the predefined exercise has been chosen to be relatively difficult compared to the classical exercises usually prescribed to post-stroke patients [143]. Indeed, since the system has been tested only on healthy participants, the exercise should ensure a sufficient level of complexity in order to make it somehow challenging. Thus, we propose an exercise that consists of rotating



(a) (b)

Figure 5.8: Figure 5.8a shows the average of the exercise accuracy for each subject. We specify that the reported value is computed as a distance and not as a percentage using equation (5.3). Figure 5.8b shows the average of the postural angle θ (in degrees) for each subject.

simultaneously the right hand and the left leg from the interior to the exterior with respect to the human body. In fact, it requires limb coordination due to the fact that we are proposing to move opposite limbs in a circular movement.

First session: Before the beginning of the session, we explain and demonstrate the proposed exercise to each participant. Then, after assimilating the exercise, the participant is asked to repeat the same exact exercise five times. During the session, all information about the 3D skeleton joint positions are acquired using the Kinect sensor and stored in a file. However, no feedback is provided to the participant in this first session.

Second session: The second training session consists of guiding the participant while performing the exercise. For that end, color-based feedback is presented to each participant. Following the same protocol described in the first session, each participant performs the same exercise five times, but now feedback is presented.

5.3.3 Experimental Results

For each participant, we average the exercise accuracy D and the postural angle θ. Figure 5.8a and Figure 5.8b respectively report these two values. In Figure 5.8a, we specify that we report the distance D and not the percentage. Thus, the higher is the distance, the worse is the exercise accuracy. Therefore, it can be noted that exercise accuracy is improved

when the color-based feedback is presented to the participant. Similarly, the postural angle θ decreases when the color-based feedback is provided (for almost all subjects, except for subject 10).

5.3.4 Discussion

The aim of this work is to develop a novel home-based system dedicated to stroke survivors using a Kinect sensor. Compared to earlier systems [9], [31], [52], [53], [149], the main clinical originality of our system resides in: 1) the provided color-based feedback, and 2) the continuous communication between the patient and the therapist, allowing an interactive personalization of the exercise, and 3) the quantification in terms of the exercise accuracy. While in computer methods point of view, our system presents innovations in the development and optimization of computer-vision based feedback to guide and monitor the patients' posture while exercising using a single RGB-D sensor. Still in the same aspect, this work also presents an abstract measurement to evaluate the correctness of the patients' movement and the full home-based training system that is composed of two dedicated applications.

In [9], feedback is provided by specifying the joint that (s)he should move as well as the recommended movement direction. This results in a complex set of instructions hardly understandable in real-time by the user. Similarly, Ferda *et al*. [31] proposed assistive feedback by manually specifying constraints on a single joint which define the correctness of the motion. In contrast, our system does not provide feedback related to a single joint but to a full body-part. This allows the definition of more simple feedback proposals which are directly translated into a simple color affecting the desired body-part (spastic limb). This intuitive message is therefore easy to interpret by the patients. Furthermore, no constraints are imposed while the patients are exercising.

Considering the second aspect, in the works [52], [149] the configuration that is done by the therapist happens when both therapist and patient are facing the game. Consequently, the therapist adapts the movement constraints based on the limitations of the patient with respect to the game environment, while in [53], the authors presented a tool to support

the evaluation of the patients. In contrast to these works, our continuous communication happens without the need for the physical presence of the therapist. This allows the therapist to configure remotely the parameters related to the patient's current condition.

The last challenge addressed in this chapter is the assessment of the quality of the movement. Compared to previous approaches [24], [47], [78], the proposed system has the advantage of being fully automated and adaptive to the patient's conditions. For instance, in [47], the authors presented a method that quantitatively evaluates muskulo-skeletal disorders of patients who suffer from Parkinson's disease. Their system may not be generalized for all kinds of exercises, this is due to their feature selection that is movement dependent (stepping time, swing level of the hands, etc). Tao *et al*. [24] also proposed a quality assessment framework. They model the dynamics of the human motion by using *Hidden Markov Models* (HMM). In addition, a manifold-based dimensionality reduction is applied to the human motion sequence. Contrary to that, we do not apply any dimensionality reduction method. Instead, based on the exercise personalization, we restrict the human body to the specific body-part and we only assess the quality with respect to that body-part. Not only that, but we also achieve real-time measurements.

In addition to the depicted novelties proposed by our system, to the best of our knowledge, we are the first proposing an experimental validation of this feedback. The obtained results on 10 healthy show an improvement of the posture and of the quality of the motion. Nevertheless, with such modest results, we were not able to draw final conclusions. More specifically, the improvement of the posture remains very slight, since the participants are healthy, do not present spastic limbs and consequently do not try to compensate with their back. Notwithstanding, these results allow us to present preliminary assessments about the interest of using such feedback proposals. As for the exercise accuracy, we noted a decrease in the measured distance for all the participants while using the home-based rehabilitation system. These first results tend to demonstrate the interest of the approach but should be applied to further participants, and more particularly to stroke survivors.

5.4 Clinical Evaluation

5.4.1 Participants

In total, 10 patients have participated in this clinical study. All the participants are chronic stroke spastic survivors with different levels of spasticity. Table 5.4.1 describes the profile of each patient. Recruitment was performed in outpatients' neurorehabilitation consultation by a physical medicine and rehabilitation doctor following a consecutive sampling method during 3 weeks. All the subjects were written and orally informed and informed consent was signed. All had the possibility to withdraw the consent to participate in the study.

Patient	Gender	Age	Years from stroke (Approx.)	MAS*	Affected side
1	F	72	9.4	3	Left
2	F	75	6.6	2	Left
3	M	52	1	1	Left
4	M	76	4.6	2	Right
5	F	51	26	2-3	Right
6	F	40	34	2-3	Right
7	M	63	8	1-2	Left
8	M	71	8.6	3	Left
9	F	57	6	2	Left
10	F	41	1.2	1	Right

Table 5.1: Profile of the post-stroke patients (gender, age, years from stroke, MAS*, affected side). *Spasticity level was measured by *Modified Ashworth Scale* (MAS) which measures the resistance during passive soft-tissue stretching. Scoring range varies from 0 (no increase in muscle tone) to 4 (affected parts are rigid in flexion or extension).

5.4.2 Procedures

The experiments have been completed within five days. During its full duration, a physiotherapist was always present to explain the exercises and intervene if necessary by helping patients in case of severe spasticity. Two different exercises were performed by the 10 patients called *scaption* and *hand to torso, face, and head*. The *scaption* exercise consists in raising the hand in the scapular plane (Exercise 1), while the *hand to torso, face, and head* exercise, more complex, implies to pass the hand in front of the torso and face before bring-

(a) (b) (c) (d) (e)

Figure 5.9: Illustration of Exercise 1 used for the experiments: (a)-(e) are organized in chronological order.

(a) (b) (c) (d) (e)

Figure 5.10: Illustration of Exercise 2 used for the experiments: (a)-(e) are organized in chronological order.

ing it to the top of the head (Exercise 2). Figure 5.9 and Figure 5.10 respectively illustrate Exercise 1 and Exercise 2. The activities were divided into three phases:

- Phase 0 (*day 1*): Before starting exercising, the physiotherapist explains and demonstrates the exercises to the patient. Then, the patient repeats the same exercises with the healthy upper-limb. While performing the exercise, a video containing the same exercise is shown to the patient. As explained earlier, the motion of the healthy limb is captured and is considered as the reference motion. This allows taking into account the anthropometry as well as the moving style of the patient.

- Phase 1 (*day 1*): During this phase, the patients are asked to run two different exercises with their spastic upper limb. Each exercise is performed 5 times with the spastic hand. In this phase, the color-based feedback proposals are not shown to the patient. Only one day has been dedicated to this phase allowing us to acquire baseline data.

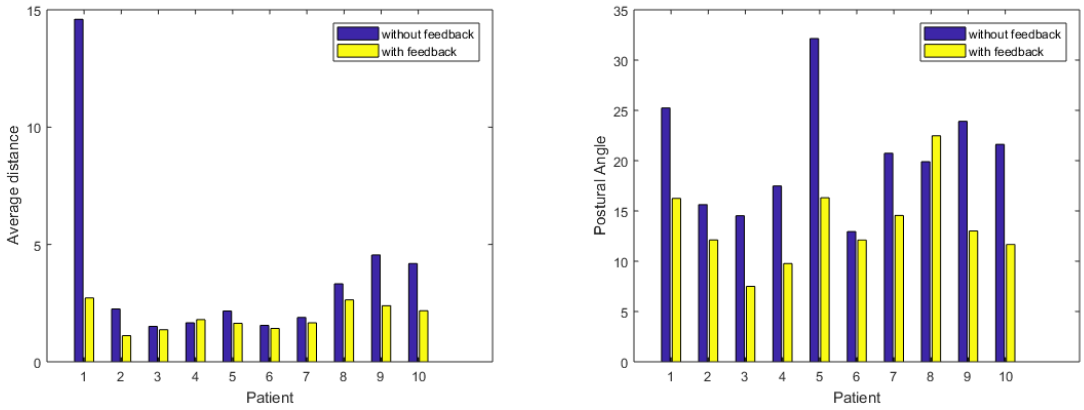
- Phase 2 (*days 2 to 5*): For the last phase of the study, the patients are also asked to carry out the same two exercises realized in Phase 0 and Phase 1 with their spastic upper-limb. In contrast to Phase 1, the feedback is provided to the patient.

For each phase, all patients are asked to perform each exercise 5 times per day. After each training session, the patient and the therapist fill the respective questionnaires presented in Appendix C and Appendix A. To avoid disturbing the patient, a very brief questionnaire has been prepared. In the last training session, a more extensive questionnaire, reported in Appendix B, is provided to the patient. The 3 questionnaires have been designed in tight collaboration with medical and human-computer interaction experts in order to consider important medical and usability patterns, respectively. They have been specifically prepared for this study and have not been used in a previous one. Most of the questions followed a Likert schema. In this case, the question is formulated as an affirmation (alternating positive and negative affirmations), and the answer can vary on a scale going from 1 to 5 (1 and 5 respectively correspond to completely agree and completely disagree), allowing a simple computation of statistics. To get better explanations, some open questions have also been formulated. The questions have been selected such that different aspects are considered, namely *technical problems*, *utility of feedback*, *psychology of the patient*. The therapist questionnaire gathers a total of 9 Likert questions (with 3 explicative questions if the answers are negative) and 3 open questions. The patient questionnaire designed for each session comprises only 4 Likert questions. Finally, the patient questionnaire designed for the last session is more extensive and is composed of 11 Likert questions and 5 open questions.

Hence, the data acquired in Phase 1 (without feedback) and Phase 2 (with feedback) are compared with respect to the reference data acquired in Phase 0.

5.4.3 Criteria of Evaluation

Similar to the previous evaluation, we use the two criteria of evaluation called TVR-based average distance and postural angle. The TVR-based average distance is a measure that



(a) (b)

Figure 5.11: Figure 5.11a shows the average of the distance $D_{\bar{F}}$ and D_F per patient obtained for Exercise 1 (respectively without and with feedback). Figure 5.11b shows the average of the postural angle $\theta_{\bar{F}}$ and θ_F (in degrees) per patient during Exercise 1 (respectively without and with feedback).

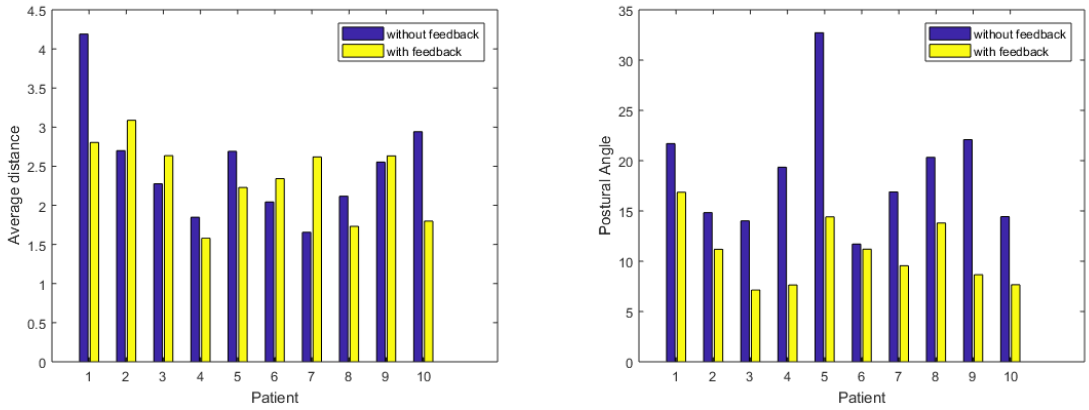
estimates the dissimilarity between two given motions, inspired by the work of [141]. Thus, in our experiments, this measure is computed for each patient between:

- The reference motion and the motion of the spastic upper limb without feedback (Phase 1). The average of this measure is reported and denoted by $D_{\bar{F}}$.
- The reference motion and the motion of the spastic upper limb with feedback (Phase 2). The average of this measure is reported and denoted by D_F.

Thus, the more similar to the reference motion the spastic motion is, the lower this distance is and the more it is considered as correct. The average postural angles without and with feedback are also computed for each patient and respectively denoted by $\theta_{\bar{F}}$ and θ_F. As previously mentioned, the higher the angle is, the worse the posture is.

5.4.4 Experimental Results

In this section, we present the obtained results with 10 stroke survivors. On the one hand, Figure 5.11a and Figure 5.11b, respectively, report the obtained average exercise accuracy and



(a) (b)

Figure 5.12: Figure 5.12a shows the average of the distance $D_{\bar{F}}$ and D_F per patient obtained for Exercise 2 (respectively without and with feedback). Figure 5.12b shows the average of the postural angle $\theta_{\bar{F}}$ and θ_F (in degrees) per patient during Exercise 2 (respectively without and with feedback).

postural angle per patient while performing the Exercise 1. On the other hand, Figure 5.12a and Figure 5.12b, respectively, report the obtained average exercise accuracy and postural angle per patient while performing Exercise 2. It can be noted that usually $D_F < D_{\bar{F}}$ for Exercise 1. Nevertheless, this is not obvious for Exercise 2, where we can observe that D_F is not always inferior to $D_{\bar{F}}$. Only the motion of 50% of the patients (Patient 1, Patient 4, Patient 5, Patient 8 and Patient 10) is improved with the use of the feedback. It is also observed that the postural angle is largely reduced in both exercises for almost all the patients. Therefore, one can say that the posture is importantly improved when using the postural feedback; consequently, avoiding musculo-skeletal injuries.

5.4.5 Discussion

The goal is to clinically evaluate the different components of the proposed home-based rehabilitation system for stroke survivors.

Posture Feedback

The reported results show the utility of the postural feedback provided by the home-based rehabilitation system. Indeed, for both exercises, the postural angle is widely reduced. 44% of the participating patients spontaneously said that the strength of this system is its posture correction. This feature can be seen as the most important one since it allows the prevention of musculoskeletal injuries and spasticity increase resulting from movement compensation.

Motion-based Feedback

In contrast to the postural feedback, the relevance of the motion-based feedback is not as clear. As mentioned earlier, while in Exercise 1, we can observe a clear improvement of the movement, this is less evident to claim in Exercise 2. This can be explained by the fact that Exercise 1 is easier to execute than Exercise 2. A more complex exercise is usually more difficult to reproduce identically than a simple one. Despite that, 100% of the patients completely agree that the color-based feedback is useful. Similarly, the physiotherapist also completely agrees with that in 100% of the training sessions. Notwithstanding, it is difficult to say if this system has a positive impact on the patient's recovery since it has been tested within a limited period of 5 days. This is confirmed by the nuanced perception of the patients: 33.3% completely agree, 22.2% agree, 33.3% do not agree or disagree and 11.1% completely disagree that their condition has improved since they are using the application. In contrast, the therapist confirmed that, in 78.3% of the cases, the application helped the patient improving his controlled movement.

Measurement Report

Thanks to the questionnaire addressed to the therapist, we can note that the measurement report received by the therapist is relevant. In 59.4% of the cases, the therapist finds that the measurements reflect the training of the patient, does not have any opinion about that in 29.7% of the cases, and thinks that the measurements are not relevant in 10.8% of the cases.

Reliability and Simplicity of the System

The proposed system has the advantage to be reliable. The therapist reports that in 97.3% of the tests, the system did not generate bugs. 88.89% of the post-stroke patients find that the capture of the movement is accurate and well reproduced on the screen. They state, in 97.1% of the cases, that the system is easy to use, while the therapist confirms this, in 100% of the cases. Finally, all the patients describe their experience as *very positive* or *good*.

Safety of the Application

100% of the patients feel in security while training with the application. They were sitting on a chair and working on the upper limb such that they felt comfortable and without any risk of falling. However, due to visual or cognitive problems or the inability that few patients have to make painless active movements, the physiotherapist agrees only in 79.3% of the cases that the patient can use the application alone at home in security; thus, avoiding the generation of abnormal movements, the exaggeration of compensatory patterns or/and pain. For this reason, the permanent presence of the therapist during the first sessions is of high importance, allowing the reinforcement of the correct movements while using system. This is also confirmed by patients' answers: while 88.9% would like to continue using it in the presence of the therapist, only 33.3% feel ready to use it alone.

Psychology of the Patient

According to the questionnaires, the application has a positive impact on the psychology of the patients. In 100% of the cases, the patients completely disagree that the training sessions are boring and that the exercises are hard to perform. Moreover, in 88.9% of the cases, they felt more motivated to exercise since they are using the application. On the other side, the therapist finds that patients were feeling comfortable with the application in 71.3% of the cases.

5.5 Conclusion

In this chapter, a novel home-based training system for the rehabilitation of stroke survivors has been introduced. Advised by clinicians, our system has been designed to answer different medical requirements. More specifically, its originality consists of: 1) a permanent remote communication between the therapist and the patient (exercise prescription and patient monitoring); 2) novel easily-understandable color-based feedback proposals to guide the patient during the training sessions; and 3) the consideration of the patient's anthropometry and specificities to evaluate the quality of the spastic body-part movement (during the personalized prescription and thanks to the calibration phase). Experimental results on 10 healthy subject and 10 stroke survivors show promising advances towards a home-based rehabilitation solution. In general, the application can be considered as reliable, simple to use, and positively impacting the psychology of the patients. In the next chapter, we address the challenge of detecting abnormalities in the context of stroke survivors rehabilitation.

Chapter 6

Deformation-Based Abnormal Motion Detection using 3D Skeletons

In line with the previous chapters, assessing and understanding motion analysis in home-based rehabilitation scenarios is essential. In this chapter, we propose a system for abnormal motion detection using 3D skeleton information, where the abnormal motion is not known a priori. To that end, we present a curve-based representation of a sequence based on few joints of a 3D skeleton and a deformation-based distance function. We further introduce a time-variation model specifically designed to assess the quality of a motion; we refer to a distance function based on such a model as *motion quality distance*. We validate our approach using a publicly available dataset. Qualitative and quantitative results show promising performance.

6.1 Introduction

It is crucial for stroke survivors to exercise to recover some autonomy in their daily life activities [28] regularly. Stroke survivors are initially subjected to physical therapy sessions under the supervision of a healthcare professional, who usually suggests activities for home-based rehabilitation [30]. In such scenarios, having an automated tool that assesses the patient's

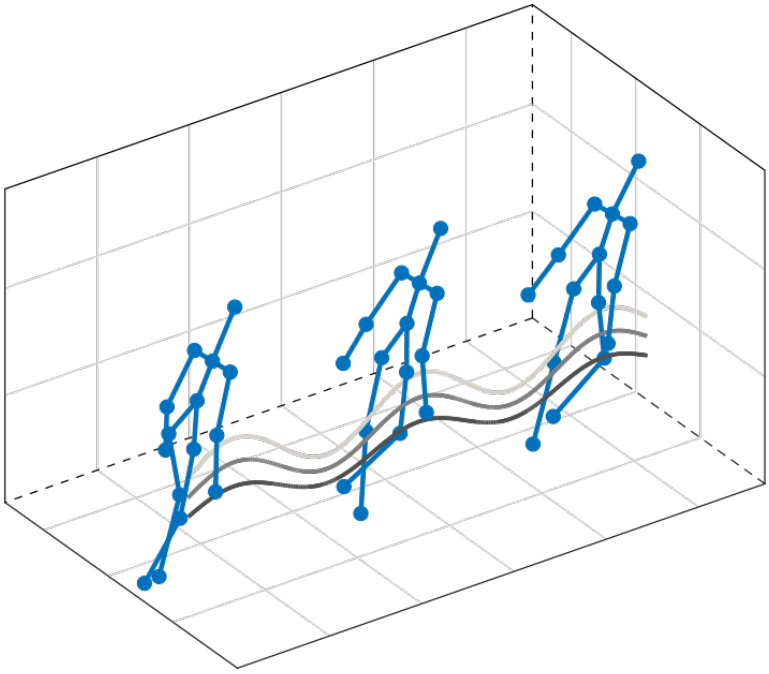


Figure 6.1: Illustration of the proposed representation: The figure shows the proposed curve-based sequence representation derived from the knee joints only.

exercise quality could be of great importance. This would facilitate the monitoring of the patients' progress by the physiotherapist, enabling an intervention in case of continuously less satisfactory reports.

In general, the quality of a given motion is estimated by measuring the deviation from what is normal. Such approaches may fall into the topic of abnormal motion detection. Abnormalities are usually detected by comparing a given motion to a normal motion model [23], [24], [54]. In order to detect abnormalities, two main approaches exist: 1) where the abnormalities are known a priori; 2) where there is no information of what an abnormality is. In [150], a binary classifier is trained to classify an observation into either normal or abnormal. In contrast, works such as [23], [24], [151], [152] presented methods where a normal motion model is learned. Then, the abnormalities are classified as deviations, given some threshold, from the model.

In this chapter, similar to [23], [24], [151], [152], we build a model from a set of normal motions and detect abnormalities based on a predefined threshold. We begin by repre-

senting a skeleton sequence as a curve by tracking few selected joints of the skeletons throughout the sequence. Contrary to [23], [24], we do not apply a manifold-based dimensionality reduction method. Instead, similar to [47], we select few joints using problem-specific knowledge to build a 2-dimensional representation from a $3 \times N$-dimensional data, for N skeleton joints, see Figure 6.1. With this curve representation, a distance function on curves could then be used to measure the dissimilarity between two sequences. We propose to adapt the *deformation-based distance* function, introduced in [153], for this purpose. Subsequently, we model normal motion sequences by computing the average of their representation.

We note that the speed, or latency variation in between sequences, contributes significantly to a distance function–mainly due to mismatching frames. Hence, the representativeness of a model depends highly on how time-variation is addressed. We thus introduce the concept of a *motion-quality* distance. This distance is based on detecting matching frames (key-frames), dividing the observed sequences into matching sub-sequences. The latest sub-sequences are then used to compute a series of sub-distances. This approach is different from a simple linearization as proposed in [154]. We define the motion-quality distance such that the variation in time is emphasized, while the optimal alignment proposed in [154] emphasizes invariance to variation in time.

6.2 Related Works

The interest in assessing the quality of human actions has been increasing very rapidly over the last years. This section briefly reviews vision-based methods to assess the quality of actions and methods related to abnormality detection in human motion analysis.

Assessment of the quality of actions: Assessing the quality of an action can be thought of as a problem of measuring how close an action is from a reference action. In other words, they are based on giving a score measured by the matching between an action and a pre-trained model [23], [24]. Pirsiavash *et al*. [9] proposed a framework for learning how to assess the quality of actions from RGB videos. In their work, a regression model is

trained from spatio-temporal pose features such that it predicts scores of actions. Training a regression model requires a large number of annotated data. Using a different type of input data, Wang *et al*. [47] presented a method to address the problem of automated quantitative evaluation of muskulo-skeletal disorders of patients who suffer from Parkinson's disease using a 3D sensor. They selected a reference motion by choosing one cycle of the motion and performing temporal alignment with the remaining motion cycles. The feature selection is movement dependent (step size, stepping time, and the swing level of the hands), making it undesirable to be generalized to other motions. Tao *et al*. [24] proposed a framework to assess the quality of actions, where they evaluate the deviation of an observation from a learned model of normal human motion. They used HMM to model the dynamics of human motion from skeleton-based samples of healthy individuals. This can be seen as an abnormal motion detection system, where the deviation from a model of normal motion is estimated.

Abnormality Detection: Abnormality detection refers mainly to the problem of finding patterns in data that do not conform to expected behavior [54]. There are two main approaches for abnormality detection: 1) the abnormality is known as prior knowledge, and 2) the abnormality is unknown beforehand. In the first approach, the work of Parra-Dominguez *et al*. [150] presented a supervised learning method to learn the abnormalities during stair descent (fall detection). A binary classifier is then trained on annotated data in order to decide whether the motion is normal or abnormal, while in the second approach, there is no knowledge of what an abnormality is. Nater *et al*. [151] proposed to learn a model of normal human behavior in an unsupervised way. The model uses a hierarchical representation of appearance and action level of normal movements to detect abnormalities (fall detection) from silhouettes. Snoek *et al*. [152] trained an HMM using sequences of normal staircase motion to detect abnormal motion during stair descent from RGB data. The closest work that is most related to ours is [23]; there, the authors presented an approach that detects abnormal events and provides an assessment of the quality of the motion on a frame-by-frame basis. The work is based on a continuous statistical model that is built from a set of normal human motion using 3D human skeleton data. A non-linear manifold

technique was used in order to reduce the dimensionality of the skeleton information.

6.3 Problem Formulation

In this section, we briefly describe the problem formulation of an abnormal motion detection system.

Let S be the skeleton pose of a human subject with N joints. Subsequently, a skeleton-based abnormal motion detection system classifies a sequence M as either normal or abnormal. In general, such a problem is formalized as

$$f(\mathsf{M} \mid \mathsf{M}_t) = \begin{cases} 0 & , \text{ if } \|\mathsf{M} - \mathsf{M}_{t_0}\| < \|\mathsf{M} - \mathsf{M}_{t_1}\| \\ 1 & , \text{ otherwise} \end{cases}, \tag{6.1}$$

where $\|\cdot\|$ denotes a distance function, and M_t denotes the set of parameters of the abnormal and normal movement models, *e.g.*, simple kernel based binary classifier [155].

In this chapter, however, we aim to estimate an abnormal motion detection system from a set of normal motion sequence examples, without any prior knowledge of what an abnormal motion is. As a result, instead of solving for a binary classifier that maximizes a decision margin as given in equation (6.1), we introduce a motion representation approach $\mathcal{G}(\cdot)$ such that the distance between two normal motions is less than a given ζ under a distance function $D(\cdot, \cdot)$ of the motion representation. Hence, equation (6.1) may be reformulated as

$$f(\mathcal{G}(\mathsf{M}) \mid \mathcal{G}(\mathsf{M}_t)) = \begin{cases} 0 & , \text{ if } D(\mathcal{G}(\mathsf{M}), \mathcal{G}(\mathsf{M}_t)) > \zeta \\ 1 & , \text{ otherwise} \end{cases}, \tag{6.2}$$

where $\mathcal{G}(\mathsf{M}_t)$ represents the model of a normal motion.
In the following section, we describe our proposed approach for representing a normal motion such that performance of equation (6.2) is maximized.

6.4 Proposed Approach

In this section, we describe the proposed approach for estimating equation (6.2) and discuss data variation due to motion velocity.

6.4.1 Motion Representation

In specific problems like abnormal motion detection, descriptive movements of a motion sequence are captured by tracking fewer joints than is needed for general problems, *e.g.*, action recognition [17], [20], [156]. Hence, a given skeleton S is a high dimensional data point with redundant information. As a result, we define a data representation approach that reduces the dimensionality of the data by using a priori problem-specific knowledge.

In abnormal motion detection problems, the motion analyzed is conditioned on a specific and fixed action, *i.e.*, every subject, in both training and testing dataset, is expected to perform a similar action. Hence, there is no variation in action class but in the manner that an action is performed. As a result, we select a subset of the joints that shows the largest variation in performing a specific action as representative joints. In this work, we select the left and right knee joints, which we denote by $\mathbf{j}_{lk}$ and $\mathbf{j}_{rk}$, respectively, as representative joints. Thus, a skeleton sequence is represented as

$$\mathcal{G}(\mathsf{M}) = \{p_i : \forall_{i\in[1,\cdots,T]}, p_i = \|\mathsf{P}\mathbf{j}_{lk_i} - \mathsf{P}\mathbf{j}_{rk_i}\|_2\}, \tag{6.3}$$

where P is defined as the projection matrix on the main direction of the motion variation. Although, the main direction can be estimated using conventional dimensionality reduction methods like *Principal Component Analysis* (PCA) [157], in this work we select the y-z plane as the main movement direction using a data normalization process. This ensures the joint's

main variation to lie on y-z plane, see upcoming section. Hence, P is defined as

$$\mathsf{P} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{6.4}$$

Subsequently, we use a deformation-based distance function, as defined in [153], [154], for measuring the difference between two given motion representations; a brief description of the considered distance function is given below.

Distance Function

Consider a curve defined by equation (6.3) and described by a set of q points that are sampled from $T \gg q$ set of points using φ defined as follows

$$\varphi : [0, T] \rightarrow [0, q]. \tag{6.5}$$

We denote such a curve as $\varphi \circ \mathcal{G}(\mathsf{M}) = (p_1, \cdots, p_q)$, where $p_i \in \mathbb{R}^2$. In [153], [154], such a curve is represented by a set of rigid-transformation matrices as $(g_1, \cdots, g_{q-1})$, so that $\varphi \circ \mathcal{G}(\mathsf{M}) = (g_1 p_1, \cdots, \prod_{i=1}^{q-1} g_i p_1)$. Henceforth, we denote rigid-matrix based representation of a curve $\varphi \circ \mathcal{G}(\mathsf{M}_a)$ as $\mathcal{C}_{a,\varphi} = (g_1, \cdots, g_{q-1})$.

Thereafter, following [153], [154], the distance between two curve representations $\mathcal{C}_{a,\varphi}$ and $\mathcal{C}_{b,\varphi}$ is defined as

$$\mathfrak{D}(\mathcal{C}_{a,\varphi}, \mathcal{C}_{b,\varphi}) = \sqrt{\sum_{i=1}^{q-1} d(g_a^i, g_b^i)^2}, \tag{6.6}$$

where $d(\cdot, \cdot)$ is the geodesic distance given by

$$d(g_a, g_b) = \sqrt{\|\mathsf{log}(\mathsf{R}_a^{\mathsf{T}}, \mathsf{R}_b)\|_F^2 + \|\mathbf{v}_b - \mathbf{v}_a\|_F^2}, \tag{6.7}$$

where T denotes the matrix transpose, $\|\cdot\|_F$ denotes the Frobenius norm, and

$$g_i = \begin{pmatrix} \mathsf{R}_i & \mathbf{v}_i \\ 0 & 1 \end{pmatrix}, \tag{6.8}$$

such that R_i and $\mathbf{v}_i$ are rotation matrices and translation vectors in $\mathbb{R}^3$, respectively. Hence, the distance between two curves $\mathcal{G}(\mathsf{M}_a)$ and $\mathcal{G}(\mathsf{M}_b)$ that are represented by $\mathcal{C}_{a,\varphi}$ and $\mathcal{C}_{b,\varphi}$, respectively, can be computed using equation (6.6), see [153], [154] for further details.

Normal Motion Model

We model a normal motion sequence using the sample mean of the training dataset. Consequently, we compute the sample mean of a curve representation dataset $\{\mathcal{C}_{1,\varphi}, \cdots, \mathcal{C}_{n,\varphi}\}$, following [153], [154], as

$$\mathcal{C}_{t,\varphi} = \underset{\mathcal{C}_{e,\varphi}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \mathfrak{D}(\mathcal{C}_{e,\varphi}, \mathcal{C}_{i,\varphi})^2, \tag{6.9}$$

where $\mathcal{C}_{t,\varphi}$ represents the estimated mean of the dataset and $\mathfrak{D}(\cdot,\cdot)$ is as defined in equation (6.6). Finally, the binary decision of classifying a motion as either normal or abnormal is made by assuming a symmetric distribution of the training data points and fixing a ζ-radius range, as defined in equation (6.2).

6.4.2 Time-Variation in Motion Analysis

Although motion variation due to speed or latency is considered as irrelevant information for tasks like action recognition, it is one of the distinctive characteristics studied in problems like motion quality assessment, since the focus is on performance variability under a predefined action. Consequently, to compare two given sequences, time-variation needs to be taken into account in both action recognition like problems and motion analysis.

In action recognition, DTW [87] is one of the widely used techniques to filter time-variation between two given motion sequences with respect to one another [21]. In principle, DTW

selects frames from both sequences such that a given cost function is minimized, which usually is the L_2-norm, refer to Chapter 2. Meanwhile, in [154] an objective function that is similar to DTW, yet flexible, is proposed and used to align curves based on deformation cost. Nevertheless, our goal is not to filter time-variation in motion but to use it for motion quality analysis. Hence, we detail an adaptation of the objective function introduced in [154] to define a distance function in the context of motion analysis.

We begin by defining the deformation-based curve alignment function introduced in [154]. Let $\bar{\varphi}$ be a function that samples uniformly spaced points as defined in equation (6.5), then a curve $\mathcal{C}_i$ is aligned to a fixed curve $\mathcal{C}_{t,\bar{\varphi}}$ by solving for

$$\varphi^* = \arg\min_{\varphi} \ \mathfrak{D}(\mathcal{C}_{t,\bar{\varphi}}, \mathcal{C}_{i,\varphi}), \tag{6.10}$$

where $\mathfrak{D}(\cdot,\cdot)$ is as defined in equation (6.6).

The solution for equation (6.10), φ^*, returns matching points which we will refer to as key-points. In other words, the distance between the two curves is minimized if $\mathcal{C}_i$ is sampled according to φ^*. Hence, the aligned curves are defined as $\mathcal{C}_{t,\bar{\varphi}}$ and $\mathcal{C}_{i,\varphi^*}$. However, we are not interested on filtering time-variation from the motion arguments, *i.e.*, minimizing the distance between two curves based on the key-points φ^*. In fact, we aim to analyze the time-variation in between key-points; in a sense, using the solution φ^* for the opposite effect than it was intended. Figure 6.2 shows an example of the distinction between the key-points and the time-variation between key-points (magnified part in Figure 6.2).

As a result, given q key-points that are identified by φ^*, we summarize the motion between two sequential key-points ℓ, j of the curve $\mathcal{C}_i$ as

$$\mathcal{F}_j(\mathcal{C}_{i,\varphi^*_{(j,\ell)}}) = \prod_{r=j}^{\ell} g_r, \tag{6.11}$$

where $j < \ell$. In contrast, the approach in [154] linearizes the deformation between the key-points (refer to Figure 6.2) while equation (6.11) attempts to preserve the observed deformation due to time-variation. Consequently, equation (6.6) is redefined to reflect the

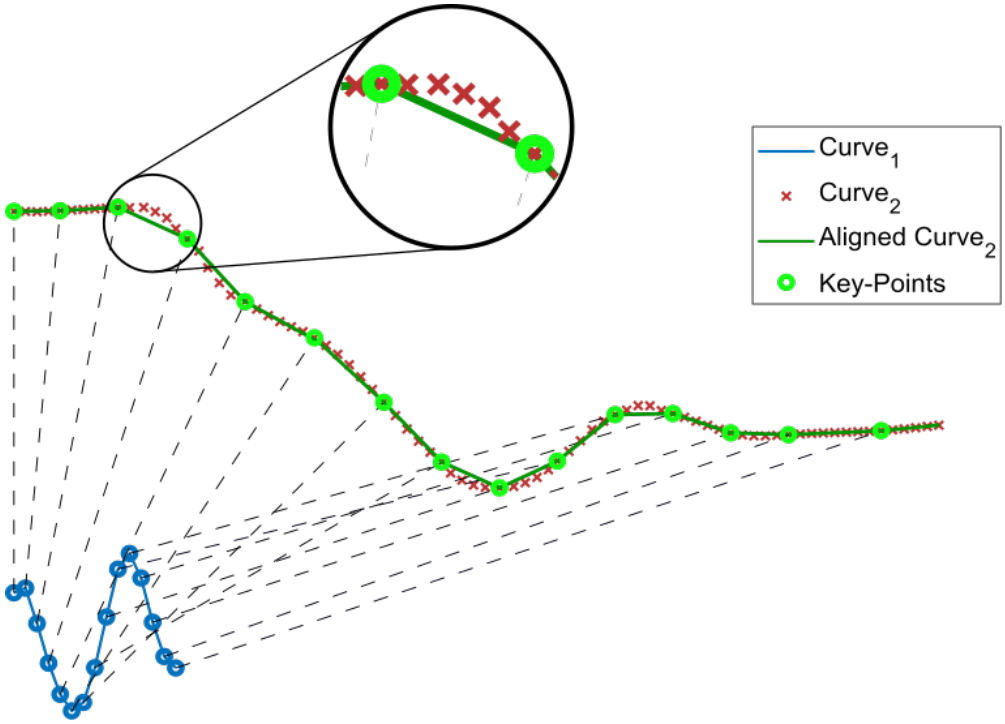


Figure 6.2: Deformation-based alignment between two curves. The blue curve is described as a set of uniformly sampled points $\bar{\varphi}$. The red curve represents the original curve to be aligned. The green line shows the linearization using the key-points φ^*. The magnified part in the figure highlights the distinction between the key-points and the time-variation between key-points.

time-variation between two curves as follows

$$\mathcal{D}(\mathcal{C}_{t,\bar{\varphi}}, \mathcal{C}_{i,\varphi^*}) = \sqrt{\sum_{j=1}^{q-1} d(\mathcal{F}_j^t, \mathcal{F}_j^i)^2}, \tag{6.12}$$

where $d(\cdot,\cdot)$ is as defined in equation (6.7). Henceforth, we will refer to equation (6.12) as motion-quality distance. For a better understanding of the proposed approach, Figure 6.3 shows an overview of the pipeline.

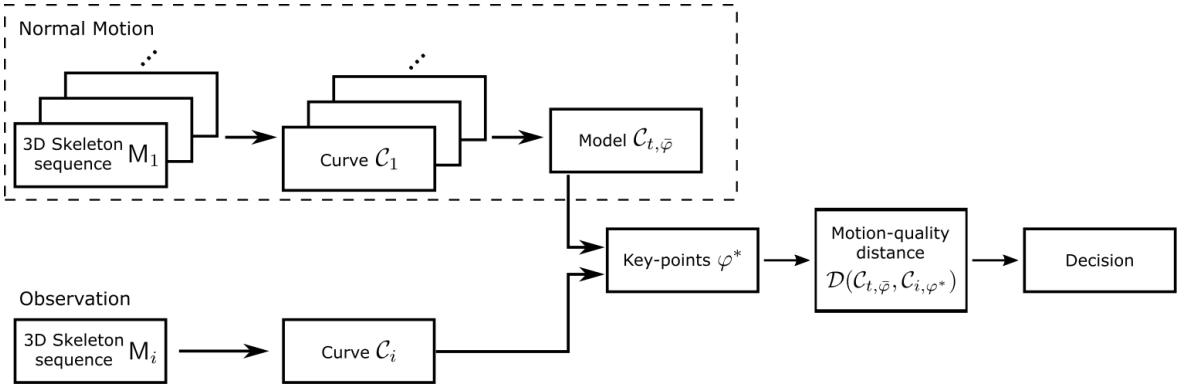


Figure 6.3: Overview of the proposed approach. In the upper part, the dashed line rectangle regards the process of achieving the normal motion model. The bottom part concerns the testing scenario, where an input 3D skeleton sequence is represented as a curve. Then, the key-points φ^* are obtained by employing a deformation-based alignment between the model and the observed curve. Consequently, the motion-quality distance is computed in order to decide if whether normal or abnormal motion.

6.5 Experiments

In this section, we present the dataset used to evaluate the proposed approach. Furthermore, we show how the skeleton data is normalized to address variations due to scale and absolute locations. Moreover, we present an analysis of the time-variation of the skeleton sequences, and finally, we show quantitative results of the abnormal motion detection system.

To evaluate the proposed approach for detecting abnormal motion using a curve-based representation from skeleton data, we use the publicly available dataset SPHERE-Staircase2014 introduced in [23]. This dataset includes 48 sequences performed by 12 subjects while walking upstairs. The sequences were captured using an RGB-D sensor placed at the top of the stairs. In the dataset, the abnormal motion events were performed by subjects under the guidance of a physiotherapist. Such a scenario is very interesting to test our approach because we are interested in home-based rehabilitation. We follow the same protocol as [23], and we use 17 sequences to build a normal motion model. The rest of the sequences were used in the testing part, where 14 sequences are considered normal motions and 17 contain abnormalities.

6.5.1 Time Variation Analysis

Generally, when comparing two temporal sequences, temporal alignment techniques are used in order to find the points that minimize a given cost function. With this in mind, we performed an analysis on the detection of the key-points, which are the points that minimize the deformation-based alignment between two curves (refer to Figure 6.2) as defined in equation (6.10). With the purpose of analyzing the impact of the time-variation between key-points, we selected a uniform sampling function $\bar{\varphi}$ for the model $\mathcal{C}_{t,\bar{\varphi}}$. Then, we tested the following scenarios for the testing curves:

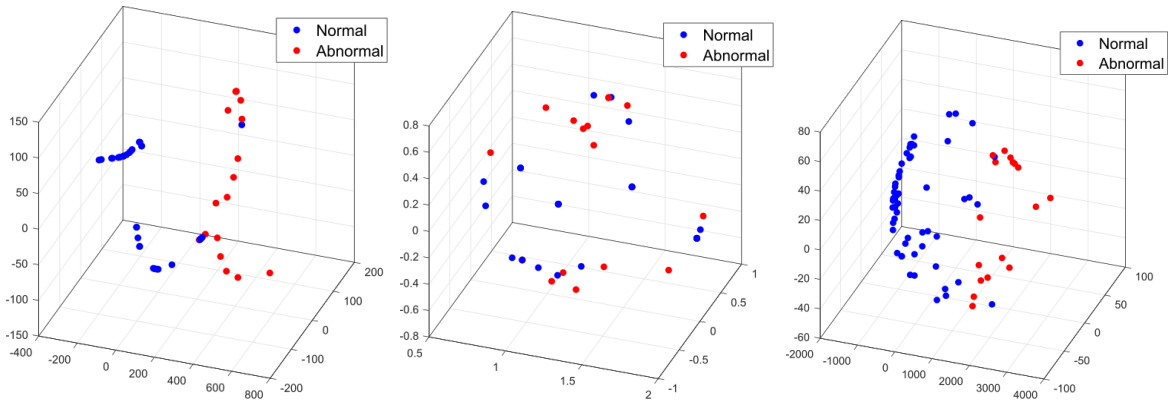
1) Uniform Sampling, US: using uniformly sampled points to compute the distance equation (6.6);

2) Optimal Sampling, OS: using the key-points estimated from equation (6.10) to compute the distance in equation (6.6);

3) Motion-Quality distance, MQ: using the key-points estimated from equation (6.10) to compute the motion-quality distance from equation (6.12).

In order to have a clear understanding, we applied a *Multi-Dimensional Scaling* (MDS) method to the data based on the corresponding cost function for each scenario. This was done to visualize the relationship between the distance functions and the ground truth labels. Figure 6.4 shows the visualization of the MDS, where Figure 6.4a refers to the US scenario, Figure 6.4b to the OS scenario, and Figure 6.4c to the proposed MQ scenario.

The conclusion that we obtain by looking at the Figure 6.4 is that for the case of the OS scenario (see Figure 6.4b), the information about time-variation is filtered. This means that while computing the distance between two curves that are sampled according to the key-points, the distance will always be the minimum possible. Consequently, the distinction between normal and abnormal motion is not that discriminative. Figure 6.4b shows that it is not possible to separate the normal from the abnormal motion for such a scenario. Hence, for our purpose, the classification using the linearization of the key-points would not be possible. Such approaches could be applied to key-pose skeleton-based detection appli-



(a) Uniform Sampling, US.

(b) Optimal Sampling, OS.

(c) Proposed Motion-Quality distance, MQ.

Figure 6.4: Multidimensional scale method applied to the data on the representation space. Each point represents a curve. Figure 6.4a shows the case of the US scenario, Figure 6.4b for the OS and Figure 6.4c using the proposed MQ scenario.

cations, *e.g.*, for action detection/recognition applications [158], [159]. On the contrary, for the US scenario (refer to Figure 6.4a), the information about time-variation is lost. Meaning that the sampling is happening independently from the model, the sampling rate is chosen a priori. In this case, and due to the small variation in the data, the classification would be considerably good. We believe that this scenario would not perform such satisfactory results with the increase of the data variation. Figure 6.4a depicts that the normal and the abnormal motion are separable from one another. In this work, we are interested in analyzing the time-variation between key-points. Unlike [154], we do not linearize the deformation between the key-points. Instead, we preserve the observed deformation due to time-variation. With such an approach, preserving the time-variation allow us to estimate the deviation of an observation from the model. Consequently, using the motion-quality distance emphasizes the time-variation between key-points, the classification of normal or abnormal motion produces encouraging results. Figure 6.4c shows that the data have similar behavior comparing to the OS scenario, but in this case, it is possible to separate the normal from the abnormal motion.

6.5.2 Abnormal Motion Detection

To demonstrate the performance of the proposed abnormal motion detection system, we compare our results with the work of [23]. In their work, they evaluate the detection per abnormal motion, while we are evaluating per sequence in our case. With this, we assume that a sequence containing at least one abnormal motion is considered as an abnormal sequence. For the classification results, we evaluated the proposed scenarios described in the previous subsection. For the normal motion model, we used 30 as the number of uniformly sampled points. For the classification of normal or abnormal motion detection, we used an empirically tested threshold $\zeta = 30.65$. Table 6.1 shows the results for the abnormal motion detection using the SPHERE-Staircase2014 dataset. Note that, as mentioned before, the results correspond with the previously described scenarios (refer to Figure 6.4). In the OS scenario, the system could not identify any kind of abnormal motion, detecting only the normal motion. Subsequently, for the other scenarios, we achieved promising results, wherein the case that we consider the motion-quality distance (MQ scenario), the results are better than the US scenario.

Methods	Accuracy (%)
Paiement *et al*. [23]	93
Uniform Sampling, US	91
Optimal Sampling, OS	50
Motion-Quality distance, MQ	**97**

Table 6.1: Results for the abnormal motion detection using the SPHERE-Staircase2014 dataset.

6.6 Conclusion

In this chapter, we presented an approach for an abnormal motion detection system. For that purpose, we proposed to represent a 3D skeleton sequence as a curve to reduce its dimensionality, yet representative. Hence, based on this curve representation, we defined a motion-quality distance, emphasizing time-variation between key-points. The proposed method showed that by highlighting time-variation, we were able to handle small variations

in the data due to the nature of the dataset (no variation in the action class, but only in the manner that an action is performed). We also presented an analysis on the time-variation for abnormal motion detection. This was done to present qualitatively the performance of the motion-quality distance, which is an important quality assessment property.

Part II

Human Motion Analysis In the Wild

Chapter 7

View-Invariant Action Recognition From RGB Data via 3D Pose Estimation

In this chapter, we propose a novel view-invariant action recognition method using a single monocular RGB camera. View-invariance remains a very challenging topic in 2D action recognition due to the lack of 3D information in RGB images. Most successful approaches make use of the concept of knowledge transfer by projecting 3D synthetic data to multiple viewpoints. Instead of relying on knowledge transfer, we propose to augment the RGB data by a third dimension by means of 3D skeleton estimation from 2D images using a CNN-based pose estimator. In order to ensure view-invariance, a pre-processing for alignment is applied followed by data expansion as a way for denoising. Finally, an LSTM architecture is used to model the temporal dependency between skeletons. The proposed network is trained to directly recognize actions from aligned 3D skeletons. The experiments performed on the Northwestern-UCLA dataset show the superiority of our approach as compared to state-of-the-art ones.

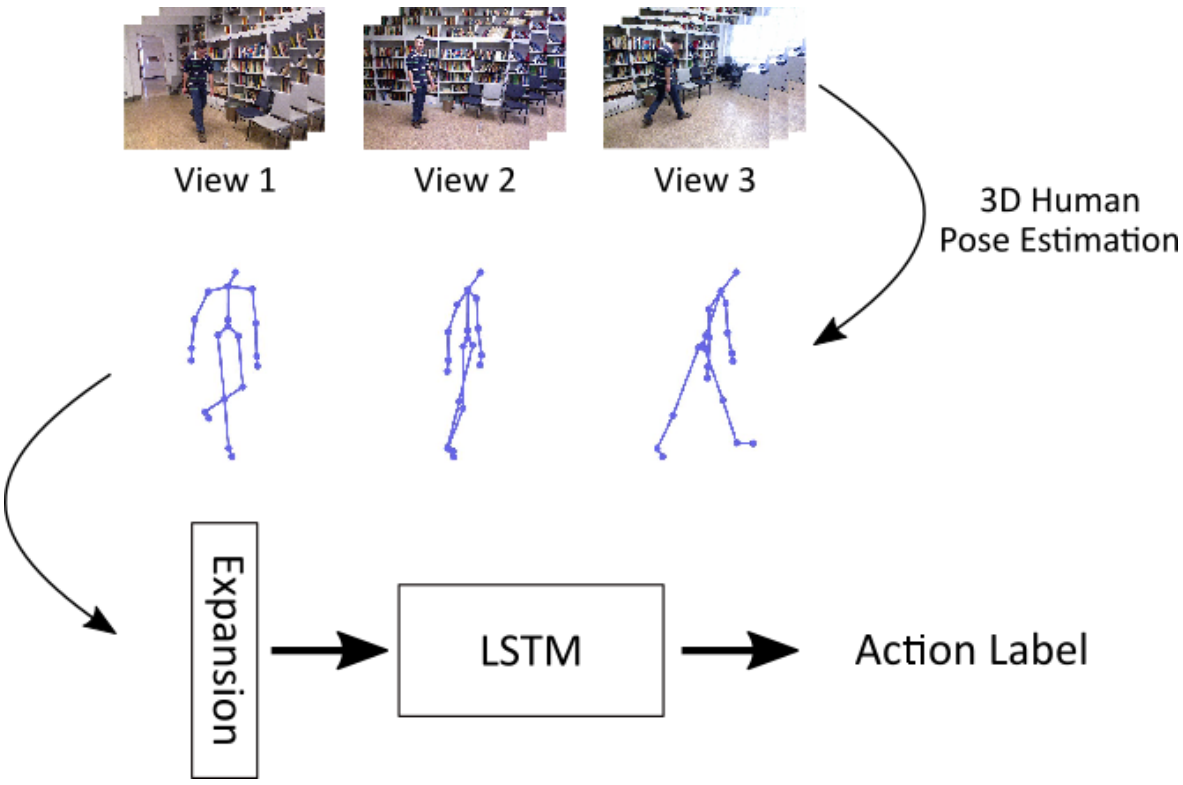


Figure 7.1: Overview of the proposed approach.

7.1 Introduction

Data acquisition with a particular camera setup depends not only on the observed scene but on the camera configuration as well. The setup leads to pixel-level data variations that are unrelated to the observed scene and subsequently poses a significant challenge in pattern recognition [160].

We distinguish two different ways of addressing the issue of viewpoint variability using skeletons provided by RGB-D sensors. The first class of methods carries out a preprocessing of alignment by estimating a transformation matrix between the skeleton and a canonical coordinate system as in [141], [161], [162]. On the other hand, the second class of approaches aims to design motion descriptors which are not affected by viewpoint variability such as: *eigen joints* [163] based on the pairwise distance between joints, *Lie Algebra Representation of body-Parts (LARP*) [21] based on transformation matrices estimated between

body-part pairs, etc. These approaches have shown great potential [57].

In this chapter, we propose a skeleton-based approach for view-invariant action recognition using a monocular camera. Our work builds on the recent effective CNN-based methods for the estimation of 3D skeletons from a single RGB image [33], [79], [93]. We propose a full system, as illustrated in Figure 7.1, where 3D skeletons are firstly extracted from RGB images using a CNN-based estimator. To achieve view-invariance, an effective pose-based data alignment is carried out. Then, the temporal dependency of the estimated pose sequences is modelled using an LSTM network. To address potentially noisy pose estimates, we train a feed-forward network together with the LSTM in an end-to-end training scheme. The feed-forward network is designed to expand the pose estimates to a higher dimensional space and potentially decouple several explanatory factors before modelling the temporal-dependency. To evaluate and validate the proposed system, experiments on the Northwestern-UCLA dataset are realized. The obtained results show that our method outperforms RGB-based state-of-the art approaches on the same dataset.

7.2 Proposed Approach

In this section, we describe the two main components of the proposed approach: 3D pose estimation from RGB and data alignment, and pose sequence modelling.

7.2.1 3D Pose Estimation and Data Alignment

Recent development in deep learning has enabled hierarchical systems to learn powerful filters from data [164]. Furthermore, filters that are learned from large datasets can effectively be used for problems with insufficient data, using what is called transfer learning. The *VNect* approach proposed in [33] is one of the systems that use transfer learning for effective 3D pose estimation directly from RGB images. VNect is based on a CNN pose regression that allows the real-time estimation of 2D and 3D skeletons using a single RGB image. For each estimated human joint, the network is trained to estimate a 2D confidence heatmap along with locations maps (for each of the three dimensions).

One of the main advantages of estimating a 3D pose is the ability to estimate the positions of corresponding 3D points in different viewpoints. In which case, 3D pose alignment can be estimated with a closed-form solution. To further explain, let $\mathbf{s}_1$ and $\mathbf{s}_2$ be the vector representation of estimates of the 3D skeletons $\tilde{\mathsf{S}}_1$ and $\tilde{\mathsf{S}}_2$ from two different viewpoints. Assuming the mean of the estimated pose is centered, the alignment of the estimated pose is performed by estimating the rotation R through the following optimization:

$$\arg\min_{\mathsf{R}} ||\mathbf{s}_1 - \mathsf{R}\mathbf{s}_2||_2^2. \tag{7.1}$$

The formulation (7.1) has a closed-form solution given as

$$\widetilde{\mathsf{R}} = \mathsf{V}\mathsf{U}^{\mathsf{T}}, \tag{7.2}$$

where $\mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}} = \mathbf{s}_1\mathbf{s}_2^{\mathsf{T}}$, with U and V being unitary matrices and Σ a diagonal matrix corresponding to the *Singular Value Decomposition* (SVD) of $\mathbf{s}_1\mathbf{s}_2^{\mathsf{T}}$. The matrix $\widetilde{\mathsf{R}}$ denotes the estimated rotation matrix. Given two sequences of T poses estimated from two different viewpoints $\tilde{\mathsf{M}}_1 = \{\mathbf{s}_1^1, \cdots, \mathbf{s}_1^T\}$ and $\tilde{\mathsf{M}}_2 = \{\mathbf{s}_2^1, \cdots, \mathbf{s}_2^T\}$, we estimate the alignment between the first corresponding poses $\mathbf{s}_1^1$ and $\mathbf{s}_2^1$ using equation (7.2). Afterwards, the estimated rotation matrix $\widetilde{\mathsf{R}}$ is used to align the rest of the subsequent poses of the sequence.

7.2.2 Pose Sequence Modelling

In general, 3D pose estimation from RGB data can be noisy depending on the estimation model and the available training dataset. We propose an LSTM-based temporal model that is suitable for estimating the temporal dependency between noisy skeletal pose estimates. Our approach has two main components: (1) a feed-forward network for expanding the data to a high-dimensional space, and (2) multi-layer LSTM units for modelling the temporal dependency, see Figure 7.2.

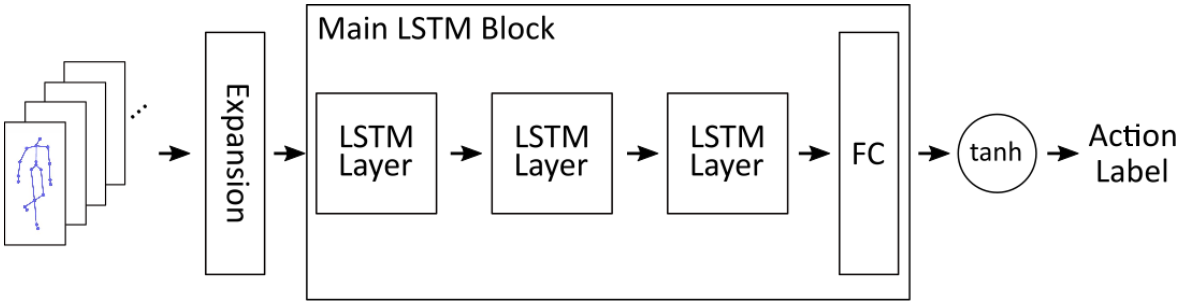


Figure 7.2: Proposed network for view-invariant action recognition. FC refers to the fully connected layer at the end of the main LSTM block.

Data expansion

An estimated 3D skeleton with N number of joints is a vector in $\mathbb{R}^{3N}$. Hence, a noisy joint estimate is directly reflected on some of the dimensions of the observed vector. One typical solution for removing noise and redundancy is to contract the data to a lower dimensional space [165]. On the contrary, in this chapter, we expand the data to a higher dimensional space. The main motivation for expanding the data is to disentangle explanatory factors that are obscured by noisy joint estimates. Consequently, the parameters of the expansion function are learned directly from the training dataset. Expansion of an observed skeleton is defined as follows

$$\tilde{\mathbf{s}} = \tanh(\mathrm{W}\mathbf{s} + \mathbf{b}), \tag{7.3}$$

where W is a $k \times 3N$ matrix with $k \gg 3N$, $\mathbf{b}$ is a bias vector in a k-dimensional space, and the $\tilde{\mathrm{s}}$ denotes the expanded pose estimate.

Temporal Model and Action Labeling

The temporal dependency between the sequential data points is modelled using layers of LSTM units [166]. An LSTM is a gated recurrent neural network that models temporal dependency as a stationary process. Although it has several components, we herein will refer to the integrated computational unit as LSTM. Subsequently, given an expanded input

data $\tilde{\mathrm{s}}$, we estimate hierarchical latent variables by layering LSTM units one on top of another, see Figure 7.2. Consequently, the inferred latent space from the i^{th} pose estimate is given as

$$h_i^L = \mathsf{LSTM}(\tilde{\mathbf{s}}_i), \tag{7.4}$$

where L denotes the index of the last LSTM layer. Finally, an action label from a set $\mathcal{Y}$, is assigned to a sequence as

$$\tilde{y} = \underset{y \in \mathcal{Y}}{\arg\max}(\tanh(\mathsf{W}h_T^L + \mathbf{b})), \tag{7.5}$$

where T is the index of the last pose estimate. The connection weights and biases of the overall network (temporal model and data expansion) are trained together by minimizing the cross-entropy between the predicted and the given probability of an action label via back-propagation and back-propagation through time [164].

7.3 Experiments

In this section, the experimental setup is presented along with the obtained results. For the evaluation of the proposed approach, our experiments are conducted on the Northwestern-UCLA Multiview Action3D dataset [167] denoted as *NW-UCLA*.

7.3.1 NW-UCLA Dataset

NW-UCLA dataset is a challenging RGB-D based datasets in multi-view action recognition. It consists of 1494 videos of 10 action classes (*pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry*) performed by 10 subjects. Each action can be repeated from 1 to 6 times per subject. These actions are captured simultaneously from 3 different viewpoints and both RGB and depth modalities are provided along with the corresponding estimated 3D skeleton sequences. For our experiments, we follow the splitting protocol suggested in [167], where two viewpoints are used for

training and the third one for testing.

7.3.2 Experimental Setup and Implementation Details

For the estimation of 3D skeleton sequences from RGB videos, we used the pre-trained VNect model[1], which provides a 3D skeleton estimate with 20 joints per frame. In addition, the skeleton sequences were temporally aligned using zero padding in an automated way.

In our experiments, we set the batch size to 2 considering the small size of NW-UCLA dataset. Moreover, using cross-validation, the optimal learning rate is set to 0.0002 and the number of epochs is chosen to be 300. The implementation of our architecture is based on PyTorch[2] using 128 hidden units per layer.

7.3.3 Experimental Results

To validate the effectiveness of the proposed RGB-based view-invariant model, we consider three scenarios: (1) Evaluation with and without expansion unit; (2) Comparison of VNect poses against RGB-D provided skeletons; and (3) Comparison against state-of-the-art.

Evaluation With and Without Expansion Unit

The expansion unit is the first layer of our proposed network, as shown in Figure 7.1 and Figure 7.2. This unit is mostly responsible for removing noise and redundancy from the input skeleton sequences. We run the experiments using both cases and the results are presented in Table 7.1. In this case, we used the provided RGB-D skeleton data and the viewpoints 1 and 2 for training and the viewpoint 3 for testing. Our proposed approach with the incorporation of the expansion module achieves 83.4% accuracy. The latter is 3.5% higher than the one reported without the utilization of this specific module mostly because of the dimensionality expansion. In this case, the abstraction of the data description increases. Thus, noisy joint estimates have lower contribution to the representation.

[1] `http://gvv.mpi-inf.mpg.de/projects/VNect/`
[2] `https://pytorch.org/`

Method	Accuracy
No expansion + LSTM	79.9
Expansion + LSTM	83.4

Table 7.1: Accuracy of recognition (%) on the NW-UCLA dataset considering the cases where the expansion module is present and not present. The results are obtained using viewpoints 1 and 2 for training and viewpoint 3 for testing.

Comparison of VNect Poses Against RGB-D Estimated Skeletons

To evaluate the reliability of VNECT poses, we conduct experiments using the provided RGB-D skeleton sequences as input to the proposed network. In this scenario, we also use the viewpoints 1 and 2 for training and viewpoint 3 for testing. The reported accuracy in Table 7.2 using the provided RGB-D skeletons is 83.4% which is 3.8% lower than the accuracy achieved with the use VNect-provided poses. Although VNect generates 3D skeleton data from RGB data, it shows robustness to partial self-occlusions compared to RGB-D sensors.

Method	Accuracy
Expansion + LSTM	83.4
VNect + Expansion + LSTM (VE-LSTM)	87.2

Table 7.2: Accuracy of recognition (%) on the NW-UCLA dataset using the provided RGB-D skeletons and the estimated skeletons from VNect. The results are obtained using viewpoints 1 and 2 for training and viewpoint 3 for testing.

Comparison Against State-of-the-art Approaches

In Table 7.3, the obtained results of our approach are presented and compared against some state-of-the-art approaches. Our network performance outperforms RGB-based approaches by more than 10%. Indeed, our approach reaches 79.9% of accuracy on NW-UCLA dataset against 69.4% using NKTM [56].

7.4 Conclusion

In this chapter, we proposed a novel view-invariant action recognition approach using a single RGB camera. This is achieved by using a 3D human pose estimator from RGB

{Source} | {Target}	{1,2}|3	{1,3}|2	{2,3}|1	Mean
Hankelets [168]	45.2	-	-	-
DVV [169]	58.5	55.2	39.3	51.0
CVP [170]	60.6	55.8	39.5	52.0
AOG [167]	73.3	-	-	-
nCTE [55]	68.8	68.3	52.1	63.0
NKTM [56]	75.8	73.3	59.1	69.4
R-NKTM [171]	78.1	-	-	-
VNect + LARP [57]	70.0	70.5	52.9	64.5
DLVIF [172]	-	-	-	77.2
VNect + KSC [57]	86.2	79.7	66.5	77.5
DeepVI (SmoothNet+DA+ST-GCN) [19]	-	-	-	78.3
VE-LSTM (ours)	**87.2**	**82.1**	**70.4**	**79.9**

Table 7.3: Accuracy of recognition (%) on the NW-UCLA dataset. The reported results are obtained using two viewpoints for training and the remaining one for testing. *Source* indicates the viewpoints used for the training step, while *Target* specifies the testing viewpoint.

images. The estimated 3D poses are used for computing a view-alignment rotation between observations. Subsequently, an LSTM-based network is proposed in order to estimate the temporal dependency between noisy skeleton pose estimates. Experimental results show the superiority of our approach when compared to existing methods. Also, the 3D skeleton estimates using VNect show higher accuracy compared to the ones provided by Kinect, showing robustness to possible occlusions that may appear on the RGB images.

3D human pose estimation on a per-frame basis shows temporal inconsistency and small fluctuations in the skeleton joint locations over time. Considering this, in the next chapter, we propose to explore a different 3D human pose estimation methodology based on a sequence-to-sequence skeleton estimation.

Chapter 8

Temporal 3D Human Pose Estimation for Action Recognition from Arbitrary and Challenging Viewpoints

In line with the previous chapter, we present a new cross-view action recognition system that is able to classify human actions by using a single RGB camera, including challenging camera viewpoints. Understanding actions from different viewpoints remains an extremely challenging problem, due to depth ambiguities, occlusion, and large variety of appearances and scenes. Moreover, using only the information from the 2D perspective gives different interpretations for the same action seen from different viewpoints. Our system operates in two subsequent stages. The first stage estimates the 2D human pose using a convolution neural network. In the next stage, the 2D human poses are lifted to 3D human poses, using temporal convolution neural network that enforces the temporal coherence over the estimated 3D poses. The estimated 3D poses from different viewpoints are then aligned to the same camera reference frame. Finally we propose to use a temporal convolution network based classifier for cross-view action recognition. The results show that we can achieve state of art view-invariant action recognition accuracy even for the challenging viewpoints by only using RGB videos, without pre-training on synthetic or *Motion Capture* (MoCap) data.

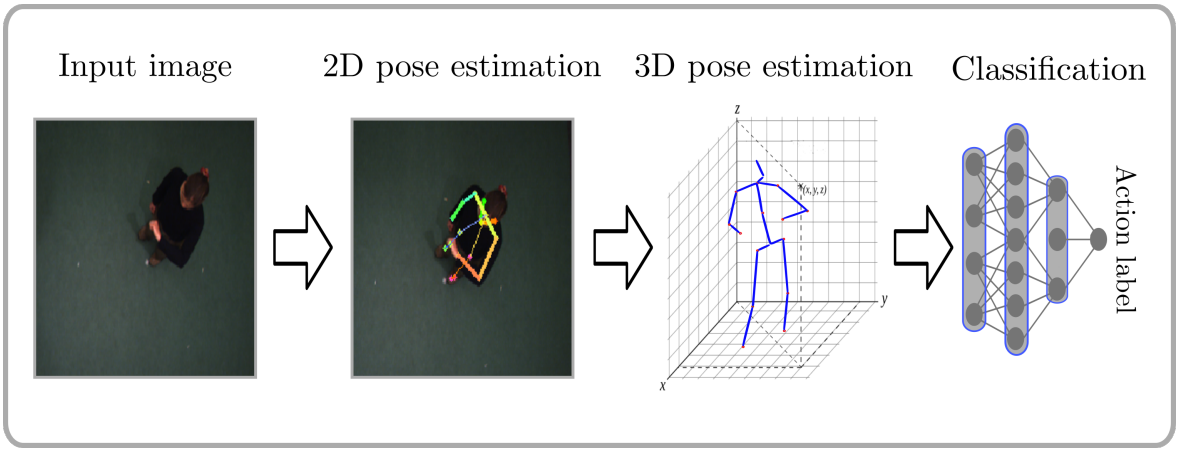


Figure 8.1: High level overview of the proposed view-invariant action recognition system using only RGB images.

8.1 Introduction

Most of the current action recognition methods assume that the subject performing the action is facing the camera [58], [59], [140], [173]. Subsequently, the performance of these methods drops while facing real-world scenarios, where camera positioning as well as the human body orientation may vary. With the introduction of RGB-D sensors, some works tackled the subject of view-invariant action recognition by directly using the 3D information provided in real-time by the depth sensors [111]. However, the usage of such sensors is not recommended in real-world scenarios due to two main limitations. First, the estimation of 3D skeletons is limited to a specific range. Second, these sensors are highly affected by lighting conditions. To overcome these limitations, recent works proposed to use only RGB information to achieve view-invariant action recognition [57], [61]. However, in scenarios where the camera is placed in challenging locations, *e.g.*, street surveillance where cameras are usually placed on top of buildings, understanding of actions is not always possible, even for humans.

In this chapter, we focus on the case of view-invariant action recognition from challenging camera viewpoints by only using RGB information. This is possible due to the recent advances of deep neural networks in estimating the 3D human pose while using a single RGB

camera [33], [36], [92]. Martinez *et al*. [92] showed that decoupling the 3D pose estimation from the 2D joint locations estimation followed by "lifting" to 3D space, gives lower error rate compared to end-to-end methods. Thus, in this work we propose a 3D skeleton-based approach for view-invariant action recognition where we first estimate the 2D locations of the human joints and then we lift the human joint locations from the 2D space to the 3D space. This is done by only using the RGB images and without any MoCap or synthetic data. Figure 8.1 shows a high level overview of the the proposed view-invariant action recognition system.

We use the TCN approach [36] in order to lift from the 2D space to the 3D space where the temporal information of the skeleton is considered. Hence, adding the temporal coherence to the estimated 3D skeletons which in turns reduces the noisy joints estimates as compared to the per-frame 3D pose estimation methods [33]. In addition, and in line with the temporal coherence of the 3D skeleton estimates, we propose to use the TCN approach at the level of the action classification task. To evaluate the proposed system, experiments on the INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [39] are conducted.

8.2 Problem Definition

In this section, we formulate the problem of view-invariant action recognition.

Let $\mathsf{V}_p = \{\mathsf{I}_{p,1}, \cdots, \mathsf{I}_{p,T}\}$ and $\mathsf{V}_q = \{\mathsf{I}_{q,1}, \cdots, \mathsf{I}_{q,T}\}$ be two sequences of RGB images corresponding to the same action label but with different and arbitrary camera viewpoints p and q capturing the same scene, where T is the total number of frames. Subsequently, the goal of this work is to estimate a function $\mathcal{V}(\cdot)$ such that we achieve view-invariant action recognition,

$$\mathcal{V}(\mathsf{V}_p) = \mathcal{V}(\mathsf{V}_q) = y, \tag{8.1}$$

where y corresponds to the action label. The objective of the function $\mathcal{V}(\cdot)$ is to map a

sequence of RGB images V_p to its corresponding label y,

$$\begin{aligned} \mathcal{V} : \mathbb{R}^{M \times T} &\mapsto \mathcal{Y} = \{1, \cdots, \ell\}, \\ \mathsf{V}_p &\mapsto y, \end{aligned} \tag{8.2}$$

where M is the image dimension, and ℓ is the total number of action labels. Considering two arbitrary camera viewpoints V_p and V_q with $p \neq q$, $\mathcal{V}(\cdot)$ is considered to be view-invariant if and only if equation (8.1) is verified.

In order to estimate $\mathcal{V}(\cdot)$, we propose a two-step based approach. First, we estimate the human body joint locations in the image plane (2D skeleton), and secondly, we lift the human joint locations from the image plane to the corresponding 3D human joints position (3D skeleton). The estimated 3D skeletons are then used to model the human motion. In fact, by lifting the action space from 2D to 3D, we obtain an action representation in 3D space that enables us to have a better understanding of the human motion and behavior. Hence, it provides better features to design a view-invariant action recognition system as compared to directly working in the 2D space.

8.3 Proposed Approach

In this section, we describe the main components of the proposed two-step 3D skeleton based view-invariant action recognition system.

8.3.1 2D Human Pose Estimation

Given an RGB image I, the goal is to map the information related to the human body present in the image I to the corresponding 2D locations of the human joints. In other words, we want to extract the 2D skeleton by applying the mapping function $g(\cdot)$ presented in Chapter 2 to the image I, such that

$$\tilde{\mathsf{S}}_{2D} = g(\mathsf{I}) \quad \text{with } \tilde{\mathsf{S}}_{2D} \in \mathbb{R}^{2N}, \tag{8.3}$$

where $\tilde{\mathsf{S}}_{2D}$ represents the estimated 2D skeleton, with N joints. In order to estimate the 2D skeleton from an RGB image, we use the state-of-art approach *AlphaPose* [174] approach as function $g(\cdot)$. Then, for a given video sequence V_p, the function $g(\cdot)$ is applied frame by frame resulting in a sequence of 2D skeletons $\mathsf{M}_{p,2D}$,

$$\mathsf{M}_{p,2D} = \{g(\mathsf{I}_{p,1}), \cdots, g(\mathsf{I}_{p,T})\}. \tag{8.4}$$

To infer the human action in a video, we first build a 2D model for the human body in every image of the video. In order to improve the results obtained with AlphaPose for the estimation of the 2D skeleton, the human body has to be detected accurately in the image plane. For this task, the pre-trained *You Only Look Once* (YOLO) object detection network [175] is used. This is done in order to obtain the bounding box containing the human subject. Consequently, this information is provided along with the image to the AlphaPose [174] resulting in the estimation of the 2D human joints in the respective image.

8.3.2 3D Human Pose Estimation and Data Alignment

For 3D pose estimation we build on the state-of-art approaches that formulate the problem as a 2D pose estimation followed by lifting to the 3D space [36], [92]. In such approaches, the low dimensional representation like 2D pose, represented by a set of joints, can be discriminative enough to estimate the 3D pose with high accuracy [176]. Such decoupling of the problem reduces the difficulty of the task at hand, and gives the possibility of human supervision on the estimated 2D poses, prior to 3D pose estimation. These two stages-approaches have been proven to be more accurate than end-to-end approaches [33]–[35].

However, estimating 3D pose from individual frames leads to temporally incoherent estimation, where independent error from each frame leads to unstable 3D pose estimation over the video sequence. Thus, we follow the same approach proposed by Pavllo *et al*. [36] where they use a fully convolutional architecture that performs temporal convolutions over 2D joints in order to estimate the 3D skeleton in a video. Given a sequence of 2D skeletons from an arbitrary camera viewpoint p, $\mathsf{M}_{p,2D} = \{\tilde{\mathsf{S}}^1_{p,2D}, \cdots, \tilde{\mathsf{S}}^T_{p,2D}\}$, the goal is to lift the 2D

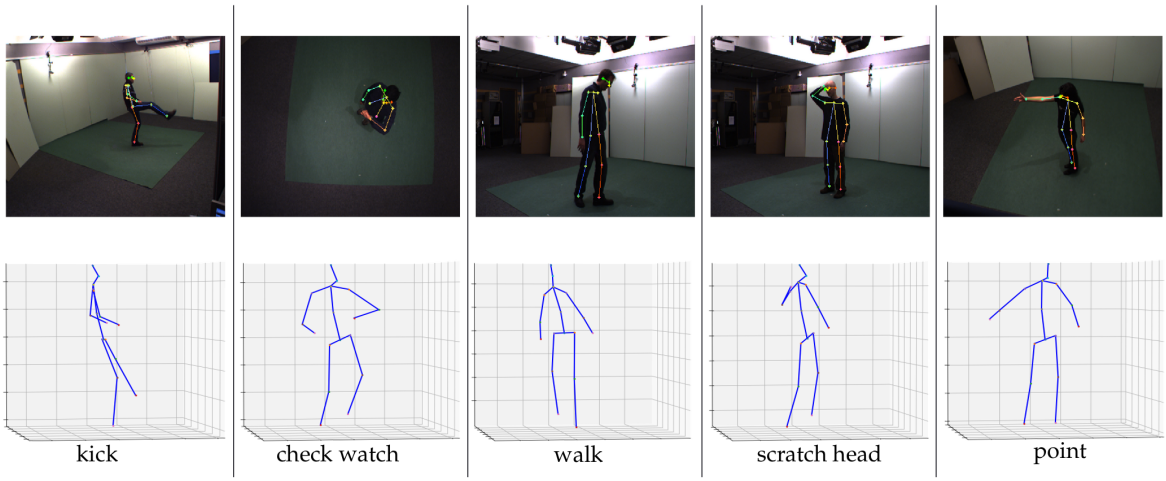


Figure 8.2: Human pose estimation of different actions acquired from different camera viewpoints. First row illustrates the 2D skeleton estimates along with the RGB images, while in the bottom row, the corresponding estimated 3D skeletons are shown.

skeleton sequence to the 3D space. To that end, we need to estimate a function $h(\cdot)$, which maps a 2D skeleton sequence to its corresponding 3D skeleton sequence, such that

$$\mathsf{M}_{p,3D} = h(\mathsf{M}_{p,2D}). \tag{8.5}$$

The function $h(\cdot)$ may be estimated using the work in [36]. Consequently, the 3D skeleton sequence can be obtained from an arbitrary viewpoint video sequence V_p by using the combination of the functions $g(\cdot)$ and $h(\cdot)$, such that

$$\mathsf{M}_{p,3D} = h(\,\{g(\mathsf{I}_{p,1}), \cdots, g(\mathsf{I}_{p,T})\}\,). \tag{8.6}$$

Examples of the 3D skeletons estimates from different camera viewpoints for different actions are shown in Figure 8.2.

Due to the variation of how every subject in the dataset preforms the action and where they are in the room, the estimated 3D skeletons of the same action can be oriented differently from one subject to another. As our focus is to infer the action, we hence normalize the estimated 3D skeletons of the same action to relate them with the same reference frame.

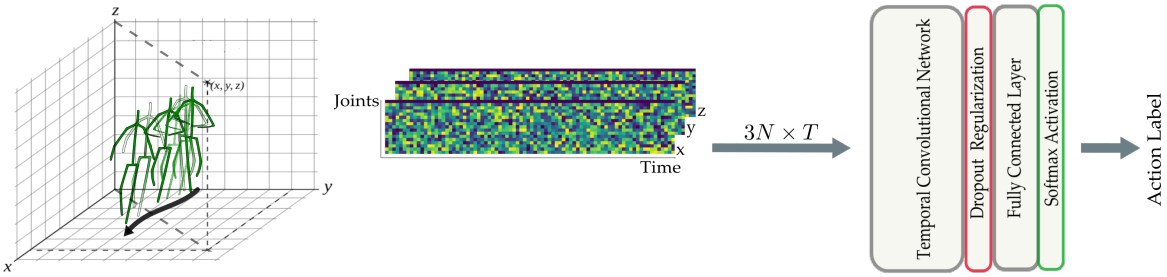


Figure 8.3: Proposed temporal classification model for view-invariant action recognition. After estimating the 3D skeletons, we use the temporal 3D joints information as input to the deep neural network. Such network consists of a TCN model as temporal feature extractor, followed by a fully connected layer with softmax activation for the multi-class classification.

To that end, we follow the same normalization process proposed in [141] to eliminate the anthropometric variability. This is done by considering the first 3D skeleton of the sequence as the rest state where the subject does not preform any action, denoted as reference pose.

8.3.3 Temporal Modeling for 3D poses

Different methods have been introduced to model the temporal evolution using deep neural networks [176], [177]. More recently Pavllo *et al*. [36] presented a fully convolutional architecture that performs temporal convolutions over the 2D human joints in order to estimate 3D poses from videos. The main idea behind is the usage of TCN, where dilated convolutions are applied along the temporal axis of the action represented by the 2D skeletons, and hence producing the desired 3D skeletons estimates. Considering the same concept adopted for the 3D skeleton estimation, we propose to use the TCN based model presented in [178] to learn the temporal features directly from the 3D skeletons and predicting the performed action. The 3D skeletons provide information about the behavior of the human subject and the different body movements over time. This is further represented by N joints locations for each 3D skeleton over time as shown on the left side of the Figure 8.3.

TCN is a variation of convolutional neural network for sequence modelling tasks. Compared to traditional *Recurrent Neural Networks* (RNNs), TCN offers more direct high-bandwidth access to past and future information. This allows TCN to be more efficient to model the

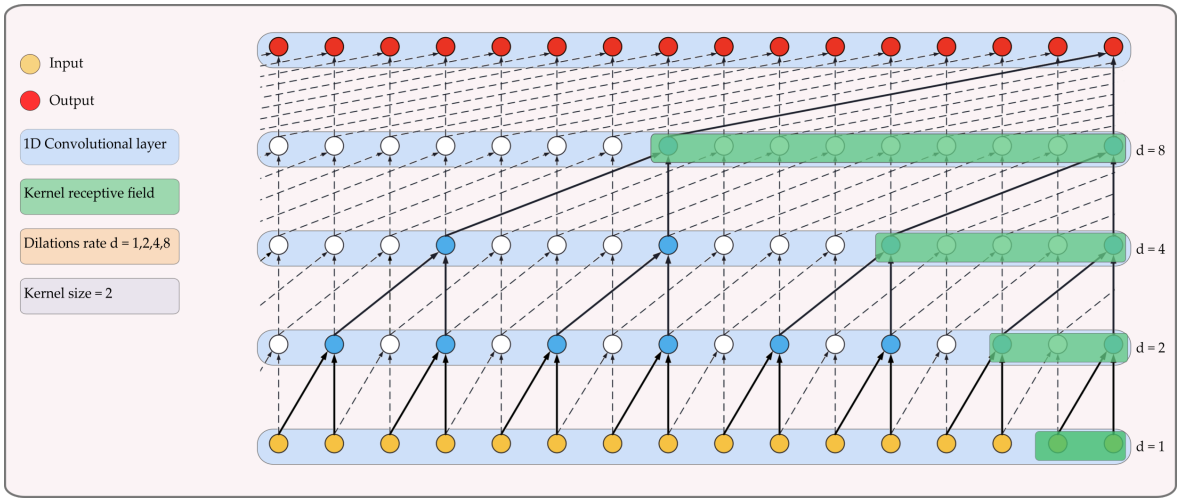


Figure 8.4: Example of the TCN model with kernel size $k = 2$ and dilation rate $d = [1, 2, 4, 8]$.

temporal information of the input data with fixed size [176]. TCN can be causal; meaning that there is no information *leakage* from future to past, or non-causal where past and future information is considered. The main critical component of the TCN is the dilated convolution [179] layer, which allows to properly treat temporal order and handle long-term dependencies without an explosion in model complexity. Figure 8.4 illustrates a TCN with different dilation factors. For simple convolution, the size of the receptive field of each unit – block of input which can influence its activation – can only grow linearly with the number of layers. In the dilated convolution, the dilation factor d increases exponentially at each layer. Therefore, even though the number of parameters grows only linearly with the number of layers, the effective receptive field of units grows exponentially with the layer depth. The dilated convolution of a 1D signal s with a kernel of size k and dilation factor d is defined as:

$$(k *_d s)_t = \sum_{\tau=-\infty}^{\infty} k_\tau \cdot s_{t-d\tau}.$$

Convolutional models enable parallelization over both the batch and the time dimension while RNNs cannot be parallelized over time [178]. Also the path of the gradient between output and input has a fixed length regardless of the sequence length. Thus, mitigating the vanishing and exploding gradients which has a direct impact on the performance of

the RNNs [178]. Architectures with dilated convolutions have been successfully used for audio generation in *Wavnet* network [180], semantic segmentation [181] , machine translation [182], and 3D pose estimation [36]. As stated in [178], TCNs generally outperform most of the commonly used networks such as LSTM [166] or *Gated Recurrent Unit* (GRU) [183] for different tasks.

8.4 Experimental Results

In this section, we present the experimental setup along with the obtained results. In order to evaluate the proposed approach, we conducted experiments on the INRIA Xmas Motion Acquisition Sequences (IXMAS) multi-view dataset [39].

8.4.1 IXMAS Dataset

IXMAS dataset is dedicated for the task of multi-view action recognition. This dataset is captured using 5 synchronized RGB cameras that are placed in 5 different locations. Such locations include four cameras placed on the side and one camera placed on top of the subject. IXMAS dataset consists of 11 different actions performed 3 times by 11 actors. The action categories are: *check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick and pick up*. This is a challenging dataset due to the fact that it contains an extreme camera location, where it is placed on top of the subject, which leads to self-occlusions. Examples of the camera viewpoints presented in this dataset are shown in Figure 8.2 top row.

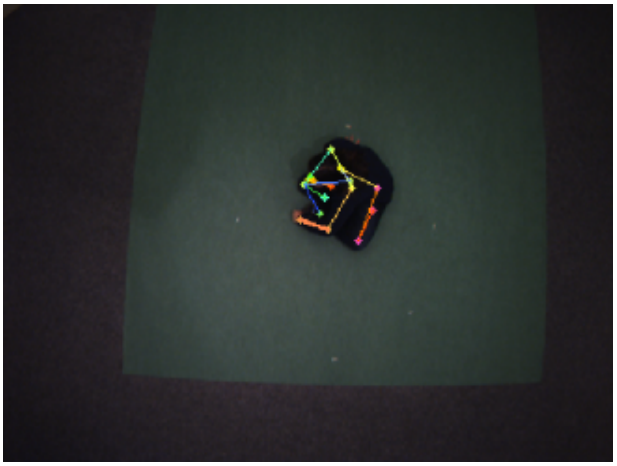
8.4.2 Implementation Details

In order to detect the human subject present in the image, we follow the same steps introduced in the AlphaPose approach [174]. We use the pretrained *YOLOv3 SPP* [175] object detection network, with spatial pyramid pooling to pool and concatenate the multi-scale local region features. Thus, the network can learn the object features more comprehensively. The

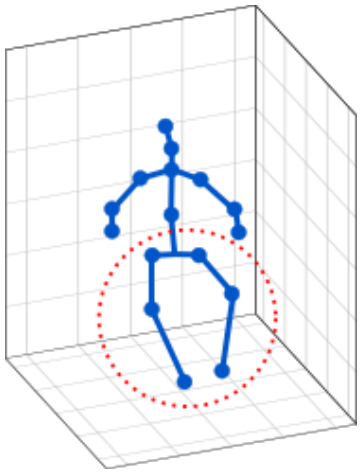
output of the object detection network is a set of bounding boxes around the human subjects (output from the YOLO network). In order to guarantee that the entire person region will be extracted, the detected bounding boxes containing the human proposals are extended by 30% along both the height and width directions. Every detected human proposal is then cropped and passed to the human pose estimator. We use the *Cascaded Pyramid Network* (CPN) [184] 2D pose estimator to estimate all the 2D skeleton sequences. The estimated 2D skeleton sequences are further passed through a TCN network in order to obtain the corresponding 3D skeleton sequences. By using a TCN network we preserve the temporal coherence present in the 2D sequences which leads, in turn, to improving the quality of the estimated 3D skeletons. Lastly, the 3D skeleton information is then used for the action classification part, where again a TCN network is used as shown in Figure 8.3. The following parameters are considered: kernel size = 6; dilation's rate = [1, 2, 4, 8, 16, 32]; number of stacks of residual blocks = 2; *Adaptive Moment Estimation* (ADAM) optimizer with learning rate = 0.001; and epochs = 50.

8.4.3 Results and Discussion

Overall, the obtained 3D skeleton estimates from an RGB video sequence resemble the human motion and its structure. Furthermore, the 3D skeleton estimates from different camera viewpoints are very close to each other, which is a desired behavior for the action recognition task. Figure 8.2 illustrates the 3D skeleton estimates from the IXMAS dataset (bottom row). However, in some cases, the lower body parts (legs, knees and ankles) of the 3D skeletons were not correctly estimated. Note that, most of the cases happened when considering the extreme camera viewpoint where the camera is located on top of the subject. This is quite challenging due to self-occlusions that may happen. Even for the human perception it is hard to understand how the lower body part of the subject present in the image behaves. Figure 8.5 illustrates a case where the 3D skeleton estimate is not correct. It is clear that the lower body part of the 3D skeleton, Figure 8.5b, does not correspond to the subject's pose seen in the RGB image, Figure 8.5a. Even though wrong 3D skeleton estimates may happen, yet and in general, the 3D skeleton estimates are good enough for conducting the



(a) 2D skeleton estimate.



(b) 3D skeleton estimate.

Figure 8.5: Example of an erroneous 3D skeleton estimate from the 2D skeleton. The red circle highlights the wrong estimate of the legs. While in Figure 8.5a the subject is sitting on the floor, in Figure 8.5b shows that the subject is standing.

{Source} | {Target}	0| 1	0| 2	0| 3	0| 4	1| 0	1| 2	1| 3	1| 4	2| 0	2| 1	2| 3	2| 4	3| 0	3| 1	3| 2	3| 4	4| 0	4| 1	4| 2	4| 3	Avg.
Hankelets [168]	83.7	59.2	57.4	33.6	84.3	61.6	62.8	26.9	62.5	65.2	72.0	60.1	57.1	61.5	71.0	31.2	39.6	32.8	68.1	37.4	56.4
DVV [169]	72.4	13.3	53.0	28.8	64.9	27.9	53.6	21.8	36.4	40.6	41.8	37.3	58.2	58.5	24.2	22.4	30.6	24.9	27.9	24.6	38.15
CVP [170]	78.5	19.5	60.4	33.4	67.9	29.8	55.5	27.0	41.0	44.9	47.0	41.0	64.3	62.2	24.3	26.1	34.9	28.2	29.8	27.6	42.16
nCTE [55]	**94.8**	69.1	**83.9**	39.1	90.6	79.7	79.1	30.6	72.1	**86.1**	77.3	62.7	82.4	79.7	70.9	37.9	48.8	40.9	70.3	49.4	67.2
NKTM [56]	92.7	**84.2**	**83.9**	44.2	**95.5**	77.6	**86.1**	40.9	82.4	79.4	85.8	**71.5**	82.4	**80.9**	82.7	44.2	57.1	48.5	**78.8**	51.2	72.5
VNect+LARP [57]	46.6	42.1	53.9	9.7	50.6	37.5	47.3	10.0	43.4	33.0	53.6	11.8	51.2	37.8	53.6	9.1	10.9	8.7	10.9	7.9	31.48
VNect+KSC [57]	86.7	80.6	82.4	15.5	91.5	79.4	81.8	15.8	**85.2**	77.0	88.5	16.4	**83.0**	77.9	82.4	12.1	28.1	24.8	29.1	24.2	58.1
Ours	92.1	77.5	**83.9**	**58.1**	90.0	**80.6**	83.6	**56.9**	80.2	79.3	**90.6**	70.8	80.9	79.0	**89.0**	**55.1**	**68.4**	**55.4**	72.4	**58.8**	**75.13**

Table 8.1: Cross-view action recognition accuracy (%) on the IXMAS dataset. Each time, one viewpoint is used for training (*Source*) and another one for testing (*Target*). Viewpoint number 4 is considered as a challenging viewpoint, where the camera is placed on top of the subject. Values in bold represent the best score for the corresponding experiment.

view-invariant action recognition task using RGB images.

Table 8.1 presents the action recognition results conducted on the IXMAS dataset. We follow the same cross-view protocol as in [55]–[57], where all sequences of the same viewpoint are used for training; then, during testing, a different viewpoint is used. The obtained results show that our method outperforms the other methods when comparing the average action classification on the IXMAS dataset. Moreover, we note that for the cases where the challenging viewpoint is considered, our method has the highest classification accuracy. The reason for this major improvement, is that the 3D skeleton estimates from the challenging viewpoint are actually better estimated using the proposed strategy when compared to

the other methods. Looking at the average column in Table 8.1, we see that we achieve the highest accuracy for the action classification task. This is achieved by only using the RGB information as input to the proposed approach, and without any additional information provided by a MoCap systems or knowledge transfer from synthetic data. In addition, the reported results are coherent regarding all testing scenarios, obtaining the highest accuracy for the majority of the testing scenarios.

Our results imply that building the 3D skeleton as a human motion model by decoupling the pose estimation into two steps (2D skeleton estimation followed by the lifting to the 3D space preserving the temporal coherence) provides a better understanding of the human action acquired from arbitrary viewpoints. This can be noted specially for the case where the camera is placed on top of the subject (in this dataset – *cam4*), where our method is achieving the highest scores.

Model selection

During the experiments, different network architectures were tested. The results lead to the conclusion that TCN based model outperforms the LSTM based models. This difference in performance can be directly attributed to the number of trainable parameters in each tested model. In the case of the LSTM model, the number of parameters is considerably higher. While LSTM based models perform well in a variety of tasks related to sequence modeling and temporal feature extraction, they are more complex and require more data to train. For this experiments, only 330 sequences are considered for training and testing.

Contrary to LSTM models, TCN based models have much less trainable parameters and can process longer sequences without an increase of the model complexity. Furthermore, using residual connection in TCN based models, do not cause vanishing gradients, like LSTM when processing longer sequences.

8.5 Conclusion

In this chapter, we proposed a new view-invariant action recognition system using 3D skeleton information via RGB images. To that end, we decoupled the 3D human pose estimation problem into two steps: 1) per-image 2D human pose estimation; and 2) per-sequence 3D human pose estimation. In addition, we proposed to use a TCN model for the action classification task. The main advantage is preserving the temporal coherency present in the 3D skeleton sequence. Experimental results show that our approach achieves the highest action recognition accuracy when compared to existing methods. Specially for the cases where the challenging viewpoint is considered.

However, we noticed that the human pose estimation for challenging camera viewpoints has some flaws, mainly due to self-occlusions. Thus, with this in mind, in the next chapter we propose to address the challenge of 3D human pose estimation from arbitrary camera viewpoints, including challenging ones.

Chapter 9

Towards Generalization of 3D Human Pose Estimation In The Wild

In this chapter, we propose *3DBodyTex.Pose*, a dataset that addresses the task of 3D human pose estimation in the wild. Generalization to in-the-wild images remains limited due to the lack of adequate datasets. Existent ones are usually collected in indoor controlled environments where motion capture systems are used to obtain the 3D ground-truth annotations of humans. 3DBodyTex.Pose offers high quality and rich data containing 405 different real subjects in various clothing and poses, and 81k image samples with ground-truth 2D and 3D pose annotations. These images are generated from 200 viewpoints among which 70 challenging extreme viewpoints. The data were created starting from high resolution textured 3D body scans and by incorporating various realistic backgrounds. Retraining a state-of-the-art 3D pose estimation approach using 3DBodyTex.Pose as data augmentation, showed promising improvement in the overall performance. Furthermore, it showed a sensible decrease in the per joint position error when testing on challenging viewpoints. The 3DBodyTex.Pose is expected to offer the research community with new possibilities for generalizing 3D pose estimation from monocular in-the-wild images.

Figure 9.1: Examples of the 3D body scans used to generate in-the-wild images with 2D and 3D annotations of humans.

9.1 Introduction

In the past couple of years, human pose estimation has received a lot of attention from the computer vision community. The goal is to estimate the 2D or 3D position of the human body joints given an image containing a human subject. This has a significant number of applications such as sports, healthcare solutions [76], action recognition [19], [26], [57], [76], [161], and animations.

Due to the recent advances in DNNs, the task of 2D human pose estimation a great improvement in results [63]–[65]. This has been mostly achieved thanks to the availability of large-scale datasets containing 2D annotations of humans in many different conditions, *e.g.*, in the wild [66]. In contrast, advances in the task of human pose estimation in 3D remains limited. The main reasons are the ambiguity of recovering the 3D information from a single image, in addition to the lack of large-scale datasets with 3D annotations of humans, specifically considering in-the-wild conditions. Existent datasets with 3D annotations are usually collected in a controlled environment using MoCap systems [185] or with depth maps [186], [187]. Consequently, the variations in background and camera viewpoints remain limited. In addition, DNNs [188] trained on such datasets have difficulties generalizing well to environments where a lot of variation is present, *e.g.*, scenarios in the wild.

Recently, many works focused on the challenging problem of 3D human pose estima-

tion in the wild [34]–[36], [67]. These works differ significantly from each other but share an important aspect. They are usually evaluated on the same dataset that has been used for training. Thus, it is possible that these approaches have been over-optimized for specific datasets, leading to a lack of generalization. It becomes difficult to judge on the generalization, and more precisely for in-the-wild scenarios where variations coming from the background and camera viewpoints are always present.

In order to address the aforementioned challenge, this chapter presents a new dataset referred to as *3DBodyTex.Pose*. It is an original dataset generated from high-resolution textured 3D body scans, similar in quality to the ones contained in the 3DBodyTex dataset introduced in [81] and later on presented in the *SHARP2020* challenge [189], [190]. 3DBodyTex.Pose is dedicated to the task of human pose estimation. Synthetic scenes are generated with ground-truth information from real 3D body scans, with a large variation in subjects, clothing, and poses (see Figure 9.1). Realistic background is incorporated to the 3D environment. Finally, 2D images are generated from different camera viewpoints, including challenging ones, by virtually changing the camera location and orientation. We distinguish extreme viewpoints as the cases where the camera is, *e.g.*, placed on top of the subject. With the information contained in 3DBodyTex.Pose, it becomes possible to better generalize the problem of the 3D human pose estimation to in-the-wild images independently of the camera viewpoint as shown experimentally on a state-of-the-art 3D pose estimation approach [34].

9.2 Related Datasets

Monocular 3D human pose estimation aims to estimate the 3D joint locations from the human present in the image independently of the environment of the scene. However, usually not all camera viewpoints are taken into consideration. Consequently, the 3D human body joints are not well estimated for the cases where the person is not fully visible or self-occluded. In order to use such images for training, labels for the position of the 2D human joints are needed as ground-truth information [66], [191]. Labeling such images from

extreme camera viewpoints is an expensive and difficult task as it often requires manual annotation. To overcome this issue, MoCap systems can be used for precisely labeling the data. However, they are used in a controlled environment such as indoor scenarios. The Human3.6M dataset [185] is widely used for the task of 3D human pose estimation and it falls under this scenario. It contains 3.6M frames with 2D and 3D annotations of humans from 4 different camera viewpoints. The HumanEva-I [84] and TotalCapture [192] datasets are also captured in indoor environments. HumanEva-I contains 40k frames with 2D and 3D annotations from 7 different camera viewpoints. TotalCapture contains approximately 1.9M frames considering 8 camera locations where the 3D annotations of humans were obtained by fusing the MoCap with inertial measurement units. Also captured within a controlled environment, Mehta *et al*. [193] proposed the MPII-INF-3DHP dataset for 3D human pose estimation which was recorded in a studio using a green screen background to allow automatic segmentation and augmentation. Consequently, the authors augment the data in terms of foreground and background, where the clothing color is changed on a pixel basis, and for the background, images sampled from the internet are used. Recently, von Marcard *et al*. [194] proposed a dataset with 3D pose in outdoor scenarios recorded with a moving camera. It contains more than 51k frames and 7 actors with a limited number of clothing style.

An alternative proposed with SURREAL [195] and exploited in [97], is to generate realistic ground-truth data synthetically. SURREAL places a parametric body model with varied pose and shape over a background image of a scene to simulate a monocular acquisition. Ground-truth 2D and 3D poses are known from the body model. To add realism, the body model is mapped with clothing texture. A drawback of this approach is that the body shape lacks details.

The 3DBodyTex dataset [81] contains static 3D body scans from people in close-fitting clothing, in varied poses and with ground-truth 3D pose. This dataset is not meant for the task of 3D human pose estimation. However, it is appealing for its realism: detailed shape and high-resolution texture information. It has been exploited for 3D human body fitting [196] and it could also be used to synthesize realistic monocular images from arbitrary viewpoints with ground-truth 2D and 3D poses. The main drawback of this dataset is the fact that it

	3DBodyTex. Pose (Ours)	HumanEva-I	Human3.6M	MPII-INF-3DHP	TotalCapture	3DPW	SURREAL
# of subjects	405	4	11	8	5	7	n/a
# of samples	81k	40k	$\sim$3.6M	$>$1.3M	$\sim$1.9M	$>$51k	$\sim$6.5M
Ground-truth pose	2D+3D	2D+3D	2D+3D	3D	3D	3D	2D+3D
Real people	Yes	Yes	Yes	Yes	Yes	Yes	No
Background	Indoor & Outdoor	Indoor	Indoor	Green Screen	Indoor	Outdoor	Indoor
Clothing	Realistic	Realistic	Realistic	Realistic$^{(\star)}$	No	Limited	No
# of total camera viewpoints	200	7	4	14	8	n/a	n/a
# of challenging viewpoints	70	0	0	3	0	n/a	n/a

Table 9.1: Comparison of datasets for the task of 3D human pose estimation. $(\star)$ indicates that clothing was synthetically added to the dataset.

contains the same tight clothing with no variations.

9.3 Proposed 3DBodyTex.Pose Dataset

In contrast with 3DBodyTex, the new 3DBodyTex.Pose dataset contains 3D body scans that are captured from 405 subjects in their own regular clothes. From these 405 subjects, 204 are females and 201 are males. Having different clothing style from different people adds more variation to the dataset when considering in-the-wild scenarios. Figure 9.1 shows a couple of examples of 3D body scans with different clothing. In this work, the goal is to use the 3D body scans to synthesize realistic monocular images from arbitrary camera viewpoints with its corresponding 2D and 3D ground-truth information for the task of 3D human pose estimation. The principal characteristics of 3DBodyTex.Pose are compared to state-of-the-art datasets in Table 9.1.

The 3DBodyTex.Pose dataset aims to address the challenges of in-the-wild images and the extreme camera viewpoints. Given that the only input is the set of 3D scans, we need to estimate the ground-truth 3D skeletons, to synthesize the monocular images from challenging viewpoints and to simulate an in-the-wild environment. These three stages are detailed below.

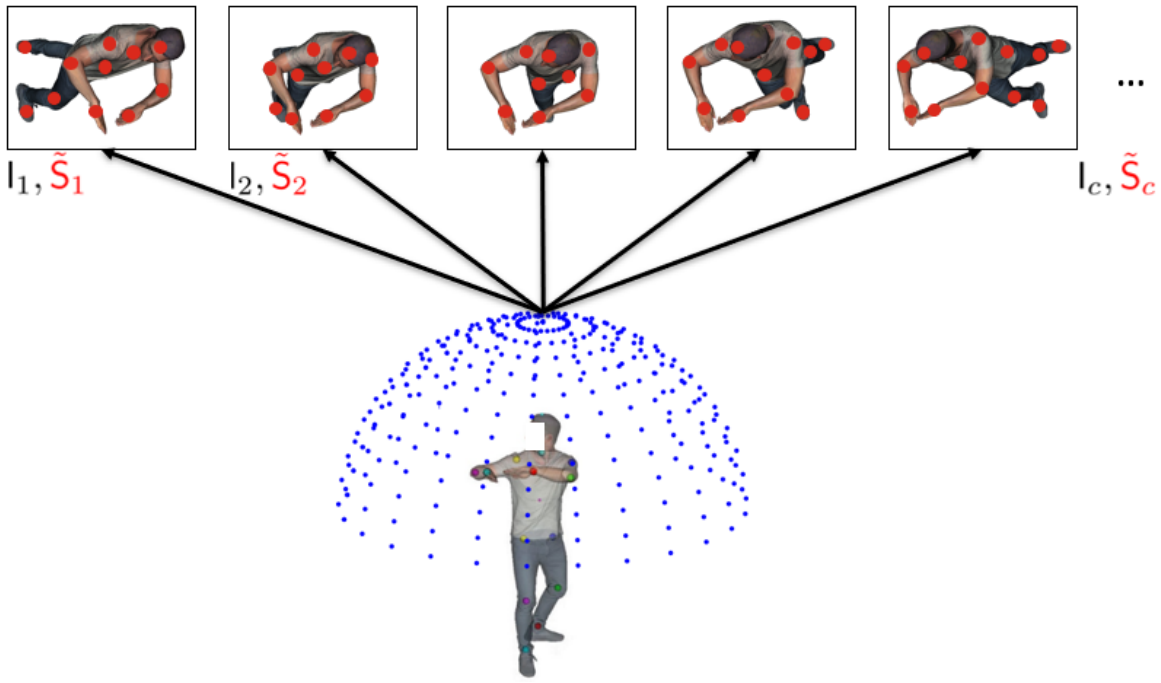


Figure 9.2: Extreme camera viewpoints images (top row) from a single 3D body scan. The blue dots represent the camera locations for each camera viewpoint.

9.3.1 Ground-truth 3D joints

To estimate the ground-truth 3D skeleton, we follow the automatic approach of 3DBodyTex [81] where body landmarks are first detected in 2D views before being robustly aggregated into 3D positions. Hence, for every 3D scan we have the corresponding 3D positions of the human body joints that is henceforth used as the ground-truth 3D skeleton.

9.3.2 Challenging Viewpoints

We propose to change the location and orientation of the camera in order to create monocular images where also extreme viewpoints are considered, see Figure 9.2. Considering a 3D body scan $\mathsf{P} \in \mathbb{R}^{3\times K}$, where K is the number of vertices of the mesh, the 3D skeleton with N joints $\mathsf{S} \in \mathbb{R}^{3\times N}$, and also the homogeneous projection matrix M_c for the camera

position c, we can back-project the 3D skeleton into the image $\mathrm{I_c}$ by

$$\tilde{\mathrm{S}}_c = \mathrm{M}_c \cdot \mathrm{S}, \tag{9.1}$$

where $\tilde{\mathrm{S}}_c \in \mathbb{R}^{3\times N}$ represents the homogeneous coordinates of the projected 3D skeleton into the image plane I_c, corresponding to a 2D skeleton. In this way, we are able to generate all possible camera viewpoints around the subject and easily obtain the corresponding 2D skeleton. In summary, each element of the 3DBodyTex.Pose is composed of image I_c, 2D skeleton $\tilde{\mathrm{S}}_c$, and 3D skeleton S in the camera coordinate system.

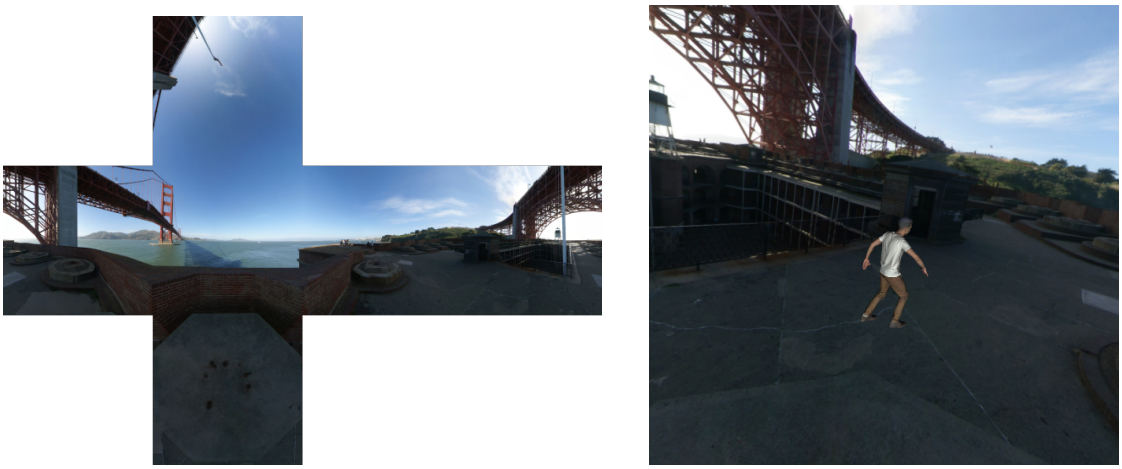
9.3.3 In-the-wild Environment

In order to address the challenge of the in-the-wild images with ground-truth information for the task of 3D human pose estimation, we further propose to embed the 3D scan in an environment with cube mapping [197] which in turns adds a realistic background variation to the dataset. An example texture cube is shown in Figure 9.3a. The six faces are mapped to a cube surrounding the scene with the 3D body scan at the center, see Figure 9.3b. Realistic textures cubes are obtained from [198].

To have variation in the data, for each image, we randomly draw a texture cube, a camera viewpoint and a 3D scan. The proposed 3DBodyTex.Pose dataset provides reliable ground-truth 2D and 3D annotations with realistic and varied in-the-wild images while considering arbitrary camera viewpoints. Moreover, it offers a relatively high number of subjects in comparison with state-of-the-art 3D pose datasets, refer to Table 9.1. It also offers richer body details in terms of clothing, shape, and the realistic texture. Figure 9.4 shows the data generation overview.

9.4 Experimental Evaluation

In what follows, we use the approach proposed by Zhou *et al*. [34] to showcase the impact of the 3DBodyTex.Pose dataset in improving the performance of 3D pose estimation in the



(a) (b)

Figure 9.3: (a) Example of an unfolded cube projection of a 3D environment (extracted from [198]). (b) Example of a 3D body scan added to the 3D environment of a realistic scene.

wild. We note that, in a similar fashion, 3DBodyTex.Pose can be used to enhance any other existent approach. Our goal is to share this new dataset with the research community and encourage (re-)evaluating and (re-)training existent and new 3D pose estimation approaches especially considering in-the-wild scenarios with a special focus on extreme viewpoints.

9.4.1 Baseline 3D Pose Estimation Approach

The work in [34] aims to estimate 3D human poses in the wild. For that, the authors proposed to couple together in-the-wild images with 2D annotations with indoor images with 3D annotations in an end-to-end framework. The authors also provide the code for both training and testing the network.

The network proposed in [34] consists of two different modules: (1) 2D pose estimation module; and (2) depth regression module. In the first module, the goal is to predict a set of N heat maps by minimizing the L^2 distance between the predicted and the ground-truth heat maps where only images with 2D annotations were used (MPII dataset [66]). Secondly, the depth regression module learns to predict the depth between the camera and the image plane by using the images where 3D annotations are provided (Human3.6M dataset [185]).

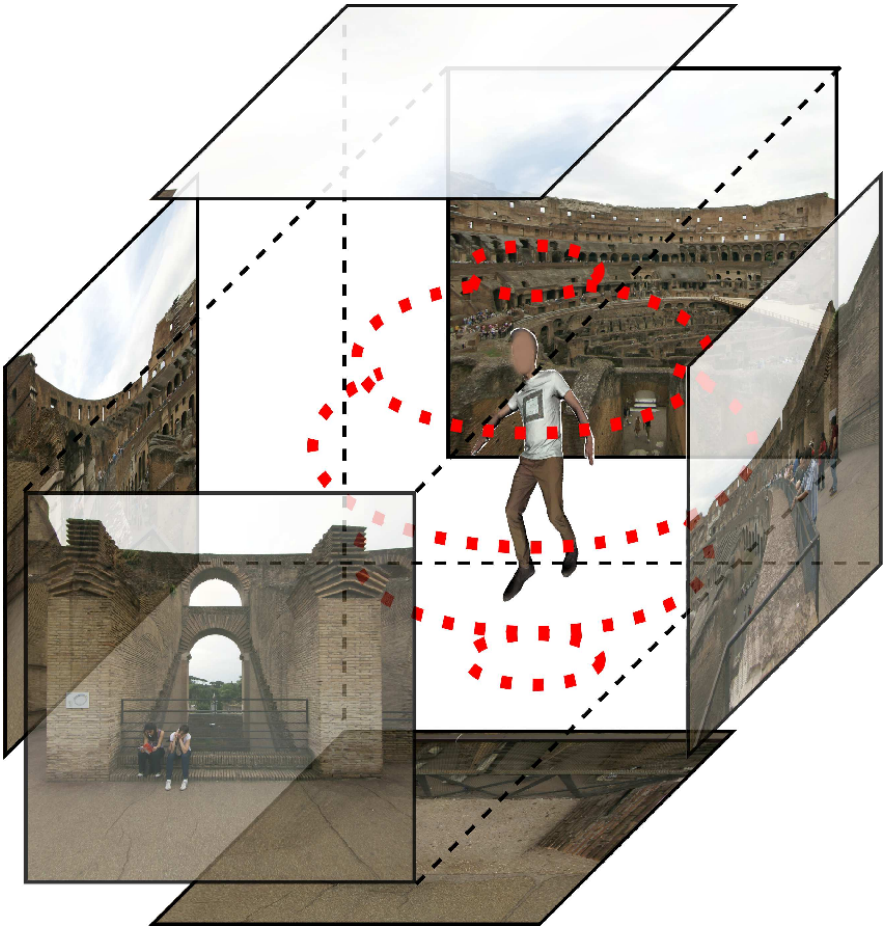


Figure 9.4: Data generation overview. The 3D body scan is placed in the center of the cube mapping environment. Different camera viewpoints (in red) are considered in order to capture the scene from multiple angles.

Also within the second module, the authors proposed a geometric constraint which serves as a regularization for depth prediction when the 3D annotations are not available. At the end, the network is built in a way that both modules are trained together.

9.4.2 Data Augmentation with 3DBodyTex.Pose

We propose to retrain the network presented in [34] by adding the 3DBodyTex.Pose data to the training set originally used in [34]. Specifically, 60k additional RGB images from 3DBodyTex.Pose and their corresponding 2D skeletons were used to increase the variation

Methods	Average (mm)
Zhou *et al*. [34]	64.9
Zhou *et al*. [34] ++ (Ours)	**61.3**
Martinez *et al*. [92]	62.9
Rogez *et al*. [35]	61.2
Yang *et al*. [67]	<u>58.6</u>

Table 9.2: Quantitative results of the MPJPE in millimeters on the Human3.6M dataset following the same protocol as in [34]. The average column represents the average error value of all actions in the validation set.

coming from realistic background and camera viewpoints.

We first follow the same evaluation protocol as in [34] by testing on the Human3.6M dataset [185], and using the *Mean Per Joint Position Error* (MPJPE) in millimeters (mm) as an evaluation metric between 3D skeletons. Table 9.2 shows the results of retraining [34] by augmenting with 3DBodyTex.Pose (**Zhou *et al*. [34] ++**) along with other reported state-of-the-art results as a reference. Without using 3DBodyTex.Pose, the average error between the estimated 3D skeleton and the ground-truth annotation is 64.9 mm, and when retrained with the addition of our proposed dataset, the error decreases to 61.3 mm. This result is a very promising step towards the generalization of 3D human pose estimation for in-the-wild images. Despite the fact that testing in Table 9.2 is on Human3.6M (indoor scenes only), retraining with 3DBodyTex.Pose helps bring the performance of [34] closer to the top performing approaches [35], [67] and even beating others, *i.e*., [92].

As one of the aims of this paper is to mitigate the effect of challenging camera viewpoints, we tested the performance of [34] on a new testing set containing challenging viewpoints only. These were selected from the 3DBodyTex.Pose dataset and reserved for testing only[1]. Table 9.3 shows that adding the 3DBodyTex.Pose to the training set in the network of [34] performs better when testing with challenging viewpoints only. Note that the relative high values of the errors, as compared to Table 9.2, are due to the fact that the depth regression module is learned with the 3D ground-truth poses of the Human3.6M dataset only.

[1]Never seen during training.

Methods	Average (mm)
Zhou *et al*. [34]	292
Zhou *et al*. [34] ++ (Ours)	**267**

Table 9.3: Results of the MPJPE while testing on challenging camera viewpoints only.

9.5 Conclusion

This chapter introduced the 3DBodyTex.Pose dataset as a new original dataset to support the research community in designing robust approaches for 3D human pose estimation in the wild, independently of the camera viewpoint. It contains synthetic but realistic monocular images with 2D and 3D human pose annotations, generated from diverse and high-quality textured 3D body scans. The potential of this dataset is demonstrated by retraining a state-of-the-art 3D human pose estimation framework. There is a significant improvement in performance when augmented with 3DBodyTex.Pose. This opens the door to the generalization of 3D human pose estimation to in-the-wild images.

Chapter 10

Conclusions

In this chapter, we summarize the core findings of our research in the different parts presented in this thesis. Moreover, we highlight interesting directions for future work.

10.1 Summary

In this thesis, we mainly focused on investigating and proposing solutions to alleviate existing constraints for the deployment of 3D skeleton-based approaches in various real-world scenarios. We presented two applications that use the 3D skeleton representation of the human body. First, we presented a solution designed for home-based rehabilitation of stroke survivors under the remote supervision of a therapist. Second, we focused on the challenging topic of cross-view action recognition using a monocular RGB camera. We further investigated and proposed different 3D human pose estimation techniques from a single RGB camera in order to take advantage of 3D skeleton-based approaches.

In the first part of this thesis, we focused on a low-cost solution to support home-based rehabilitation of stroke survivor. Our first contribution mainly addresses the challenge of providing feedback on how to improve a movement being performed. Thus, we proposed a novel manner to present guidance feedback proposals by showing easily human-understandable feedback messages. Besides that, we further proposed to extend the feedback proposals with the objective of monitoring postural defects over-time, and providing

guidance feedback. Our next contribution combines the feedback proposals into one solution targeting the home-based rehabilitation of stroke survivors. Moreover, we proposed to adapt the feedback proposals to the specific body condition of the stroke survivor. In more detail, we proposed the concept of exercise personalization which considers the physical conditions of the stroke survivor to better and most importantly, the feedback proposals are provided based on the anthropometry of the stroke survivor. Our next contribution tackles the challenge of detecting abnormalities in the gait pattern of the stroke survivor. To that end, a curve-based representation of the 3D skeleton is proposed and the quantification of time variation is analyzed.

In the second part of this thesis, motivated by the investigation of solutions to mitigate existing constraints for the deployment of 3D human pose estimation to the real-world conditions, we proposed two different techniques of human pose estimation in the context of cross-view action recognition. First, we proposed to use a per-frame pose estimation technique followed by an LSTM-based network to effectively model the temporal dependency. Considering the limitations of the per-frame pose estimation, we further proposed a two-stage human pose estimation on a sequence-to-sequence basis; thus, incorporating the temporal aspect of the skeleton sequences already into the estimation stage. Furthermore, we adopted a TCN to model the temporal information and classify the skeleton sequences into human actions. Nevertheless, we noted an important case where the human pose estimation presents a major failure. The estimation of 3D skeletons is extremely difficult when the subject is observed from challenging camera viewpoints. Considering this, we introduced a novel dataset to address the challenge of camera viewpoint variation, including challenging ones. In addition, we presented an approach to add realistic background variation to the dataset with the objective of addressing the challenge of 3D human pose estimation in the wild.

10.2 Future Directions

10.2.1 Home-based Rehabilitation via 3D Human Pose Estimation Using a Single RGB Camera

In Chapter 3, Chapter 4, Chapter 5 and Chapter 6 different methodologies were presented with the objective of improving and supporting the rehabilitation of stroke survivors at home. However, healthcare solutions are highly dependent on the acquisition system, which in this case is the Microsoft Kinect v2 [32]. The discontinuity of such sensor is already ongoing and new approaches based on 3D human pose estimation from RGB cameras could be the next step. In addition, it would standardize the home-based rehabilitation to any RGB camera, *e.g.*, laptop's webcam, and consequently, reduce the cost. Therefore, an interesting research direction is investigating solutions to improve the quality of the 3D skeleton estimates from a single RGB camera, targeting healthcare criterion.

Moreover, the proposed guidance feedback proposals for improving the movement being performed and correcting the posture of the patient present some limitations. They are are not perfect and they are highly linked to the quality of the 3D skeleton estimates. Considering this, an interesting research direction is the investigation of learning-based approaches to better present guidance feedback proposals based on the patient's movement history. Consequently, guidance feedback proposals would be even more tailored to the patient's condition and they would be automatically presented based on the patient's preference.

10.2.2 Generalization of 3D Human Pose Estimation In The Wild and View-Invariant Action Recognition

In Chapter 7 and Chapter 8 we presented solutions to mitigate existing constraints for the deployment of 3D human pose estimation methods to more real-world scenarios. We proposed to use a 3D skeleton representation of the human to model view-invariant features by aligning the 3D skeleton sequences with respect to a canonical 3D skeleton. However, if the 3D skeleton estimate is not accurate, it will impact the performance of the methodology and

inherently affect action recognition accuracy. Therefore, an interesting future direction is the investigation of end-to-end approaches where the 3D human pose estimation and the action classification are learned together from RGB input information. In that context, the RGB information could be fused together with the 3D skeleton information to possibly incorporate the advantages of both modalities, enhancing the human motion analysis.

However, 3D human pose estimation approaches need to provide a more accurate skeleton estimate and a more robust to occlusions, viewpoint variability, and outdoor conditions. With that in mind, in Chapter 9 we proposed 3DBodyTex.Pose dataset to address the challenges of 3D human pose estimation in the wild. Nevertheless, in the current state, the dataset is lacking realism. Therefore, an attractive research direction is the addition of realism to the dataset. Not only in terms of viewpoint variability or outdoor conditions, but also adding temporal information to the dataset. Subsequently, this would generalize the human pose estimation task for a better applicability for human motion analysis in the wild.

Appendices

A: Therapist Questionnaire

For questions from 1 to 9, choose a number from 1 to 5 to answer such that:

Completely Agree	Agree	Do not agree, do not disagree	Disagree	Completely disagree
1	2	3	4	5

Technical Problems

1. The application is stable (no bugs).
2. The application is simple to use.
3. The latency of the movement capture is not too long.

Utility of Feedback (Color-based Feedback)

4. The feedback is reliable and correspond to our expectations. For an answer from 3 to 5, explain your choice.
5. In your opinion, the feedback helped the patient to improve his movement. For an answer from 3 to 5, explain your choice.
6. (To answer this question, please open the therapist application and visualize the training data), the curve indicating the quality of the movement reflects the quality of the movement during the session. For an answer from 3 to 5, explain your choice.

Patient Psychology

7. You feel that the patient was comfortable when using the application.
8. You have the feeling that the patient would like to reuse the application.
9. The patient can use the application at home safely.

Open Questions

10. In your opinion, what are the strengths of the application?
11. In your opinion, what are the weaknesses of the application?
12. What features would like to see in a future release?

B: Patient Questionnaire (Last Session)

For questions from 1 to 9, choose a number from 1 to 5 to answer such that:

Completely Agree	Agree	Do not agree, do not disagree	Disagree	Completely disagree
1	2	3	4	5

Technical Problems

1. The application ”Self-Training” is easy to use.
2. I find that the application capture reliably my motion and reproduce them correctly on the screen.
3. The movement capture latency in the application is important.

Utility of Feedback (Color-based Feedback)

4. I have the feeling that the color-based feedback does not vary a lot.

5. During the training, I find the feedback useful.

6. I find that my condition got better since I am using this application.

Patient Psychology

7 . I feel more motivated to exercise since I am using the application.

8. I would to continue using it in the presence of a therapist.

9. I find the training boring.

10. I would like to continue using it at home without the presence of a therapist.

11. I find the exercises difficult to do.

Open Questions

12. Do you feel safe when suing it?

13. In your opinion, what are the strengths of the application?

14. In your opinion, what are the weaknesses of the application?

15. How would you describe your user experience?

16. What are the features that you would like to see in a future release?

C: Patient Questionnaire (Each Session)

For questions from 1 to 4, choose a number from 1 to 5 to answer such that:

Completely Agree	Agree	Do not agree, do not disagree	Disagree	Completely disagree
1	2	3	4	5

Technical Problems

1. The application “Self-Training” is easy to use.

Utility of Feedback (Color-based Feedback)

2. During the training, I found the color-based feedback useful.

Patient Psychology

3. I feel more motivated while training since I use this application.

4. I find the training boring.

References

[1] Z. Duric, W. D. Gray, R. Heishman, F. Li, A. Rosenfeld, M. J. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1272–1289, 2002.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[3] S. Hongeng, F. Brémond, and R. Nevatia, "Bayesian framework for video surveillance application," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, IEEE, vol. 1, 2000, pp. 164–170.

[4] S. Saxena, F. Brémond, M. Thonnat, and R. Ma, "Crowd behavior recognition for video surveillance," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, Springer, 2008, pp. 970–981.

[5] Y.-J. Chang, S.-F. Chen, and J.-D. Huang, "A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," *Research in developmental disabilities*, vol. 32, no. 6, pp. 2566–2570, 2011.

[6] M. Trombetta, P. P. B. Henrique, M. R. Brum, E. L. Colussi, A. C. B. De Marchi, and R. Rieder, "Motion rehab ave 3d: A vr-based exergame for post-stroke rehabilitation," *Computer methods and programs in biomedicine*, vol. 151, pp. 15–20, 2017.

[7] A. Thangali, J. P. Nash, S. Sclaroff, and C. Neidle, “Exploiting phonological constraints for handshape inference in asl video,” in *CVPR 2011*, 2011, pp. 521–528. DOI: 10.1109/CVPR.2011.5995718.

[8] H. Cooper and R. Bowden, “Large lexicon detection of sign language,” in *International Workshop on Human-Computer Interaction*, Springer, 2007, pp. 88–97.

[9] H. Pirsiavash, C. Vondrick, and A. Torralba, “Assessing the quality of actions,” in *Computer Vision–ECCV 2014*, Springer, 2014, pp. 556–571.

[10] P. Parmar and B. Tran Morris, “Learning to score olympic events,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[11] R. Poppe, “Vision-based human motion analysis: An overview,” *Computer vision and image understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[12] J. Aggarwal and M. Ryoo, “Human activity analysis: A review,” *ACM Comput. Surv.*, vol. 43, no. 3, 16:1–16:43, Apr. 2011, ISSN: 0360-0300. DOI: 10.1145/1922649.1922653. [Online]. Available: http://doi.acm.org.proxy.bnl.lu/10.1145/1922649.1922653.

[13] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[14] G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[15] *Johansson: Motion perception part 1 - accessed on november, 2020*, https://www.youtube.com/watch?v=1F5ICP9SYLU&ab_channel=BioMotionLab.

[16] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, “Real-time human pose recognition in parts from single depth images,” *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[17] M. Antunes, D. Aouada, and B. Ottersten, “A revisit to human action recognition from depth sequences: Guided svm-sampling for joint selection,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2016, pp. 1–8.

[18] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3d skeletal data: A review,” *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.

[19] K. Papadopoulos, E. Ghorbel, O. Oyedotun, D. Aouada, and B. Ottersten, “Deepvi: A novel framework for learning deep view-invariant human action representations using a single rgb camera,” in *IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires*, 2020, pp. 18–22.

[20] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297. DOI: 10.1109/CVPR.2012.6247813.

[21] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[22] C.-Y. Chang, B. Lange, M. Zhang, S. Koenig, P. Requejo, N. Somboon, A. A. Sawchuk, A. A. Rizzo, *et al.*, “Towards pervasive physical rehabilitation using microsoft kinect.,” in *PervasiveHealth*, 2012, pp. 159–162.

[23] A. Paiement, L. Tao, M. Camplani, S. Hannuna, D. Damen, and M. Mirmehdi, “Online quality assessment of human motion from skeleton data,” in *Proceedings of the British Machine Vision Conference*, BMVA Press, 2014.

[24] L. Tao, A. Paiement, D. Aldamen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock, “A comparative study of pose representation and dynamics modelling for online motion quality assessment,” *Computer Vision and Image Understanding*, vol. 11, 2016.

[25] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” *arXiv preprint arXiv:1801.07455*, 2018.

[26] K. Papadopoulos, E. Ghorbel, D. Aouada, and B. Ottersten, “Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition,” *arXiv preprint arXiv:1912.09745*, 2019.

[27] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, “Skeleton-based action recognition with shift graph convolutional network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 183–192.

[28] G. Kwakkel, B. J. Kollen, and H. I. Krebs, “Effects of robot-assisted therapy on upper limb recovery after stroke: A systematic review,” *Neurorehabilitation and neural repair*, 2007.

[29] P. Andlin-Sobocki, B. Jönsson, H.-U. Wittchen, and J. Olesen, “Cost of disorders of the brain in europe,” *European Journal of Neurology*, 2005.

[30] P. Langhorne, G. Taylor, G. Murray, M. Dennis, C. Anderson, E. Bautz-Holter, P. Dey, B. Indredavik, N. Mayo, M. Power, *et al.*, “Early supported discharge services for stroke patients: A meta-analysis of individual patients’ data,” *The Lancet*, 2005.

[31] F. Ofli, G. Kurillo, S. Obdrzálek, R. Bajcsy, H. B. Jimison, and M. Pavel, “Design and evaluation of an interactive exercise coaching system for older adults: Lessons learned,” *IEEE J. Biomedical and Health Informatics*, 2016.

[32] *Microsoft kinect v2 - accessed on november, 2020*, `https://developer.microsoft.com/en-us/windows/kinect/`.

[33] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, “VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera,” 4, vol. 36, 2017. DOI: `10.1145/3072959.3073596`. [Online]. Available: `http://gvv.mpi-inf.mpg.de/projects/VNect/`.

[34] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, “Towards 3d human pose estimation in the wild: A weakly-supervised approach,” in *ICCV*, 2017.

[35] G. Rogez, P. Weinzaepfel, and C. Schmid, “Lcr-net++: Multi-person 2d and 3d pose detection in natural images,” *IEEE TPAMI*, 2019.

[36] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3d human pose estimation in video with temporal convolutions and semi-supervised training,” in *CVPR*, 2019.

[37] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, “Learnable triangulation of human pose,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7718–7727.

[38] F. Huang, A. Zeng, M. Liu, Q. Lai, and Q. Xu, “Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image,” in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 429–438.

[39] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *CVIU*, vol. 104, no. 2-3, pp. 249–257, 2006.

[40] H. Zhou and H. Hu, “Human motion tracking for rehabilitation—a survey,” *Biomedical Signal Processing and Control*, vol. 3, no. 1, pp. 1–18, 2008.

[41] L. E. Sucar, R. Luis, R. Leder, J. Hernandez, and I. Sanchez, “Gesture therapy: A vision-based system for upper extremity stroke rehabilitation,” in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010.

[42] H. M. Hondori, M. Khademi, L. Dodakian, S. C. Cramer, and C. V. Lopes, “A spatial augmented reality rehab system for post-stroke hand rehabilitation.,” in *MMVR*, 2013.

[43] H. Mousavi Hondori and M. Khademi, “A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation,” *Journal of Medical Engineering*, vol. 2014, 2014.

[44] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, “A review on vision techniques applied to human behaviour analysis for ambient-assisted living,” *Expert Systems with Applications*, 2012.

[45] P. M. Kato, “Video games in health care: Closing the gap.,” *Review of General Psychology*, 2010.

[46] J. W. Burke, M. McNeill, D. Charles, P. J. Morrow, J. Crosbie, and S. McDonough, "Serious games for upper limb rehabilitation following stroke," in *Games and Virtual Worlds for Serious Applications, 2009. VS-GAMES'09. Conference in*, IEEE, 2009.

[47] R. Wang, G. Medioni, C. Winstein, and C. Blanco, "Home monitoring musculo-skeletal disorders with a single 3d sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013.

[48] D. Sadihov, B. Migge, R. Gassert, and Y. Kim, "Prototype of a vr upper-limb rehabilitation system enhanced with motion-based tactile feedback," in *World Haptics Conference (WHC), 2013*, IEEE, 2013, pp. 449–454.

[49] X. H. Bao, Y. Mao, Q. Lin, Y. Qiu, S. Chen, L. Li, R. S. Cates, S.-F. Zhou, and D. Huang, "Mechanism of kinect-based virtual reality training for motor functional recovery of upper limbs after subacute stroke," in *Neural regeneration research*, 2013.

[50] J. M. I. Zannatha, A. J. M. Tamayo, Á. D. G. Sánchez, J. E. L. Delgado, L. E. R. Cheu, and W. A. S. Arévalo, "Development of a system based on 3d vision, interactive virtual environments, ergonometric signals and a humanoid for stroke rehabilitation," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 2, pp. 239–249, 2013, SI: Computer Assisted Tools for Medical Robotics, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2013.04.021.

[51] J.-A. Lozano-Quilis, H. Gil-Gomez, J.-A. Gil-Gómez, S. Albiol-Perez, G. Palacios, H. M. Fardoum, and A. S. Mashat, "Virtual reality system for multiple sclerosis rehabilitation using kinect," in *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013, pp. 366–369.

[52] S. Simmons, R. McCrindle, M. Sperrin, and A. Smith, "Prescription software for recovery and rehabilitation using microsoft kinect," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, IEEE, 2013, pp. 323–326.

[53] S. Spasojević, N. V. Ilić, A. Rodić, and J. Santos-Victor, “Kinect-based application for progress monitoring of the stroke patients,” in *Proceedings of IcETRAN conference, vol. ROI2*, vol. 6, 2017, pp. 1–5.

[54] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Comput. Surv.*, vol. 41, no. 3, 15:1–15:58, Jul. 2009, ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. [Online]. Available: http://doi.acm.org.proxy.bnl.lu/10.1145/1541880.1541882.

[55] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, “3D Pose from Motion for Cross-View Action Recognition via Non-linear Circulant Temporal Encoding,” in *CVPR*, IEEE, 2014. DOI: 10.1109/cvpr.2014.333. [Online]. Available: https://doi.org/10.1109/cvpr.2014.333.

[56] H. Rahmani and A. Mian, “Learning a non-linear knowledge transfer model for cross-view action recognition,” in *CVPR*, IEEE, 2015. DOI: 10.1109/cvpr.2015.7298860. [Online]. Available: https://doi.org/10.1109/cvpr.2015.7298860.

[57] E. Ghorbel, K. Papadopoulos, R. Baptista, H. Pathak, G. Demisse, D. Aouada, and B. Ottersten, “A view-invariant framework for fast skeleton-based action recognition using a single rgb camera,” in *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Prague, 25-27 February 2018*, 2019.

[58] K. Papadopoulos, M. Antunes, D. Aouada, and B. Ottersten, “Enhanced trajectory-based action recognition using human pose,” in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, pp. 1807–1811.

[59] ——, “A revisit of action detection using improved trajectories,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 2067–2071.

[60] K. Papadopoulos, G. Demisse, E. Ghorbel, M. Antunes, D. Aouada, and B. Ottersten, “Localized trajectories for 2d and 3d action recognition,” *Sensors*, vol. 19, no. 16, p. 3503, 2019.

[61] R. Baptista, E. Ghorbel, K. Papadopoulos, G. Demisse, D. Aouada, and B. Ottersten, "View-invariant action recognition from rgb data via 3d pose estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019*, 2019.

[62] K. Papadopoulos, E. Ghorbel, R. Baptista, D. Aouada, and B. Ottersten, "Two-stage rgb-based action detection using augmented 3d poses," in *Computer Analysis of Images and Patterns*, M. Vento and G. Percannella, Eds., Cham: Springer International Publishing, 2019, pp. 26–35, ISBN: 978-3-030-29888-3.

[63] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017.

[64] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv:1812.08008*, 2018.

[65] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.

[66] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.

[67] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *CVPR*, 2018.

[68] M. Vergara and A. Page, "Relationship between comfort and back posture and mobility in sitting-posture," *Applied Ergonomics*, vol. 33, no. 1, pp. 1–8, 2002, ISSN: 0003-6870. DOI: http://dx.doi.org/10.1016/S0003-6870(01)00056-4. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003687001000564.

[69] L. Zhang, M. G. Helander, and C. G. Drury, "Identifying factors of comfort and discomfort in sitting," *Human Factors*, vol. 38, no. 3, pp. 377–389, 1996. DOI: 10.1518/001872096778701962. eprint: http://dx.doi.org/10.1518/001872096778701962. [Online]. Available: http://dx.doi.org/10.1518/001872096778701962.

[70] S. Konz and S. Johnson, *Work Design: Occupational Ergonomics*. Holcomb Hathaway, 2008, ISBN: 9781890871796. [Online]. Available: https://books.google.lu/books?id=%5C_7nCJwAACAAJ.

[71] M. Lehto and S. Landry, *Introduction to Human Factors and Ergonomics for Engineers, Second Edition*, ser. Human Factors and Ergonomics. CRC Press, 2012, ISBN: 9781466584167. [Online]. Available: https://books.google.lu/books?id=lGjOBQAAQBAJ.

[72] M. Antunes, G. Demisse, and D. Aouada, "Physical activity feedback," WO2017207802A1, 2017. [Online]. Available: https://patents.google.com/patent/WO2017207802A1/en.

[73] M. Antunes, R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, "Visual and human-interpretable feedback for assisting physical activity," in *European Conference on Computer Vision (ECCV) Workshop on Assistive Computer Vision and Robotics Amsterdam*, 2016.

[74] R. Baptista, M. Antunes, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Flexible feedback system for posture monitoring and correction," in *IEEE International Conference on Image Information Processing (ICIIP)*, 2017.

[75] R. Baptista, E. Ghorbel, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Key-skeleton based feedback tool for assisting physical activity," in *2018 Zooming Innovation in Consumer Technologies Conference (ZINC)*, IEEE, 2018, pp. 175–176.

[76] R. Baptista, E. Ghorbel, A. E. R. Shabayek, F. Moissenet, D. Aouada, A. Douchet, M. André, J. Pager, and S. Bouilland, "Home self-training: Visual feedback for assisting physical activity for stroke survivors," *CMPB*, 2019.

[77] E. Ghorbel, R. Baptista, A. E. R. Shabayek, D. Aouada, M. G. Oramaeche, J. O. Lago, and L. O. Fernandez, "Home-based rehabilitation system for stroke survivors: A clinical evaluation," *Journal of Medical Systems*, 2020.

[78] R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, “Deformation-based abnormal motion detection using 3d skeletons,” in *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2018, pp. 1–6. DOI: 10.1109/IPTA.2018.8608143.

[79] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt, “Monocular 3D Human Pose Estimation Using Transfer Learning and Improved CNN Supervision,” *CoRR*, vol. abs/1611.09813, 2016. arXiv: 1611.09813. [Online]. Available: http://arxiv.org/abs/1611.09813.

[80] M. Adel Musallam, R. Baptista, K. Al Ismaeil, and D. Aouada, “Temporal 3d human pose estimation for action recognition from arbitrary viewpoints,” in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 253–258. DOI: 10.1109/CSCI49370.2019.00052.

[81] A. Saint, E. Ahmed, A. E. R. Shabayek, K. Cherenkova, G. Gusev, D. Aouada, and B. Ottersten, “3dbodytex: Textured 3d body dataset,” *3DV*, 2018.

[82] R. Baptista, A. Saint, K. A. Ismaeil, and D. Aouada, “Towards generalization of 3d human pose estimation in the wild,” in *IEEE International Conference on Pattern Recognition (ICPR) Workshop on 3D Human Understanding*, 2020.

[83] ——, “Towards generalization of 3d human pose estimation in the wild,” *arXiv preprint arXiv:2004.09989*, 2020.

[84] L. Sigal, A. O. Balan, and M. J. Black, “Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,” *IJCV*, vol. 87, no. 1-2, p. 4, 2010.

[85] A. Bulat and G. Tzimiropoulos, “Human pose estimation via convolutional part heatmap regression,” in *European Conference on Computer Vision*, Springer, 2016, pp. 717–732.

[86] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

[87] M. Müller, “Dynamic time warping,” *Information retrieval for music and motion*, pp. 69–84, 2007.

[88] A. Agarwal and B. Triggs, “Recovering 3d human pose from monocular images,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 1, pp. 44–58, 2005.

[89] X. K. Wei and J. Chai, “Modeling 3d human poses from uncalibrated monocular images,” in *2009 IEEE 12th International Conference on Computer Vision*, IEEE, 2009, pp. 1873–1880.

[90] M. Andriluka, S. Roth, and B. Schiele, “Monocular 3d pose estimation and tracking by detection,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, pp. 623–630.

[91] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, “Sparseness meets deepness: 3d human pose estimation from monocular video,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4966–4975.

[92] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A simple yet effective baseline for 3d human pose estimation,” in *ICCV*, 2017.

[93] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, “Coarse-to-fine volumetric prediction for single-image 3D human pose,” in *CVPR*, 2017, pp. 1263–1272.

[94] S. Li and A. B. Chan, “3d human pose estimation from monocular images with deep convolutional neural network,” in *Asian Conference on Computer Vision*, Springer, 2014, pp. 332–347.

[95] S. Li, W. Zhang, and A. B. Chan, “Maximum-margin structured learning with deep networks for 3d human pose estimation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2848–2856.

[96] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, “Direct prediction of 3d body poses from motion compensated sequences,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 991–1000.

[97] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, “Learning to estimate 3d human pose and shape from a single color image,” in *CVPR*, 2018.

[98] G. Mori and J. Malik, “Recovering 3d human body configurations using shape contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1052–1062, 2006.

[99] I. Akhter and M. J. Black, “Pose-conditioned joint angle limits for 3d human pose reconstruction,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1446–1455.

[100] F. Moreno-Noguer, “3d human pose estimation from a single image via distance matrix regression,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2823–2832.

[101] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, “Learning to fuse 2d and 3d image cues for monocular body pose estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3941–3950.

[102] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[103] F. Sun, I. J. Norman, and A. E. While, “Physical activity in older people: A systematic review,” *BMC Public Health*, 2013.

[104] J. M. Veerbeek, E. van Wegen, R. van Peppen, P. J. van der Wees, E. Hendriks, M. Rietberg, and G. Kwakkel, “What is the evidence for physical therapy poststroke? a systematic review and meta-analysis,” *PloS one*, 2014.

[105] I. R. Spremolla, M. Antunes, D. Aouada, and B. E. Ottersten, “Rgb-d and thermal sensor fusion-application in person tracking.,” in *VISIGRAPP (3: VISAPP)*, 2016, pp. 612–619.

[106] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, “Bio-inspired dynamic 3d discriminative skeletal features for human action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[107] C. Wang, Y. Wang, and A. Yuille, “An approach to pose-based action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[108] I. Lillo, A. Soto, and J. Niebles, “Discriminative hierarchical modeling of spatio-temporally composable human activities,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[109] L. Tao and R. Vidal, “Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition,” in *ChaLearn Looking at People Workshop 2015*, 2015.

[110] K. D. Cicerone, D. M. Langenbahn, C. Braden, J. F. Malec, K. Kalmar, M. Fraas, T. Felicetti, L. Laatsch, J. P. Harley, T. Bergquist, *et al.*, “Evidence-based cognitive rehabilitation: Updated review of the literature from 2003 through 2008,” *Archives of physical medicine and rehabilitation*, 2011.

[111] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2012, pp. 20–27.

[112] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in *Workshop on Human Activity Understanding from 3D Data*, 2010.

[113] R. M. C. Aroeira, B. Estevam, A. E. M. Pertence, M. Greco, and J. M. R. Tavares, “Non-invasive methods of computer vision in the posture evaluation of adolescent idiopathic scoliosis,” *Journal of Bodywork and Movement Therapies*, 2016.

[114] D. J. Gladstone, C. J. Danells, and S. E. Black, “The fugl-meyer assessment of motor recovery after stroke: A critical review of its measurement properties,” *Neurorehabilitation and neural repair*, 2002.

[115] H.-F. Mao, I.-P. Hsueh, P.-F. Tang, C.-F. Sheu, and C.-L. Hsieh, “Analysis and comparison of the psychometric properties of three balance measures for stroke patients,” *Stroke*, 2002.

[116] D. Singla and Z. Veqar, “Methods of postural assessment used for sports persons,” *Journal of clinical and diagnostic research: JCDR*, 2014.

[117] S. J. Ray and J. Teizer, “Real-time construction worker posture analysis for ergonomics training,” *Advanced Engineering Informatics*, 2012.

[118] J. R. Wilson and S. Sharples, *Evaluation of human work*. CRC press, 2015.

[119] D. J. Magee, *Orthopedic physical assessment*. Elsevier Health Sciences, 2014.

[120] V. E. Wilson and E. Peper, “The effects of upright and slumped postures on the recall of positive and negative thoughts,” *Applied psychophysiology and biofeedback*, vol. 29, no. 3, pp. 189–195, 2004.

[121] *American posture institute*, `http://americanpostureinstitute.com/`.

[122] *Kinetisense*, `https://kinetisense.com/`.

[123] *Physicaltech*, `http://www.physicaltech.com/adibas-posture/`.

[124] *Zflo motion*, `https://www.zflomotion.com/`.

[125] f. t. S. A. f. E. King’s College London, “Report the burden of stroke in europe,” 2017.

[126] R. Li, B. Lu, and K. D. McDonald-Maier, “Cognitive assisted living ambient system: A survey,” *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[127] M. C. Domingo, “An overview of the internet of things for people with disabilities,” *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 584–596, 2012.

[128] M. A. Musen, B. Middleton, and R. A. Greenes, “Clinical decision-support systems,” in *Biomedical informatics*, Springer, 2014, pp. 643–674.

[129] D. González-Ortega, F. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, “A kinect-based system for cognitive rehabilitation exercises monitoring,” *Computer methods and programs in biomedicine*, vol. 113, no. 2, pp. 620–631, 2014.

[130] P. Wang, I. A. Kreutzer, R. Bjärnemo, and R. C. Davies, “A web-based cost-effective training tool with possible application to brain injury rehabilitation,” *Computer methods and programs in biomedicine*, vol. 74, no. 3, pp. 235–243, 2004.

[131] C.-S. Lin, C.-C. Huan, C.-N. Chan, M.-S. Yeh, and C.-C. Chiu, “Design of a computer game using an eye-tracking device for eye’s activity rehabilitation,” *Optics and lasers in engineering*, vol. 42, no. 1, pp. 91–108, 2004.

[132] R. M. E. M. da Costa and L. A. V. de Carvalho, “The acceptance of virtual reality devices for cognitive rehabilitation: A report of positive results with schizophrenia,” *Computer Methods and Programs in Biomedicine*, vol. 73, no. 3, pp. 173–182, 2004.

[133] J. A. Edmans, J. R. Gladman, M. F. Walker, *et al.*, “Validity of a virtual environment for stroke rehabilitation,” *Stroke*, 2006.

[134] D. Rand, R. Kizony, and P. L. Weiss, “Virtual reality rehabilitation for all: Vivid gx versus sony playstation ii eyetoy,” in *5th Intl. Conf. On Disability, Virtual Environments and Assoc. Technologies*, 2004, pp. 87–94.

[135] C.-S. Lin, T.-C. Wei, A.-T. Lu, S.-S. Hung, W.-L. Chen, and C.-C. Chang, “A rehabilitation training system with double-ccd camera and automatic spatial positioning technique,” *Optics and Lasers in Engineering*, vol. 49, no. 3, pp. 457–464, 2011.

[136] M. I. Vousdoukas, P. Perakakis, S. Idrissi, and J. Vila, “Svmt: A matlab toolbox for stereo-vision motion tracking of motor reactivity,” *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 318–329, 2012.

[137] A. Mirelman, B. L. Patritti, P. Bonato, and J. E. Deutsch, “Effects of robot-virtual reality compared with robot alone training on gait kinetics of individuals post stroke,” in *2007 Virtual Rehabilitation*, 2007, pp. 65–69. DOI: 10.1109/ICVR.2007.4362132.

[138] G. Saposnik, R. Teasell, M. Mamdani, J. Hall, W. Mcilroy, D. Cheung, K. E. Thorpe, L. G. Cohen, and M. Bayley, “Effectiveness of virtual reality using wii gaming technology in stroke rehabilitation: A pilot randomized clinical trial and proof of principle,” *Stroke*, vol. 41, no. 7, pp. 1477–1484, 2010.

[139] R. A. Clark, Y.-H. Pua, A. L. Bryant, and M. A. Hunt, "Validity of the microsoft kinect for providing lateral trunk lean feedback during gait retraining," *Gait & Posture*, vol. 38, no. 4, pp. 1064–1066, 2013, ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2013.03.029. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966636213001884.

[140] A. E. R. Shabayek, R. Baptista, K. Papadopoulos, G. Demisse, O. Oyedotun, M. Antunes, D. Aouada, B. Ottersten, M. Anastassova, M. Boukallel, S. Panëels, G. Randall, M. André, A. Douchet, S. Bouilland, and L. O. Fernandez, "Starr - decision support and self-management system for stroke survivors vision based rehabilitation system," in *European Project Space on Networks, Systems and Technologies - Volume 1: EPS Porto 2017,*, INSTICC, SciTePress, 2017, pp. 69–80, ISBN: 978-989-758-310-0. DOI: 10.5220/0007902400690080.

[141] E. Ghorbel, R. Boutteau, J. Boonaert, X. Savatier, and S. Lecoeuche, "Kinematic spline curves: A temporal invariant descriptor for fast action recognition," *Image and Vision Computing*, 2018.

[142] D. D. N. Natta, E. Alagnide, G. T. Kpadonou, G. G. Stoquart, C. Detrembleur, and T. M. Lejeune, "Feasibility of a self-rehabilitation program for the upper limb for stroke patients in benin," *Annals of Physical and Rehabilitation Medicine*, vol. 58, no. 6, pp. 322–325, 2015, ISSN: 1877-0657. DOI: https://doi.org/10.1016/j.rehab.2015.08.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877065715004996.

[143] C. Bonnyaud, P. Gallien, P. Decavel, P. Marque, C. Aymard, F. Pellas, M.-E. Isner, F. C. Boyer, F. Muller, J.-C. Daviet, P. Dehail, B. Perrouin-Verbe, N. Bayle, E. Coudeyre, D. Perennou, I. Laffont, J. Ropers, N.-Y. Domingo-Saidji, D. Bensmail, and N. Roche, "Effects of a 6-month self-rehabilitation programme in addition to botulinum toxin injections and conventional physiotherapy on limitations of patients with spastic hemiparesis following stroke (adju-tox): Protocol study for a randomised controlled, investigator blinded study," *BMJ Open*, vol. 8, no. 8, B. Parratte, T. Maulet, P. Aegerter, P.

Velou, J. San, Y. Omri, and M. Carrier, Eds., 2018, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2017-020915. eprint: https://bmjopen.bmj.com/content/8/8/e020915.full.pdf. [Online]. Available: https://bmjopen.bmj.com/content/8/8/e020915.

[144] F. Spyridonis, J. Gawronski, G. Ghinea, and A. O. Frank, "An interactive 3-d application for pain management: Results from a pilot study in spinal cord injury rehabilitation," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 356–366, 2012.

[145] M. F. Levin, J. A. Kleim, and S. L. Wolf, "What do motor "recovery" and "compensation" mean in patients following stroke?" *Neurorehabilitation and Neural Repair*, vol. 23, no. 4, pp. 313–319, 2009, PMID: 19118128. DOI: 10.1177/1545968308328727. eprint: https://doi.org/10.1177/1545968308328727. [Online]. Available: https://doi.org/10.1177/1545968308328727.

[146] M. F. Levin, S. M. Michaelsen, C. M. Cirstea, and A. Roby-Brami, "Use of the trunk for reaching targets placed within and beyond the reach in adult hemiparesis," *Experimental Brain Research*, vol. 143, no. 2, pp. 171–180, 2002. DOI: 10.1007/s00221-001-0976-6. [Online]. Available: https://doi.org/10.1007/s00221-001-0976-6.

[147] A. Roby-Brami, S. Jacobs, N. Bennis, and M. F. Levin, "Hand orientation for grasping and arm joint rotation patterns in healthy subjects and hemiparetic stroke patients," *Brain Research*, vol. 969, no. 1, pp. 217–229, 2003, ISSN: 0006-8993. DOI: https://doi.org/10.1016/S0006-8993(03)02334-5. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0006899303023345.

[148] J. L. McGinley, R. Baker, R. Wolfe, and M. E. Morris, "The reliability of three-dimensional kinematic gait measurements: A systematic review," *Gait & posture*, vol. 29, no. 3, pp. 360–369, 2009.

[149] A. E. F. D. Gama, T. M. Chaves, L. S. Figueiredo, A. Baltar, M. Meng, N. Navab, V. Teichrieb, and P. Fallavollita, "Mirrarbilitation: A clinically-related gesture recognition interactive tool for an ar rehabilitation system," *Computer Methods and Programs in Biomedicine*, vol. 135, pp. 105–114, 2016, ISSN: 0169-2607. DOI: https://doi.org/

10.1016/j.cmpb.2016.07.014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260716300542.

[150] G. S. Parra-Dominguez, B. Taati, and A. Mihailidis, "3d human motion analysis to detect abnormal events on stairs," in *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission*, 2012, pp. 97–103. DOI: 10.1109/3DIMPVT.2012.34.

[151] F. Nater, H. Grabner, and L. Van Gool, "Exploiting simple hierarchies for unsupervised human behavior analysis," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2014–2021. DOI: 10.1109/CVPR.2010.5539877.

[152] J. Snoek, J. Hoey, L. Stewart, R. S. Zemel, and A. Mihailidis, "Automated detection of unusual events on stairs," *Image and Vision Computing*, vol. 27, no. 1-2, pp. 153–166, 2009. DOI: 10.1016/j.imavis.2008.04.021.

[153] G. G. Demisse, D. Aouada, and B. Ottersten, "Similarity metric for curved shapes in euclidean space," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[154] ——, "Deformation based curved shape representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1338–1351, 2018, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2017.2711607.

[155] B. Schölkopf and A. J. Smola, "A short introduction to learning with kernels," in *Advanced lectures on machine learning*, Springer, 2003, pp. 41–64.

[156] E. Ghorbel, G. Demisse, D. Aouada, and B. Ottersten, "Fast adaptive reparametrization (far) with application to human action recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 580–584, 2020.

[157] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[158] L. Miranda, T. Vieira, D. Martínez, T. Lewiner, A. W. Vieira, and M. F. M. Campos, "Online gesture recognition from pose kernel learning and decision forests," *Pattern Recognition Letters*, vol. 39, pp. 65–73, 2014.

[159] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 8–13. DOI: 10.1109/CVPRW.2012.6239231.

[160] D. Mumford, "Pattern theory: A unifying perspective," in *Fields Medallists' Lectures*, World Scientific, 1997, pp. 226–261.

[161] G. G. Demisse, K. Papadopoulos, D. Aouada, and B. Ottersten, "Pose encoding for robust skeleton-based action recognition," *CVPRW: Visual Understanding of Humans in Crowd Scene, Salt Lake City, Utah, June 18-22, 2018*, 2018.

[162] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *Transactions on Cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015.

[163] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *CVPRW*, 2012, pp. 14–19.

[164] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[165] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," *JMLR*, vol. 10, pp. 66–71, 2009.

[166] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[167] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *CVPR*, 2014, pp. 2649–2656.

[168] B. Li, O. I. Camps, and M. Sznaier, “Cross-view activity recognition using hankelets,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 1362–1369.

[169] R. Li and T. Zickler, “Discriminative virtual views for cross-view action recognition,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 2855–2862.

[170] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi, “Cross-view action recognition via a continuous virtual path,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2690–2697.

[171] H. Rahmani, A. Mian, and M. Shah, “Learning a deep model for human action recognition from novel viewpoints,” *PAMI*, vol. 40, no. 3, pp. 667–681, 2018.

[172] Y. Kong, Z. Ding, J. Li, and Y. Fu, “Deeply learned view-invariant features for cross-view action recognition,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 3028–3037, 2017.

[173] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, “Action recognition by dense trajectories,” in *CVPR 2011-IEEE Conference on Computer Vision & Pattern Recognition*, IEEE, 2011, pp. 3169–3176.

[174] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “Rmpe: Regional multi-person pose estimation,” in *ICCV*, 2017.

[175] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv*, 2018.

[176] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, “Meta-learning with temporal convolutions,” *CoRR*, vol. abs/1707.03141, 2017. arXiv: `1707.03141`. [Online]. Available: `http://arxiv.org/abs/1707.03141`.

[177] K. Lee, I. Lee, and S. Lee, “Propagating lstm: 3d pose estimation based on joint interdependency,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–135.

[178] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *arXiv:1803.01271*, 2018.

[179] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, “A real-time algorithm for signal analysis with the help of the wavelet transform,” in *Wavelets*, Springer, 1990, pp. 286–297.

[180] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[181] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” *arXiv preprint arXiv:1511.07122*, 2015.

[182] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu, “Neural machine translation in linear time,” *arXiv preprint arXiv:1610.10099*, 2016.

[183] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *CoRR*, vol. abs/1409.1259, 2014. arXiv: 1409.1259. [Online]. Available: http://arxiv.org/abs/1409.1259.

[184] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, “Cascaded pyramid network for multi-person pose estimation,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. DOI: 10.1109/cvpr.2018.00742. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2018.00742.

[185] 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 TPAMI*, vol. 36, no. 7, pp. 1325–1339, 2014.

[186] A. D’Eusanio, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, “Manual annotations on depth maps for human pose estimation,” in *International Conference on Image Analysis and Processing*, Springer, 2019, pp. 233–244.

[187] S. Pini, A. D'Eusanio, G. Borghi, R. Vezzani, and R. Cucchiara, "Baracca: A multi-modal dataset for anthropometric measurements in automotive," in *International Joint Conference on Biometrics (IJCB)*, 2020.

[188] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *ECCV*, 2016.

[189] A. Saint, A. Kacem, K. Cherenkova, and D. Aouada, "3dbooster: 3d body shape and texture recovery," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2020.

[190] A. Saint, A. Kacem, K. Cherenkova, K. Papadopoulos, J. Chibane, G. Pons-Moll, G. Gusev, D. Fofi, D. Aouada, and B. Ottersten, "Sharp 2020: The 1st shape recovery from partial textured 3d scans challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[191] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation.," in *BMVC*, Citeseer, vol. 2, 2010, p. 5.

[192] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3d human pose estimation fusing video and inertial sensors.," in *BMVC*, vol. 2, 2017, p. 3.

[193] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017.

[194] 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 *ECCV*, 2018.

[195] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *CVPR*, 2017.

[196] A. Saint, A. E. R. Shabayek, K. Cherenkova, G. Gusev, D. Aouada, and B. Ottersten, "Bodyfitr: Robust automatic 3d human body fitting," *ICIP*, 2019.

[197] N. Greene, “Environment mapping and other applications of world projections,” *IEEE CG&A*, vol. 6, no. 11, pp. 21–29, 1986.

[198] *Humus Cubemap*, `http://www.humus.name/index.php?page=Textures`, Accessed: 2020-01-29.