Deep state-space modeling for explainable representation, analysis, and forecasting of professional human body dynamics in dexterity understanding and computational ergonomics

Brenda Olivas Padilla

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ PSL

Préparée à Mines Paris - PSL

Deep state-space modeling for explainable representation, analysis, and forecasting of professional human body dynamics in dexterity understanding and computational ergonomics

Modélisation profonde espace-état pour la représentation explicable, l'analyse et la prédiction des dynamiques du corps humain dans la compréhension de la dextérité et l'ergonomie computationnelle

Soutenue par

Brenda Elizabeth OLIVAS PADILLA

Le 17 mars 2023

Dirigée par

Sotiris MANITSARIS

École doctorale n°621

Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité

Informatique temps réel, Robotique et Automatique

Composition du jury :

Amel BOUZEGHOUB Professeur, Télécom SudParis	*Présidente*
Nasser REZZOUG HDR, INRIA Bordeaux	*Rapporteur*
Bill V. BALTZOPOULOS Professeur, Liverpool John Moores University	*Rapporteur*
Brigitte D'ANDRÉA-NOVEL Professeur, UPMC	*Examinatrice*
Patrick HÉNAFF Professeur, MINES Nancy	*Examinateur*
Frederic BEVILACQUA HDR, IRCAM	*Examinateur*
Alina GLUSHKOVA Chercheuse, MINES Paris	*Examinatrice*
Sotiris MANITSARIS HDR, MINES Paris	*Directeur de thèse*

MINES PARIS | PSL

Acknowledgements

I would like to begin by expressing my gratitude to my advisor, Prof. Sotiris MANITSARIS, for his great dedication and guidance, as well as for allowing me to join his team and benefit from their research expertise. I'm also grateful to my reviewers for reading and evaluating the work I've done over the past few years. They provided insightful comments and suggestions to improve this work and the clarity and readability of the manuscript.

I would like to thank everyone at the Centre for Robotics, especially my lab colleagues Alina GLUSHKOVA, Gavrielita SENTERI, Dimitraki MAKRYGIANNIS, Dimitraki MENYCHTAS, Dimitraki PAPANAGIOTOU, Ioanna THANOU, and Agnès AUBERT, for the cherished time we spent together in the lab, social settings, and wonderful trips, as well as for their assistance with various aspects of my thesis.

Many thanks go to my friends for all the great times we shared during this thesis and pandemic. I will always be grateful to José David for his unwavering support and for agreeing to travel with me to the other side of the globe. I am really thankful to my parents, Elizabeth and Adolfo, and my sisters, Jazmín and Anahí, for constantly motivating me to be a better person, tolerating me through my tough periods, and always being there for me. Last but not least, I am grateful to God for the courage, patience, and blessings he has given me.

Abstract

The analysis of human movements has been extensively studied in the past due to its wide variety of practical applications, such as human-robot interaction, human learning applications, clinical diagnosis, and monitoring of human activities. Nevertheless, the state-of-the-art still faces scientific challenges while modeling human movements. Firstly, to model the spatial and temporal dynamics of human movement and accurately predict the evolution of motion descriptors over time, the stochasticity of human movement and the physical body structure must be considered. Second, the explainability of existing deep learning algorithms regarding their predictions still needs to be improved as they lack human-comprehensible representations of human movement. This thesis studies and introduces machine learning approaches for the automatic analysis and representation of human movement. Human movement is formulated as a state-space model of a dynamic system whose parameters are estimated using deep learning and statistical algorithms. The models adhere to the structure of the Gesture Operational Model (GOM), which incorporates spatial and temporal dynamics assumptions in the mathematical representation of human movement. Two novel deep state-space models are presented that model a variety of human movements using nonlinear network parameterization and provide interpretable predictions using the GOM representation. The encoder-decoder structure of the models not only allows them to simulate full-body human movement, but also to disentangle variation factors across distinct movements, cluster related motion descriptors, and identify joint dynamics across sequences. The third method estimates GOM representations using Maximum Likelihood Estimation via Kalman Filters. In contrast to the deep state models, the statistical approach is sufficiently accurate to generate specific human movements utilizing one-shot training. This training strategy enables users to model single human movements and estimate their mathematical representation using simple procedures that require less computational power than data-driven methods. Finally, two applications of the generated models are described. The first is for dexterity analysis of professional movements, where dynamic associations between body joints and meaningful motion descriptors are identified. The second application is for implementing an ergonomically effective task delegation methodology to optimize human-robot collaboration frameworks.

Keywords : Automatic movement analysis; Human motion modeling; State-space modeling; Data-driven learning; Motion capture.

Résumé

L'analyse des mouvements humains a été étudiée de manière approfondie dans le passé en raison de sa grande variété d'applications pratiques, telles que l'interaction homme-robot, les applications d'apprentissage humain, le diagnostic clinique et la surveillance des activités humaines. Néanmoins, l'état de l'art reste confronté à des défis scientifiques lors de la modélisation des mouvements humains. D'abord, pour modéliser la dynamique spatiale et temporelle des mouvements humains et prédire avec précision l'évolution des descripteurs de mouvement, il faut considérer la stochasticité des mouvements humains et la structure physique du corps. Ensuite, l'explicabilité des algorithmes d'apprentissage profond existants concernant leurs prédictions doit encore être améliorée car ils manquent pour la plupart de représentations du mouvement humain compréhensibles par les humains. Par conséquent, cette thèse étudie et présente des méthodes d'apprentissage machine pour l'analyse automatique et la représentation du mouvement humain. Le mouvement humain est formulé comme un modèle d'espace d'état d'un système dynamique dont les paramètres sont estimés à partir d'algorithmes d'apprentissage profond et de statistiques. Les modèles adhèrent à la structure du Gesture Operational Model (GOM), qui intègre des hypothèses sur la dynamique spatiale et temporelle dans la représentation mathématique du mouvement humain. Deux nouveaux modèles profonds d'espace d'état sont présentés, qui modélisent une variété de mouvements humains à partir d'une paramétrisation non linéaire et fournissent des prédictions interprétables en utilisant la représentation GOM. La troisième méthode estime les représentations GOM en utilisant l'estimation par maximum de vraisemblance avec des filtres de Kalman. Contrairement aux modèles d'état profond, l'approche statistique est suffisamment précise pour produire des mouvements humains précis en utilisant des procédures d'entraînement simples qui nécessitent moins de puissance de traitement que les méthodes d'apprentissage profond.Enfin, deux applications des modèles créés sont décrites. La première est destinée à l'analyse de la dextérité des mouvements professionnels, où les associations dynamiques entre les articulations du corps et les descripteurs de mouvement significatifs sont identifiées. La seconde application concerne la réalisation d'une méthodologie de délégation de tâches pour optimiser l'ergonomie des structures de collaboration humain-robot.

Mots clés : Analyse automatique du mouvement; Modélisation du mouvement humain; Modèles d'espace d'état; Apprentissage profond; Capture du mouvement.

Contents

List of figures

List of Tables

Acronyms

Adam Adaptive Moment Estimation

AE Autoencoder

AI Artificial Intelligence

ANN Artificial Neural Network

BVH Biovision Hierarchy format

CNN Convolutional Neural Network

DTW Dynamic Time Warping

EAWS European Assembly Worksheet

ELBO Variational Lower Bound

FC Fully-connected network

GOM Gesture Operational Model

GRU Gated Recurrent Unit

HMM Hidden Markov Model

HRC Human-Robot Collaboration

IMUs Inertial Measurement Units

ISB International Society of Biomechanics

KF Kalman Filters

KL Kullback-Leibler

LSTM Long Short-Term Memory

MLE Maximum Likelihood Estimation

MoCap Motion Capture

PCA Principal Component Analysis

RMSProp Root Mean Square Propagation

RNN Recurrent Neural Network

RULA Rapid Upper Limb Assessment

Seq2Seq Sequence-to-Sequence

SGD Stochastic Gradient Descent

SSM State-Space Model

VAE Variational Autoencoder

WMSDs Work-related Musculoskeletal Disorders

Chapter 1

Introduction

"A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!."

— Lewis Carroll, *Alice in Wonderland*

Contents

1.1 Overview

Movement is an essential component of human life. Through their movements, humans are continually exchanging information and interacting with their surroundings. Human movement is the result of the complex and highly coordinated mechanical interaction between bones, muscles, ligaments, and joints within the musculoskeletal system. Through the study of this interaction and its effects, the structure, function, and motion of human bodies can be examined, and the resulting knowledge be used to improve the quality of life.

In the last decade, the study of human movement has been one of the most interesting and active research areas in various major fields of Artificial Intelligence (AI), including machine learning, robotics, and automated reasoning. Human movement analysis is any method that involves acquiring a quantitative or qualitative measurement of human movements. Biomechanical descriptors, such as force distribution, joint angles, and spatiotemporal parameters, are measured in the quantitative analysis. The qualitative analysis, on the other hand, concentrates on evaluating the technical quality of the movement in order to offer the most appropriate feedback or intervention to enhance performance. The automatic analysis of human movements based on Motion Capture (MoCap) data as a research domain has increased in significance due to the emergence of numerous applications, such as: (a) health, for detecting movement abnormalities; (b) sports, for improving athletes' performances; (c) ergonomic studies, for assessing operational conditions for comfort and productivity; (d) motion-driven user interfaces, for creating intuitive human-machine interfaces; (e) intelligent surveillance, that automatically monitors individuals and detects irregular activity; and (f) virtual reality, to animate virtual characters. The above examples illustrate human motion research's impacts on society and the economy.

Computer-based motion analysis serves the same purpose as a trainer, ergonomist, or other specialist who objectively examines motions. In order to do this, motion data must first be segmented. Then the tracked motion data must be mapped into meaningful motion descriptions that a scientist, specialist, or user can interpret. Depending on the motion-related application, statistical models or data-driven approaches, like machine learning or deep learning algorithms, can be used to model (or map) human motion data. However, it remains complex and requires overcoming scientific challenges to design an accurate and versatile automatic analysis tool to describe human motion dynamics based on MoCap data. The complexity of motion data has often led scientists to seek new approaches that can capture the spatial and temporal dynamics of the human body by accounting for the stochastic nature of human movement and the physical structure of the human body. By learning latent spatiotemporal representations from the data, current data-driven methods have successfully modeled human movements. Although they can simulate human movements accurately, their usefulness is limited by their inability to be debugged and to explain their results in a way that is understandable to humans. Fitting analytical models to motion data may be a more beneficial approach for modeling human movements, as they do not suffer from this limitation. Analytical models represent systems using a set of mathematical equations that define parametric relationships and their corre-

sponding parameter values as a function of time, space, and other system parameters. Insights into the system's dynamics may be derived from a close examination of the model's content and the effect of the model's parameters.

Analytical models can incorporate assumptions about the stochasticity of human movement and the mediations of body joints to properly simulate and explain the evolution of human motion descriptors across time, enabling proactive use of this information. For instance, in human-centered AI technologies where the physical embodiment of humans is the central focus (human-robot collaboration, risk monitoring, or dexterity analysis). Understanding and capturing the dependencies between the motion of different joints is crucial not only for creating more realistic human motion simulations, but also for investigating how diverse and intricate full-body human movements are performed. Knowledge of the neurophysiological mechanisms behind complicated dexterity and motor learning may be gleaned from the models. Eventually, the use of such analytic models may enable the development of interdisciplinary frameworks for the research of the process of learning and skill acquisition while performing professional tasks in the industrial or craft sectors, in dancing, or while playing an instrument. In addition, they might facilitate research into the key factors that lead to musculoskeletal disorders in ergonomics. Some possible applications that could be developed with these models are, for instance, while teaching the craft of glassblowing, an interface or collaborative robot that could teach an apprentice the movements and sub-processes for creating a particular glass piece. It can automatically assess the learner's movements to predict the likely outcome of the piece. Therefore, the interface can suggest adjusting the posture or motion of a single body part or the entire body if it is determined that the current performance may affect the desired outcome. Likewise in the industry, for example when assembling airplanes, a collaborative robot can hold the airplane structure while an operator sets the rivets. Meanwhile, the robot can also continuously assess the operator's movements and place the structure in an ergonomic position so that the operator does not adopt uncomfortable postures while working, reducing the risks of musculoskeletal disorders in the long run.

Using analytical models makes it relatively straightforward to build a representation of human movement that considers the human body's biomechanical structure and the stochastics of motion. However, fitting these models to motion data is far more challenging. One of the most prominent analytical models in analyzing different time-series data, including MoCap data, is State-Space Models (SSMs). SSMs combine a transition model, a hypothetical mechanistic description of human movement, with an observation model. The observation model provides the probability of obtaining a certain observation given the human's actual posture. For the training of these models, one of the various Maximum-Likelihood Estimation (MLE) methods or Bayesian simulation is commonly used. The Kalman filter is one of the most effective and common methods for analytically estimating MLE. This approach is computationally efficient, but the data distribution is assumed to be normal and linear, which does not often apply to human movements. In addition, as stated before, MoCap data can be complex and high-dimensional, making conventional estimation methods inadequate for many applications and complicated to apply to large datasets. Deep neural networks may be a viable alternative

for estimating nonlinear and non-Gaussian SSMs and can handle large MoCap datasets. An additional major benefit of artificial neural networks over conventional statistical methods is their greater modeling capacity and ability to extract higher-order features. This allows for the identification of intricate patterns inside and across time series, as well as the usage of raw time series, thereby minimizing the amount of human effort necessary for feature selection when dealing with statistical approaches.

Consequently, this thesis focused on the analytical modeling of human motion dynamics. Exploring the estimation of movement parameters with the end goal of developing a generalized motion understanding approach. The analysis is done over global motion patterns (full-body) rather than only local patterns such as hand gestures or facial expressions. For a realistic simulation of human movement, the appropriate transition model was investigated, which should also describe the phenomenon or relationships between spatial and temporal assumptions. The spatial assumptions must account for the potential interdependencies between the linked joints within the articulated skeletal structure. On the other hand, temporal assumptions involve the fundamental principle that most time series, such as human movements, inherently exhibit that is the dependency between adjacent observations. Being said, the proposed model must be a sufficiently precise representation of the dynamic system that is the human movement to meet the previous goals (accurate human motion simulation and understanding of movement performance). The following two hypotheses were formulated to guide this research:

Hypothesis 1 Human motion dynamics can be modeled analytically by taking into account the interdependencies between joints as well as the dependencies between their prior values.

Hypothesis 2 The cooperation of body joints and their contribution during the performance of a human movement can be learned and represented through interpretable models.

This investigation resulted in the creation of novel deep state-space models to represent and simulate human movement. In these models, the nonlinear parameterization of SSMs was accomplished using data-driven approaches with encoder-decoder architectures. These architectures have demonstrated their ability to process time-series data and accurately model and forecast object or human motion trajectories. Comparisons were made between the human motion representations generated from deep state-space models and those estimated using one-shot training (MLE via Kalman filters), along with their simulation performance. The advantages and disadvantages of each approach were examined, as well as the potential applications of the proposed analytical models.

Lastly, even though MoCap technology adds significant value to the analysis by providing measurements that cannot be identified from observation, such as detailed information about the movement's biomechanics, using a full-body MoCap suit or precise optical systems is costly and impractical to implement in real-world workplace scenarios. For this reason, the application of the proposed analytical models to identify the minimal number of body parts to capture for proper modeling and analysis of specific sets of human movements is investigated. Continuous monitoring of human movements in an everyday context using a minimal number of inertial

sensors might enable the measurement of valuable and complementary information to that gained through laboratory experiments.

1.2 Objectives

To address each of the formulated hypotheses, specific objectives were defined. For the first hypothesis:

1. **Investigate statistical and data-driven approaches for parameterizing mathematical representations of human movement:** The analytical model that would be trained with full-body movements is designed first. This should simplify the musculoskeletal system for straightforward interpretation, while incorporating relevant assumptions regarding human body dynamics. Then, statistical and data-driven approaches would be explored to capture full-body motion patterns and estimate the model's parameters.

2. **Evaluate the capability of the proposed approach to model and simulate specific human movements statically and dynamically:** The performance of the analytical model to simulate statically or dynamically realistic human movements is evaluated via a series of experiments. The static simulation implies that all inputs of the model are real data samples, whether, in the dynamic simulation, the input corresponds to previous predictions.

Second hypothesis:

1. **Estimate the significance of various body dynamics occurring during a human movement:** The trained human motion models are subjected to a statistical analysis to uncover relevant associations between joint motion descriptors (spatial dynamics) and their dependency on previous transitions (temporal dynamics). Then, based on the significant assumptions, it is deduced how body joints collaborate to accomplish a specific movement.

2. **Identify the optimal joints to measure for maximizing the recognition of human movements:** On the basis of the trained human motion models, a methodology is developed for determining the optimal set of sensors for accurately recognizing a set of human movements. The purpose is to validate the possibility of identifying a minimal set of sensors that could be more practical for everyday motion-related applications.

Several choices have been made in order to restrict the scope of this thesis while still attempting to achieve its objectives. First, there are two alternative approaches for studying human movement: kinetic analysis and kinematic analysis. A kinetic analysis investigates the forces and joint torques involved in the performance of a human movement, as well as the manner in which they cause this movement. In contrast, the kinematic analysis concerns more about the movement itself and explains it in terms such as acceleration, velocity, joint positions, or joint angles. As this work focuses on creating approaches for explaining professional

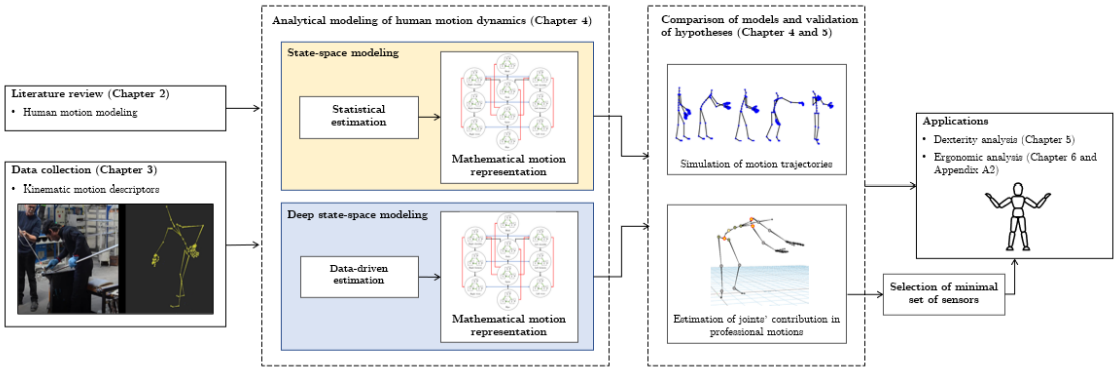


Figure 1.1: Flow diagram illustrating the research conducted towards the creation and validation of motion-based methods to simulate and describe human movement.

movements, virtually simulating them, and identifying the minimal number of sensors for their accurate analysis in real-world scenarios, the kinematic analysis was selected. Furthermore, kinematic analysis permits using only simple technologies, such as inertial sensors, contrary to kinetic analysis, which requires force sensors or plates for accurate analysis.

Figure 1.1 illustrates a diagram of how this thesis proceeded to achieve the defined objectives, beginning from the literature review and generation of datasets to applications of the proposed approaches. Industrial operators, craftsmen, and subjects were recorded using inertial-based motion capture technology, from which relevant kinematic motion descriptors were extracted. The extracted motion descriptors were used for the modeling and automatic analysis of human motion dynamics. The purpose of creating new datasets was to test the models using professional movements captured from real-world scenarios. In this thesis, three novel approaches for generating interpretable human motion models were developed, which are subsequently presented in Chapter 4.

In addition to accurately simulating human movement, the generated models can be utilized for other applications, two of which are detailed and evaluated in this dissertation. The first is for the dexterity analysis of professional tasks performed in industrial settings and traditional crafts. The second use is for computational ergonomics, in which the findings of the presented methods are utilized to construct a pipeline for ergonomically optimizing a workplace scenario.

1.3 Contributions

The following is a summary of this thesis' major contributions.

Interpretable models for analyzing and simulating human movement

This thesis proposes novel approaches for creating explainable human motion representations. Current data-driven methods have been effectively trained to produce realistic human motion simulations. However, there is a lack of methods that can as well explain the reasoning behind their predictions in a way that a human can interpret. Therefore, three approaches were developed that follow a state-space representation and incorpo-

rate assumptions about the stochasticity of human movement and the mediations of body joints. For the parametrization of the SSMs, the first method does one-shot training using the MLE via Kalman filters. The other two use data-driven approaches with encoder-decoder architectures. One encodes the time series into a latent space with a form of a Gaussian distribution for probabilistic prediction, and the other utilizes an attention mechanism to capture state dynamics. The statistical approaches are the most computationally efficient and do not require the use of a large dataset in order to model and simulate single human movements accurately. Regarding the data-driven, these can be scaled to create representations of a greater variety of human movements, as they can process large datasets of human movements. Further contributions are presented and discussed in Chapter 4.

Analysis of the dexterity of industrial operators and skilled craftsmen

This thesis analyzes the human movements performed in the industrial and craft sectors by interpreting the proposed motion representations. The method involves examining the learned parameters of the temporal and spatial assumptions incorporated into motion representations to gain insight into how experts perform a movement. In addition, a method for identifying the most significant joint motion descriptors for modeling and recognizing a set of human movements is proposed. This knowledge can then be utilized to determine the ideal sensor configuration for human motion recognition problems. Chapter 5 presents and discusses further contributions, such as the creation of tolerance intervals for evaluating human motion performance based on expert performers.

Computational ergonomics for task delegation in industrial settings

This thesis presents a methodology for effective task delegation while integrating Human-Robot Collaboration (HRC) frameworks in a manufacturing cell. The task delegation is based on the automatic ergonomic analysis of the professional tasks, where the ergonomic scores of the tasks are estimated based on the postural risk factors detected. In order to be able to be integrated into real industrial applications, the proposed algorithm can accurately compute ergonomic scores of human movements using a minimal set of sensors. Additional contributions are presented and discussed in Chapter 6.

Motion capture benchmark of industrial tasks and European historic crafts

This thesis presents a motion capture benchmark composed of seven inertial-based MoCap datasets. These include movements performed by actual industrial operators and skilled artisans. According to the datasets found, most are composed of movements executed during everyday activities, sports, or dances and are captured in a laboratory setting. Consequently, these seven datasets are among the few that contain MoCap recordings of professional motions captured in real workplaces. Chapter 3 includes further information regarding the development of the benchmark.

1.4 Thesis outline

This manuscript consists of six chapters, the conclusions, and one appendix:
Introduction describes the context of this thesis by providing a brief overview of the research on human movement analysis as well as its current challenges. The formulated hypotheses and objectives are then presented, along with a summary of the contributions and the thesis outline.

Chapter 2 provides the required context for the rest of this thesis. First, relevant research in human motion modeling is reviewed, starting with motion capture technologies and motion descriptors. Next, an overview of the various methodologies for modeling human motion and their principal applications is presented. These methods may be divided into four categories: biomechanical modeling, stochastic modeling, hybrid stochastic-biomechanical models, and data-driven approaches for sequential modeling. Stochastic-biomechanical models have demonstrated their ability to improve the simulation of human movements, over simple stochastic and biomechanical models, and to generate interpretable mathematical representations. However, they lack the robustness and scalability of data-driven approaches for applications requiring the analysis and modeling of multiple human movements. For nearly all human motion modeling problems, data-driven approaches have supplanted traditional methods such as biomechanical or stochastic, where feature selection is crucial. Nonetheless, studies of data-driven approaches that accurately simulate human movements and are explicable in terms of their models and results are still scarce.

Chapter 3 presents the motion capture benchmark with the datasets used in the experiments reported in this thesis. These datasets were collected using inertial-based motion capture and are composed of movements performed by industrial operators and skilled craftsmen. The professional movements were collected with the intention of being used for research in human movement analysis and modeling. The recording and processing procedures are described in detail, as well as the movements captured.

Chapter 4 introduces the three novel approaches for modeling human movements through interpretable mathematical representations. The first approach uses statistical modeling (KF-RGOM) to estimate the motion parameters of a state-space system, whereas the second and third apply data-driven approaches (VAE-RGOM and ATT-RGOM). The motion representations follow the hybrid stochastic-biomechanical structure of GOM to model the dynamics of human movements. All three approaches estimate time-varying motion representations, which can be used to simulate human movements statically or dynamically. Moreover, the motion representations can be used to obtain insights into the dynamic relationship between body joints during the execution of a movement. Experiments proved Hypothesis 1 and revealed that employing time-varying representations and adding extra exogenous variables increases the models' robustness for accurately simulating various human movements. In addition, ac-

cording to a sensitivity analysis of the generated models, the models exhibited tolerance to external perturbations.

Chapter 5 details the application of the motion representations estimated in Chapter 4 for dexterity analysis in order to demonstrate Hypothesis 2. The models are statistically analyzed, and findings are utilized to assess the importance of each model's assumptions regarding the body part associations specified within the motion representation. Calculating the statistical significance of joint motion descriptors allowed identifying the most meaningful for the modeled human movement. A set of sensors that provide significant motion descriptors was then selected using the motion representation estimated per each approach. The selected sets were validated based on their ability to improve the recognition performance of gesture vocabularies from seven distinct datasets. The recognition performance using motion data from the selected sensors was compared to that attained using all sensor data and data from a minimal setup of two hand-picked sensors. In conclusion, the motion representations demonstrated their ability to capture and describe human movements based on their assumptions. Moreover, the performance of each approach's selected sensors exceeded or matched that of the set comprising all sensor data. Thus, it was feasible to determine the motion descriptors that solved each recognition problem most effectively.

Chapter 6 describes the proposed methodology for task delegation to design HRC frameworks that improve ergonomics in manufacturing applications. The formulated hypothesis is that operators' movements can be properly evaluated using the motion data captured with a minimal set of sensors. This enables a more thorough ergonomic analysis of their activities and facilitates task delegation when implementing HRC frameworks. First, a system for detecting four postural risk factors is created. The automated posture evaluation system is composed of Hidden Markov Models (HMMs) that learned to recognize the motion patterns caused by exposure to the postural risk factors. Based on the detected risk factors, an ergonomic risk score is calculated according to the European Assembly Worksheet (EAWS). The potentially dangerous tasks (high ergonomic scores) are then proposed to be delegated to a collaborative robot, and the rest safe tasks to the operator. The methodology was evaluated by examining professional tasks performed in a television manufacturing process.

Conclusions closes with a discussion of the thesis's main contributions and open questions and directions for further research.

Appendix A outlines web-based and Android-based applications developed for automated ergonomic evaluation using MoCap data. The proposed applications evaluate body segment rotations collected by inertial sensors and provide simple, intuitive, and meaningful feedback in the form of ergonomics scores, color visualizations, and limb angles. The scores are based on the Rapid Upper Limb Assessment (RULA), one of the most popular observational methods for assessing occupational risk factors for upper-extremity musculoskeletal disorders. By

automating RULA, an interesting perspective is created for extracting posture analytics for ergonomic evaluation and incorporating complementary features. Future work consists of implementing other human motion analyses, such as the dexterity analysis described in Chapter 4, and making the applications compatible with data from optical motion capture systems.

Chapter 2

Background and Related Work

"Study hard what interests you the most in the most undisciplined, irreverent and original manner possible."

— Richard Feynman

Contents

2.1 Introduction

This chapter reviews the fundamental concepts and relevant studies in human motion modeling. To start, the main motion capture technologies utilized for the automatic analysis of human movements are presented in Section 2.2. Over the last two decades, MoCap technologies have advanced significantly, particularly those designed for recording human movement. With these advancements, it has become more practicable to collect human MoCap data. MoCap or motion data can be defined as time series samples describing the spatial configuration of a set of physical features of interest. These physical features can be expressed as kinetic or kinematic motion descriptors. Section 2.2.1 describes the most basic motion descriptors used for modeling human movement. In addition, feature extraction strategies that have been applied to MoCap data to improve modeling performance are discussed.

Following is an overview of classical methodologies for human motion modeling. Although a more extensive description of previous works relevant to the approaches studied in this dissertation (stochastic and hybrid biomechanical-stochastic modeling) is offered, recent state-of-the-art works on biomechanical modeling are also presented to give a general idea about how the field has evolved. An introduction to data-driven approaches for sequence modeling is provided next in Section 2.4. First, approaches that have effectively processed human motion data, such as deep temporal and generative models, are presented. Then, the most relevant architectures are explained in detail to build a ground for this thesis work. Finally, Section 2.5 closes with the general conclusions of the reviewed state-of-the-art, which led to the development of the proposed models in Chapter 3.

2.2 Motion capture technologies

Extracting accurate and unbiased data based on quantitative measurements is critical for accurate human motion analysis. The ability to provide analysts with quantitative measures of human motion performance represents an added value to subjective observation measurements. Diverse disciplines have utilized motion capture systems to record and reconstruct the human body's posture and movement in order to study it. Some of the most popular applications of acrshortMoCap are for sport and medical sciences, filmmaking, and human-computer interaction.

Systems based on optical markers or markerless are among the most commonly used for capturing full-body human movements. In optical marker systems, markers are placed on anatomical parts to track movement, as illustrated in Figure 2.1a. Some examples of these specialized systems are Vicon[1] or Optitrack[2]. These optical marker systems can offer high positional precision in controlled recording environments, e.g., multiple fixed cameras calibrated and correlated in a specific area and capturing configuration. In contrast, markerless systems (Figure 2.1b)are more ambulatory systems that identify body segments' motion using image features. These systems can be used in relatively uncontrolled environments; however, their

[1]Vicon Motion Systems Ltd., Oxford, UK
[2]NaturalPoint Inc., Corvallis, USA

(**a**) (**b**)

Figure 2.1: Examples of MoCap recordings using optical systems. (a) Marker system [Feldmann, 2019]; (b) Markerless system, where the body posture tracking is done using the pose estimation algorithm OpenPose [Cao, 2019].

positional precision depends greatly on the light conditions, have a restricted field of view, and can suffer from occlusions [Busch, 2017; Manghisi, 2017; Von Marcard, 2016; Sharma, 2019]. With the arrival of microelectromechanical systems (MEMS), miniature Inertial Measurement Units (IMUs), composed of a tri-axial accelerometer, gyroscope, and a three-axis magnetometer, have emerged as wearable MoCap devices for both localization and posture tracking [Kok, 2014]. The accelerometer measures the external specific force exerted on the sensor, composed of gravity and the sensor's acceleration. The gyroscope measures the sensor's angular velocity, which is the rate at which its orientation changes. The magnetometer is used to determine the direction of magnetic fields to provide a stable heading (yaw) angle over time. Combining the data from these three components makes it feasible to extract information about the pose and orientation of any item or body region to which the inertial sensor is rigidly attached to. However, localization based on IMUs has the issue that, in some environments, it exhibits a drift over time due to the integration of non-constant accelerometer errors resulting from the presence of magnetic disturbances during the recording. The drifting can be removed by recalibrating the system or post-processing the motion data. Currently, exits MoCap suits based on IMUs that already post-process the inertial data and provide a biomechanical skeleton from which joint angles and positions may be easily extracted. Figure 2.2 depicts a recording using an inertial-based MoCap suit[3], in which the system precisely captures the person's whole posture.

In order to develop robust and practical human movement analysis methods, the ideal is to employ basic technologies that are applicable for daily monitoring, unobtrusive, and reasonably affordable. Nearly all of these requirements are satisfied by inertial MoCap systems for applications that do not demand precise positional measures of the user but require accurate body posture measurements. IMUs have been successfully used in clinical practice for gait analysis [Takeda, 2009; Martinez-Hernandez, 2018], Parkinson disease screening and diagnosis [Caramia, 2018], fall detection [Zhu, 2015], and analysis of motor impairments [Otten, 2015]. They have also been proven useful for automating the evaluation of body postures and activities in ergonomic studies, which required the accurate measure of the full-body posture to detect

[3]The Nansense Inc.'s full-body suit (Baranger Studios, Los Angeles, CA, USA)

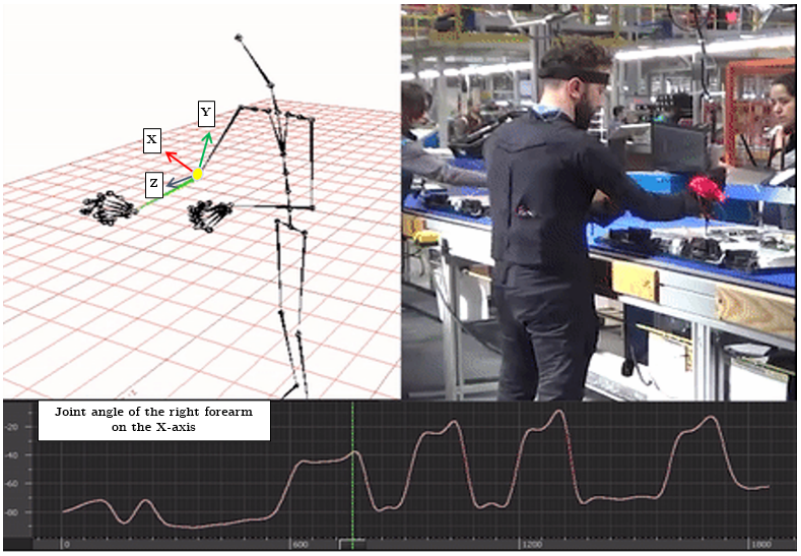


Figure 2.2: Example of an inertial-based MoCap suit recording. On the left is shown the person's captured posture. At the bottom, it is a plot of the joint angle captured from the right forearm on the X-axis.

ergonomic risks [Vignais, 2013; Nath, 2017; Caputo, 2019]. Because of the previous remarks, this thesis focused on developing methods that use motion descriptors derived from inertial sensors. The following section details some motion descriptors or features that can be extracted from motion data collected with inertial sensors.

2.2.1 Human motion descriptors

Past studies have extracted diverse descriptors of human movement from inertial data in order to model and analyze human movement. Depending on the goal of the analysis, either kinematic or kinetic descriptors were utilized. The two sections that follow provide examples of motion descriptions that have been used for human movement analysis. Then, given MoCap data is often multidimensional since the human body has many degrees of freedom, Section 2.2.1.3 discusses feature extraction techniques that have been used to minimize the dimensionality of MoCap data, extract new features, and enhance the modeling of human movements.

2.2.1.1 Kinematic

Motion capture systems provide a broad set of kinematic features that can be used as input for human movement analysis. In order to estimate body segment kinematics using IMUs, the subject's movement is first recorded using a sensor network comprised of several synchronized IMUs linked to a biomechanical model. The placement of the sensor is determined by the objectives of the study (relevant segments or joints to be modeled) and the biomechanical modeling criteria (trunk inclination, upper or lower body posture estimation). Prior to the MoCap recording, a calibration method is applied to reduce sensor placement variations and

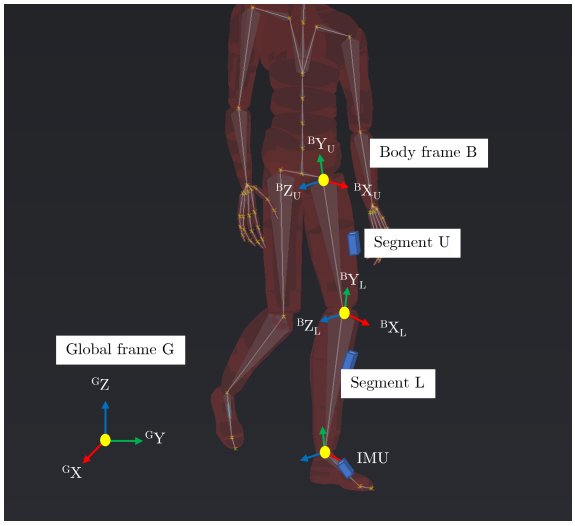


Figure 2.3: Body segment coordinate systems in the global coordinate system.

align the sensor's axis with the anatomical axis of the body segment. Calibration methods entail instructing participants to perform various postures (I-pose, T-pose, bending forward) for a specified period of time in order to establish a benchmark signal reading. The average of the IMUs signals captured over that time period is then subtracted from the raw sensor data received.

Next, depending on the sensor fusion technique (extended Kalman filter and variations) [Alatise, 2017; Bancroft, 2011; Luinge, 2005], a global (earth-fixed) reference coordinate system G is used to calculate the position, velocity, acceleration, orientation, angular velocity, and angular acceleration of each body segment. As an illustration, Figure 2.3 is provided along with a brief explanation of the computation of kinematic descriptors for the body frame B (shown in Figure 2.3). Firstly are defined the body segments' coordinate system, where most MoCap systems use the standards of the International Society of Biomechanics (ISB) [Ferrari, 2002; Wu, 2005; Roetenberg, 2009]. This means that when a person is standing in the anatomical posture, the Y-axis of a body segment points up. All joint angle representations (Euler [Woltring, 1994], joint coordinate system [Grood, 1983], helical angle [Challis, 1995]) can be then derived from the joint rotation matrices (or quaternions). The rotation matrices contain the coordinates of the rotated axes in the reference frame axes. In the case of B, this joint rotation matrix is denoted as ${}^{B}r_{UL}$ and is often described as the orientation of a distal segment ${}^{GB}r_{L}$ relative to a proximal segment ${}^{GB}r_{U}$:

$$ {}^{B}r_{UL} = {}^{GB}r_{U}^{*} \otimes {}^{GB}r_{L} \tag{2.1} $$

The rotation sequence can be based on the ones indicated by the ISB for the lower [Wu, 2002] and upper body segments [Wu, 2005]. The segmental lengths are determined using an anthropometric database, with the subjects' height as input [Leva, 1996; Dumas, 2007]. Velocities and accelerations may then be calculated based on the positions and angles of each body segment.

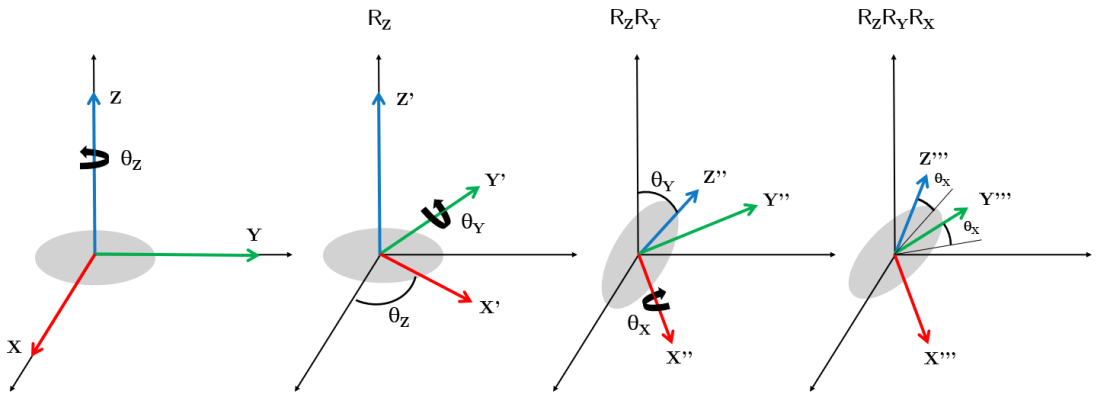


Figure 2.4: Rotation using Euler angles and with the convention ZYX.

Rotation parameterizations The most frequent rotation parameterizations in the literature on modeling and analyzing human movements are the Euler angles, quaternions, and axis-angle (sometimes referred to as exponential map) [Du, 2016; Zhou, 2019; Pavllo, 2020]; hence they are described in further detail next. The most popular parameterization of orientation space is Euler angles. Additionally, it is the simplest to interpret intuitively by visualizing the movement of a joint. A general Euler rotation of a body joint is defined as a sequence of rotations around three mutually orthogonal coordinate axes fixed in the space (X, Y, and Z). The Euler angle values are calculated in accordance with the rotation convention ("ZYX", "ZXY", "ZYZ", or "XYZ") specified. Figure 2.4 depicts an example of rotation using Euler angles, utilizing rotations $R(\theta_X, \theta_Y, \theta_Z)$ and the convention ZYX. First, the original Z-axis is turned by θ_Z, then the Y-axis is rotated by θ_Y, and lastly, the X-axis is rotated by θ_X. There are two well-known restrictions associated with the use of Euler angles. The first is the Euler Angle singularity, which happens when the angle of the second rotation is 180° or -180°. In certain cases, the first and third Euler angles may change independently, controlling the same degree of freedom, resulting in an unlimited number of potential combinations for defining a single orientation. The second issue is the gimbal lock, which also arises for the same values during the second rotation. Due to the alignment of two rotating axes, a degree of freedom is lost, preventing the system from executing predetermined movements. Depending on the intended uses, these constraints may be either rectified or constitute a major problem.

Quaternions provide an alternate measurement technique that is not susceptible to gimbal lock. These have mostly been used in computer graphics to represent rotation [Vince, 2011]. Quaternions are hypercomplex numbers with a real component and three imaginary components that describe a rotation in three degrees of freedom. The definition of a quaternion is as follows:

$$q = [q_w, q_x, q_y, q_z] \tag{2.2}$$

In the previous equation, q is the quaternion vector, q_w is the quaternion's real component, and q_x, q_y, and q_z are the quaternion's imaginary components. Due to the fact that unit quaternions are free from gimbal lock, quaternions must be normalized to attain this property.

Nonetheless, they continue to have the same singularity issue as the Euler angles. The axis-angle or exponential map is, in general, a reparametrization of a quaternion that maps a three-dimensional rotation vector v into a unit quaternion:

$$q = exp(v) = \begin{cases} [0,0,0,1]^T & \text{if } v = 0; \\ \left[sin(\frac{1}{2}\theta)\frac{v}{\theta}, cos(\frac{1}{2}\theta)\right] & \text{if } v \neq 0. \end{cases} \tag{2.3}$$

where $v \in \mathbb{R}^3$ and $\theta = \|v\|$. The advantage of exponential map is that it linearizes quaternions. But, because an exponential map involves a quaternion conversion, the singularity problem persists. Compared to Euler angles, quaternions and exponential maps are less intuitive, and their math can be a bit more challenging. This thesis uses Euler angles to create more comprehensible motion representations in which links between the modeled movement and its dependencies on joint movements along particular axes can be more easily interpreted. On the other hand, it is acknowledged that the performance of these representations for regression or simulation of human movement may be severely affected by singularity and gimbal lock issues that can arise on certain joints, particularly when measuring shoulder motion. Finally, it is important to note that there is no ideal rotation parameterization for all applications, and in certain ways, all are comparable since each has an equivalent rotation matrix representation.

Euler joint angles are frequently employed as motion descriptors in human movement analysis. In previous studies, joint angles were utilized to identify uncomfortable postures and mitigate ergonomic risks [Vignais, 2013; Álvarez, 2016; Lee, 2017; Yan, 2017]. The joint angles were evaluated in terms of their deviation from neutral posture or used to assess work-related movements (e.g., lifting, carrying, dropping, pushing, and pulling). Other descriptors, including joint positions, IMUs accelerations, angular velocities, as well as their statistical descriptions, have also been used to classify everyday activities [Hsu, 2018; Sousa Lima, 2019], or movements with varying levels of ergonomic risk [Nath, 2018; Malaisé, 2019].

2.2.1.2 Kinetic

Forces acting on (or produced by) a worker or athlete can provide valuable information regarding their performance and risk of injury. Combined with kinematic measurements, force data have been used to analyze the mechanical loading of joints. In the ergonomic analysis of human movements, for instance, studies have examined the mechanical loading on the L5/S1 joint to determine the level of risk associated with a lifting activity in relation to a person's capacity limits. Usually, in a motion analysis laboratory, force plates and optical MoCap systems are used to determine the moment (or torque) and force on the L5/S1 joint. However, due to the impracticality of using this approach in a real-world work setting, earlier studies have tried to estimate the L5/S1 joint's moment and force using solely IMUs data [Muller, 2020; Faber, 2016; Shojaei, 2016].

The basic process for calculating the moment and force on the L5/S1 is illustrated in Figure 2.5, and is explained next as an example for the computation of any joint's moment and force. To begin, body segment parameters are computed for the top-down inverse dynamic algorithm

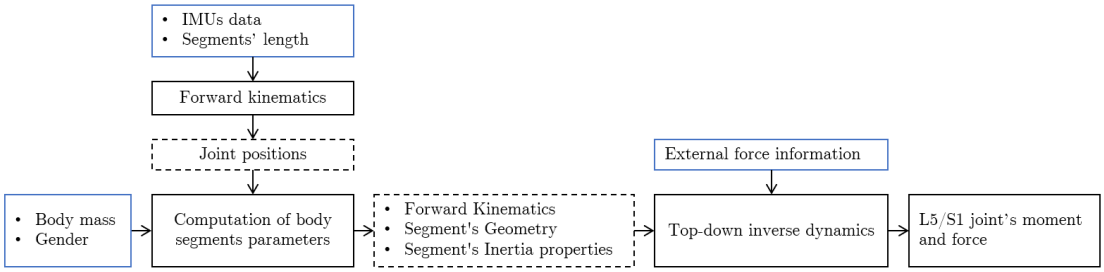


Figure 2.5: Methodology for the estimation of L5/S1 joint's moment and force.

used to predict the L5/S1 joint's moment and force. A forward kinematics model is defined to obtain the joint positions for the computation of the body segment parameters. Given the subject's joint positions, gender, and total body mass, all body segment parameters such as segment mass, position of the centre of mass, and inertia tensors are estimated using values from anthropometric tables [Leva, 1996; Dumas, 2007]. The length of a segment i is defined as the Euclidean distance between its distal and proximal joints l_i. The proximal joint is the one closest to the torso, whereas the distal joint is the one that is further away from the torso. For example, for the forearm, the distal joint is the wrist and the proximal joint is the elbow. Using the overall mass of the subject M, the mass of the segment m_i is calculated as follows:

$$m_i = \overline{r}_i^m \times M \tag{2.4}$$

where $\overline{r}_i^m$ is the mean relative mass, obtained from anthropometric tables [Leva, 1996; Dumas, 2007]. The center of mass CoM_i is located on the link that connects the corresponding distal $p_{ds(i)}$ and proximal $p_{pr(i)}$ joints, and it is calculated based on the mean longitudinal distance of the CoM_i from its proximal joint ($\overline{r}_i^{cm}$):

$$CoM_i = p_{pr(i)} + \overline{r}_i^{cm} \times \left(p_{ds(i)} - p_{pr(i)}\right) \tag{2.5}$$

Finally, the inertial tensor of the segment T_i is calculated in the following equation:

$$T_i = m_i \times (l_i \times \overline{r}_i)^2 \tag{2.6}$$

where $\overline{r}_i = \left[\overline{r}_i^x, \overline{r}_i^y, \overline{r}_i^y\right]$ is the mean relative radius of gyration on each axis [Leva, 1996; Dumas, 2007]. By having the segments' geometry and inertia properties, a top-down inverse dynamic model is used to compute the joint kinetics from the joint kinematics extracted from the IMUs data (angular velocities and accelerations). A global equation of motion is applied to estimate net forces F_{L5S1} and moments M_{L5S1} at L5/S1 joint in the global coordinate system:

$$F_{L5S1} = -F_r - \sum_{i=1}^{k} m_i g + \sum_{i=1}^{k} m_i a_i \tag{2.7}$$

$$M_{L5S1} = -(r_r - r_{L5S1}) \times F_r - \sum_{i=1}^{k} [(r_r - r_{L5S1}) \times m_i g] + \sum_{i=1}^{k} [(r_r - r_{L5S1}) \times m_i a_i] + \sum_{i=1}^{k} T_i \epsilon_i \quad (2.8)$$

where r_r and r_{L5S1} denote the vectors pointing to the external force and L5/S1 joint positions, respectively; g is gravity; F_r denotes the external force vector; r_i corresponds to the vector to the CoM_i; k is the number of segments of the upper body up to L5/S1 joint (e.g., head, trunk, upper arms, and forearms); finally, a_i and ϵ_i are the linear and angular acceleration vectors of the CoM_i. As demonstrated in the preceding two equations, external force data is required to calculate F_{L5S1} and M_{L5S1}. External force information can be estimated using the top-down or bottom-up models. The top-down approach can include the mass and acceleration of the object or tool the subject is carrying during the movement. In the bottom-up model, on the other hand, force plate data can be used to determine external forces and their application points. In an on-site biomechanical study, the top-down model can be more practical than a bottom-up model for applying inverse dynamics, as force plates are not required.

2.2.1.3 Manual and automatic feature extraction

Dimensionality is one of the most fundamental issues with generic solutions based on statistical models or machine learning when dealing with MoCap data. In the most general case, all joint angles in the human body, which have numerous degrees of freedom, are predicted while simulating or predicting human movements. This makes the approach computationally demanding as well as potentially unstable due to the high-dimensional spaces. The most straightforward solution to this issue is to reduce the number of degrees of freedom in the model, which has been accomplished through diverse methods.

Research has consistently demonstrated an awareness of the problems of utilizing only summary metrics (e.g., mean acceleration or velocity) to characterize human movements, as they are insufficient to capture the variability in MoCap data. An alternative to preserve the variability of multivariate MoCap datasets while reducing dimensionality is the Principal Components Analysis (PCA). PCA has become an increasingly used analysis method in the movement domain to identify patterns in the MoCap data and compress its dimensions [Barbič, 2004; Halilaj, 2018; Olivas-Padilla, 2019]. PCA computes new variables known as principle components, which are produced through linear combinations of the initial variables. This method has been used to analyze movement kinematics [Haid, 2019; Portnova-Fahreeva, 2020] and kinetics [Chang, 2020; Yoshida, 2022]. For the extraction of new features of less dimensionality for the study of human movements, other works applied Information Theory [Drotar, 2015; Peng, 2022], in which features (motion descriptors) are ranked and selected based on their distribution similarities between classes. Fast Fourier Transform [Ahlrichs, 2016] and Wavelet decomposition [Nielsen, 2011; Hasan, 2020] were used to extract Fourier or Wavelet coefficients for the analysis of human movements in the frequency or frequency-time domain, respectively. Dynamic Time Warping (DTW) has been used to calculate similarity measures

between human movements [Wang, 2010; Switonski, 2019; Mohammadzade, 2021].

Recently, data-driven approaches have been used to process multidimensional MoCap data, as they can automatically extract features without requiring manual feature extraction and selection. Convolutional Neural Networks (CNNs) [Shaheen, 2016; Jogin, 2018] and Autoencoders (AE) [Alo, 2020; Jun, 2020] are architectures with automatic feature extraction due to their numerous processing layers, which are composed of multiple linear and non-linear transformations. Further details of these approaches are provided later in Section 2.4.

2.3 Methodologies for human movement modeling

In this section are presented diverse methodologies for the modeling and analysis of full-body human movements. These are organized based on the goal of their analysis. Firstly are introduced the biomechanical models, which are utilized to study the continuum mechanics of the human body. This includes the forces that multiple body parts exchange internally or externally, as well as the effects motions and forces have on organs and the tissues that form them. The stochastic models are presented next, designed to learn the movement patterns produced on motion descriptors either for predictive modeling, action recognition, or describing the phenomenon and relationships between motion descriptors. The basic principle underlying stochastic models is that human movement is a dynamic system. This implies that there is a change in time and also in the state of the system, with future states being defined by a probabilistic rule based on the current state. Determining these rules for particular human motion systems is the central challenge of this area of research [Stergiou, 2018].

Lastly, recent works on the development of hybrid stochastic-biomechanical methods are discussed. The field of complex biomechanical modeling has begun to rely on stochastic models to investigate the effects of parameters variability and measurement uncertainty on model's outputs. Typically, biomechanical models are used to simulate human movement in a deterministic fashion. Applying stochastic models that consider the body's biomechanical structure has allowed the search for optimal parameter combinations and establishing model limitations for better simulations. Simple stochastic models often ignore the constraints of the human musculoskeletal system and solely rely on the MoCap data, causing distorted motion simulations.

2.3.1 Biomechanical modeling

Biomechanical models have generally been used to simulate human movement and the changes that occur as a result of internal and external action forces. These models represent the human body as a set of articulated links in a kinetic chain, with joint torques and forces calculated using anthropometric, postural, and hand load data [Lu, 2012]. As illustrated in the subsection 2.2.1.2, inertial data like accelerations and velocities and information regarding external forces, such as ground reaction forces measured by force plates, are utilized as input to biomechanical models [Muller, 2020]. Inverse dynamics is applied to extract quantitative information about the mechanics of the musculoskeletal system during the performance of a motor task. This

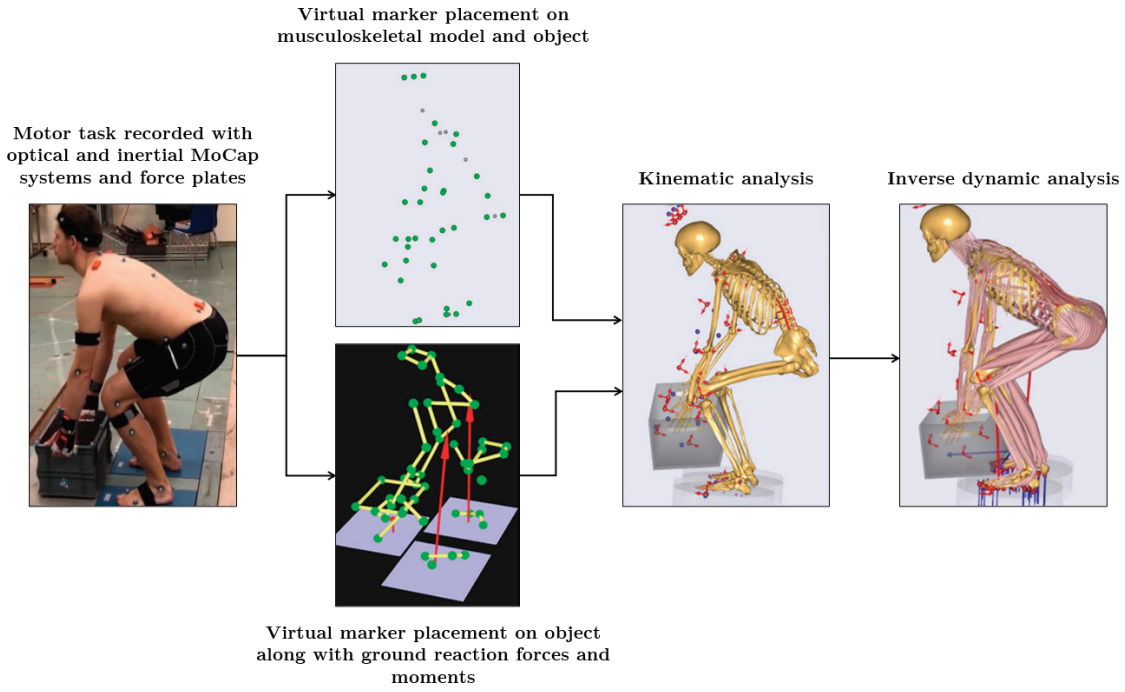


Figure 2.6: Flowchart of the process for the biomechanical modeling of human movements [Larsen, 2020].

process is depicted in Figure 2.6. In a study conducted by Larsen et al. [Larsen, 2020], optical and inertial MoCap systems were utilized to measure the subject's posture as well as the position of an object. Then, along with the measurements of ground reaction forces with force plates, a kinematic analysis was done. The outcomes are afterward utilized by the inverse dynamic algorithm. Previous research has employed biomechanical modeling to extract the joint's kinematic and kinetic contributions to a variety of motor tasks and then examined the joint's mechanical loading and reaction for ergonomic interventions. Menychtas et al. [Menychtas, 2020] used the Newton-Euler algorithm to compute upper body joint torques in order to study the ergonomic impact of various positions on human joints. After that, the normalized integral of joint angles and joint torques was computed to describe the kinematic and kinetic contribution of the body joints in uncomfortable positions. The research determined which joints moved the most during tasks and were subjected to the greatest strain when making ergonomically risky movements. Faber et al. [Faber, 2016] estimated 3D L5/S1 moments and ground forces using a spanned inverse dynamics model and then compared symmetric, asymmetric, and rapid trunk bending movements for ergonomic analysis. Similarly, Shojaei et al. [Shojaei, 2016] evaluated the lower back reaction forces and moments during manual material handling tasks in order to determine age-related changes in trunk kinematics and mechanical stresses on the lower back.

Developing precise and noninvasive methods for analyzing human movements remains challenging in biomechanical modeling. For an accurate study, current biomechanical models utilize measurements from optical MoCap systems (which are only available in specialized laboratories) or from several inertial sensors placed throughout the body. Additionally, force plates are

required to quantify forces or loads. The need for new methods that can analyze movements captured outside of laboratories arises from the fact that laboratory recordings lack authenticity as they are not done on the actual scenarios where the movements are performed, probably leading to inaccurate measurements.

2.3.2 Stochastic modeling

Stochastic modeling has been used to learn the random behavior of human movement. These models utilize the variance information contained in body motion trajectories to predict and identify human intentions and actions. A major challenge of human motion prediction arises from the intrinsic stochastic nature of the problem: Multiple future motions are possible given an observed sequence of postures. The high dimensionality and complexity of human motion dynamics compound this issue. Among the most successful methods for dealing with the temporal variations of human movements are generative models, in which time series are reorganized by sequential states. Thus, the temporal dynamics of movements are trained as a series of transitions between these states [Rabiner, 1989]. A common approach to describing human movement in this way is by using state-space modeling or a state-based model, such as Hidden Markov Models (HMMs).

Previous research applied state-space models based on Kalman filters for representing kinematic models, such as constant velocity (CV) and constant acceleration (CA), to forecast pedestrian position trajectories [Barth, 2008; Binelli, 2005]. The Kalman filter (KF) was mostly used to track the position of pedestrians based on their estimated velocity or acceleration. Caramiaux et al. [Caramiaux, 2015] introduced the GestureVariationFollower (GVF), an adaptive SSM based on particle filtering that recognizes and continuously tracks the variations of gestures. The system monitored gesture variations, allowing users to control ongoing actions via offset position, size, and direction of two-dimensional gestures.

HMMs have proven successful in modeling the temporal evolution of gestures globally, which is more robust to sequence warping. HMMs have mostly been used to model and recognize human gestures, with each gesture being associated with a single HMM. Additionally, since HMMs can process inputs as a sequence of successive values, they can recognize gestures regardless of their temporal duration. Glushkova et al. [Glushkova, 2018] and Manitasaris et al. [Manitsaris, 2020; Manitsaris, 2014] used HMMs to recognize gestures associated with professional tasks performed in the crafts and manufacturing industries. Malaisé et al. [Malaisé, 2018] recognized elementary manual material handling tasks using trained HMMs with joint angle sequences.

In other works that have used stochastic approaches to model full-body movements, Wang et al. [Wang, 2013] proposed the Intention-Driven Dynamics Model (IDDM) based on Gaussian processes. The dynamic model presupposes that human behavior is directed by a goal, which means that the dynamics change when actions are motivated by various intentions. The study established that integrating human dynamics into the modeling process improves the accuracy of forecasting human movements. Agarwal et al. [Agarwal, 2004] trained a mixture of Gaussian auto-regressive processes, using joint angles and position trajectories to represent the motion

patterns that emerge during common human activities (e.g., walking and running). The dynamic models exploit local joint-motion correlations to successfully track complex movements (turns in several directions) using only 2D body measurements (joint positions and angles). Devanne et al.[Devanne, 2017] used a Dynamic Naive Bayes model to capture the dynamics of motion primitives and continuously segment distinct human behaviors in extended sequences.

Despite these encouraging advances in recent years, accurate human motion modeling in unconstrained environments remains challenging. Unresolved difficulties in modeling human movements concern spatiotemporal dynamics. For example, even the same movement done by the same individual can have varying speeds and starting/ending positions, let alone in scenarios involving multiple performers. As a result, the variance in a category of human behavior can be quite large, and if either spatial or temporal dynamics are ignored, the modeling accuracy could suffer greatly [Stergiou, 2018]. Hybrid biomechanical-stochastic models have been developed in an effort to overcome the aforementioned challenges and improve the modeling performance of simple stochastic models. Besides capturing human motion stochastics, these also encapsulate the kinematic correlations or dependencies among different skeletal joints. These models are later introduced in Section 2.3.3.

Two previously mentioned probabilistic models, the State-Space Model and Hidden Markov Models, are described in the following subsections due to their advantages in modeling human movement and relevance to the work presented in this thesis. These methods permit the mathematical representation of human movements, where the parameters of the assumptions may be examined to obtain information about how they are executed, a feature that is highly relevant to the objectives of this thesis. Moreover, as they are based on a mechanistic movement model, they are superior to classical analytical approaches, such as linear models, for purposes of extrapolation, like predicting movements in novel environments [Patterson, 2008; Manitsaris, 2020].

2.3.2.1 State-Space Modeling

State-space modeling refers to a type of probabilistic graphical model that depicts the probabilistic dependence between a latent state variable and an observed measurement [Koller, 2009]. In the 1960s, the term "state space" was created in the field of control engineering [Kalman, 1960]. State-space modeling provides a framework for understanding both deterministic and stochastic dynamical systems that are measured or observed via a stochastic process. Moreover, it offers an unifying methodology of solving a variety of time series analysis challenges, including human motion modelling [Barth, 2008; Binelli, 2005; Manitsaris, 2014; Manitsaris, 2020; Caramiaux, 2015]. The Kalman filter is the most well-researched SSM because it defines the optimal algorithm for inferring linear Gaussian systems using the Normal distribution as a working model. The subsequent subsections describe this approach.

State-Space representation Firstly, the time series is expressed as an SSM, with the state equation (or transition model) specifying how the system evolves from one time point to the next and the observation equation specifying how the underlying state is transformed

(with noise added) into what is directly measured. Assume that there exists an initial state $s_0 \sim \mathcal{N}(s_{0|0}, P_{0|0})$ from which subsequent states are estimated. Each time data x_t is observed, it is incorporated into the calculation of s_t. The following is the formulation of the observation equation:

$$x_t = As_t + V_t \tag{2.9}$$

and the state equation:

$$s_t = \Theta s_{t-1} + W_t \tag{2.10}$$

where x_t is a $d \times 1$ vector, s_t is a $k \times 1$ vector, A is a $d \times k$ matrix and Θ is $k \times k$ matrix, supposing $V_t \sim \mathcal{N}(0, S)$ and $W_t \sim \mathcal{N}(0, R)$ and are the measurement noise processes. The parameters A, Θ, S, and R are considered to be known or used as tuning parameters to generate an estimate of s_t for every t of interest.

Essentially while modeling human movements, the state equation predicts the future state of the human, given its current state, an assumption known in mathematics as the Markov condition. The observation equation then weights these predictions based on the likelihood of data, thereby connecting the state equation to the observations. Maximum likelihood estimates of parameters can be calculated analytically using the Kalman filter, which is explained next.

Maximum Likelihood Estimation via Kalman filters The Kalman filter is used to calculate the observed data's log-likelihood for a given set of parameters. Therefore, the MLE approach can be applied using the KF each time to compute the log-likelihood.

To evaluate and maximize the likelihood function of the data, the joint density of the observed data is required $p(x_{1:T})$, which may then be factored into the following:

$$\begin{aligned} p(x_1, x_2, \dots, x_T) &= p(x_1)p(x_2, \dots, x_n \mid x_1) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3, \dots, x_T \mid x_1, x_2) \\ &\vdots \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_T \mid x_1, \dots, x_{T-1}) \end{aligned} \tag{2.11}$$

It can be integrated $p(x_1)$ using the conditional probability formula in 2.12, where it is included in the state variable s_1:

$$p(x_1, s_1) = p(x_1|s_1)\, p(s_1) \tag{2.12}$$

$$p(x_1) = \int p(x_1, s_1) ds_1 = \int p(x_1|s_1) p(s_1)\, ds_1 \tag{2.13}$$

$p(x_1|s_1)$ is the density for the observation equation, which in this case is $\mathcal{N}(A_1 s_1, S)$, or in the SSM corresponds to Equation 2.9. It begins at time $t = 1$ with the observation of x_1, assuming an initial state $s_0 \sim \mathcal{N}(s_{0|0}, P_{0|0})$. Following that, the marginal distribution of $p(s_1)$ must be determined. Because there is no x_0 yet, it cannot be conditioned by any observed

data. So, $p(s_1)$ is calculated as follows:

$$\begin{aligned} p(s_1) &= \int p(s_1|s_0)\, p(s_0)\, ds_0 \\ &= \int \mathcal{N}(\Theta s_0, R) \times \mathcal{N}\left(s_{0|0}, P_{0|0}\right) ds_0 \\ &= \mathcal{N}(\Theta s_{0|0}, \Theta P_{0|0}\Theta' + R) \\ &= \mathcal{N}(s_{1|0}, P_{1|0}) \end{aligned} \tag{2.14}$$

Note that $s_{1|0} \triangleq \Theta s_{0|0}$, representing Equation 2.10 in the SSM, and $P_{1|0} \triangleq \Theta P_{0|0}\Theta' + R$, being $s_{1|0}$ the initial prediction. By integrating $\int p(x_1|s_1)p(s_1)\, ds_1$:

$$p(x_1) = \mathcal{N}(A_1 s_{1|0}, A_1 P_{1|0} A_1' + S) \tag{2.15}$$

Assuming that A is known (tuning parameter), then the quantities $s_{1|0}$ and $P_{1|0}$ are all routinely computed in implementing the KF algorithm. The KF algorithm is divided into two steps: update and prediction. In the update step, given a new observation x_1, it is utilised to estimate s_1. To do so, the conditional distribution $p(s_1|x_1)$, also known as the filter density, is required. The filter density is calculated using Bayes' rule:

$$p(s_1|x_1) \propto p(x_1|s_1)\, p(s_1) \tag{2.16}$$

From the observation equation, it is known that $p(x_1|s_1) = \mathcal{N}(A_1 s_1, S)$ and $p(s_1)$ is computed in 2.16. Therefore, by using the basic properties of the normal distribution:

$$\begin{aligned} p(s_1|x_1) &= p(x_1|s_1)\, p(s_1) \\ &= \varphi(x_1|A_1 s_1, S) \times \varphi\left(s_1 \middle| s_{1|0}, P_{1|0}\right) \\ &= \mathcal{N}(s_{1|0} + K_1(x_1 - A_1 s_{1|0}), (I - K_1 A_1) P_{1|0}) \end{aligned} \tag{2.17}$$

K_1 is the Kalman gain coefficient calculated as follows:

$$K_1 = \frac{P_{1|0} A_1'}{A_1 P_{1|0} A_1' + S} \tag{2.18}$$

Then in the prediction step for $t = 1$, the estimates are:

$$s_{1|1} = \mathbb{E}[s_1|x_1] = s_{1|0} + K_1(x_1 - A_1 s_{1|0}) \tag{2.19}$$

$$P_{1|1} = Var(s_1|x_1) = (I - K_1 A_1) P_{1|0} \tag{2.20}$$

So, the filter density is $p(s_1|x_1) = \mathcal{N}(s_{1|1}, P_{1|1})$.

In general, the estimate of s_t for each t is the mean of the filter density $p(s_t \mid x_1, \ldots, x_t)$ and the filter density is a product of the observation density and the predicted density:

$$p(s_t|x_1, \ldots, x_t) \propto p(x_t|s_t) p(s_t|x_1, \ldots, x_{t-1}) \tag{2.21}$$

Then it is iteratively applied the Kalman filtering algorithm to each t, calculating the necessary quantities for each step of the likelihood function:

$$p(x_t \mid x_1, \ldots, x_{t-1}) = \mathcal{N}(As_{t|t-1}, AP'_{t|t-1} + S). \tag{2.22}$$

When $t = T$ is reached, the joint likelihood function can be computed for the MLE. By representing the vector of unknown parameters as β, calculating the log-likelihood would require computing the following sum:

$$\ell(\beta) = \sum_{t=1}^{T} \log p(x_t \mid x_1, \ldots, x_{t-1}) \tag{2.23}$$

In Equation 2.23, it is defined $p(x_1|x_0) = p(x_1)$ since there is no x_0. Next, it is maximised $\ell(\beta)$ with respect β using standard non-linear maximisation routines like Newton's method or quasi-Newton approaches [Olsson, 2007]. Note that β represents all parameters to tune, $\beta = (A, \Theta, S, R)$. Additionally, initial values for $s_{0|0}$ and $P_{0|0}$ must be defined, which we can be either assumed as known or included in the vector of unknown parameters.

2.3.2.2 Hidden Markov Models

An HMM is a time series statistical model in which the observed data is assumed to be a noisy measurement of a system that can be modeled as a Markov process [Rabiner, 1989]. The density of sequences is modeled by including a first-order Markov dependency between latent variables. Thus, an HMM associates an observation model with a hidden discrete-time discrete-state Markov chain. HMMs have demonstrated efficacy in a variety of fields, including automatic speech recognition [Rabiner, 1989], gesture recognition [Glushkova, 2018; Malaisé, 2018; Manitsaris, 2014; Manitsaris, 2020], and movement generation [Calinon, 2011; Sato, 2019; Kitzig, 2018; Samadani, 2020]. This section briefly examines the representation, learning, and inference techniques for HMMs.

Representation Consider a movement that is captured as a succession of observations $x = [x_1, x_2, ..., x_T]$, where $x_t \in \mathbb{R}^D$ is a D-dimensional vector, a stream of motion descriptors obtained from sensors or a biomechanical model. The joint distribution of an HMM can be expressed in terms of the hidden states $s_t = [s_1, s_2, ..., s_N]$:

$$p(x_{1:T}, s_{1:T}) = p(s_{1:T})\, p(x_{1:T}|s_{1:T}) = \underbrace{\left[p(s_1) \prod_{t=2}^{T} p(s_t|s_{t-1})\right]}_{\text{Markov process}} \underbrace{\left[\prod_{t=1}^{T} p(x_t|s_t)\right]}_{\text{Observation model}} \tag{2.24}$$

The first part of Equation 2.24 embodies first-order Markov properties, which indicate that the state at time t is dependent exclusively on the state at time $t-1$. The second part, referred to as the observation model, defines the state-conditional observation density distribution. When dealing with discrete observations, the observation model can be reduced to a matrix. Thus, an N-state HMM is described by a set of parameters $\lambda = (A, B, \Pi)$ consisting

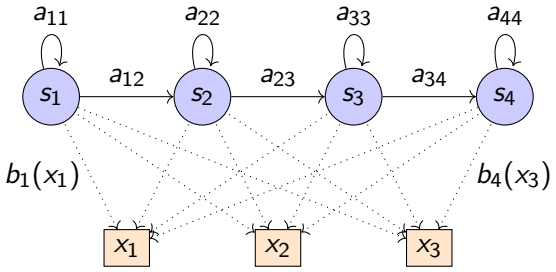
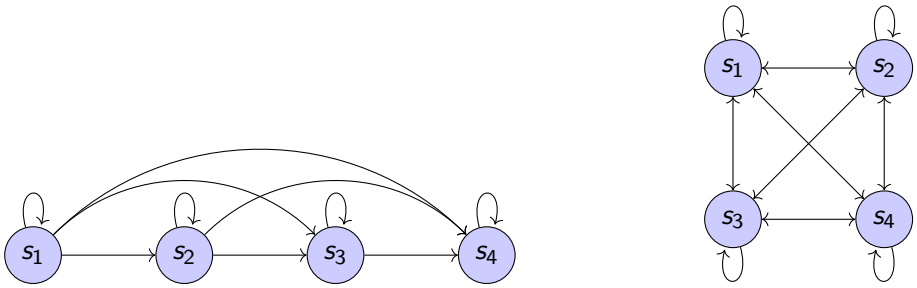


Figure 2.7: An HMM with four states that can emit four observations: x_1, x_2, or x_3. a_{ij} is the probability to transition from state s_i to state s_j. $b_j(x_t)$ is the probability to emit x_t in state s_j. In this particular HMM, states can only reach themselves or the adjacent state.



(**a**) (**b**)

Figure 2.8: Two four-state HMMs with (a) left-to-right topology and (b) ergodic topology.

of a state transition matrix $A = \{a_{ij}\}$, an observation probability distribution $B = \{bj(x_t)\}$, and a prior vector $\Pi = \{\pi_i\}$. Figure 2.7 shows an example of an HMM with four states, illustrating these probabilities. The states are connected via probabilities known as transition probabilities a_{ij}, which represent the probability of transiting from state i to state j:

$$a_{ij} \triangleq p\left(s_t = j | s_{t-1} = i\right), \; a_{ij} \geq 0 \; and \; \sum_{i=1}^{N} a_{ij} = 1 \tag{2.25}$$

Then, the observation probability distribution is given by:

$$b_j\left(x_t\right) \triangleq p\left(x_t | s_t = j\right), \; b_j\left(x_t\right) \geq 0 \; and \; \int_{x_t} b_j\left(x_t\right) dx_t = 1 \tag{2.26}$$

Finally the prior probability of the ith state:

$$\pi_i \triangleq p\left(s_1 = i\right), \; \pi_i \geq 0 \; and \; \sum_{i=1}^{N} \pi_i = 1 \tag{2.27}$$

The transition matrix can specify a variety of Markov chain topologies. The left-to-right and ergodic (fully-connected) are two common topologies and are represented in Figure 2.8.

The model's complexity can be changed by adjusting the number of hidden states. This parameter specifies how precisely the model segments the training samples. Utilizing a small

number of hidden states implies that the movement's information is embedded in a lower-dimensional space, lowering the accuracy of the movement's temporal modeling. However, using a few states can assist in ensuring the model's generalization. As a result, the recognition will be tolerant of input variances. On the other hand, increasing the number of states improves the temporal structure's accuracy. Nevertheless, once states begin to incorporate random variability, this might result in overfitting. The number of hidden states to use is mostly determined by the application. Cross-validation or automatic model selection are used to address this issue in the HMM literature.

Training and inference The Baum-Welch algorithm, which is an expectation–maximization (EM) algorithm, is used to estimate the HMMs' transition probabilities, observation probabilities, and initial state probability [Rabiner, 1989; Françoise, 2015]. Baum-Welch estimation is an iterative procedure in which two steps, estimation and maximization, alternate. The estimation step uses the current model parameters to calculate the smoothed and edged marginals that quantify the contribution of states and transitions to each data frame. Then, the parameters are re-estimated in the maximization step utilizing these intermediate values. The Baum-Welch algorithm's equations ensure that the log-likelihood of the data increases with each iteration, assuring convergence.

After the training of the HMMs, the algorithm Viterbi is used to determine the state sequence most probable and the Baum's "forward" algorithm to compute the probability of the sequence according to the observation. It is defined as the forward variable $\alpha_t(j) = p\left(s_t = j | x_{1:t}\right)$, that is calculated recursively in a prediction-update cycle:

$$\alpha_t(j) = \frac{1}{Z_t} \underbrace{\left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\right]}_{\text{Prediction}} \underbrace{b_j\left(x_t\right)}_{\text{Update}} \tag{2.28}$$

where $\alpha_1(j) = \frac{1}{Z_1} \pi_j b_j\left(x_t\right)$ and Z_t is a normalisation constant defined by:

$$Z_t \triangleq p\left(x_t | x_{1:t-1}\right) = \sum_{j=1}^{N} \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\right] b_j\left(x_t\right) \tag{2.29}$$

Z_t can be used to compute the likelihood of the observed data given the model's parameters, which is represented in log form as follows:

$$log\ p\left(x_{1:t}\right) = log\left[p\left(x_t | x_{1:t-1}\right) p\left(x_{1:t-1}\right)\right] = \sum_{\tau=1}^{t} log\ Z_\tau \tag{2.30}$$

This is the critical formula for classification. Because a single HMM is trained with all samples from a specific class, whenever a gesture has to be identified, the class of the gestures used to train the HMM that provided the highest likelihood is then given to the unknown gesture.

2.3.3 Hybrid biomechanical-stochastic modeling

The analysis of the random outcomes of human movement has been improved by developing hybrid methodologies that consider both the human biomechanical structure and the stochastic nature of human movement [Shi, 2003; Lin, 2009; Bologna, 2020]. This type of model has been extensively used to study musculoskeletal pathologies. By human motion modeling, the deviations from normal movement in terms of altered kinematic or kinetic patterns are identified and then utilized to evaluate neuromusculoskeletal conditions, to aid in subsequent treatment planning, or to analyze the efficacy of treatment in different patient groups. Some recent works on hybrid modeling for medical applications are reviewed in the following Subsection 2.3.3.1.

Biomechanical-stochastic modeling has also been employed to accurately simulate human motions and explain the interaction between joints for achieving the learned motion trajectories. The Gesture Operational Model (GOM) presented by Manitsaris et al. [Manitsaris, 2020] has proven successful for this purpose. As one of the primary goals of this research is to build innovative approaches for explaining human movements in a clear and reconstructible manner, the Subsection 2.3.3.2 presents an overview of GOM and its functional human motion representations, which are pertinent to this thesis work.

A limitation of hybrid models, as well as biomechanical and stochastic models, is that their processing requirements grow exponentially as the number of model parameters increases. They are not practical for analyzing large datasets of human movements or high-dimensional MoCap data as data-driven approaches (described next in Section 2.4), requiring the use of feature selection or extraction algorithms. Moreover, to adequately design the hybrid models and their assumptions, it is necessary to have prior knowledge of the data that would be used for the training. For example, if the objective is to understand muscle coordination, a model that omits joints and muscles is unlikely to be helpful. Data-driven approaches do not require this prior knowledge since they generate their own internal representations based on the training data.

2.3.3.1 Hybrid modeling for medical applications

Biomechanical-stochastic modeling has been effectively used to research human movement variability and the prevention of a wide range of musculoskeletal system injuries [Davidson, 2004; Langenderfer, 2006; Santos, 2004]. A hybrid model designed to predict the probability of injury and identify factors contributing to the risk of non-contact anterior cruciate ligament (ACL) injuries, has been proposed by Lin et al. [Lin, 2012]. A biomechanical model of the ACL estimated the lower leg kinematics and kinetics. In turn, the means and standard deviations of the number of simulated non-contact ACL injuries, injury rate and female-to-male injury rate, were calculated in Monte Carlo simulations of non-contact ACL injury and non-injury trials. T-tests revealed the biomechanical characteristics of the simulated injury trials. In another work, Donnell et al. [Donnell, 2014] used a two-state Markov chain model to represent the survival of surgical repair from a torn rotator cuff. The load applied to the shoulder and the structural capacity of tissue were defined as the random variables. The analysis was

based on the biomechanical application of structural reliability modeling [Saraygord Afshari, 2022]. By introducing this new modeling paradigm for explaining clinical retear data, the model successfully predicted the probability of rotator cuff repair retears and contributed to understanding their causes.

2.3.3.2 The Gesture Operational Model

Modeling and simulating human movement is a critical component of human movement analysis. Anticipating how human' motion descriptors will evolve over time enables proactive integration of this knowledge, for example, in human-robot interaction or risk prevention. However, as stated in Section 2.3.2, human motion is a stochastic process characterized by a high level of uncertainty, complicating its modeling. Yet, the prediction of joint position sequences of a 2D skeleton has been adequately addressed by a biomechanical-stochastic model called the Gesture Operational Model, presented by Manitsaris in [Manitsaris, 2020]. GOM describes how humans move based on assumptions about the dynamic association of body joints, their synergies, their serial and non-serial mediations, as well as their transition through time from one state to another. Through state-space modeling, the GOM assumptions are converted into a simultaneous equation system for each body joint. The model is capable of simulating human movements and generates a confidence-bounding box for each joint's motion descriptor, which represents the tolerance for its spatial variance over time. Additionally, the joint motion representations that comprise GOM can be interpreted in such a way that by examining their learned coefficients, information regarding how a performer's body joints are organized to execute a specific motion trajectory can be deduced.

The Gesture Operational Model is constructed from auto-regressive models that are used to learn the dynamics of each body joint. Each representation makes different assumptions regarding the dynamic interaction of body joints. These assumptions include:

The time-dependent transitions (H1): Current values are dependent on their predecessors.

The intra-joint association (H2): A bidirectional relationship is assumed between variables in which the motion is decomposed, e.g., joint angles on the X-axis, Y-axis, and Z-axis.

The inter-limb synergies (H3): A bidirectional relationship between body joints that cooperate to accomplish a motion trajectory, such as using both hands to make a specific gesture.

The serial (H4.1) and non-serial (H4.2) intra-limb mediations: The mediation between joints, whether they are directly or indirectly connected; for example, the wrist is directly connected to the elbow (serial mediation) and indirectly connected to the shoulder (non-serial mediation).

These assumptions are illustrated in Figure 2.9. The number of models that compose GOM is equal to the number of dimensions associated with each body joint motion descriptor, multiplied by the number of joints defined. The transitioning assumptions (H1) are the lagged

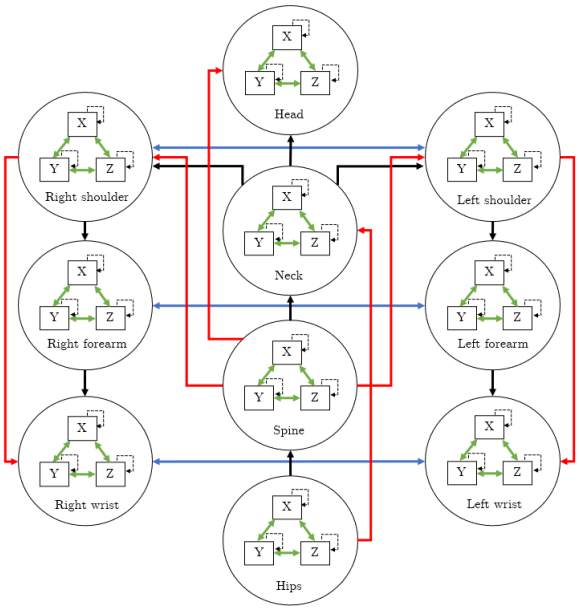


Figure 2.9: An example of a Gesture Operational Model with only upper-body assumptions. Dashed arrows show time-dependent transitions; green arrows represent intra-joint associations; blue arrows suggest inter-limb synergies; black arrows indicate serial intra-limb mediation, and red arrows indicate non-serial intra-limb mediation.

endogenous variables, whose lag is determined by the model's order. The rest of the assumptions (H2, H3, and H4) are the exogenous variables.

An example of a mathematical representation of the assumptions is shown in Equation 2.31, for the motion on the X-axis ($Px_{1,t}$) of the body joint $P_{1,t}$. The movement of $P_{1,t}$ only decomposes on XY axes, $Px_{1,t}$ and $Py_{1,t}$, and it is assumed that its motion has an association with the movement of other two body joints, $Px_{2,t-1}$ and $Px_{3,t-1}$.

$$Px_{1,t} = \underbrace{Px_{1,t-1} + Px_{1,t-2}}_{\text{H1}} + \underbrace{Py_{1,t-1}}_{\text{H2}} + \underbrace{Px_{2,t-1} + Px_{3,t-1}}_{\text{H3 or H4}} \tag{2.31}$$

Through the use of state-space models, these representations are subsequently converted to simultaneous equations (Section 2.3.2.1). To illustrate the application of state-space modeling in GOM, it is illustrated the initial state-space representation in equations 2.32 and 2.33. This representation sets as the observation y_t, which for Equation 2.31 corresponds to the prediction of $Px_{1,t}$ at time t. Then, x_t is the exogenous (pre-determined) variables which consist of our assumptions H2, H3, and H4. The state variable is defined as s_t, consisting of endogenous variables (assumption H1).

$$y_t = As_t + Bx_t \tag{2.32}$$

$$s_t = \Theta s_{t-1} + W_t \tag{2.33}$$

The observation equation is 2.32, in which the time derivative of the state vector s_t is used along with the input vector x_t to compute the output y_t. A is the output matrix and B is the feed-through matrix. Equation 2.33 is the state equation, a first-order Markov process where

Θ is the transition matrix.

To model the GOM representation of the Equation 2.31 using second-order SSM, first, the state variable is substituted with the subtraction of two previous values of the body joint to model, each multiplied by one coefficient of the transition matrix:

$$s_t = \Theta s_{t-1} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} Px_{1,t-1} \\ -Px_{1,t-2} \end{bmatrix} = \begin{bmatrix} \alpha_1 Px_{1,t-1} \\ -\alpha_2 Px_{1,t-2} \end{bmatrix} \tag{2.34}$$

As mentioned earlier, the input vector x_t corresponds to the exogenous variables for the observation equation. For the case of Equation 2.31, it consists of intra-joint associations (H2), inter-limb synergies (H3) or intra-limb mediations (H4):

$$Px_{1,t} = \begin{bmatrix} 1 & 1 \end{bmatrix} s_t + \alpha_3 Py_{1,t-1} + \alpha_4 Px_{2,t-1} + \alpha_5 Px_{3,t-1} \tag{2.35}$$

Finally, by merging Equations 2.34 and 2.35, the state-space representation is obtained:

$$Px_{1,t} = \alpha_1 Px_{1,t-1} - \alpha_2 Px_{1,t-2} + \alpha_3 Py_{1,t-1} + \alpha_4 Px_{2,t-1} + \alpha_5 Px_{3,t-1} \tag{2.36}$$

The tuning parameters of the equation system, including the constant coefficients α, are estimated using MLE via Kalman filtering (Section 2.3.2.1).

2.4 Data-driven approaches for sequence modeling

The biological neural networks that exist in the human brain inspired the creation of Artificial Neural Networks (ANNs). ANNs were initiated by McCulloch and Pitts with their simple artificial neuron model called "perceptron" [Mcculloch, 1943]. Research in this area led to the development of multi-layer neural networks. Now, ANNs are composed of multiple layers of mathematical functions that gradually map an input X to an output Y through a series of intermediary representations known as hidden states. These layers can be fully connected, convolutional, or recurrent. A typical fully-connected layer is composed of inputs that are linearly combined. A convolutional layer scans the input in terms of its dimensions using filters (kernels) that execute convolution operations. In the case of the recurrent layer, it updates its current output by including prior outputs and hidden states. Gradient descent algorithms such as the Stochastic Gradient Descent (SGD) [Bottou, 2012], Root Mean Square Propagation (RMSProp) [Kurbiel, 2017], Adaptive Gradient Algorithm (Adagrad) [Duchi, 2011], or Adaptive Moment Estimation (Adam) [Kingma, 2014] are commonly used to train ANNs. The backpropagation method calculates the gradient of the loss with regard to the ANN's parameters. Then these are adjusted in such a way that the ANN's output deviation is minimized. Accordingly, all applied operations in the ANN, particularly the loss function, should be differentiable. Selecting a differentiable loss function is critical for enforcing the desired behavior on the ANN.

The main advantage of ANNs over other conventional machine learning and statistical approaches is their great modeling capacity and considerable flexibility in designing architectures.

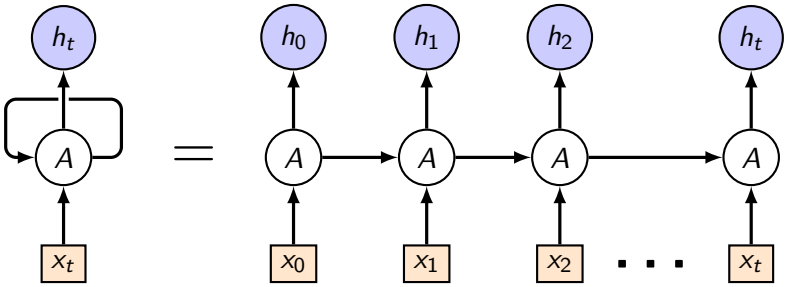


Figure 2.10: Diagram of an unrolled RNN.

By incorporating deep temporal architectures, ANNs have shown promising performance for sequence modeling of time series data, such as human movements. These account for the temporal aspects of the data, which is critical for modeling human movement. In the following subsections, the most relevant deep temporal models are reviewed to build a ground for the work done in this thesis. Furthermore, it is discussed how these approaches were applied to the problem of human motion modeling and their limitations.

2.4.1 Recurrent Neural Networks

The modeling of human movements has been an active area of research in deep learning, as it requires approaches capable of capturing the temporal dependencies contained in MoCap data. One of the earliest methods for sequence modeling of human movements was the HMMs, as discussed in Section 2.3.2.2. HMMs are able to capture the motion data distribution using multinomial latent variables. These models condition every data point at time t on a hidden state at time $t-1$. The observation and transition probability distributions, $p(x_t|s_t)$ and $p(s_t|s_{t-1})$, are then learned and are the same for all time series. In Recurrent Neural Networks (RNNs), a similar concept of parameter sharing is utilized. A vanilla RNN is a feedforward neural network that contains recurrent connections in which weights are shared for every time step in the sequential data. Consequently, it enables the RNN to learn the temporal dynamics of the sequential data, in this case, human movement. RNNs have a recurrent hidden state whose activation at each time t depends on the previous. Figure 2.10 illustrates a simple one-layer recurrent network unrolled over time.

RNNs are capable of generating motion sequences by predicting the next immediate data point given all preceding data points. Given a variable-length movement $x = [x_1, x_2, ..., x_T]$, where x_t is the measured posture at time t, the RNN updates its recurrent hidden state h_t, serving as memory of the system:

$$h_t = \begin{cases} 0, & t = 0 \\ f_\phi(h_{t-1}, x_t) & otherwise \end{cases} \tag{2.37}$$

f_ϕ is a feedforward neural network and ϕ its parameters. A basic RNN formulation can be written as follows, where there is an update of the recurrent hidden state h_t and prediction y_t,

obtained by a projection of the latent state with a weight matrix V:

$$h_t = g(Wx_t + Uh_{t-1} + b_h) \tag{2.38}$$

$$y_t = d(Vh_t + b_y) \tag{2.39}$$

g and d are non-linear functions (logistic sigmoid or hyperbolic tangent), b_h and b_y are the biases, and W and U are the weights of the current input sequence x_t and the previous recurrent hidden state h_{t-1}, respectively. Multiple recurrent layers can be piled on top of one another for deeper temporal architectures. In this situation, the output activations serve as the input to the following layer. Backpropagation Through Time (BPTT) is used to train RNNs [Mozer, 1989], with the prediction y_t being compared to the ground truth target in the loss function.

Exploding and vanishing gradients are known issues when training RNN by backpropagation [Ribeiro, 2020]. The exploding of gradients can be prevented by clipping their magnitudes [Pascanu, 2012; Ribeiro, 2020]. The latter problem, though, is more challenging to resolve. Some studies suggest skipping connections to build loss functions that are smoother and simpler to train. Another way is the use of Truncated BPTT [Aicher, 2020]. Adding normalization layers and recurrent dropout are additional methods for improving the training performance of RNNs [Zaremba, 2014].

2.4.1.1 Gated Recurrent Neural Networks

Using a gated activation function has been found to be one of the most effective approaches for improving training performance in RNN so far. RNNs were enhanced with the introduction of the LSTM cell and variants such as the GRU. In comparison to the standard recurrent unit, which computes only the weighted sum of x_t and h_{t-1} and applies a non-linear function, each LSTM unit maintains a memory c_t [Chung, 2014]. Similarly to the LSTM unit, the GRU employs gating units to control the flow of information inside the unit, but without the need for distinct memory cells [Cho, 2014]. Figure 2.11 shows the diagrams of an LSTM and GRU cell.

As shown in Figure 2.11a,the LSTM cell has four gates: the forget gate f_t, input gate i_t, memory cell c_t, and output gate o_t. The output h_t of the LSTM cell is formulated as follows:

$$h_t = o_t \tanh(c_t) \tag{2.40}$$

The output gate o_t modulates the amount of memory content exposure, and is calculated by:

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \tag{2.41}$$

In Equation 2.41, σ is a logistic sigmoid function. The memory cell c_t is updated by partially forgetting the existing memory and adding a new memory content, $\widetilde{c}_t$:

$$c_t = f_t c_{t-1} + i_t \widetilde{c}_t \tag{2.42}$$

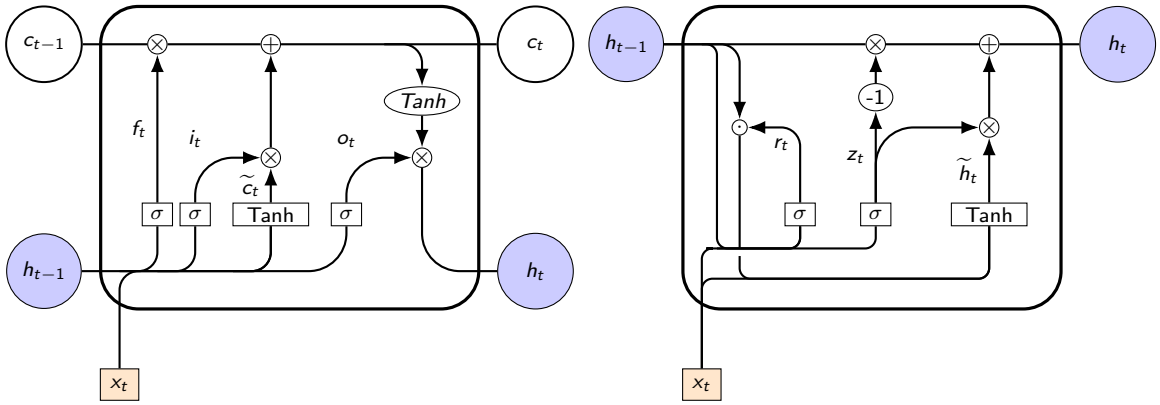


Figure 2.11: Diagram of the structure of an (a) LSTM cell and (b) GRU cell.

The $\widetilde{c}_t$ is computed as follows:

$$\widetilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1}\right) \tag{2.43}$$

The forget gate f_t modulates the extent to which the existing memory is forgotten, and the input gate i_t modulates the degree to which new memory $\widetilde{c}_t$ is added to the memory cell c_t. These gates are computed by:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1}\right) \quad i_t = \sigma\left(W_i x_t + U_i h_{t-1}\right) \tag{2.44}$$

Thus, unlike ordinary recurrent cells, which erase their content each time step, an LSTM cell can determine whether to retain or discard existing memory based on the inserted gates. Additionally, if a significant feature is detected in an input posture sequence, the LSTM retains this information in subsequent iterations, capturing any long-term dependencies.

In the case of the GRU cell (Figure 2.11b), the recurrent hidden state h_t is a linear interpolation between the previous state h_{t-1} and a candidate state $\widetilde{h}_t$:

$$h_t = (1 - z_t)\, h_{t-1} + z_t \tag{2.45}$$

In Equation 2.45, the update gate z_t decides how much the cell updates its content. The z_t is computed as follows:

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \tag{2.46}$$

As the LSTM cell, a linear sum between the existing state and the newly computed state is done. However, the GRU does not control the degree to which its state is exposed to new information. The candidate state $\widetilde{h}_t$ is then calculated as the traditional recurrent unit but includes a set of reset gates r_t:

$$\widetilde{h}_t = \tanh\left(W x_t + U\left(r_t \odot h_{t-1}\right)\right) \tag{2.47}$$

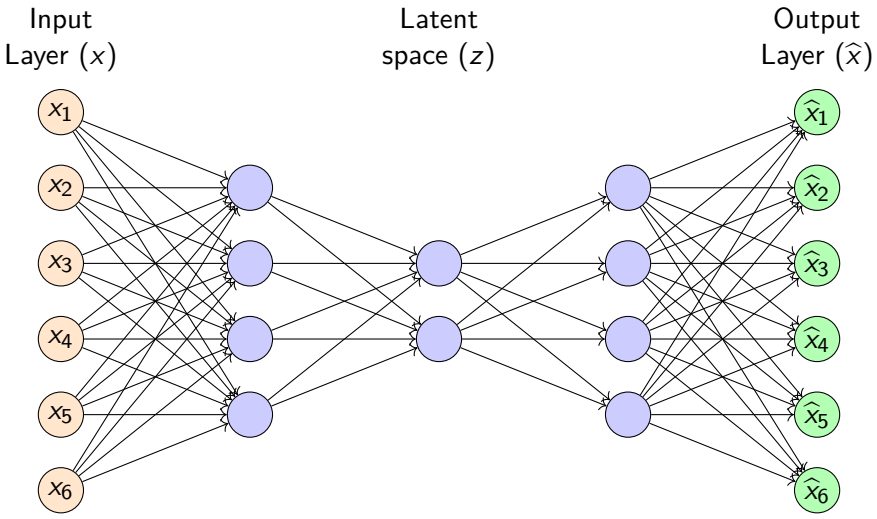


Figure 2.12: Diagram of the structure of an Autoencoder.

where $\odot$ is an element-wise multiplication. When r_t is close to zero, the unit forgets the previously computed state, as if it is reading the first value of an input sequence x_t. The reset gate r_t is computed similarly to the update gate:

$$r_t = \sigma\left(W_r h_t + U_r h_{t-1}\right) \tag{2.48}$$

2.4.2 Encoder-decoder architectures

For modeling and generating time series, the encoder-decoder architecture with RNNs has become a standard and efficient approach. These architectures are capable of learning disentangled representations of sequential data [Yan, 2018; Zhu, 2020]. This latent representation learning allows for extracting semantically meaningful information from the sequential data, improving the modeling performance. Because of their success in neural machine translation (NMT) and sequence-to-sequence (Seq2Seq) prediction, many approaches based on these architectures have been developed to model and predict objects or human motion trajectories [Martinez, 2017; Sun, 2017; Hernández, 2019; Slaton, 2020]. The subsections that follow explain the vanilla Autoencoder (AE), Variational Autoencoder, and AE with attention mechanism, as well as how these have been applied to model human movements.

2.4.2.1 Vanilla Autoencoder

The vanilla autoencoder is a neural network composed of an encoder and a decoder. A basic structure of an autoencoder using fully-connecter layers is shown in Figure 2.12. In this figure, AE is intended to map the encoder's input data x to an internal latent representation h. Then, the decoder generates an output $\widehat{x}$ similar to the input data (a reconstruction of the input). This AE was first introduced by [Rumelhart, 1986] for the purpose of non-linear dimension reduction, performing linear operations to achieve a latent representation similar to that of PCA.

Nowadays, the encoder and decoder can be neural networks with custom architectures. Increasing the complexity of the architecture allows the autoencoder to learn more complex latent encoding. Traditional autoencoders have the primary goal of reconstructing their input as accurately as possible. During the network's training, a specific loss function is used to achieve this objective. This function, known as reconstruction loss, is typically the squared mean error between the output and input. Throughout the training, the loss function penalizes the network for producing outputs that deviate from the input:

$$Loss = \|x - \widehat{x}\|^2 \tag{2.49}$$

For resolving Seq2Seq problems, where both the input and output are sequences, the encoder and decoder are either single-layer RNNs, LSTMs, GRUs, or multi-layer stacks of them. In the case of generating human movements, the input x_t at time t corresponds to the current pose described by n motion descriptors, $x_t = [x_{t,1}, x_{t,2}, ..., x_{t,n}]$. Hereby the autoencoder outputs the most probable pose y_t given x_t. Firstly, the encoder generates an internal representation of the input pose sequence $x = [x_1, x_2, ..., x_t]$, consisting of a context vector c of fixed length:

$$h_t = f(x_t, h_{t-1}) \tag{2.50}$$

$$c = q(h_1, ..., h_t) \tag{2.51}$$

where f and q are non-linear functions, h_t is the hidden state at time t, and c is the context vector generated from the sequence of hidden states. Afterwards, the context vector is used to initialize the decoder. In the case of LSTM, the decoder uses the encoder's internal state to initialise its own and then estimates the output pose sequence $y = [y_1, ..., y_t]$, step by step. In other words, the decoder models the following conditional probability, where g is a non-linear function that outputs the probability of y_t, and s_t is the hidden state of the RNN:

$$p(y_t \mid y_1, \dots, y_{t-1}, c) = g(y_{t-1}, s_t, c). \tag{2.52}$$

Figure 2.13 illustrates a diagram of a Seq2Seq model. The decoder is implemented as an autoregressive model that uses previous steps as input to improve the accuracy of its predictions. Both RNNs are trained together in a Seq2Seq model to maximise the likelihood $p(y_1, ..., y_t|x_1, ..., x_t)$, averaged over all of the training set's input and output pose sequences. The RNN decoder is rolled forward by recursively feeding back its own predictions as inputs for the next time-steps when making predictions [Lopez Pinaya, 2020].

2.4.2.2 Variational Autoencoder

The Variational Autoencoder has significantly enhanced the autoencoders' capacity to represent information. VAEs are generative models that attempt to describe data through a probabilistic distribution based on Variational Bayes Inference [Kingma, 2013]. In a VAE, it is assumed that a generative model for each sample of x is conditioned by an unobserved random latent variable z, where the generative distribution is parameterised by θ. The decoder defines this

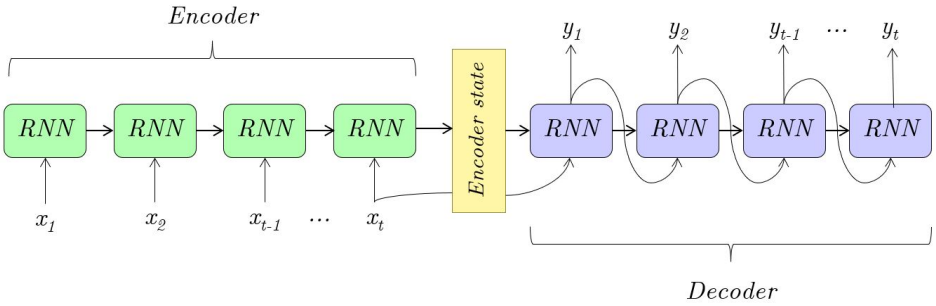


Figure 2.13: Diagram of an Seq2Seq network.

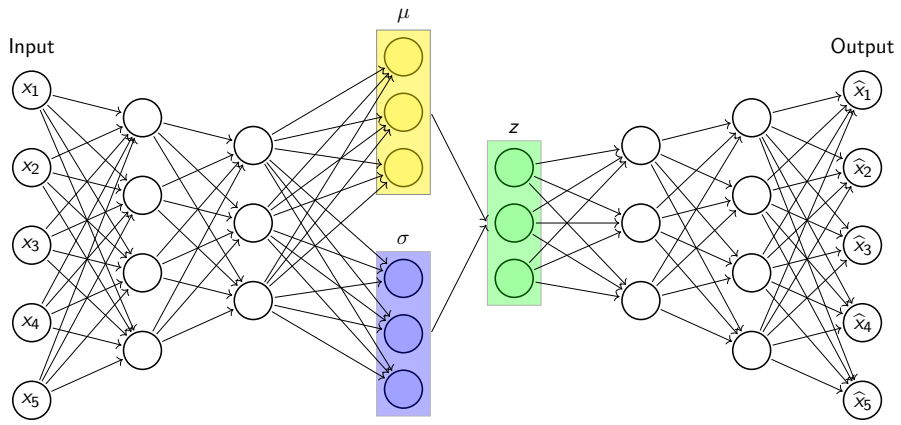


Figure 2.14: Diagram of the conventional architecture of a VAE.

conditional distribution of the observation $p_\theta(x|z)$, which takes a latent sample as input and outputs the parameters for the observation's conditional distribution. The encoder defines the approximate posterior distribution $q_\varphi(z|x)$, parameterized by φ, which receives an observation as input and outputs the set of parameters for specifying the conditional distribution of the latent representation z. A Gaussian prior distribution is assumed, $p_\theta(z)$, for the latent variables z. Figure 2.14 shows an example of the structure of a VAE, where the encoder network outputs the mean and log-variance parameters of a diagonal Gaussian. For numerical stability, the output of the encoder is the log of the variance rather than the variance directly. The parameters θ and φ are undetermined and must be derived from the data.

For training of the ANNs that compose VAE, a sample z is obtained from the latent distribution defined by the parameters φ outputted by the encoder, given an input observation x. But, as this sampling operation causes a bottleneck since backpropagation cannot flow through a random node, it is applied a reparameterization trick (explained next) and a stochastic gradient optimization [Bank, 2020; Baldi, 2021].

There are two variables, μ and σ, for each sample z that determine the mean and standard deviation of the Gaussian distribution corresponding to z. The accumulation of all Gaussian distributions in the integration domain yields the original distribution $p_\theta(x)$:

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz \tag{2.53}$$

In equation 2.53, $p_\theta(z) = \mathcal{N}(0,1)$ and $p_\theta(x|z) = \mathcal{N}(\mu(z), \sigma(z))$. Because $p_\theta(z)$ is known and $p_\theta(x|z)$ is unknown, the expressions of μ and σ must be solved. However, because $p_\theta(x)$ complexity, μ and σ are difficult to calculate. Therefore, the ANNs that comprise VAE are trained to tune and achieve the marginal log-likelihood $\log p_\theta(x)$ instead:

$$Maximum\ \ell(\theta, \varphi) = \sum_x \log p_\theta(x) \tag{2.54}$$

Next, Equation 2.54 is rewritten so that the distribution with respect to which the gradient is taken is independent of parameter θ. To achieve this, $\log p_\theta(x)$ is decomposed into the following, where the stochastic element $q_\varphi(z|x)$ independent of θ is integrated:

$$\begin{aligned}
\log p_\theta(x) &= \int q_\varphi(z|x) \log p_\theta(x) dz = \int q_\varphi(z|x) \log \left(\frac{p_\theta(z,x)}{p_\theta(z|x)} \right) dz \\
&= \int q_\varphi(z|x) \log \left(\frac{q_\varphi(z|x)}{p_\theta(z|x)} \right) dz + \int q_\varphi(z|x) \log \left(\frac{p_\theta(z,x)}{q_\varphi(z|x)} \right) dz \\
&= KL(q_\varphi(z|x)||p_\theta(z|x)) + \int q_\varphi(z|x) \log \left(\frac{p_\theta(z,x)}{q_\varphi(z|x)} \right) dz
\end{aligned} \tag{2.55}$$

The first term in Equation 2.55 corresponds to the Kullback-Leibler (KL) divergence, whose value may be greater than or equal to zero. Consequently, the second term is the Variational Lower Bound (ELBO) of $\log p_\theta(x)$, thus:

$$\log p_\theta(x) \geq ELBO = \int q_\varphi(z|x) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{q_\varphi(z|x)} \right) dz = \mathbb{E}_{q_\varphi(z|x)} \left[\log \frac{p_\theta(z,x)}{q_\varphi(z|x)} \right] \tag{2.56}$$

By adjusting $q_\varphi(z|x)$ to increase ELBO, KL divergence decreases. When $q_\varphi(z|x)$ is adjusted so that $q_\varphi(z|x)$ and $p_\theta(z|x)$ are equal, KL divergence gets closer to 0 and ELBO and $\log p_\theta(x)$ are fully consistent. Accordingly, ELBO can be adjusted to be equal to $\log p_\theta(x)$, and since ELBO is the lower bound of $\log p_\theta(x)$, solving for maximum $\log p_\theta(x)$ is equivalent to solving for maximum ELBO. Therefore:

$$\begin{aligned}
\ell(\theta, \varphi) &= \int q_\varphi(z|x) \log \left(\frac{p_\theta(z,x)}{q_\varphi(z|x)} \right) dz = \int q_\varphi(z|x) \log \left(\frac{p_\theta(x|z) p_\theta(z)}{q_\varphi(z|x)} \right) dz \\
&= \int q_\varphi(z|x) \log \left(\frac{p_\theta(z)}{q_\varphi(z|x)} \right) dz + \int q_\varphi(z|x) \log p_\theta(x|z) dz \\
&= -KL(q_\varphi(z|x)||p_\theta(z)) + \int q(z|x) \log p_\theta(x|z) dz \\
&= -KL(q_\varphi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\varphi(z|x)} [\log p_\theta(x|z)]
\end{aligned} \tag{2.57}$$

Maximizing ELBO is equivalent to minimising $KL(q_\varphi(z|x)||p_\theta(z))$ and maximising the second term in the equation above for all data points with respect to θ and φ.

Extending this VAE network to sequential data, such as human movements, would require utilizing as encoder and decoder temporal networks, such as RNNs, that capture the dependencies between latent variables at various time steps. The loss function would thus be the

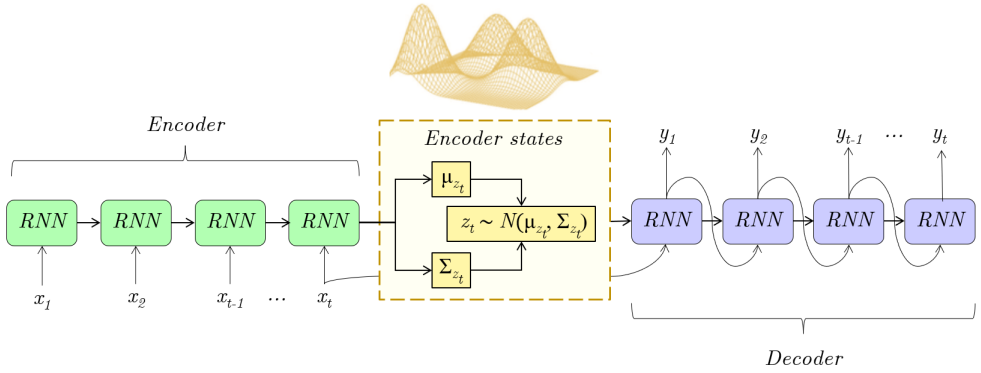


Figure 2.15: Diagram of a variational Seq2Seq network.

following [Fraccaro, 2016; Li, 2019]:

$$\ell(\theta, \varphi) = \sum_{t=1}^{T} \underbrace{-KL(q_\varphi(z_t|z_{t-1}, x_t, y_t)||p_\theta(z_t|z_{t-1}, x_t))}_{\text{Regularization loss}} + \underbrace{\mathbb{E}_{q_\varphi(z_t|z_{t-1}, x_t, y_t)}[\log p_\theta(y_t|z_t, x_t)]}_{\text{Prediction loss}} \tag{2.58}$$

Given a sequence of body poses $x = [x_1, x_2, ..., x_t]$, the encoder outputs distribution parameters from which the KL divergence is calculated. The z latent representation is then sampled from μ and σ. The decoder receives the latent vector z as input and returns the predicted pose sequence $y = [y_1, y_2, ..., y_t]$ against which the prediction loss is computed. A diagram of a variational Seq2Seq network is presented in Figure 2.15.

2.4.2.3 Autoencoder with attention mechanism

The attention mechanism in autoencoders was first introduced by Bahdanau et al. [Bahdanau, 2014] for Seq2Seq models. This approach is intended to overcome the bottleneck issue that arises with the use of a fixed-length encoding vector, in which the decoder has limited access to the entire input sequence. This limitation is especially problematic when modeling lengthy or complex sequences, as the dimensionality of their representation is constrained to match that of shorter or simpler sequences. The Bahdanau attention, also known as Additive attention, and the Luong attention [Luong, 2015], often referred to as multiplicative attention, are the two most prominent approaches.

In Bahdanau attention, similarly to the traditional AE for modeling of human movements, the conditional probability modeled by the decoder is:

$$p(y_t \mid y_{1:t-1}, c_t) = g(y_{t-1}, s_t, c_t). \tag{2.59}$$

However, the probability is conditioned on a distinct context vector c_t for each target output y_t. The context vector is determined by a sequence of hidden states or annotations ($h = [h_1, ..., h_T]$) to which the encoder maps the input pose sequence x of length T. Each h_i

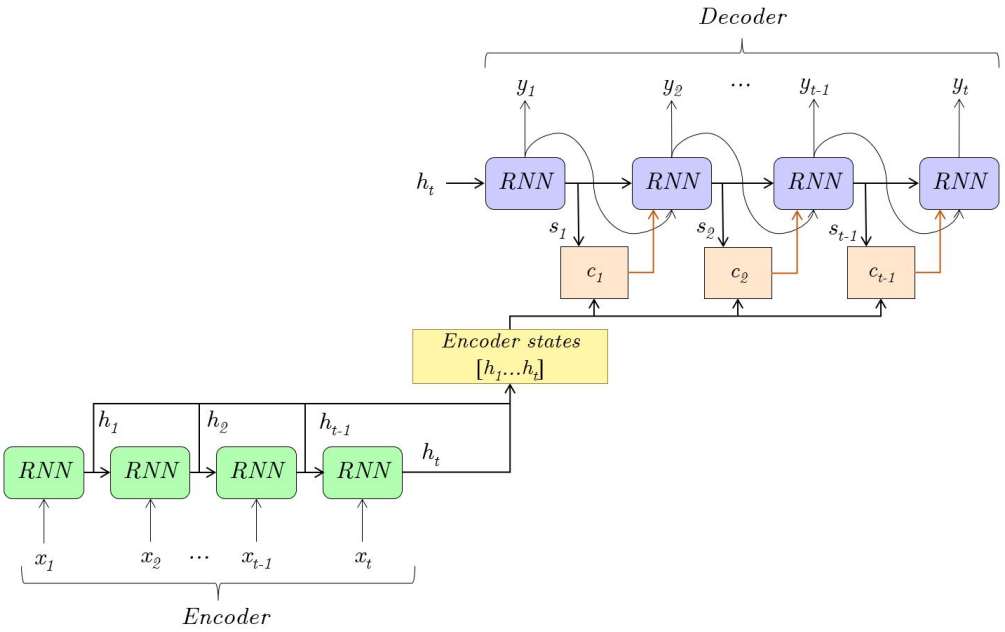


Figure 2.16: Diagram of an Bahdanau attention mechanism on a Seq2Seq network.

comprises information about the entire input sequence, with a particular emphasis on the sections surrounding the i-th value. The context vector is computed as a weighted sum of these h_i:

$$c_t = \sum_{i=1}^{T} w_{t,i} h_i \tag{2.60}$$

The weights $w_{t,i}$ for each h_i are calculated through a softmax as follows:

$$w_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{T} e_{t,j}} \tag{2.61}$$

$e_{t,i}$ are the alignment models which scores how well the input sequence elements align with the current output at the position t:

$$e_{t,i} = a(s_{t-1}, h_i) \tag{2.62}$$

The alignment model is represented by a function a, which can be implemented by a neural network. Figure 2.16 shows a diagram of the overall process for Bahdanau Attention on a Seq2Seq network.

The Luong attention brought various enhancements to the Bahdanau attention, including introducing two new classes of attention mechanisms: a global approach that attends to the entire input sequence and a local approach that only attends to a subset of the input when predicting the output. The global attention resembles Bahdanau's in terms of attending to the entire input sequence, but the architecture is simplified. Local attention was primarily influenced by the attention mechanisms outlined by Xu et al. [Xu, 2015], in which only a small portion of the sequence is attended. The whole process of Luong attention in an AE model

begins similarly to that of Bahdanau. However, the computation of the context vector varies, and it is then used to compute an attentional hidden state $\widetilde{s}_t$. An attentional hidden state is calculated by a weighted combination of the context vector and the current decoder hidden state:

$$\widetilde{s}_t = tanh(W_c\,[c_t; s_t]) \tag{2.63}$$

The decoder then generates a final output by passing a weighted attentional hidden state:

$$y_t = softmax(W_y \widetilde{s}_t) \tag{2.64}$$

Note that W_c and W_y are the parameters to be learned by the RNNs.

The global attention mechanism, as mentioned, considers all of the values in the input pose sentence when calculating the alignment scores and, ultimately, the context vector. For calculating alignment scores, there are three alternate methods:

$$a(s_t, h_i) = \begin{cases} s_t^\top h_i \\ s_t^\top W_a h_i \\ v_a^\top tanh(W_a\,[s_t; h_i]) \end{cases} \tag{2.65}$$

The first and second employ a multiplicative attention, and the third is comparable to Bahdanau's since it is based on the concatenation of s_t and h_i.

By attending to the entirety of the input posture sequence, global attention can become computationally expensive and impractical for lengthy movements. Similar to the soft and hard attention mechanisms suggested by Xu et al. [Xu, 2015], the local attention mechanism is designed to address these limitations by focusing on smaller sets of postures and their descriptors. The local attention mechanism proposed by Luong [Luong, 2015] constructs a context vector by calculating a weighted average over the set of annotations h_i, within a window $[p_t - D, p_t + D]$ centred on an aligned position p_t, and where D is empirically selected. The value of p_t can be determined using either monotonic alignment or predictive alignment. The monotonic alignment presupposes that both the input and output sequences are monotonically aligned, so $p_t = t$. In the predictive alignment, the mechanism predicts an aligned position based on learned parameters W_p, v_p, and the length of the sequence input T:

$$p_t = T \cdot sigmoid(v_p^\top tanh(W_p, s_t)) \tag{2.66}$$

As a result of sigmoid, $p_t \in [0, T]$. When determining the alignment scores, a Gaussian distribution centred around p_t is used to prioritize input poses near p_t. Figure 2.17 shows a diagram of the Luong Attention on a Seq2Seq network.

2.4.3 Recent works and challenges

Conventional methods based on HMMs, Gaussian Processes, restricted Boltzmann machine, and dynamic random forest have been outperformed by data-driven approaches in simulating human movement [Liu, 2017; Kulsoom, 2022; Rudenko, 2020]. The underlying similarity of

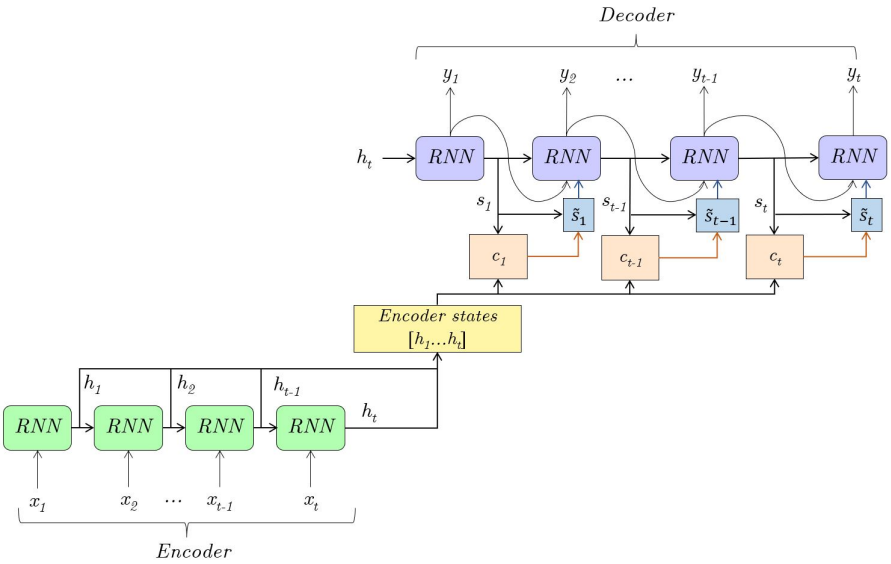


Figure 2.17: Diagram of an Luong attention mechanism on a Seq2Seq network.

motion forecasting and sequence-to-sequence prediction tasks has led research in this domain toward encoder-decoder architectures. Previous research has employed RNNs as encoders and decoders, networks that have become the standard in sequential human movement analysis.

To predict pedestrians' 2D position and orientation, Sun et al. [Sun, 2017] used an LSTM autoencoder. First, the encoder received spatial and temporal context information of diverse pedestrians (3D positions of pedestrians in real-world coordinates). Then, it learned the human activity patterns generated in different environments at different times of the day. Xue et al. [Xue, 2018] proposed the Social-Scene-LSTM (SS-LSTM), which uses three LSTM encoders to capture individual, social, and scene scale information. RGB images of the scenario are used to extract various features (position trajectories, occupancy maps), as well as feature maps from convolutional layers. The encoder's output was then utilized by an LSTM decoder to predict the pedestrian trajectory coordinates.

For the prediction of full-body human movements, recent works have integrated skeletal representations to include spatial correlations among joints. Pavllo et al. [Pavllo, 2018] trained a two-layer GRU using quaternion sequences to predict 3D human postures. The loss function utilized forward kinematics with Euler joint angles to compute the joints' positions. Then, the Euclidean distance between the predicted and real positions of the joints was calculated. Liu et al. [Liu, 2019] introduced the Hierarchical Motion Recurrent (HMR) model to predict upcoming motion sequences. By representing skeletal frames using the Lie algebra representation, the authors were able to capture spatial correlations and model the global and local motion contexts using hierarchical LSTMs. Tang et al. [Tang, 2018] presented an AE with an attention mechanism, in which the AE focuses on the human's moving joints for motion prediction. Shu et al. [Shu, 2022] employed a similar concept to learn the spatial coherence and temporal evolution of joints using a co-attention mechanism. Mao et al. [Mao, 2019] suggested a feedforward model for forecasting 3D postures. The Discrete Cosine Transform (DCT) was

used to model the temporal dependencies of human movement. Then, a Graph Convolutional Network (GCN) was utilized to represent the joints of the human body as a graph to capture the spatial structure information of the human body. Similarly, Cai et al. [Cai, 2020] utilized DCT to transform the motion into the frequency domain. The frequency components were then processed using a transformer-based architecture (global attention mechanism) in order to capture spatio-temporal correlations of the human pose. Instead of modeling attention on the full body alone, another work by Mao et al. [Mao, 2021b] proposes combining the predictions of three AE with attention. Each processes movement on distinct levels: the whole body, body parts, and single joints.

VAE has been applied for stochastic motion prediction, where they predict multiple and diverse motion sequences in the future from a single input sequence [Yan, 2018; Petrovich, 2021; Aliakbarian, 2021; Mao, 2021a]. Aliakbarian et al. [Aliakbarian, 2021] accomplished this through conditional VAE, whereas Mao et al. [Mao, 2021a] used a VAE to generate the motion of various body parts in a sequential manner. In order to enhance advances in ANNs for probabilistic time series forecasting, Chung et al. [Chung, 2015] and Fraccaro et al. [Fraccaro, 2016] used RNNs to build connections between SSMs and VAEs. This led to the emergence of deep state-space models, in which ANNs are used to parameterize the non-linear observation and transition models (shown in Section 2.3.2.1). Deep Kalman filters (DKFs) initially introduced exogenous input to SSMs [Krishnan, 2017]. Li et al. [Li, 2019] parameterized a deep SSM using a VAE framework and RNNs to capture long-term dependencies. The method trained an Automatic Relevance Determination (ARD) network, which included exogenous variable information in the predictions. The ARD ultimately assisted in identifying valuable exogenous variables and suppressing those that were irrelevant for forecasting. In another work, deep SSMs used RNNs to generate the parameters of a linear-Gaussian state-space model (LGSSM) at each time step for forecasting [Rangapuram, 2018]. For probabilistic forecasting, Salinas et al. [Salinas, 2017] proposed the DeepAR, which uses auto-regressive RNNs with mean and standard deviation as output. For probabilistic prediction of human movements, Liu et al. [Liu, 2020b] introduced a deep SSM. The deep SSM utilized CNNs as encoders and decoders as part of a AE architecture and used joint positions, velocities, and accelerations as motion descriptors. The deep SSM provided a unified formulation for multiple human motion systems and enabled the accurate prediction of 3D human postures.

Even though there has been a lot of progress in modeling human movements recently, with ANNs that make impressive predictions and simulations of human movements, as these approaches get more complicated, they become harder to understand and interpret their results. They can learn highly complex non-linear relationships from large datasets and surpass humans and other methods at many tasks. Nevertheless, their obscurity restricts their applicability and inspires little confidence among scientists and analysts who, for example, undertake the prognosis of movement disorders. Suppose, for instance, that an ANNs predicts with high confidence that an operator will do an ergonomically dangerous action based on their body motion patterns, but provides no insight into the specific features of the action that make it harmful. In this scenario, it is unclear how this knowledge could be utilized to proactively

protect the operator from the risk. So, while complex networks can handle activity recognition and event detection problems that put predictive accuracy above interpretability, models that can be intuitively interpreted, like analytical models, are better for applications that help people learn and improve their skills in handicrafts, industry, or sports, as well as for medical diagnostic and prognostic tools.

2.5 Conclusion of the chapter

This chapter reviewed MoCap technologies, the diverse kinematic and kinetic descriptors that can be obtained from each technology, and techniques for extracting new features that have been used to model human movements. Selecting the MoCap system would largely depend first on the human motion application, whether precise human tracking or human posture measures are required. Secondly, if the application is meant to be used outside MoCap-dedicated laboratories to monitor everyday activities. As this dissertation aims to develop methods that can be used to analyze whole-body movements performed in diverse scenarios, IMUs were selected for motion capture.

Joint angles can be used as motion descriptors as they can be accurately estimated using inertial MoCap data. For that, it is necessary to select the appropriate rotation representation according to the intended application, whether it be regression, computer graphics, or human movement analysis. Because the main goal of the thesis is to develop interpretable models for the comprehension of human movement, the Euler angles were selected for modeling. These can be intuitively interpreted and help illustrate easier how human movements are executed. Nevertheless, it is critical to consider the issues associated with employing Euler angles, such as the gimbal lock and the potential of singularities. These must be addressed in either the data processing or the design of the proposed method. Note that although feature extraction and selection techniques have aided in improving modeling performance, none of these techniques will be used in this research since the proposed models are intended to describe full-body movements with Euler angles.

Different approaches for human motion modeling were briefly discussed. These include biomechanical models, stochastic models, hybrid biomechanical-stochastic models, and conventional and cutting-edge data-driven architectures for sequence modeling. The fundamental idea behind the analytical models (biomechanic, stochastic, and hybrid models) is to represent human motion systems by means of a set of analytical equations. These typically rely on simplifying assumptions about the behavior of the human musculoskeletal system. Thus, the analytical model's accuracy depends on the formulation and selection of these assumptions. In contrast, data-driven approaches infer representations by observing their performance in predicting or detecting human movements. Consequently, the accuracy of data-driven approaches is reliant on the representativeness of the training dataset. The disadvantage of these learned representations is that they cannot be mapped to physical quantities, which makes it difficult to intuitively understand how they function.

It was observed that data-driven approaches are being used for nearly all human motion

modeling challenges, with a trend to replace traditional methods, such as biomechanical or stochastic methods, where feature selection is essential. Nevertheless, this migration procedure is still ongoing. The applicability of data-driven approaches in motion-based applications is still limited due to their low ability to explain their outcomes. For example, to generate human motion representations that can serve both to predict human postures accurately and explain how these are performed. Several works have accomplished this utilizing hybrid biomechanical and stochastic models (Section 2.3.3), but their inference methods are not as powerful or scalable for human motion applications that require the analysis and modeling of multiple human movements.

Deep state-space models have shown potential for developing novel methods for estimating interpretable human motion representations, as they make use of the advantages of both the state-space theory (from which motion representations can be defined) and ANNs (strong modeling ability). Deep state-space models have proven that they can learn complex, high-dimensional sequential data distributions; however, their deployment in human movement analysis has not yet been extensively studied. Lastly, attention mechanisms in ANNs have attracted interest in human motion modeling due to their substantial performance gains in Natural Language Processing (NLP). The ability of the attention mechanism to select and focus on specific components of their input while ignoring other available information could be leveraged in new approaches for modeling human movement, since meaningful spatial and temporal dependencies can be captured simultaneously.

Chapter 3

Building databases to study the movements of real industrial operators and skilled craftsmen

"A mind that is stretched by a new experience can never go back to its old dimensions."

— Oliver Wendell Holmes

Contents

3.1 Introduction

This chapter describes the datasets created for the training and evaluation of the models proposed in this thesis. As stated in Chapter 2, previous studies on human movement have helped scientists comprehend body dynamics and their stochastic nature. Due to the fact that these studies and developed algorithms rely on motion data for analysis, a systematic quantitative evaluation is necessary to establish how well a method performs compared to the present state of the art. For this purpose, motion capture datasets have been made public to serve as a baseline. However, the majority of the datasets available were recorded inside a laboratory, causing inaccurate measurements since they lack authenticity.

New datasets were created in this thesis for the analysis of actual operators' and artisans' motions. These were captured in actual workplace settings, and their analysis aims to develop models that aid in comprehending how these experts accomplish all of these diverse and complex full-body movements. The recorded professional movements can be divided into two categories: industrial and crafts.

In the industrial sector, there is significant interest in developing systems that enable the tracking and prediction of operators' motion descriptors in order to use this knowledge proactively, for example, to enhance human-robot collaboration or ergonomic risk prevention. Therefore, operators from factories that manufacture televisions and airplane parts were recorded within the framework of the European Project Collaborate[1]. In addition, to promote research in ergonomics analysis, subjects in a laboratory were recorded performing 28 movements with varying ergonomic risk levels as defined by the European Assembly Worksheet [Schaub, 2013]. For the digitalization of knowledge of heritage crafts, the movements of skilled artisans from silk-weaving and glassblowing, and mastic farmers were captured in their natural environments as part of the European project Mingei[2]. These recordings were made to analyze the experts' gestural knowledge and dexterity while practicing their crafts. All datasets are accessible in Zenodo[3], following the General Data Protection Regulation (GDPR).

The existing MoCap datasets are first reviewed in Section 3.2, which mainly involve recordings made in laboratory settings. The data acquisition procedures followed to build the datasets are then described in Section 3.3, where descriptions of each recorded task are also provided. Next, the data processing and segmentation of the recordings are explained in Section 3.4. Lastly, it is provided a brief conclusion of the chapter.

3.2 Existing motion capture datasets

Widely used datasets in the literature are the Human3.6M [Ionescu, 2014], the Carnegie Mellon University (CMU) MoCap dataset [Carnegie Mellon University,], and the Archive of Motion Capture As Surface Shapes (AMASS) dataset [Mahmood, 2019]. The Human3.6M is an indoor dataset consisting of RGB videos and 2D and 3D body annotations of seven participants

[1]EU Horizon 2020 Research and Innovation Programme under Grant Agreement No. 820767
[2]EU Horizon 2020 Research and Innovation Programme under Grant Agreement No. 822336
[3]Benchmark website: https://doi.org/10.5281/zenodo.5356992

executing a variety of actions. These actions include ordinary actions such as talking on the phone, conversing, smoking, etc. The CMU dataset contains the movements of 144 participants who perform a wide variety of complex movements (Sports, everyday activities, communication gestures for Human-Computer Interaction). This was recorded with 12 infrared MX40 cameras in a MoCap laboratory. The AMASS dataset is a big collection of different and random human movements. This one was captured with optical MoCap systems and represents the MoCap data within SMPL body model parameters. HumanEva and MoVi [Sigal, 2010; Ghorbani, 2021] are two existing datasets that contain video and marker-based MoCap data of a single person performing ordinary activities and sports movements. For movements in multiperson interactions and scenarios, Van der Aa et al. [Van Der Aa, 2011] presented the UMPM benchmark. Some popular action-recognition datasets containing multiple daily actions and sportive activities are NTU RGB+D [Liu, 2020a] and Action3D [Xia, 2012], recorded with Microsoft Kinect V2 cameras. Another is the Penn Action dataset [Zhang, 2013], which consists of 2326 video sequences of people performing 15 distinct sports-related motions.

A dataset of human movements in industry-like activities was created by Maurice [Maurice, 2019], where subjects performed assembly activities and were recorded with a full-body inertial MoCap suit. For analyzing the tasks performed by construction workers, there are the datasets VTT-ConIot [Mäkela, 2021] and DeTECLoad [Lee, 2020], which were recorded using IMUs. The first was created for the evaluation of activity recognition algorithms using a small set of inertial sensors. The second is for the generation of algorithms to analyze and detect excessive load-carrying tasks with different carrying modes performed in construction. Lastly, there is the IKEA ASM dataset [Ben-Shabat, 2021], which contains 371 furniture assembly videos of 48 subjects from three camera views, including 3D depth.

Besides these available datasets, there is still a need for MoCap data that includes a greater diversity of movements, particularly professional movements captured in real-world scenarios.

3.3 Data acquisition

This section begins with a description of the MoCap system used for recording, followed by information on the subjects captured for each dataset. The description of each movement in the datasets is presented next, followed by some illustrations of the recordings. Note that the descriptions and images offered in this section will be referenced in Chapters 4 and 5 when discussing the results of each chapter.

3.3.1 Motion capture technology

The BioMed bundle motion capture system from Nansense Inc.[4] was utilized to capture the movement of industrial operators and craftsmen. The system is composed of a full-body suit with 52 IMUs strategically positioned across the torso, limbs, and hands. At a rate of 90 frames per second, the sensors measure the orientation and acceleration of body segments on the articulated spine chain, shoulders, arms, legs, and fingertips. After a recording, the Euler

[4]Baranger Studios, Los Angeles, CA, USA

local joint angles on the X, Y, and Z axes are automatically calculated through the Nansense Studio's inverse kinematics solver and stored in a Biovision Hierarchy format (BVH). A BVH file is a text file comprised of two parts. The first part provides a hierarchical description of the skeleton, beginning with the root (hips) and proceeding to the extremities of each limb. The second part of the file contains, for each frame of the recording, the absolute position of the root of the skeleton and the angles of the joints defined in the first part of the BVH file.

3.3.2 Subjects recruited

For the creation of each dataset, industrial operators and skilled artisans agreed to be recorded in their actual workplace while wearing the Nansense suit. Firstly, industrial operators from a television plant in Istanbul, Turkey, and an aerospace company in Bucharest, Romania, were captured as they carried out their professional tasks. Four healthy people, three men and one woman, participated in the MoCap recording session at the television plant. Their average age was 31.5±6.2 years, their height was 167.8±4.6 cm, and their average weight of 65.3±9.9 kg. Two male subjects participated in the MoCap session for the recordings in the aerospace company. They had an average age of 50±5 years, a height of 170±2 cm, and a weight of 77±1.4 kg.

Ten healthy individuals consented to participate in MoCap recordings of potentially dangerous ergonomic postures in a neutral environment laboratory. The subjects consisted of three women and seven men. The average age was 28.7±4.6 years, with an average height of 172.9±9.2 cm, and the average weight was 70.5±12.9 kg. None of them sustained musculoskeletal injuries, and they all completed all trials in under one hour.

Motions of skilled artisans performing in three different crafts were recorded. The first is a master silk weaver recorded at a traditional jacquard workshop in Krefeld, Germany. The expert's height was 168 cm, and his weight was 62 kg. The second artisan is a master glassblower who was recorded in action during a glassblowing workshop. The glassblower's height was 177 cm, and his weight was 73 kg. Finally, two mastic farmers were recorded at a mastic cultivation field in Chios, Greece. Their average age was 30.5±5.5 years, their height was 178.8±8.5 cm, and their average weight was 69.3±8.0 kg.

3.3.3 Recording of the professional tasks

Next, the procedure followed for each recording is outlined, as well as each captured movement. Before recording, a calibration procedure was done. The subject assumed different postures, such as I-pose or T-pose, and performed different movements, like walking or touching his fingertips, each for 10 seconds. In order to facilitate the later annotation and segmentation of the data, only operators and artisans were asked to explain each component of the task prior to the recording.

3.3.3.1 Industrial-related tasks

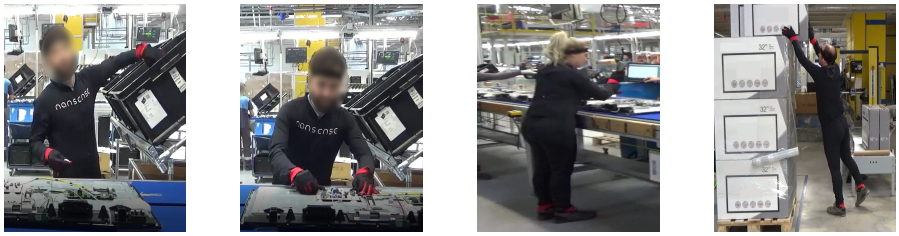
The movements performed in two industrial settings have been recorded for CoLLaboratE, delivering natural movements while operators execute industrial tasks. The tasks were captured on-site during regular production by actual operators.

3.3.3.1.1 Televising manufacturing Two tasks were recorded at a television manufacturing plant related to assembly and packaging. The set of movements involved in each task is designated by the abbreviations **TVA** (assembly) and **TVP** (packaging). The television assembly task consists of mounting electronic circuit boards to a television chassis and using a power tool to drive screws into the boards to secure them firmly. For this procedure it was defined the following motion vocabulary:

1. TVA_1: Reaching high with one hand, above shoulder level, to pick one component (circuit board) from a container.
2. TVA_2: Reaching low with the other empty hand, below the knee level, to pick up the second component (wire) from a second container.
3. TVA_3: Connecting the components and placing the board on the chassis to be screwed.
4. TVA_4: Drilling four screws on the circuit board by holding the driller with the right hand and placing the screws with the left.

The final operation required stacking the completed, boxed televisions on wooden pallets and wrapping them in a plastic membrane for shipping (TVP). The following set of movements were recorded for this procedure:

1. TVP_1: Placing eight TVs on a wooden pallet (bottom level).
2. TVP_2: Preparing to wrap the bottom level with a membrane.
3. TVP_3: Wrapping the bottom level.
4. TVP_4: Placing eight TVs on top of the bottom level (second level).
5. TVP_5: Wrapping the second level with a plastic membrane.
6. TVP_6: Placing eight TVs on top of the second level (third level).
7. TVP_7: Wrapping the third level with a plastic membrane.
8. TVP_8: Placing eight TVs on top of the third level (fourth level).
9. TVP_9: Wrapping the fourth level with a plastic membrane.



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

Figure 3.1: Professional movements in television manufacturing. (a) Grab the circuit board from a container (TVA_1); (b) Connect the circuit board and wire and place them on the TV chassis (TVA_3).(c) Drilling four screws on the circuit board (TVA_4).(d) Placing a television box on top of the third level (TVP_8).

Boxes are given to the operator through a conveyor belt. He places one box at a time onto the pallet using both hands. After stacking eight boxes on a single level, he grabs the plastic membrane with both hands and wraps them by going around them with it. After wrapping them properly, the operator proceeds to stack boxes on top of the previous one wrapped, repeating the process. The task is complete when there are four levels of boxes on the pallet.

All tasks associated with television assembly were recorded over the course of an eight-hour shift, with one subject recorded installing the circuit boards during the first half of the shift and another recorded drilling the circuit boards to the television chassis during the second half. Three subjects were recorded separately for the packaging tasks during one shift. In Figure 3.1 is illustrated some of the movements recorded in television assembly and packaging.

3.3.3.1.2 Airplane floater assembly The complete riveting procedure for an airplane floater was captured in an aerospace company. The floater is a plane component that enables planes to float when they land on water. The set of movements recorded from this procedure is denoted as **APA**. Collaboration between two operators is essential for this activity. Therefore, their data were collected sequentially; one person wore the MoCap suit to capture their movement while collaborating, and then donned it to the second person and continued the activity. The following movements were recorded, which are also illustrated in Figure 3.2:

1. APA_1: Rivet with the pneumatic hammer.
2. APA_2: Prepare the pneumatic hammer and grab rivets.
3. APA_3: Place the bucking bar to counteract the incoming rivet.

One iteration of rivet assembly consisted of the first operator placing a rivet in one hole (Figure 3.2a). The second operator from the opposite side of the floater then positions the bucking bar to counter the rivet (Figure 3.2c). After precisely positioning the bucking bar, the second operator signals the first operator to activate the pneumatic hammer. The first operator verifies the proper placement of the assembled rivet by touching it, then moves on to

(**a**) (**b**) (**c**)

Figure 3.2: Example of airplane assembly movements. (a) Rivet with the pneumatic hammer (APA_1); (b) Prepare the pneumatic hammer and grab rivets (APA_2); (c) Place the bucking bar to counteract the incoming rivet (APA_3).

the next hole and the process is repeated. After completing one line of rivets, the first operator grabs additional rivets and prepares the pneumatic hammer for the second line (Figure 3.2b).

The motion of the fingers during the riveting with the pneumatic hammer was not recorded because the operator could not work realistically while wearing the MoCap gloves. The operator needed to touch with his bare hands the rivet to determine whether it was positioned correctly.

3.3.3.1.3 Motion primitives with varying ergonomic risk level A recording protocol consisting of 28 distinct motion primitives was designed to capture postures with varying ergonomic risk levels based on EAWS. Each motion was repeated three times, giving a total of 84 MoCap recordings per subject. The recorded motions were neutral as they were not associated with a specific activity but rather served solely to demonstrate several ergonomically incorrect postures. The motions can be divided into three main categories: those performed standing, those performed while in a chair, and those executed while kneeling. The motions are progressing from comfortable postures to increasingly more uncomfortable but never dangerous ones. All postures were held for six seconds, and no particular discomfort was reported. This set of 28 motions with different ergonomic risk levels is denoted as **ERGD**.

Initially, the subject is standing with a straightened back. The subject then assumes the following three postures:

- $ERGD_1$: The subject remains standing straight up, with the arms relaxed (I-pose).
- $ERGD_2$: The subject rotates their torso to the left as far as they can for six seconds.
- $ERGD_3$: The subject bends laterally the torso to the left for six seconds.

For the next three postures, the torso is slightly bent forwards:

- $ERGD_4$: The subject remains in the bending position for six seconds.
- $ERGD_5$: While the subject is bending forward, they rotate their torso to the left and hold this position for six seconds.
- $ERGD_6$: While the subject bends forward and rotates their torso to the left, they extend their arm as if trying to reach something that is on the ground.

The next three tasks have the torso bending forward at a large angle ($> 60°$):

- $ERGD_7$: The subject remains in the bending position for six seconds.
- $ERGD_8$: While the subject has bent forwards, they rotate their torso to the left and hold this position for six seconds.
- $ERGD_9$: While the subject bends forward and rotates their torso to the left, they extend their arm as if trying to reach something that is on the ground.

In the next few tasks, the position of the arms will change, and the torso motions will be repeated:

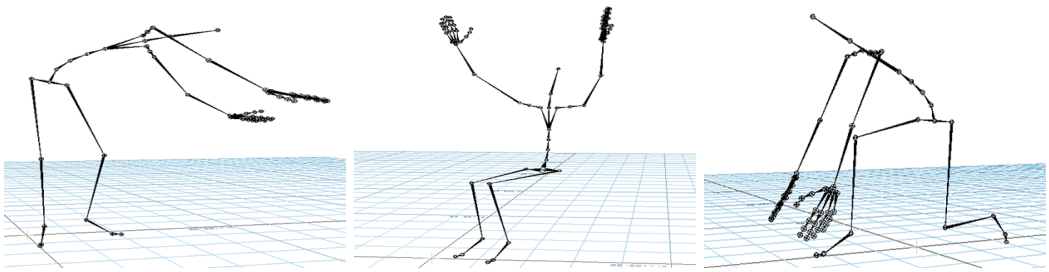
- $ERGD_{10}$: The subject is standing upright with the forearms bend at $90°$ and the arms raise at the shoulder level, perpendicular to the floor.
- $ERGD_{11}$: With the arms at the same position as P10, the subject rotates their torso, and laterally bends to the left.
- $ERGD_{12}$: The participant raises their arms perpendicular to the ground while the forearms are fully extended. They proceed by rotating and laterally bending their torso to the left.
- $ERGD_{13}$: The subject raises their arms above the head for six seconds.
- $ERGD_{14}$: With the arms above the head level, the subject rotates and laterally bends to the left for six seconds.

These were all the postures that were assumed from a standing position. The next part describes the postures that will be recorded while the person is seated on a chair.

- $ERGD_{15}$: The person is sitting on a chair with the arms relaxed (neutral position).
- $ERGD_{16}$: While seated, the subject bends forward at an angle of $60°$ or more.
- $ERGD_{17}$: The subject bends forwards at an angle of $60°$ or more while rotating their torso and bending laterally to the left.
- $ERGD_{18}$: The subject repeats P17 but has their arms extended in front of them.
- $ERGD_{19}$: The subject raises their arms above the head level while they are fully extended.
- $ERGD_{20}$: With the arms above the head level, the participant will rotate and laterally bend their torso to the left.

Finally, the remaining tasks will be performed while the subject is kneeling on their right knee. These are the most ergonomically uncomfortable postures. Beyond that, the upper body options will be the same as before:

- $ERGD_{21}$: The subject stays upright.



(**a**) (**b**) (**c**)

Figure 3.3: Example of motion primitives based on EAWS contained in ERGD. (a) $ERGD_7$: Standing while bending forward and rotating the torso; (b) $ERGD_{19}$: Sitting while raising arms above shoulder level; c) $ERGD_{28}$: Kneeling while bending forward.

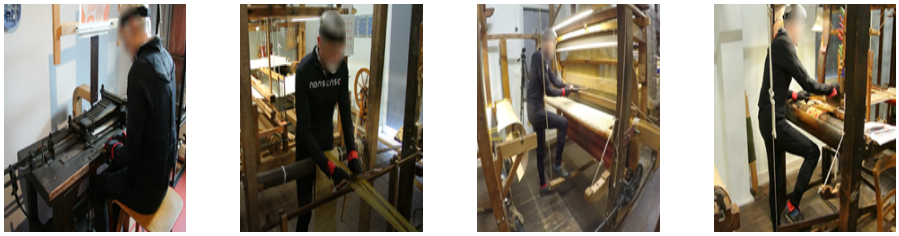
- $ERGD_{22}$: The subject rotates their torso to the left as far as they can, they remain in that position for six seconds.
- $ERGD_{23}$: The subject laterally bends their torso to the left.
- $ERGD_{24}$: The subject bends forward at an angle larger than 60°.
- $ERGD_{25}$: While bending the torso at an angle larger than 60°, the participant rotates and laterally bends their torso to the left.
- $ERGD_{26}$: The P25 task is repeated, but this time, the person's arms are extended as if to pick something up from the ground.
- $ERGD_{27}$: The subject raises their arms to be perpendicular to the ground.
- $ERGD_{28}$: With the arms raised, the subject rotates and laterally bends their torso to the left.

After completing the recordings, ERGD has examples from the most comfortable positions to some of the most ergonomically improper according to the risk factors defined by EAWS. Though those motions are not in the context of any specific goal, they can act as a baseline to test different methods of an ergonomic assessment. An example of three postures assumed by the subjects are shown in Figure 3.3.

3.3.3.2 Crafts-related tasks

Master artisans and mastic farmers were captured doing their professional tasks in their realistic workplaces for the Mingei project. An additional MoCap session was conducted to capture the simulation of the procedure for cultivating mastic without using any material or tools.

3.3.3.2.1 Silk weaving In a jacquard loom workshop in Krefeld, Germany, the movements of on skilled silk weaver were captured. This set of movements is referenced as **SLW**. Throughout three days, the expert was recorded performing the following silk weaving-related tasks:



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

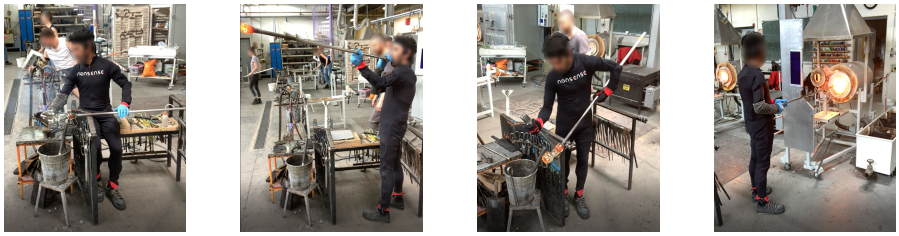
Figure 3.4: Examples of the jacquard weaving tasks recorded. (a) Creation of the punch cards (SLW_1); (b) Preparation of the beam (SLW_3); (c) Weaving on a large size loom ($SLW_{4,3}$).(d) Weaving on a small size loom ($SLW_{4,1}$).

1. SLW_1: The creation of the punch cards.
2. SLW_2: Wrapping of the beam.
3. SLW_3: Preparation of the beam.
4. $SLW_{4,1:3}$: Jacquard weaving with looms of different sizes (small, medium, and large).

Figure 3.4 illustrates some examples of the movement recorded. On the first day, the silk weaver was recorded performing SLW_1, SLW_2, and SLW_3 continuously. The creation of the punch cards was recorded for one hour. Due to the complexity and length of the tasks, the wrapping and preparation of the silk beams were recorded only once, taking about four hours to record. The next two days consisted of continuous recordings of the expert weaving using looms of three different sizes. The recording only stopped when the weaver switched to a different loom. The task of waiving with a loom can be divided into three main movements ($SLW_{4,1}$,$SLW_{4,2}$, and $SLW_{4,3}$). Firstly, the expert pushes the pedal down with his right leg at the same time that he pushes away the threads with his left hand (the initial posture of the weaver is shown in figures 3.4c an 3.4d). Then, by controlling the shuttle that passes the thread horizontally with the right hand, he sends the shuttle to the other side with a quick pulling gesture. Finally, he pulls back the threads with the left hand while simultaneously releasing the pedal with the right leg. This process is repeated up to the end of the piece.

3.3.3.2.2 Glassblowing The creation of a glass decanter was recorded four times at Vannes-le-Châtel, France, in a European center for research and training in glasswork. Because the temperature of the glass had to be maintained throughout the process, each trial was recorded without pausing between movements. This resulted in one motion file for each attempt, which starts with collecting the molten glass and finishes when the decanter is left to cool down. The set of movements composing the process of creating one decanter is denoted as **GLB**.

The glass decanter was created in three stages. To begin, inflate and shape the molten glass inside the decanter's main body (container). The base was created next, followed by the handle. Next, the expert rolled and shaped the decanter throughout the task to prevent the glass from deforming due to gravity. Finally, an assistant was necessary to blow into the glass while the expert shaped the decanter's main body. Figure 3.5 shows some of the tasks that



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

Figure 3.5: Example of gestures captured in a glassblowing workshop. (a) Shape the decanter's curves (GLB_4); (b) Blow through the blowpipe (GLB_3); (c) Shape the decanter's neck with pliers (GLB_8); (d) Laying the cord on the decanter (GLB_9).

were recorded during the decanter's fabrication. For shaping the molten glass, the glassblower constantly rotated with his left hand the blowpipe while shaping the glass with his right hand. He utilized various tools with his right hand, including a block (Figure 3.5a), jacks (Figure 3.5c), soffietta, shears, and metal pencils. These were employed to give the glass the form of the decanter and to add further decorative details. The block is used to maintain the glass's round shape. The jacks are used to shape the decanter's cervix. The shears were utilized to cut the glass and form the decanter's peak. The soffietta forms the decanter's top. Metal pencils were then used to add the handle and extra glass details (cord around the neck) and make the foot (base) of the decanter. Manipulating the tools required constant movement of the right shoulder, right arm, and right forearm. At the same time, the glassblower was seated, rotating back and forth with the left hand the blowpipe on a metal structure. Moving the blowpipe on the metal structure required a small bending to keep the grip of the blowpipe. Placing the handle or shaping the cervix with the jacks required at some times for the glassblower to stand up, but he kept moving the blowpipe with the left hand.

While forming the glass, the artisan frequently put the glass on the blowpipe into the furnace (Figure 3.5d). He also continuously blew through the blowpipe while holding it horizontally at shoulder height with both arms to maintain the decanter's round shape (Figure 3.5b). After finishing the decanter, it was passed to a punty to cool down.

3.3.3.2.3 Mastic cultivation The cultivation of mastic was recorded in the span of three days in Chios, Greece. The first and second days' recordings were made outside, in front of a mastic tree. The recordings of the last day were simulated inside a room. Each movement was divided into separate recordings due to the nature of the cultivation process. This resulted in separate MoCap files for each part of the process. In general, the cultivation of mastic was recorded realistically. However, specific tasks are, in reality, done days or weeks apart or take hours to be completed. As such, the expert was required to demonstrate the gestures briefly while remaining realistic. The movements recorded from this cultivation process are denoted as **MSC**. Some movements that were captured from the mastic farmer are shown in Figure 3.6.

The process begins with the preparation of the soil beneath the trees. So that dripping mastic can be easily collected, the earth surrounding the tree is cleaned and the terrain around

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

Figure 3.6: Example of movements captured related to the cultivation of mastic. (a) Sweep the soil under the plant (MSC_3); (b) Cover the area under the three with calcium carbonate (MSC_4); (c) Harvesting the tree with a razor (MSC_5); (d) Collect the mastic (MSC_8).

the tree trunk is leveled. The farmer was recorded using two distinct tools to scrape the soil. The first is an antique agricultural tool (Amia) with a metal head and wooden handle, similar to a trowel. With this one, the farmer scraped the soil on his knees, holding the tool with his right hand. The second tool is a shovel, which allows the farmer to scrape the soil while standing. The farmer then swept the ground with a short broom (Figure 3.6a). After preparing the soil, the farmer evenly distributed calcium carbonate ($CaCO_3$) on the ground to create a flat surface. For this task, the farmer knelt and spread the white dust with his right hand while holding the container with his left (Figure 3.6b).

The tree is then cut in order to obtain mastic. There are three different tools to do incisions in the tree. The first is a small tool with sharp points at the ends (Kenditiri), the second is another small tool called Timitiri, and the third is a small axe. The farmer was standing while using each tool, but he had to lean over to make the incisions in the tree. The tools were held with the right hand. The next step recorded was the gathering and harvesting of the mastic that had emerged from the tree's wounds. The farmer picked the fallen mastic using a small basket and tweezers (Figure 3.6d), and then harvested more resin off the tree with a razor (Figure 3.6c). Both movements required the farmer to bend and manipulate the tool with his right hand.

The farmer wiped the soil to collect it on a metal mesh with a brush. In order to remove dust from the mastic, the mesh is continuously moved (or shifted). The use of two types of mesh was recorded. For all variants, the farmer knelt and moved the mesh with both hands. Finally, a third method for removing the dust from the mastic was recorded: throwing the mastic and dust while standing into the wind.

3.4 Data processing and segmentation

The processing of the MoCap consisted of two steps. To begin, a low pass filter was applied, followed by the correction of incorrect postures caused by electromagnetic interference or sensors drifting when the recording lasted too long, and calibration was required. A low-pass Butterworth filter was applied to the raw MoCap data to eliminate high-frequency noise. To avoid over smoothing the data, the cut-off frequency was selected using the power spectrum

Table 3.1: Segmentation of the television assembly tasks.

Task	Motion	Repetitions
Television Assembly	TVA_1	107
	TVA_2	107
	TVA_3	108
	TVA_4	157
Packaging	TVP_1	8
	TVP_2	2
	TVP_3	7
	TVP_4	5
	TVP_5	12
	TVP_6	7
	TVP_7	7
	TVP_8	4
	TVP_9	2

Table 3.2: Segmentation of the riveting procedure.

Task	Motion	Repetitions
Riveting	APA_1	6
	APA_2	5
	APA_3	8

density of the signal.

The MoCap system's sensors may drift or be influenced by magnetic disturbances from surrounding metallic objects during the recording process. As a result, occasionally erroneous joint angles were recorded during otherwise precise motion capture. The recordings were adjusted to correct this error using a 3D character animation software[5]. The software was used to adjust the unrealistic movements based on common sense and video feedback. After adjusting and removing noise from the MoCap data, it was segmented by movements. Firstly, recordings were collected per task, with one recording representing a whole task; however, these recordings were later segmented by gestures. A task may contain a single gesture that is performed numerous times, or it may contain additional gestures that are repeated throughout the task.

The segmentation of the television assembly and packaging was based on repetitions of the movements given in Section 3.3.3.1. The repetitions segmented from the recordings are shown in Table 3.1. For the riveting task, the segmentation of the first movement consisted of riveting and completing an entire line. The second movement was to set up the pneumatic hammer for the next line of rivets. Lastly, the final gesture involved placing a bucking bar for an entire line of rivets. Table 3.2 illustrates the final segmentation. The recordings of movements with different ergonomic risk levels were segmented into repetitions. Given that ten subjects were recorded assuming 28 poses three times, segmentation produced 840 files containing one repetition of each pose.

The tasks recorded from the silk weaving, glassblowing, and mastic cultivation procedures, were segmented by single movements (as there were repetitions). The resulting segmentation is displayed in tables 3.3, 3.4, and 3.5.

In order to facilitate the training of the models described in the next chapters, the discontinuities of the Euler joint angles present in part of the MoCap files were reduced manually. These discontinuities are dramatic shifts between the values 180° and -180° in only certain local joint angles. By examining each MoCap file, it was determined to transform the time series with discontinuities to a data of range $[-250°, 250°]$. Note that this transformation may

[5]MotionBuilder, Autodesk Inc., San Rafael, CA. USA

Table 3.3: Segmentation of the silk weaving tasks.

Task	Motion	Repetitions
Creating a card	SLW_1	110
Beam preparation	$SLW_{2,1}$	3
	$SLW_{2,2}$	2
	$SLW_{2,3}$	4
	$SLW_{2,4}$	1
	$SLW_{2,5}$	1
Wrapping the beam	SLW_3	2
Weaving with small size loom	$SLW_{4,1,1}$	11
	$SLW_{4,1,2}$	11
	$SLW_{4,1,3}$	11
Weaving with medium size loom	$SLW_{4,2,1}$	35
	$SLW_{4,2,2}$	35
	$SLW_{4,2,3}$	35
Weaving with large size loom	$SLW_{4,3,1}$	16
	$SLW_{4,3,2}$	16
	$SLW_{4,3,3}$	15

Table 3.4: Segmentation of the glassblowing tasks.

Task	Motion	Repetitions
Beak cutting	GLB_1	11
	GLB_2	6
Blowing and shaping	GLB_3	5
	GLB_4	8
	GLB_5	15
	GLB_6	7
	GLB_7	35
Cervix refining	GLB_8	6
Cord laying	GLB_9	2
	GLB_{10}	8
	GLB_{11}	4
Finish details	GLB_{12}	5
Handle laying	GLB_{13}	4
	GLB_{14}	5
	GLB_{15}	4
Transfer to punty	GLB_{16}	4
Leg and foot laying	GLB_{17}	6
	GLB_{18}	7

Table 3.5: Segmentation of the mastic cultivation procedure.

Task	Motion	Repetitions
Scrapping (New tool)	MSC_1	3
Scrapping (Old tool)	MSC_2	9
Sweeping	MSC_3	9
Dusting	MSC_4	9
Embroidery A	MSC_5	9
Embroidery B	MSC_6	3
Embroidery with an axe	MSC_7	3
Gathering	MSC_8	8
Harvesting	MSC_9	7
Wiping	MSC_{10}	6
Shifting A	MSC_{11}	6
Shifting B	MSC_{12}	3
Cleaning with the wind	MSC_{13}	3

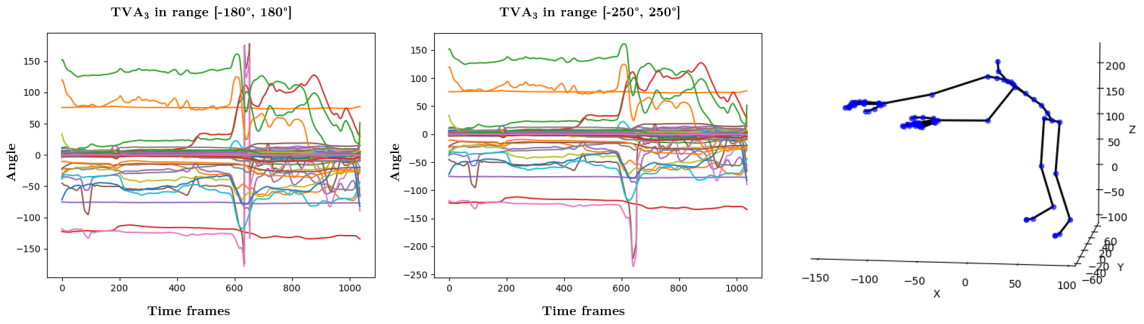


Figure 3.7: Transformation of MoCap data representing the movement TVA_3, which is the placement of a circuit board on a television frame.

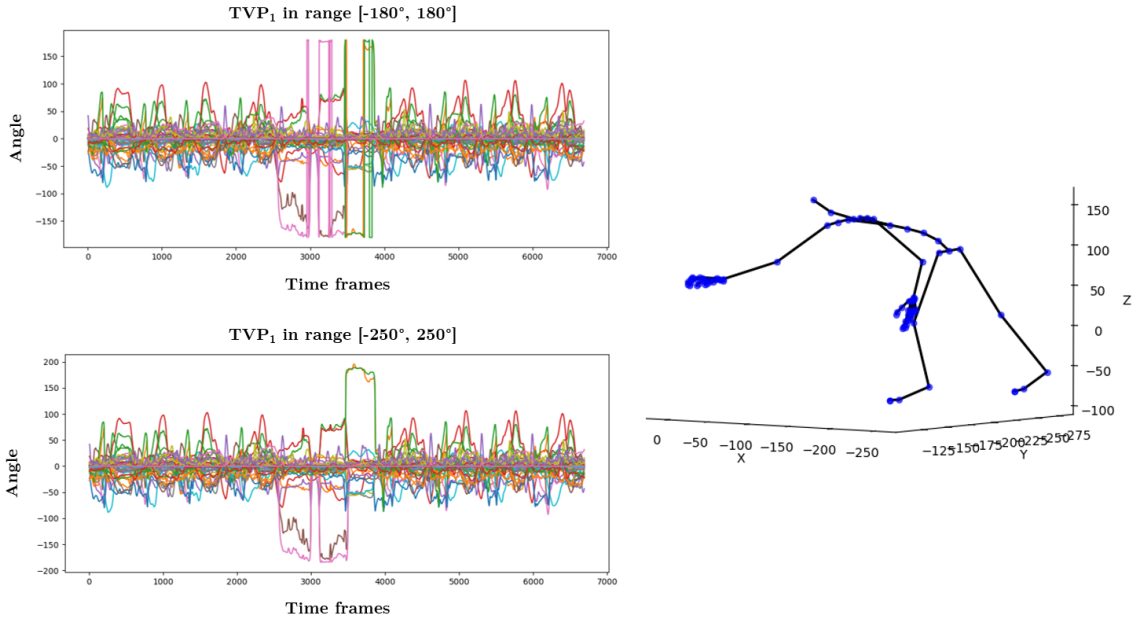


Figure 3.8: Transformation of MoCap data representing the movement TVP_1, which is the placement of eight television boxes on the first level of a pallet.

not be appropriate for new movements recorded with IMUs. Nonetheless, it was sufficient to eliminate most discontinuities in the datasets presented in this chapter. Each transformation was documented so that the transformed data may be inversed to Euler angles. Some examples of these transformations are represented in figures 3.7 to 3.13. These figures illustrate the MoCap data before and after the modifications, as well as the reconstructed skeleton.

The angles from the arms and forearms and one angle of the Hips were mainly the local angles with discontinuities. The angle of the Hips on the Y axis (pointing up, measuring torso rotation) was the most problematic and prone to drifting. The explanation for this could be related to the sensor's position. If the suit is loose, the sensor can produce inaccurate readings. Another factor is that after the suit is turned on and connected to the computer for recording, the subjects must move their entire body to "wake up" the sensors. This sensor was most likely still in an idle state while performing calibrations. Any MoCap file with a distortion caused by drifting or poor calibration was removed from the datasets. The total size of the seven

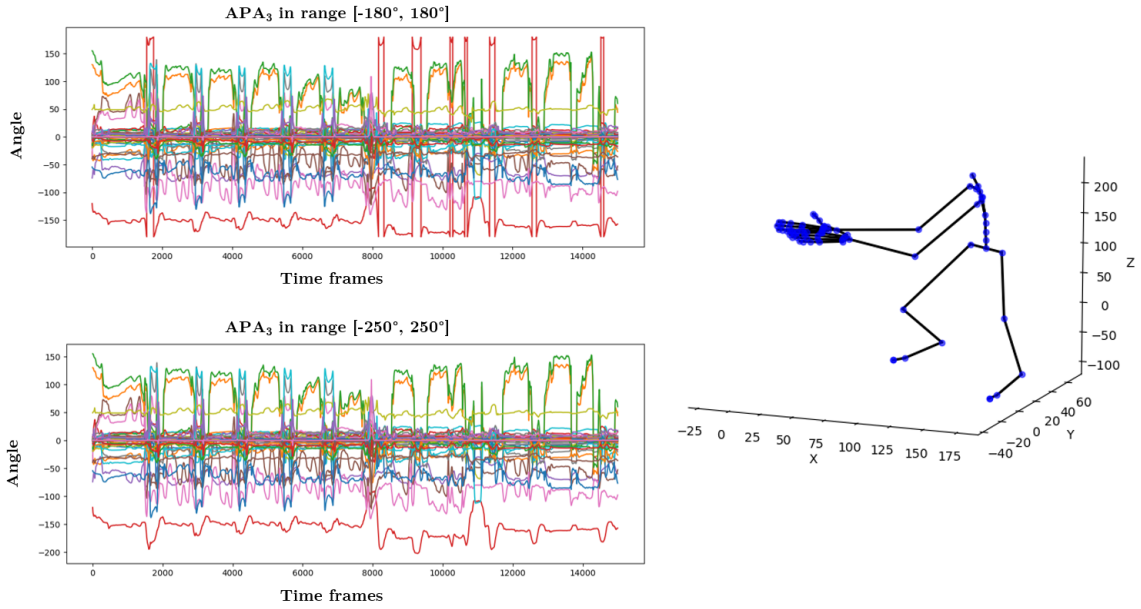


Figure 3.9: Transformation of MoCap data representing the movement APA_3, which is riveting a full line of an airplane float structure.

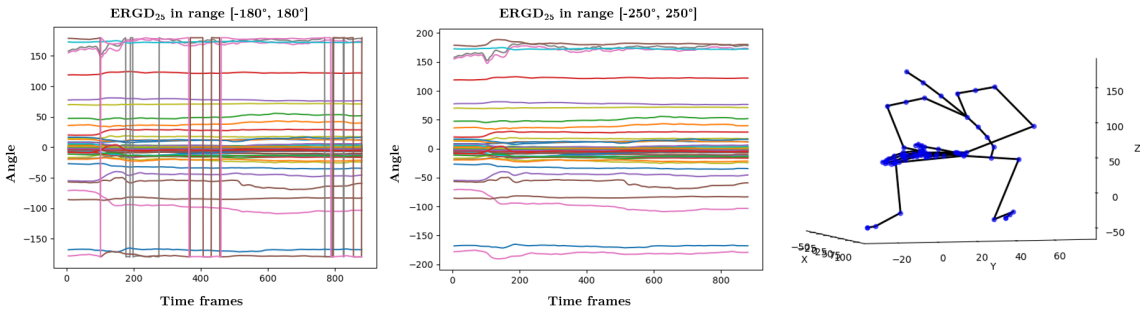


Figure 3.10: Transformation of MoCap data representing the movement $ERGD_{25}$. In this movement primitive, the subject rotates and laterally bends their torso to the left while kneeling and bending the torso at an angle larger than 60°.

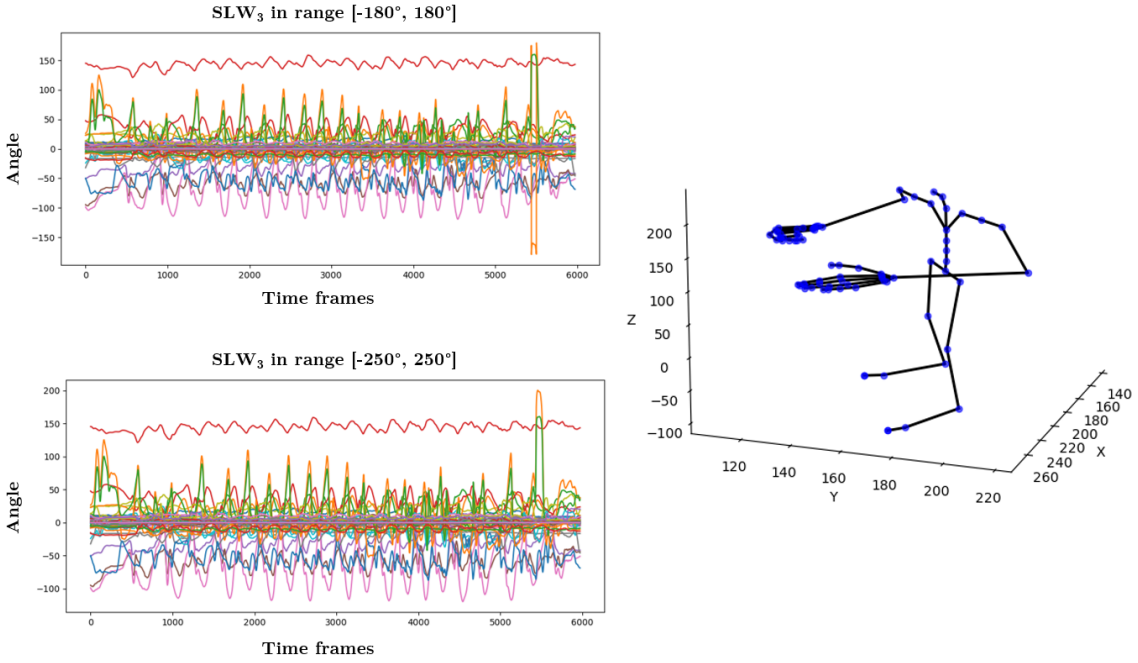


Figure 3.11: Transformation of MoCap data representing the movement $SLW_{3,4}$, which is a set of movements performed while preparing the silk beam.

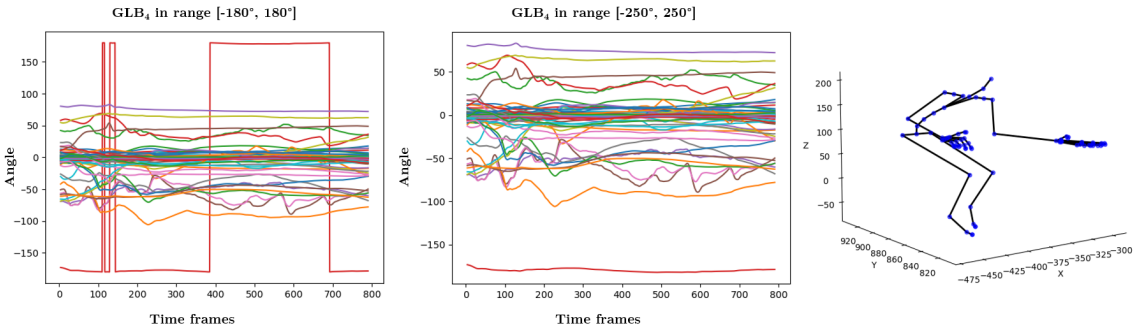


Figure 3.12: Transformation of MoCap data representing the movement GLB_4, which is the movement of shaping the molten glass with a block while simultaneously rotating the blowpipe.

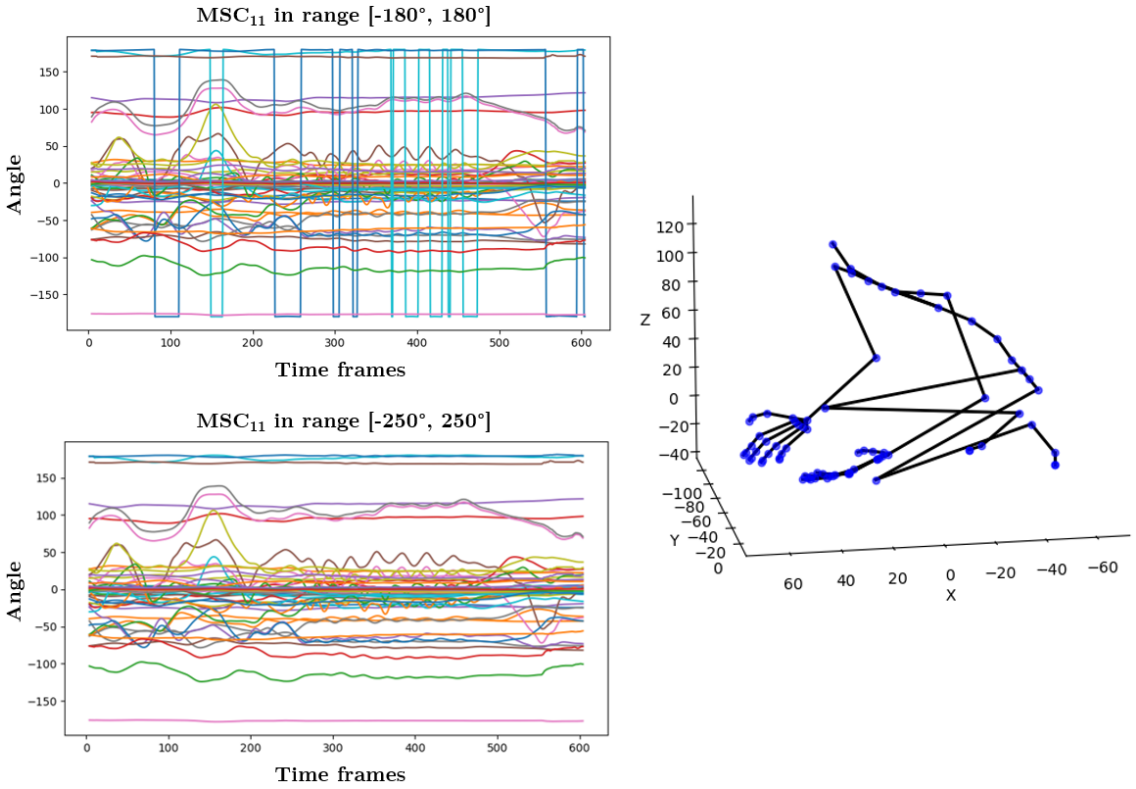


Figure 3.13: Transformation of MoCap data representing the movement MSC_{11}, which is the movement of separating the mastic from dust and stones using a metal mesh.

datasets utilized in the following chapters is 5GB. A total of 163,4776 frames, or 5 hours and 2 minutes, make up the segmented movements with 156 local joint angles measured.

3.5 Conclusion of the chapter

This chapter presented seven datasets: TVA, TVP, APA, ERGD, SLW, GLB, and MSC. Most publicly available datasets contain simulated movements performed in a laboratory and related to everyday activities or sports. Therefore, new datasets were created containing movements performed in professional tasks either from the industry or crafts workshops. These were recorded with actual operators and experts in their real workplace scenarios using an inertial full-body suit of 52 sensors. The aim was to test the proposed analytical models with these complex movements in the following chapters and extract information regarding the dexterity, skill, and know-how related to the adequate use of tangible elements such as materials and tools. Each professional task was segmented by repetitions, and discontinuities were reduced to improve the modeling of the movements.

Chapter 4

Modeling and simulation of human movements using one-shot training and data-driven strategies

"Life can only be understood backwards; but it must be lived forwards."

— Søren Kierkegaard

Contents

4.1 Introduction

Any voluntary movement of the body segments is accomplished via the musculoskeletal system. The musculoskeletal system is an intricate structure comprised of bones, muscles, ligaments, and tendons. Thus, modeling a structure with such complexity is not an easy task. However, despite the fact that the musculoskeletal system is primarily responsible for the complexity of human locomotion, it can be acceptable to represent human movements using analytical models that include relevant assumptions about body joint associations and their temporal dependencies. This thesis hypothesizes that human motion dynamics can be modeled by analytical models that can be interpreted and whose assumptions take into account the stochasticity of human motion and physical body structure. Consequently, given the nature of the hypotheses defined in this thesis and its specific objectives, a hybrid stochastic-biomechanical approach based on kinematic descriptors was selected to model the dynamics of human movements and create interpretable representations for human motion trajectories. This approach consists of GOM, detailed in Section 2.3.3.2.

Representing human motion as a state-space model of a dynamic system provides a simplified mathematical formalization of the motion phenomenon and approximates it to reality, for instance, through static or dynamic simulation [Manitsaris, 2020]. In addition, the mathematical representation of GOM permits a more intuitive description of how body joints cooperate (spatial dynamics) and evolve over time (temporal dynamics). GOM has been demonstrated to be effective at simulating human joint position trajectories. Furthermore, due to the usage of a transition function, it performs well with observations obtained from varied environments and subjects without requiring extensive training datasets [Olivas-Padilla, 2021]. This generalization capability is essential for applications requiring rapid and accurate analysis of varied human movements.

This chapter presents three novel approaches for modeling human movements using full-body motion representations of GOM. The first method estimates the motion parameters using statistical modeling, whereas the second and third methods employ ANNs. Each approach trains time-varying motion representations that can be utilized to simulate realistic human movements. The simulation performance achieved with each approach is evaluated and compared with that attained with the first GOM version that uses constant motion representations. The seven datasets described in the preceding chapter are utilized for training. Regarding the capability of the trained analytical models to explain human movements, it is later evaluated in Chapter 5.

The following section describes the used constant and time-varying mathematical representations of GOM. Additionally, it expands on the applicability of GOM for human movement analysis. The parameterization of constant and time-varying GOM representations using one-shot training is then described in Section 4.3. The statistical estimation of models with constant coefficients is referred to as KF-GOM, whereas that of models with time-varying coefficients is called KF-RGOM. Section 4.4 introduces the two methods for learning time-varying GOM representations using data-driven approaches. These methods consist of deep state-space models

based on a variational autoencoder and an autoencoder with Luong attention. Each method is designated as VAE-RGOM and ATT-RGOM, respectively.

The results of the evaluation of each model's simulation capabilities to support the first hypothesis of this thesis are then presented in Section 4.5. In addition, the outcomes of a sensitivity analysis are provided to examine the stability and behavior of the trained models when their input is affected by external stimuli. The chapter concludes with a discussion of the results in Section 4.6 and general conclusions in Section 4.7.

4.2 Definition of the motion representation based on GOM

As described in Section 2.3.3.2, the first version of GOM consists of an equation system of autoregressive models with constant coefficients estimated using MLE via KF. When modeling full-body movements, there is an equation system of autoregressive models, each modeling one of the motion descriptors measured. This equation system corresponds to GOM. Suppose a human movement is depicted as a sequence of human postures $P_t = [P_1, P_2, ..., P_T] \in \mathbb{R}^{T \times N}$. T is the length of the posture sequence and $N = J \times D$, where J is the number of joints measured, and D is the number of dimensions that the joint's motion descriptor is decomposed. The number of models in the equation system is equal to the number of dimensions associated with a given body joint (D), multiplied by the number of body joints (J) captured with the MoCap system. Inside these N models are defined four different assumptions of variables that account for the dynamic relationship between body joints and their temporal dependencies. These correspond to the transitioning assumptions (H1), intra-joint associations (H2), inter-limb synergies (H3), and intra-limb serial and non-serial mediations (H4). Each assumption consists of a specific set of variables (motion descriptors) that are parametrized and depict a particular relationship between body joints or a temporal dependency. By examining the generated coefficients and statistical significance of each variable, it can be gleaned how relevant these are according to the movement modeled and the predicted trajectory.

Human postures are expressed as 3D Euler joint angles in order to generate movements with subjects of various morphologies. Unlike joint positions, Euler joint angles are unaffected by identity-specific body shape. Moreover, Euler angles can be intuitively interpreted in the analytical model and clearly illustrate how human movements are conducted. Figure 4.1 depicts the sensors' placement, labeling, and orientation. For the purposes of this work, only measurements from 19 inertial sensors were used for the modeling. Discarding MoCap data from the fingers and feet to simplify the human motion representation. Thus, 57 joint angles were modeled in GOM.

State-space modeling was performed to create the mathematical representations, where a second-order model is designed for each motion descriptor that incorporates the assumptions as endogenous and exogenous data. Second order due to the correlation between lag values (auto-correlation) in the time series. For example, while modeling the Euler angle trajectory of the body joint P_t on the X-axis (Px_t), whose movement is decomposed on XYZ axes (Px_t, Py_t, and Pz_t) and has an association with j body parts. The two prior values are integrated

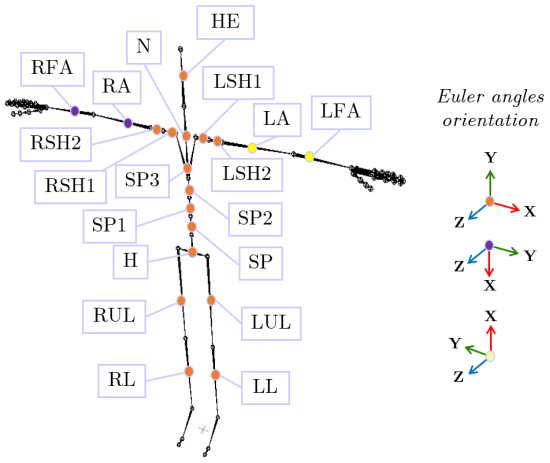


Figure 4.1: Location and Euler angle orientation of the sensors that provide the *XYZ* joint angles included in GOM.

into the transition model as shown in Equation 4.1, where s_t corresponds to the state variable at time t. Then, exogenous data (u_t), corresponding to the variables from H2, H3, and H4, are included in the observation model as illustrated in Equation 4.2.

$$s_t = As_{t-1} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} Px_{1,t-1} \\ -Px_{1,t-2} \end{bmatrix} \tag{4.1}$$

$$\begin{aligned} Px_{1,t} = \begin{bmatrix} 1 & 1 \end{bmatrix} s_t + Bu_t = \\ \begin{bmatrix} 1 & 1 \end{bmatrix} s_t + \beta_1 Py_{1,t-1} + \beta_2 Pz_{1,t-1} + \beta_3 Px_{2,t-1} + \cdots + \beta_n Px_{j,t-1} \end{aligned} \tag{4.2}$$

By merging equations 4.1 and 4.2, the state-space representation of the motion descriptor is obtained:

$$\begin{aligned} Px_{1,t} = \underbrace{\alpha_1 Px_{1,t-1} - \alpha_2 Px_{1,t-2}}_{\text{H1}} + \underbrace{\beta_1 Py_{1,t-1} + \beta_2 Pz_{1,t-1}}_{\text{H2}} + \\ \underbrace{\beta_3 Px_{2,t-1} + \cdots + \beta_n Px_{j,t-1}}_{\text{H3 or H4}} \end{aligned} \tag{4.3}$$

The motion representation is parametrized by $\alpha \in \mathbb{R}^{1\times 2}$ and $\beta \in \mathbb{R}^{1\times N}$. Overall, GOM comprises 57 models as Equation 4.3, each of which represents a joint motion trajectory.

The first version of GOM, denoted as KF-GOM, utilizes constant coefficients (α and β), implying that the SSM assumes that observations are created linearly from hidden states using a linear dynamical model that does not vary over time [Luttinen, 2014]. As a result, the relationship between endogenous and exogenous variables is consistent across all time series. In this version, the model is simple to analyze and efficient to learn due to the assumptions of linearity and constant dynamics. Moreover, it can be trained with small data sets (one reference movement for training), and it is relatively easy to detect irrelevant assumptions with the trained model. The majority of real-world processes cannot be effectively described by linear Gaussian models, yet in many cases, processes behave approximately linearly within specific

restrictions [Manitsaris, 2014]. KF-GOM offers a benchmark by which other more advanced and complex models can be evaluated. Nonetheless, the constant GOM has certain drawbacks. Primarily, it cannot distinguish the variance unique to specific periods of the time series. During training, the model is unable to directly identify this contribution because the error cannot be disaggregated beyond a single error structure for the entire time series.

Given that human movement is a stochastic process, presuming that the relationship between the dependent variable and its assumptions will be the same for all time series is implausible. Therefore, the use of time-varying coefficients in GOM is proposed in this thesis. This chapter introduces three distinct estimation methods for these coefficients. The first is by applying the MLE via KF, as was done for computing constant coefficients, but now it is inferred values for the time sequence of vectors of unknown parameters $\alpha_t = [\alpha_1, ..., \alpha_T]$ and $\beta_t = [\beta_1, ..., \beta_T]$, having motion representations such as the following:

$$Px_{1,t} = \underbrace{\alpha_{t,1} Px_{1,t-1} - \alpha_{t,2} Px_{1,t-2}}_{\text{H1}} + \underbrace{\beta_{t,1} Py_{1,t-1} + \beta_{t,2} Pz_{1,t-1}}_{\text{H2}} + \underbrace{\beta_{t,3} Px_{2,t-1} + \cdots + \beta_{t,n} Px_{j,t-1}}_{\text{H3 or H4}} \tag{4.4}$$

The second and third are data-driven approaches, further detailed in Section 4.4.

4.2.1 Potential applications of GOM

The trained GOM, which contains the previous mathematical representations, can be utilized for two main applications. First, the joint angle trajectories of each modeled movement can be simulated by solving the simultaneous equation system that composes GOM. The simulation can be either static or dynamic, with models predicting a single time frame per iteration. The static simulation implies that all endogenous and exogenous variables are real data samples. On the other hand, in the dynamic simulation, the endogenous variables are not real samples of motion data but rather the model's previous prediction values.

Accurate simulation of human movement has a variety of applications, including professional training in technical skills, monitoring, animations, and movement analysis for ergonomics, music, and medical purposes. From the machine learning perspective, the proposed models can be used for data augmentation, an approach frequently used in supervised learning to increase the models' robustness. The second application of GOM is that, as an analytical model, the trained models can be used to obtain insights into the dynamic relationship between body joints during the performance of a movement. In Chapter 5, the use of GOM for analyzing body dexterity is explained and validated. This dexterity analysis corresponds to describing the movement of each joint based on its estimated mathematical representation. This information can be utilized to teach technical motor skills, such as by comparing the motion representations trained using an expert craftsman's movements to those trained with a beginner's movements. The next chapter details other applications of GOM, such as feature selection and the creation of tolerance intervals representing the acceptable range of motion for reproducing a specific movement

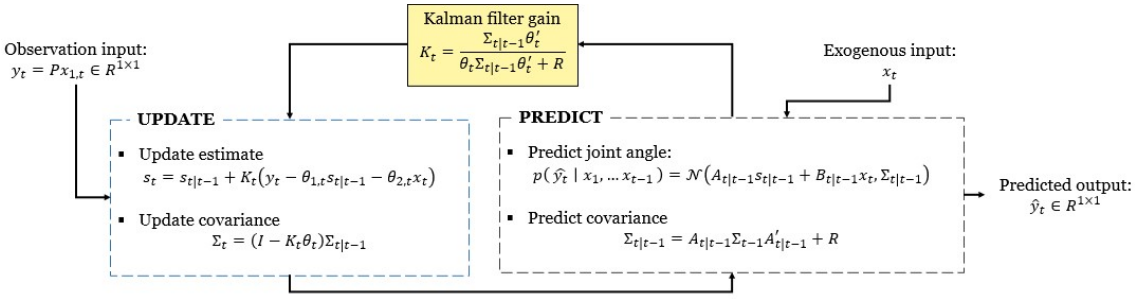


Figure 4.2: Flow chart of the iterative process of the KF while doing Maximum Likelihood Estimation (MLE) for estimating GOM's coefficients.

4.3 Learning of constant and time-varying GOM representations using one-shot training

The first approach trains GOM representations using one-shot training with Kalman Filters (KF-RGOM). The fundamental concept is to formulate each motion representation as a separate SSM and then use KF to compute the log-likelihood of the observed motion descriptor for the given set of parameters. Suppose that the joint angle on the X-axis of a body part, $Px_{1,t}$, is modeled. Its GOM representation is the Equation 4.3. The observation would be the real joint angle, $y_t = Px_{1,t}$, where $Px_{1,t} \in \mathbb{R}^{1\times1}$, and the variables from the assumptions H2, H3, and H4 correspond to the exogenous input x_t. The following log-likelihood is then maximized concerning all time-varying coefficients θ and α and β, utilizing the KF to calculate the log-likelihood for each time t:

$$\ell(\theta, \alpha, \beta) = \sum_{t=1}^{T} \log p_\theta(y_1, \dots y_{t-1} | x_1, \dots x_{t-1}) \tag{4.5}$$

θ corresponds to the tuning parameters of the KF. The diagram of the iterative process of the Kalman filter for calculating the likelihood in 4.5 for every time t is illustrated in Figure 4.2. Note that $\widehat{y}_t$ corresponds to the prediction of the motion descriptor using the motion representation in 4.3 with the estimated coefficients α and β. The preceding approach is repeated for every model in GOM. This results in 57 motion representations according to the motion descriptors captured with the inertial MoCap system.

Because this approach employs one-shot training, just one movement sample per class is used to train the motion representation. This reference movement was determined using the Dynamic Time Warping (DTW) [Wang, 2010]. This algorithm measures the similarity between two time-series. Therefore, the movement sample that was closest to all other movement samples of the same class was chosen for one-shot training.

4.4 Data-driven strategies for estimating time-varying GOM representations

This section introduces two new deep state-space models for modeling human movements in which a time-varying GOM representation is estimated. These models are referred to as deep SSMs, given that the non-linear observation and transition models are parameterized using data-driven approaches. The data-driven approaches constitute encoder-decoder frameworks containing RNNs to capture temporal dependencies. The first possesses a variational autoencoder architecture (VAE-RGOM), whereas the second has integrated a Luong attention mechanism (ATT-RGOM).

Large datasets of human movements with long-term dependencies and non-linear relationships between their descriptors cannot be effectively modeled using conventional state-space models like linear-Gaussian models or HMMs. Thereby, the parameterization of SSMs based on these ANNs, with the ability to encode data through a probabilistic distribution (VAE) or accurately process long input sequences (AE with attention), can permit more accurate modeling of a broad range of data distributions and simulations of human movement. Furthermore, the incorporation of interpretable motion representations into the architecture of these deep SSMs constitutes a step toward explaining the predictions of human motion trajectories made by data-driven approaches.

The state-space representation used by both approaches, including the observation and transition models, is detailed in the next subsection. Then, it is explained how GOM is incorporated into their architecture. Lastly, a thorough description of the VAE-RGOM and ATT-RGOM architectures is given, along with the techniques utilized for their training.

4.4.1 Integration of time-varying GOM representations

By taking advantage of the modeling power of ANNs, two approaches are proposed for training all motion representations of GOM simultaneously. To this end, now the observations are defined as $Y_t = P_t \in \mathbb{R}^{1\times N}$, meaning the N joint angles at time t that compose the whole body posture P_t, and $X_t = [P_{t-1}, P_{t-2}] \in \mathbb{R}^{2\times N}$. Due to their advantages in sequence-to-sequence tasks, both frameworks use Autoencoders, where the decoders have the full-body GOM mathematical representations as the output layer. The decoder then calculates the coefficient matrix $A_t \in \mathbb{R}^{N\times 2\times N}$:

$$A_t = \left\{ \begin{bmatrix} \alpha_{1,1,1,t} & \cdots & \beta_{1,1,N,t} \\ \alpha_{1,2,1,t} & \cdots & \beta_{1,2,N,t} \end{bmatrix}, \cdots, \begin{bmatrix} \alpha_{N,1,1,t} & \cdots & \beta_{N,1,N,t} \\ \alpha_{N,2,1,t} & \cdots & \beta_{N,2,N,t} \end{bmatrix} \right\} \tag{4.6}$$

This is then utilized by the GOM equation system to produce the prediction $\widehat{Y}_t$. Since GOM employs a second-order equation system, each element in A_t corresponds to a 2D tensor with the shape $(2, N)$. N because all joint angles are included in each GOM equation as an assumption (H2, H3, and H4), and two vectors as it also computed the coefficients of the transition assumptions (H1). Thus, being X_t and A_t tensors of shape $(2, N)$ and $(N, 2, N)$,

respectively, the procedure for generating $\widehat{Y}_t$ utilizing the GOM representations in the decoder is as follows:

$$M_t = A_t \circ \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \underbrace{\begin{bmatrix} P_{1,t-1} & \cdots & P_{N,t-1} \\ P_{1,t-2} & \cdots & P_{N,t-2} \end{bmatrix}}_{X_t} \tag{4.7}$$

$$\widehat{Y}_t = \sum_{w=1}^{2} \sum_{k=1}^{N} M_{t,i,w,k} \tag{4.8}$$

where $\circ$ is an element-wise product, $\widehat{Y}_t \in \mathbb{R}^{1\times N}$, and $M_t \in \mathbb{R}^{N\times 2\times N}$. M_t corresponds to GOM in a matrix form, consisting of the 57 joint angle models as Equation 4.4.

Both Autoencoders are composed of RNNs for the encoder and decoder, representing human movements similarly to KF-RGOM, conditioning every data point at time t on a hidden state at time $t-1$ as a state-space model. The observation and transition probability distributions, $p(\widehat{Y}|Z,X)$ and $p(Z|X)$, are then learned maximizing the following likelihood, which approximates $\widehat{Y}$ to the observed Y:

$$p_\theta\left(\widehat{Y}_{1:T}|X_{1:T}\right) = \int \underbrace{p_\theta(\widehat{Y}_{1:T}|Z_{1:T},X_{1:T})}_{\text{Observation model}} \underbrace{p_\theta(Z_{1:T}|X_{1:T})}_{\text{Transition model}} dZ_{1:T} \tag{4.9}$$

where $Z_{1:T}$ represents the states of the system and $X_{1:T}$ is the input that, with $Z_{1:T}$, generates the outputs $\widehat{Y}_{1:T}$. In this generative model the observation model and transition model are calculated as follows:

$$p_\theta(\widehat{Y}_{1:T}|Z_{1:T},X_{1:T}) = \prod_{t=1}^{T} p_\theta(\widehat{Y}_t|Z_t,X_t) \tag{4.10}$$

$$p_\theta(Z_{1:T}|X_{1:T}) = \prod_{t=1}^{T} p_\theta(Z_t|Z_{t-1},X_t) \tag{4.11}$$

The parameters θ from the observation and transition models are learned during training by the decoder of each framework. However, the encoder has a different function in each approach, taking advantage of their very specific encoder-decoder architecture. In the first deep learning approach, denoted as VAE-RGOM, an architecture of a VAE is used, meaning the encoder functions as an inference network. In the second approach, designated as ATT-RGOM, the Autoencoder has incorporated a Luong attention mechanism (global) which initializes the system's state as a selected sequence of observed motion descriptors.

The subsequent two sections describe in more detail the architecture and loss of each encoder-decoder network. Figure 4.3 provides a high-level overview of the three approaches proposed for estimating time-varying GOM representations.

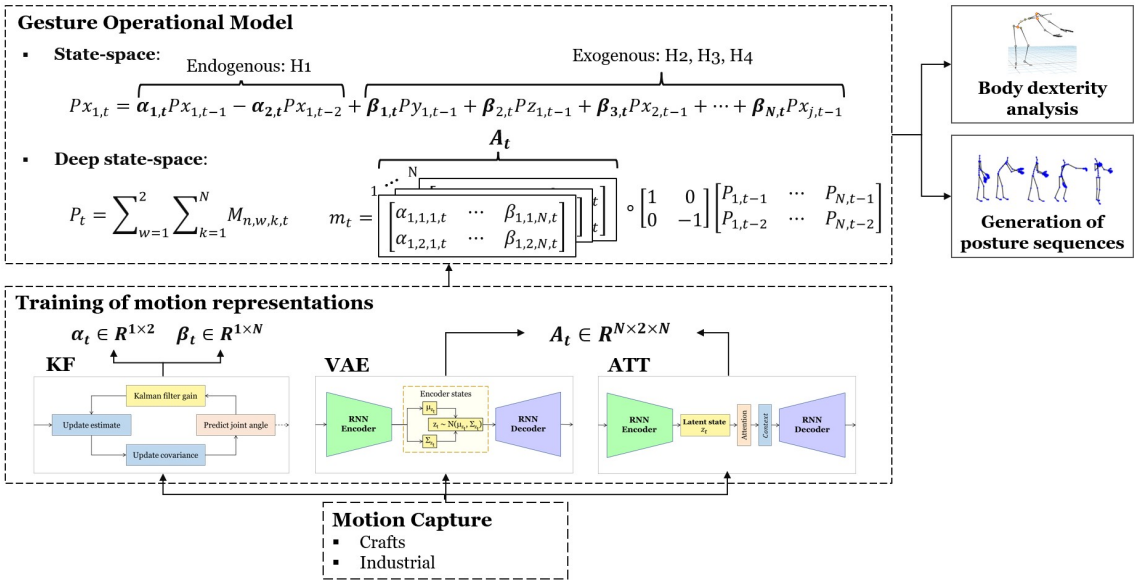


Figure 4.3: Methodology for creating explainable motion representations for body dexterity analysis and generation of human posture sequences. The motion data of industrial operators and artisans is utilized for training time-varying motion representations. Three methods are proposed for training: one-shot training with Kalman Filters to estimate the coefficients α_t and β_t of a single motion representation ($Px_{1,t}$); two methods that use deep learning with either a VAE or an Autoencoder with global attention (ATT) to automatically calculate the matrix A_t, which contains the coefficients of the full-body motion representations (P_t).

4.4.2 Deep state-space modeling based on a Variational Autoencoder

As a probabilistic generative model, VAE is typically trained to achieve the marginal log-likelihood $\log p_\theta(x_{1:T})$:

$$Maximum\ \ell(\theta,\varphi) = \sum_{t=1}^{T} \log p_\theta(X_{1:T}) \tag{4.12}$$

On the other hand, as stated in Equation 4.9, VAE-RGOM seeks to maximize the marginal log-likelihood below:

$$Maximum\ \ell(\theta,\varphi) = \sum_{t=1}^{T} \log p_\theta\left(\widehat{Y}_{1:T}|X_{1:T}\right) \tag{4.13}$$

An LSTM decoder learns the observation and transition models. Similar to a conventional VAE, the encoder estimates the stochastic latent states z by approximating $q_\varphi(Z_t|X_t, Y_t)$ to the true posterior distribution $p_\theta(Z_t|Z_{t-1}, X_t)$ defined by a mean μ_{Z_t} and a log covariance Σ_{Z_t}.

Stochastic gradient optimization was used to train the networks. This entails first sampling Z_t, subsequently estimating the ELBO, then the gradients for θ, φ, and A, and lastly, updating these parameters. The loss of Equation 4.13 is thus equivalent to the maximum ELBO with

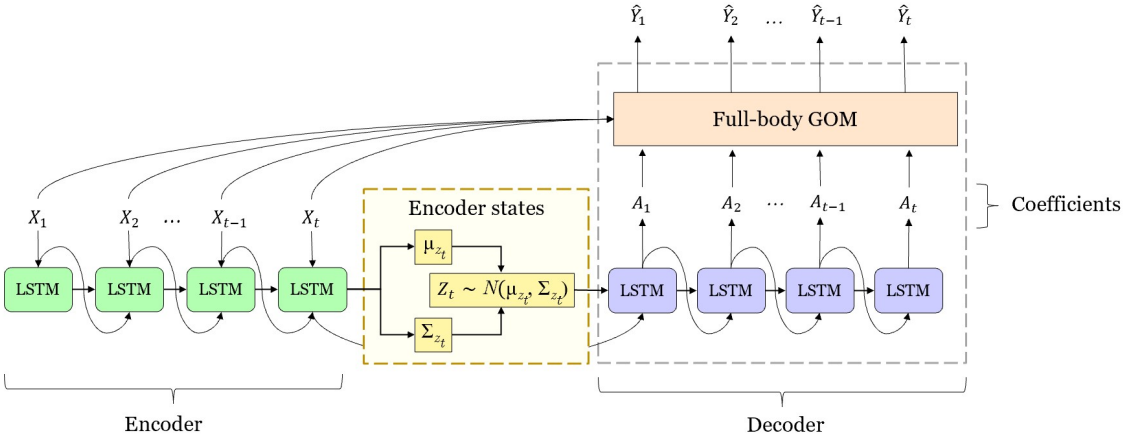


Figure 4.4: Overview of the Variational Autoencoder network for estimating GOM's coefficients.

respect to θ, φ, and A that results:

$$\ell(\theta, \varphi, A) = \sum_{t=1}^{T} \underbrace{-\beta_{\text{VAE}}\ KL(q_\varphi(Z_t|Z_{t-1}, X_t, Y_t)|\,|p_\theta(Z_t|Z_{t-1}, X_t))}_{\text{Regularization loss}} + \underbrace{\beta_{\text{GOM}}\mathbb{E}_{q_\varphi(Z_t|Z_{t-1},X_t,Y_t)|}\left[\log p_\theta(Y_t|Z_t, X_t)\right]}_{\text{Prediction loss}} \tag{4.14}$$

In Equation 4.14, KL denotes the Kullback-Leibler divergence that captures the complexity of the data; the prediction loss measures the accuracy of the model in the prediction; β_{VAE} and β_{GOM} correspond to tuning hyperparameters. Instead of directly sampling from q_φ at each time step, $Z_t = \mu_{z_t} + \epsilon \odot \Sigma_{z_t}$ is re-parametrized using samples from a normal random variable $\epsilon \sim \mathcal{N}(0, I)$. Consequently, the gradients relative to the parameters θ, φ, and A can be back-propagated through the encoder via the sampled Z_t. As the prediction loss, the mean squared difference of all motion descriptors is used:

$$\ell_{euler} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{D}\sum_{d=1}^{D} P_{t,j,d} - \widehat{P}_{t,j,d}$$

[2](4.15)

Figure 4.4 depicts a conceptual diagram of VAE-RGOM, where Long Short-Term Memory networks (LSTM) are used for both the encoder and decoder. In order to tune the framework's hyperparameters, a Bayesian optimization was carried out based on the loss achieved on a validation set [Snoek, 2012]. In the Bayesian hyperparameter optimization, the most promising hyperparameters are chosen using a stochastic model (Gaussian process) of the objective function [Dewancker, 2016]. A set of hyperparameters is used by the objective function, which then returns the validation loss. In the optimization process, past evaluation results are tracked to determine the next set of hyperparameters for evaluation that may provide the best performance for a surrogate function p(loss|hyperparameters). The optimized hyperparameters

Table 4.1: VAE-RGOM architecture.

Layer	Type	Output shape	Activation	Dropout	Recurrent dropout	Input layer
1	Input	(2,57)	-	-	-	-
2	LSTM	(2,32)	Softsign	0.2	0.2	1
3 (μ_{z_t})	FC	2	Linear	-	-	2
4 (Σ_{z_t})	FC	2	Linear	-	-	2
5 (Z_t)	Sampling	2	-	-	-	3,4
6	LSTM	(2,32)	Softsign	0.2	0.2	5
7	Dropout	(2,32)	-	0.8	-	6
8	Time distributed (FC)	(2,3249)	Linear	-	-	7
9 (A_t)	Reshape	(57,2,57)	-	-	-	8
10 (GOM)	Lambda	(1,57)	-	-	-	1,9

included the number of units of the LSTM decoder and LSTM encoder, their learning rate, activation function, dropout rate, and recurrent dropout rate.

The best architecture is described in Table 4.1. Here, the encoder is comprised of layers 2 to 5, and the input consists of the two previous whole-body postures (X_t). First, the input is processed by an LSTM, whose output is then used as the input for two fully-connected networks (FC), which model the μ_{z_t} and Σ_{z_t}. The Sampling layer represents the sampling from the normal random variable for the reparametrization of Z_t. The sampled Z_t is passed to the decoder along with the last hidden states and cell states of the encoder's LSTM. The decoder consists of layers 6 to 9. The sampled Z_t is fed into an LSTM whose states are initialized with the last hidden states and cell states of the encoder's LSTM. The output of the LSTM is then sent to a fully connected and time-distributed layer. The GOM coefficients at time t (A_t) are obtained from the time-distributed layer and reshaped before passing to the lambda layer, where they are multiplied with the input X_t to obtain the predicted $\widehat{Y}_t$.

The training of the model parameters was performed using backpropagation with the Adam optimizer. Adam may be compared to a hybrid of RMSProp and stochastic gradient descent. It scales the learning rate using squared gradients, similar to RMSProp, and leverages momentum by utilizing the gradient's moving average rather than the gradient itself, similar to SGD with momentum [Kingma, 2014]. The initial learning rate of the optimizer was set to 1×10^{-3}, β to 0.99, and the Adam parameters were $b_1 = 0.90$ and $b_2 = 0.99$.

VAE-RGOM was trained and validated using all seven datasets detailed in Chapter 3. These include a wide range of movements used in both industry and handcrafted professions. Later in Chapter 5, the generated motion representations are used to analyze the operators' or artisans' gestural knowledge and dexterity while practicing their tasks. A 5-fold cross-validation was performed to prevent overfitting.

4.4.3 Deep state-space modeling based on an Autoencoder with Luong Attention

In this approach, the state-space system is parametrized by using the LSTM encoder to initialize the system's state Z_t as a context vector C_t of the observed joint angles. The context vector is determined by the sequence of hidden states $H_{1:T}$ to which the encoder maps the input sequence $X_{1:T}$ of length T. The LSTM decoder then models the state transition $p_\theta(Z_t|Z_{t-1}, C_t)$

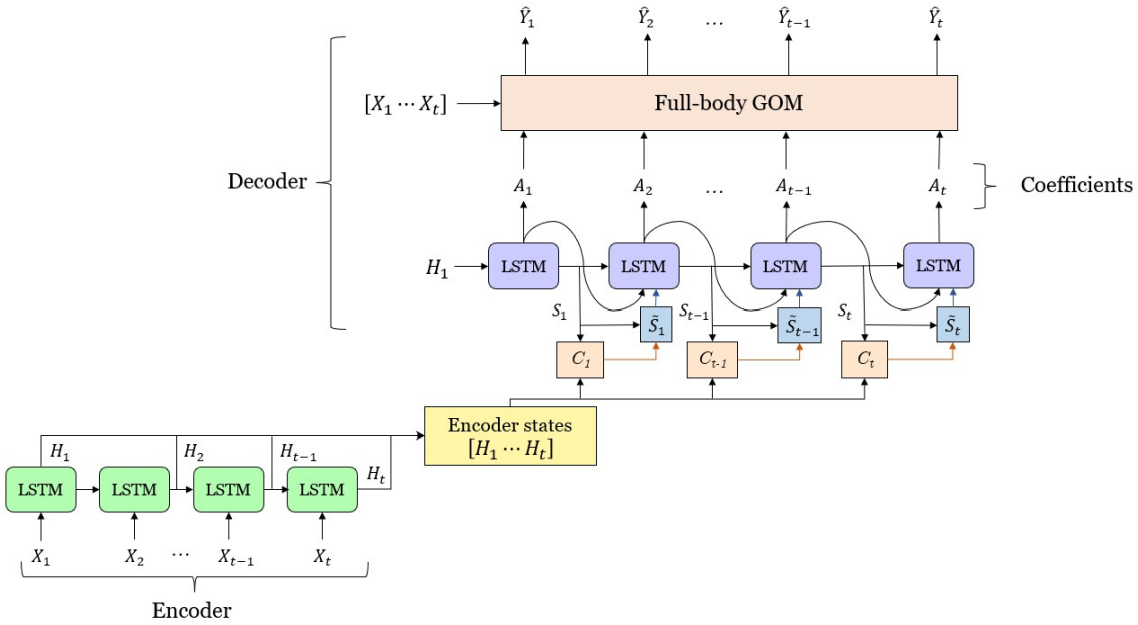


Figure 4.5: Overview of the Autoencoder with Luong Attention for estimating GOM's coefficients.

in order to update the system's state and generate the future posture $p_\theta(Y_t|Z_t, X_t)$. The Luong attention mechanism is integrated in order to capture state dynamics. The attention mechanism takes previous postures into account and maps them to attention weights (W_t), computed using a dot product alignment. The attention weights determine the degree to which previous hidden states $H_{1:T}$ influence future state transitions. This influence is indicated in the context vector, which is the weighted sum of the H_i:

$$C_t = \sum_{i=1}^{T} W_{t,i} H_i \tag{4.16}$$

In accordance with the model structure of an Autoencoder with Luong attention mechanism [Luong, 2015], the context vector is first used to compute the attentional hidden state $\widetilde{S_t}$. The decoder then uses this state to generate the GOM's coefficients, followed by $\widehat{Y}_t$. Figure 4.5 illustrates a diagram of the ATT-RGOM structure. The network is trained to maximize the log-likelihood in Equation 4.13, considering only a prediction loss, which as VAE-RGOM, is the mean absolute difference of all motion descriptors (Equation 4.4.2).

A Bayesian optimization was applied for tuning ATT-RGOM's hyperparameters: the number of units of the LSTMs, learning rate, activation function, dropout rate, and recurrent dropout rate. Table 4.2 provides specifics on the optimized architecture. Layers 2 through 4 make up the encoder, where two batch normalization layers with a momentum of 0.99 were added to normalize the hidden states and cell states of the encoder's LSTM. The layers from 5 to 12 correspond to the decoder and the attention mechanism. The LSTM encoder receives the input X_t and produces H_t, a fixed-size representation of X_t, together with its last hidden states and cell states. The last states of the LSTM encoder are used as an initial state

Table 4.2: ATT-RGOM architecture.

Layer	Type	Output shape	Activation	Dropout	Recurrent dropout	Input layer
1	Input	(2,57)	-	-	-	-
2	LSTM	2.1 Output state:(2,32) 2.2 Hidden state: 32 2.3 Cell state: 32	Softsign	0.2	0.2	1
3	Batch Normalization	32	-	-	-	2.2
4	Batch Normalization	32	-	-	-	2.3
5	LSTM	(2,32)	Softsign	0.2	0.2	3, 4
6	Dot	(2,2)	-	-	-	2.1, 5
7 (W_t)	Softmax	(2,2)	-	-	-	6
8 (C_t)	Dot	(2,32)	-	-	-	2.1, 7
9	Concatenate	(2,64)	-	-	-	5, 8
10	Time distributed (FC)	(2,3249)	Linear	-	-	9
11 (A_t)	Reshape	(57,2,57)	-	-	-	10
12 (GOM)	Lambda	(1,57)	-	-	-	1, 11

of the LSTM decoder, and the hidden state as the first input of the LSTM decoder. The W_t, C_t, and $\widetilde{S_t}$ are calculated using the LSTM encoder's and decoder's outputs. The GOM coefficients (A_t) are generated from the time-distributed layer, reshaped, and then passed to the lambda layer, where they are multiplied with the input X_t to produce the predicted $\widehat{Y}_t$.

The Adam optimizer was used for the training of ATT-RGOM. The initial learning rate was set to 5×10^{-3}, and the Adam parameters were $b_1 = 0.90$ and $b_2 = 0.99$. ATT-RGOM was trained and evaluated using all seven datasets presented in Chapter 3, for comparison with VAE-RGOM and KF-RGOM. A 5-fold cross-validation was performed during training to avoid overfitting.

4.5 Static and dynamic simulation

This section evaluates KF-GOM, KF-RGOM, VAE-RGOM, and ATT-RGOM in terms of their ability to simulate realistic human movements. The human movements involved in professional tasks related to industrial processes, such as television assembly (TVA), packaging (TVP), and airplane assembly (APA), were modeled to assess each method. Additionally, motion data from craftsmen engaged in glassblowing (GLB), silk weaving (SLW), and mastic cultivation (MSC), as well as movements of different ergonomic risk levels performed by subjects in a laboratory (ERGD), were used for the evaluation. Besides using available MoCap datasets, new ones were created that included a greater diversity of movements, particularly professional movements captured in real-world scenarios. These recordings allowed analyzing the operators' and artisans' gestural knowledge and dexterity while practicing their tasks with the train GOMs in Chapter 5. In addition, the analysis of the ERGD dataset inspired the development of new methodologies for detecting ergonomic risk factors, which are later presented in Chapter 6. Chapter 3 provides more information regarding the preprocessing, segmentation, and description of these human motion datasets.

Section 4.5.1 examines the simulation performance of KF-GOM, which represents human movements with constant coefficients and incorporates all full-body assumptions in the model of each joint motion descriptor. The following section evaluates the static and dynamic simu-

lations of KF-RGOM, VAE-RGOM, and ATT-RGOM, which employ a time-varying full-body GOM representation. In a static simulation, all endogenous and exogenous variables are actual data samples. In contrast, in the dynamic simulation, the endogenous variables are not actual samples of motion data but the model's past predicted values. Each approach predicted one time step per iteration. After predicting all the time frames of a movement, the simulated movement was compared with the original for evaluation. The simulation performance was measured using the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2} \tag{4.17}$$

Additionally, the Theil's inequality coefficient U_1, along with its decompositions: bias proportion U_B, variance proportion U_V, and covariance proportion U_C, were included in the metrics and are calculated as follows:

$$U_1 = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} y_t^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T} \widehat{y}^2}} \tag{4.18}$$

$$U_B = \frac{(\mu_{y_t} - \mu_{\widehat{y}_t})^2}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2} \qquad U_V = \frac{(\sigma_{y_t} - \sigma_{\widehat{y}_t})^2}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2} \qquad U_C = \frac{2(1-\rho)\sigma_{y_t}\sigma_{\widehat{y}_t}}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2} \tag{4.19}$$

In Equation 4.19, μ_{y_t} corresponds to the mean of the original movement, $\mu_{\widehat{y}_t}$ the mean of the simulated movement, σ_{y_t} the standard deviation of the original movement, $\sigma_{\widehat{y}_t}$ the standard deviation of the simulated movement, and ρ is the correlation between the simulation and original movement. The U_B measures the relationship between the means of the original and the simulated movement, U_V considers the prediction's ability to match the variation in the original movement, and U_C examines the residual unsystematic element of prediction errors. By definition, $U_B + U_V + U_C = 1$, hence the optimal outcome for these statistics would be for U_B and U_V to be as close to zero as possible and U_C to be as close to one. For U_1, the closer it is to zero, the greater the quality of the forecast. All the results of the four approaches, KF-GOM, KF-RGOM, VAE-RGOM, and ATT-RGOM, in the simulation of each movement of the seven datasets, are presented in the Appendix A.

A sensitivity analysis was conducted to investigate the stability of each approach after a shock occurred in one of the variables that compose their input. For this analysis, a disturbance of 80% was applied only in the first two frames of the movement simulated before the entire movement was predicted.

4.5.1 Static simulation with constant coefficients

This section summarizes the results of KF-GOM's static stimulation and sensitivity analysis. Initially, the models were trained using a reference movement for each class, which was determined using the DTW algorithm [Wang, 2010]. For the training of KF-GOM, it was used an

Table 4.3: Static simulation performance of KF-GOM.

Motion	**Joint angle**	*RMSE*	U_1	U_B	U_V	U_C
TVA_3	$LSH1_X$	0.095	0.017	0.249	0.003	0.748
	$LSH1_Y$	0.007	0.006	0.000	0.002	0.999
	$LSH1_Z$	0.008	0.014	0.001	0.001	0.999
APA_1	$RSH2_X$	0.064	0.093	0.001	0.076	0.923
	$RSH2_Y$	0.007	0.014	0.000	0.000	0.999
	$RSH2_Z$	0.009	0.024	0.001	0.001	0.998
GLB_8	$LSH2_X$	0.213	0.206	0.278	0.027	0.694
	$LSH2_Y$	0.182	0.395	0.232	0.003	0.764
	$LSH2_Z$	0.632	0.366	0.491	0.172	0.336
MSC_5	$LSH2_X$	0.172	0.134	0.169	0.080	0.751
	$LSH2_Y$	0.750	0.041	0.001	0.263	0.737
	$LSH2_Z$	1.0831	0.0850	0.419	0.040	0.640
$ERGD_8$	$SP2_X$	0.074	0.007	0.029	0.018	0.953
	$SP2_Y$	0.143	0.035	0.090	0.210	0.700
	$SP2_Z$	0.077	0.0115	0.069	0.059	0.871

Intel Core i7-8750H CPU. Figure 4.6 shows five examples of simulated movements and their original joint angle sequences, with confidence bounds of 95%.

The evidence indicates that KF-GOM is capable of capturing the patterns present in the sequences of joint angles for each movement. For instance, the action of buckling a rivet generated sequential patterns of bending on the spine, which were captured by KF-GOM and are seen in Figure 4.6b. As seen in Figure 4.6c, it is also depicted in the precise rotations of the forearm made by the glassblower as he turns the blowpipe to shape the melting glass. Table 4.3 presents the simulation performance on three Euler angles for movements of five datasets. By observing the examples in Figure 4.6 and Table 4.3, KF-GOM is able to reproduce the majority of movements within the confidence intervals and in close proximity to the original movement. For the most complex and lengthy movements in TVA, GLB, and MSC, however, KF-GOM cannot accurately simulate the time series for all axes XYZ, as shown in Table 4.3 for the movements TVA_3, GLB_8, and MSC_5. The computed constant coefficients may be insufficient to describe the variance in long time series of human movements, or it may be necessary to supplement GOM with new assumptions. The next section evaluates the usage of time-varying coefficients for modeling long human movements.

Figure 4.7 depicts three examples of shocks given to different variables for the sensitivity analysis. Figures 4.7a and 4.7b illustrate the forecasting behavior of the model of the joint angle LA_X for the movement of raising the hands above the level of the shoulders ($ERGD_{13}$). In Figure 4.7a, a shock was given to the joint angles of LSH2, and in Figure 4.7b, it was applied to the joint angles of RSH2. It is evident that giving a shock to the left shoulder affected the motion of the left arm significantly more than applying it to the right shoulder due to the strong mediation of the left shoulder over the motion of the left arm.

Figure 4.7c depicts the simulated movement of $SP2_Y$ when the subjects twisted their torso to the left ($ERGD_2$). In this instance, the shock was applied to the joint angles derived from the sensor located at the hips, H. The fact that the model was able to adapt in less than one second

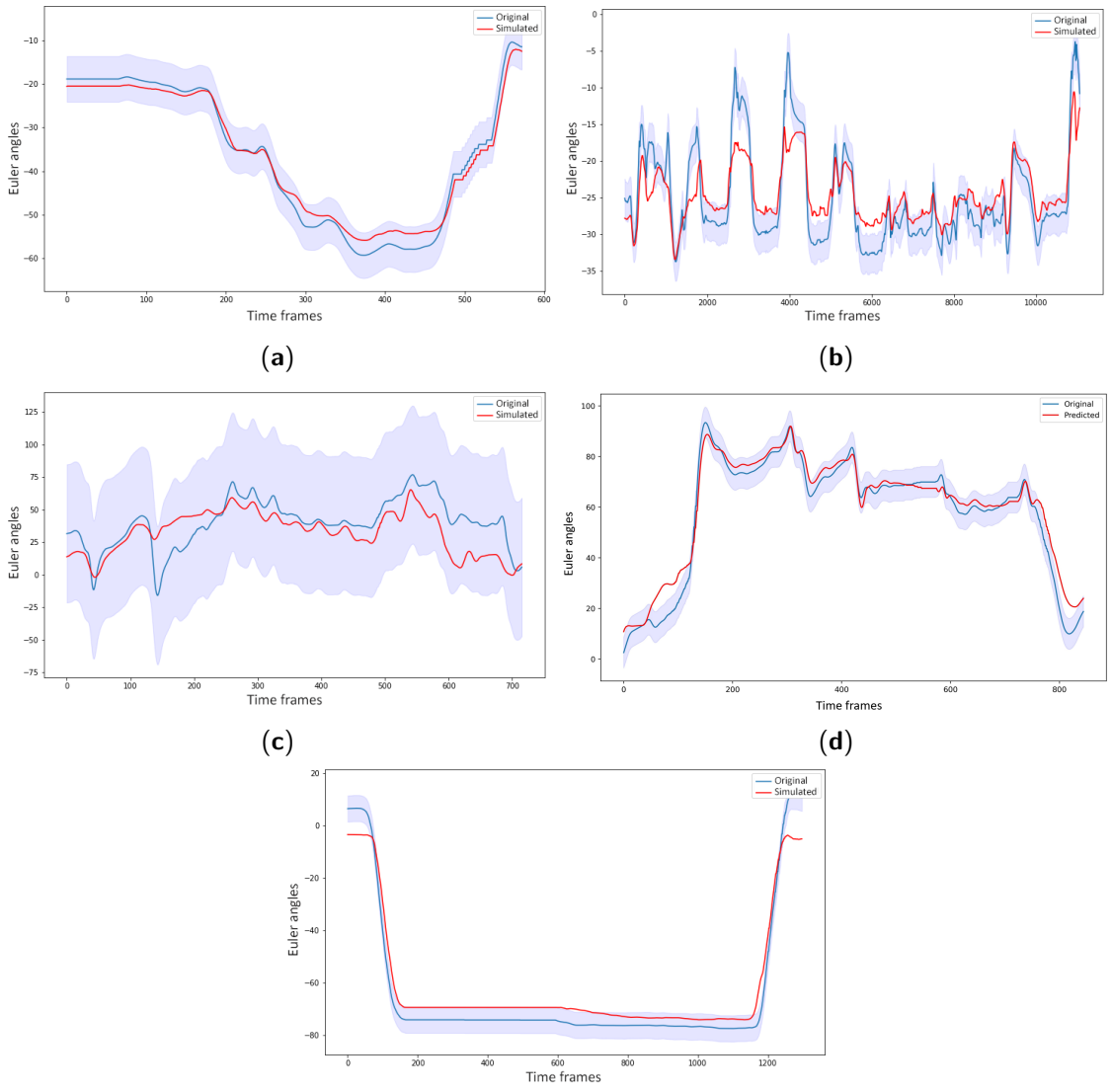


Figure 4.6: Examples of simulated movements by KF-GOM. (a) Simulation of the movement TVA_3 on the joint angle LA_X; (b) The simulated joint angle sequence of $SP1_Z$ for APA_3; (c) Simulation of LFA_Y for the movement GLB_3; (d) The simulated joint angle sequence of the right forearm on the Y-axis RFA_Y, for the gesture MSC_5; (e) Simulation of RA_X for $ERGD_{19}$, which consists of raising the forearms above the shoulder level.

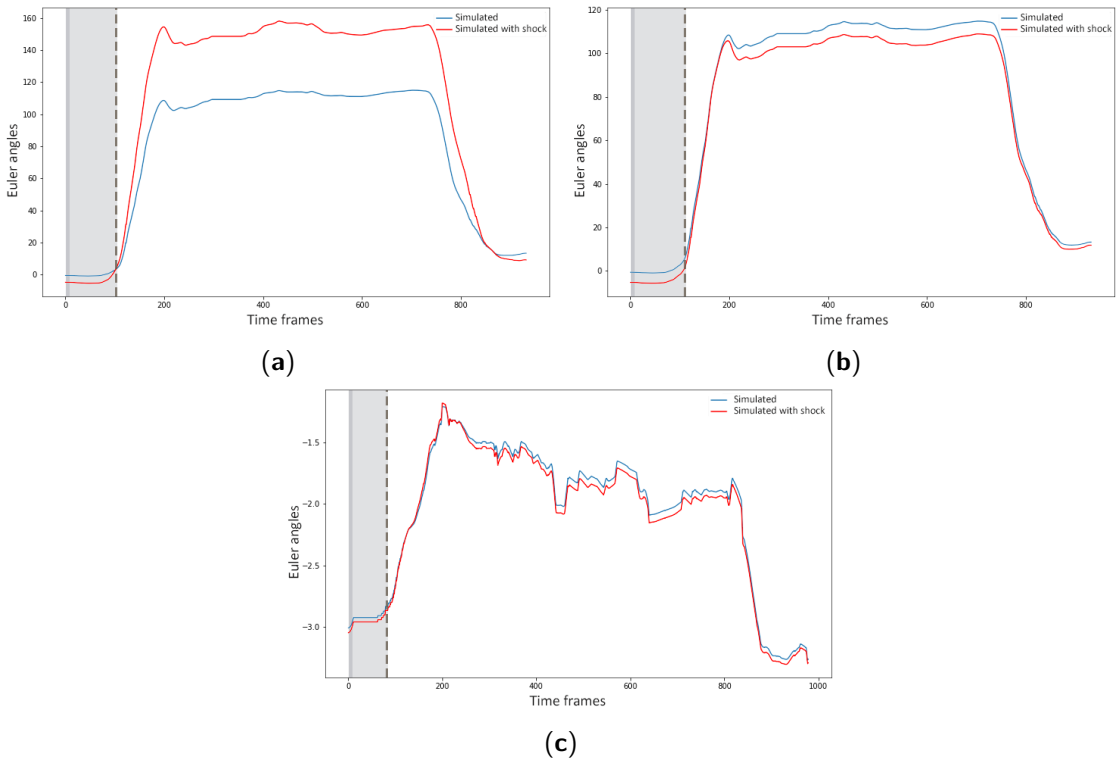


Figure 4.7: Simulated joint angles with and without disturbance of 80% on the two initial time frames. (a) Simulation of the joint angle LA_X with a disturbance on the joint angles of LSH2 (blue) and without (red); (b) Simulated joint angle sequence of LA_X with a disturbance on the joint angles of RSH2 (blue) and without (red); (c) Simulation of the joint angle $SP2_Y$ with a disturbance on the joint angles of H (blue) and without (red).

(90 frames) indicates the model's low sensitivity to external perturbations. However, there was still a minor deviation in the forecasting if compared to the simulated movements predicted without shocks. This is due to the association of H to SP2 in its estimated representation.

4.5.2 Static and dynamic simulation with time-varying coefficients

This section presents the results of the models KF-RGOM, VAE-RGOM, and ATT-RGOM, which construct time-varying representations of human movements. These results are later discussed in the next section. As with the training of the KF-GOM, a reference movement was utilized for the training of the KF-RGOM. Regarding the training of VAE-RGOM and ATT-RGOM, these were trained with all seven datasets using an NVIDIA GPU RTX 2060, and applying 80-10-10 sets (80% for training, 10% for validation, and 10% for testing). The validation set was used to estimate the neural networks' hyperparameters, while the test set was utilized for evaluation. Figures 4.9 to 4.14 illustrate the static simulation (blue line) in comparison to the real values (orange line) and with 95% confidence bounds. The metrics calculated for these simulated movements are presented from Table 4.5 to Table 4.10. To complement the analysis of the full-body simulations generated with the deep SSMs, visual comparisons between the quality of the generated sequences are offered in figures 4.15, 4.16,

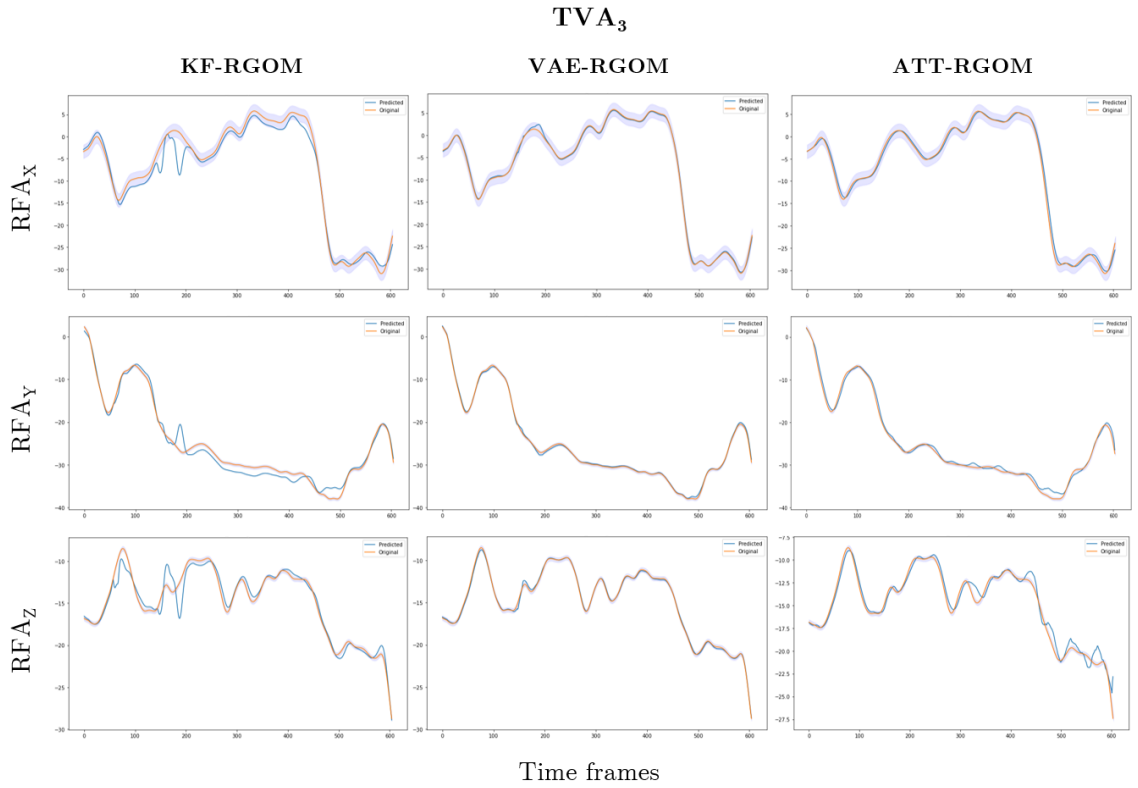


Figure 4.8: Static simulation of RFA for the movement TVA_3. In this movement, the operator connects a circuit board and a wire and then places the board on a television chassis to be screwed.

and 4.17. The red boxes show variations in the movements at different temporal windows.

Table 4.4: Quantitative comparison of the models for TVA_3.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	RFA_X	1.994	0.070	0.106	0.092	0.802
	RFA_Y	1.532	0.028	0.041	0.011	0.948
	RFA_Z	1.059	0.034	0.049	0.022	0.929
VAE-RGOM	RFA_X	0.786	0.020	0.017	0.006	0.977
	RFA_Y	**0.329**	**0.006**	**0.011**	**0.002**	**0.987**
	RFA_Z	**0.211**	**0.007**	**0.017**	**0.012**	**0.971**
ATT-RGOM	RFA_X	**0.703**	**0.025**	**0.002**	**0.010**	**0.988**
	RFA_Y	0.688	0.015	0.041	0.011	0.948
	RFA_Z	1.073	0.035	0.040	0.034	0.926

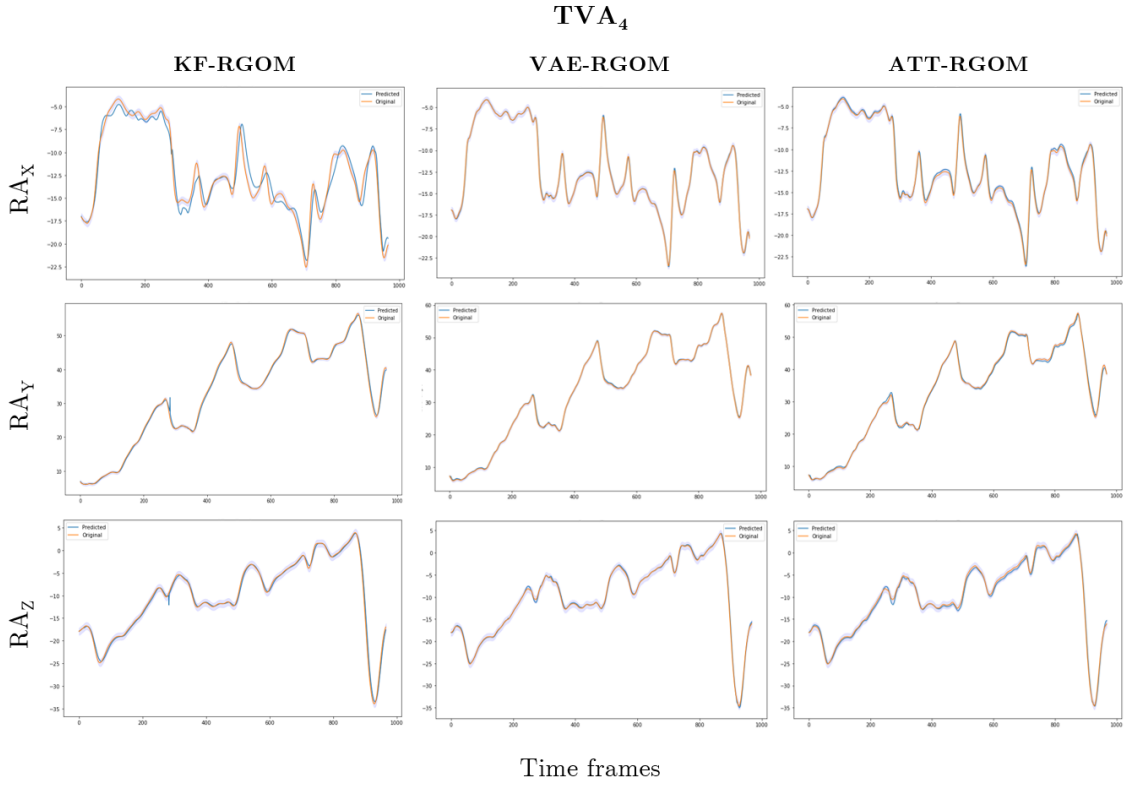


Figure 4.9: Static simulation of RA for the movement TVA_4. The movement consists of drilling a circuit board into the chassis of a television.

Table 4.5: Quantitative comparison of the models for TVA_4.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	RA_X	1.203	0.075	0.003	0.039	0.958
	RA_Y	0.788	0.025	0.021	0.011	0.968
	RA_Z	0.730	0.035	0.020	0.014	0.966
VAE-RGOM	RA_X	**0.169**	**0.020**	**0.017**	**0.006**	**0.977**
	RA_Y	**0.222**	**0.006**	**0.011**	**0.002**	**0.987**
	RA_Z	**0.290**	**0.007**	**0.017**	**0.012**	**0.971**
ATT-RGOM	RA_X	0.513	0.068	0.021	0.010	0.969
	RA_Y	0.332	0.034	0.029	0.002	0.969
	RA_Z	0.483	0.050	0.007	0.026	0.967

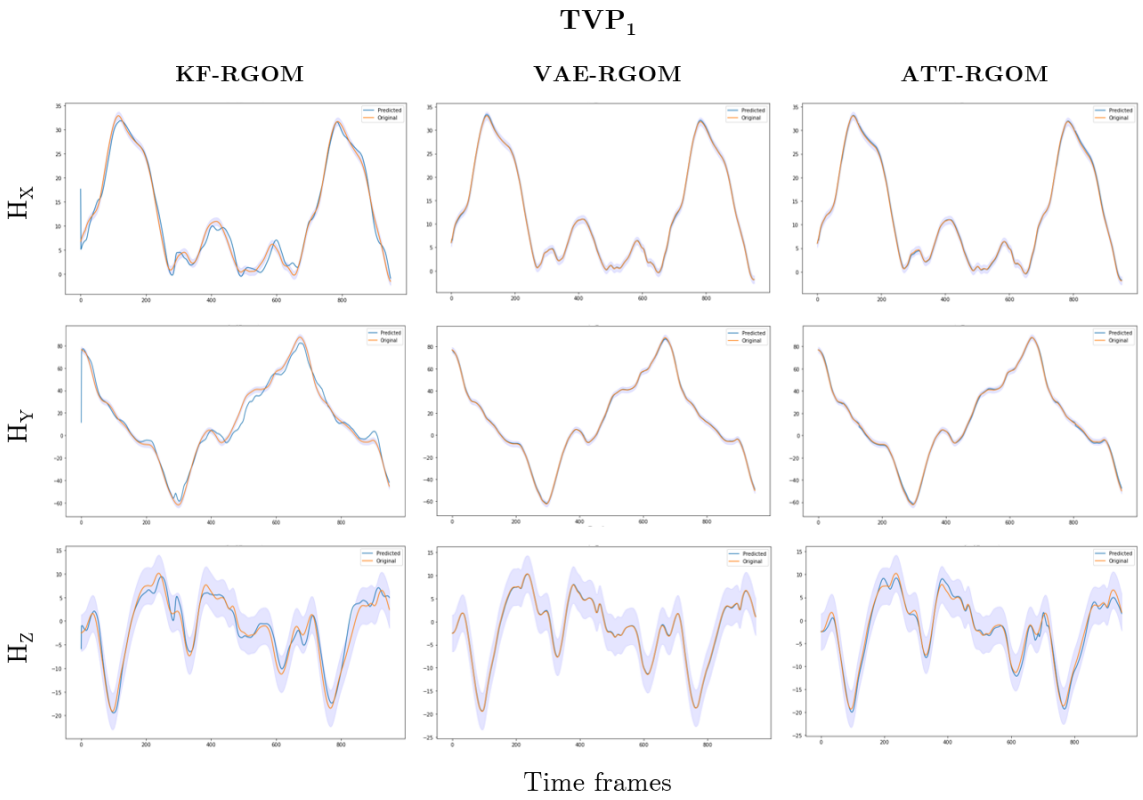


Figure 4.10: Static simulation of H for the movement TVP_1. The movement consists of placing a television box on the first level of a wooden pallet.

Table 4.6: Quantitative comparison of the models for TVP_1.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	H_X	1.245	0.088	0.029	0.082	0.889
	H_Y	0.810	0.068	0.026	0.050	0.924
	H_Z	1.068	0.110	0.035	0.030	0.935
VAE-RGOM	H_X	**0.160**	**0.006**	**0.001**	**0.000**	**0.999**
	H_Y	**0.290**	**0.012**	**0.001**	**0.001**	**0.998**
	H_Z	**0.219**	**0.017**	**0.000**	**0.012**	**0.988**
ATT-RGOM	H_X	**0.180**	**0.033**	**0.002**	**0.001**	**0.997**
	H_Y	**0.262**	**0.015**	**0.001**	**0.001**	**0.998**
	H_Z	0.661	0.066	0.001	0.023	0.976

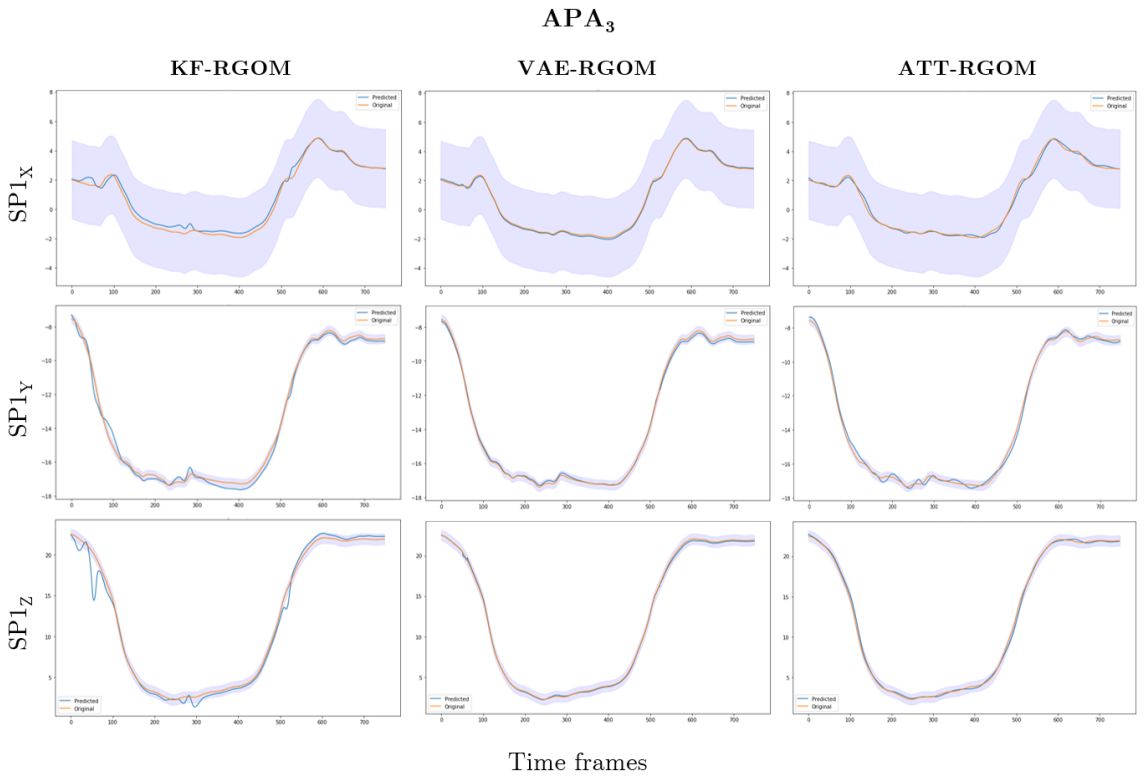


Figure 4.11: Static simulation of SP1 for the movement APA_3. In this movement, the operator places a bucking bar to counteract the incoming rivets while assembling an airplane structure.

Table 4.7: Quantitative comparison of the models for APA_3.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	$SP1_X$	0.399	0.057	0.044	0.036	0.920
	$SP1_Y$	0.481	0.070	0.082	0.012	0.906
	$SP1_Z$	1.301	0.031	0.087	0.084	0.829
VAE-RGOM	$SP1_X$	**0.082**	**0.017**	**0.001**	**0.020**	**0.979**
	$SP1_Y$	**0.113**	**0.004**	**0.011**	**0.005**	**0.984**
	$SP1_Z$	0.195	0.007	0.001	0.014	0.985
ATT-RGOM	$SP1_X$	0.243	0.033	0.030	0.005	0.965
	$SP1_Y$	0.432	0.041	0.024	0.011	0.965
	$SP1_Z$	**0.168**	**0.009**	**0.002**	**0.007**	**0.991**

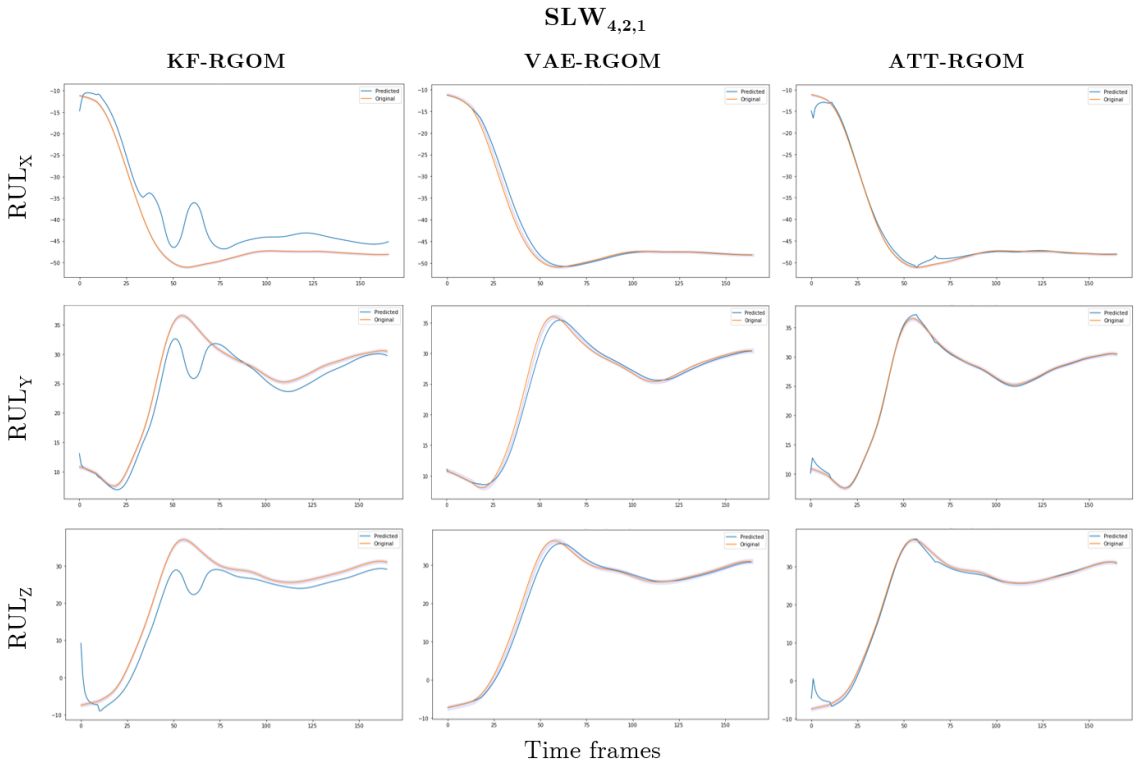


Figure 4.12: Static simulation of RUL for the movement $SLW_{4,2,1}$. This movement consists of the first step while weaving with a silk loom. The expert weaver pushed the pedal down with his right leg while pushing the threads with his left hand.

Table 4.8: Quantitative comparison of the models for $SLW_{4,2,1}$.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	RUL_X	5.677	0.067	0.324	0.003	0.673
	RUL_Y	2.946	0.056	0.294	0.010	0.696
	RUL_Z	4.892	0.097	0.358	0.003	0.639
VAE-RGOM	RUL_X	0.795	0.011	0.053	0.004	0.943
	RUL_Y	**0.533**	**0.010**	**0.021**	**0.009**	**0.970**
	RUL_Z	**0.864**	**0.024**	**0.039**	**0.000**	**0.961**
ATT-RGOM	RUL_X	**1.112**	**0.023**	**0.028**	**0.001**	**0.971**
	RUL_Y	1.121	0.025	0.023	0.010	0.967
	RUL_Z	1.420	0.038	0.051	0.009	0.940

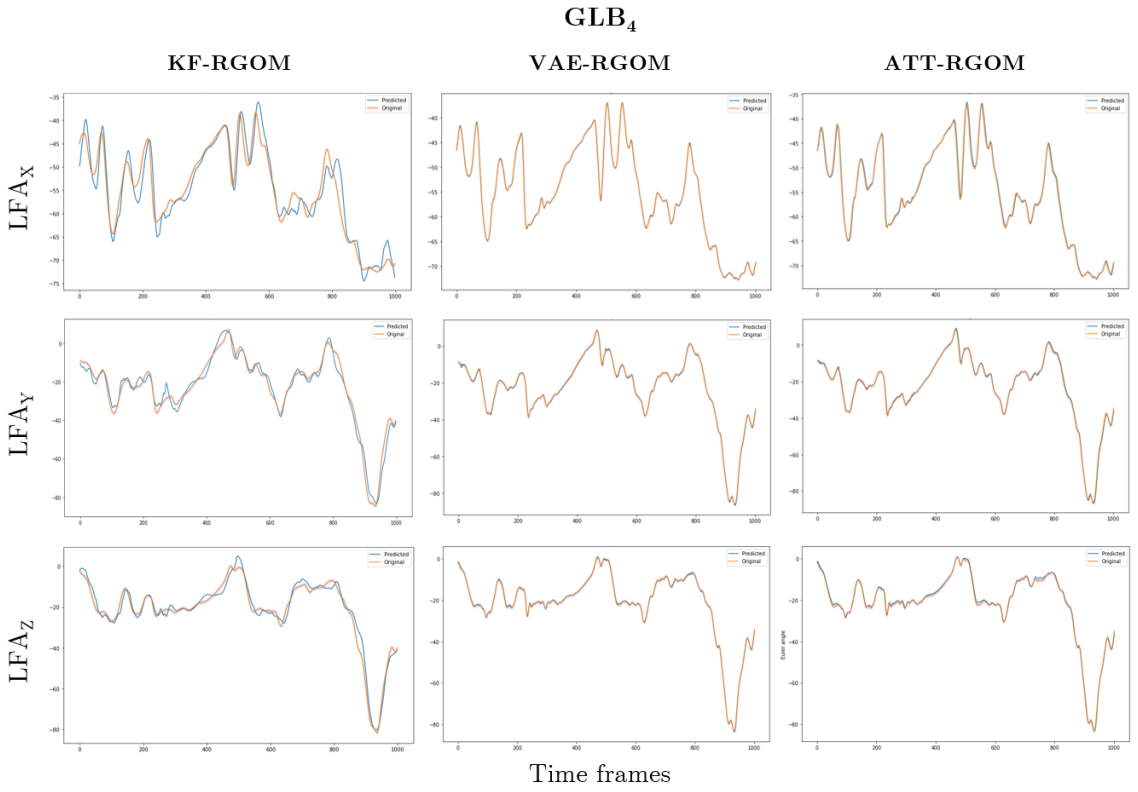


Figure 4.13: Static simulation of LFA for the movement GLB_4. In this movement, the glass-blower rotated the blowpipe with the left hand while shaping the glass with the right hand using a block.

Table 4.9: Quantitative comparison of the models for GLB_4.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	LFA_X	3.630	0.051	0.203	0.088	0.709
	LFA_Y	1.757	0.074	0.103	0.042	0.855
	LFA_Z	1.389	0.063	0.119	0.014	0.867
VAE-RGOM	LFA_X	**0.184**	**0.002**	**0.003**	**0.005**	**0.992**
	LFA_Y	**0.254**	**0.008**	**0.019**	**0.001**	**0.98**
	LFA_Z	**0.162**	**0.007**	**0.008**	**0.004**	**0.988**
ATT-RGOM	LFA_X	**0.185**	**0.005**	**0.010**	**0.002**	**0.988**
	LFA_Y	**0.194**	**0.013**	**0.001**	**0.015**	**0.984**
	LFA_Z	0.205	0.013	0.029	0.006	0.965

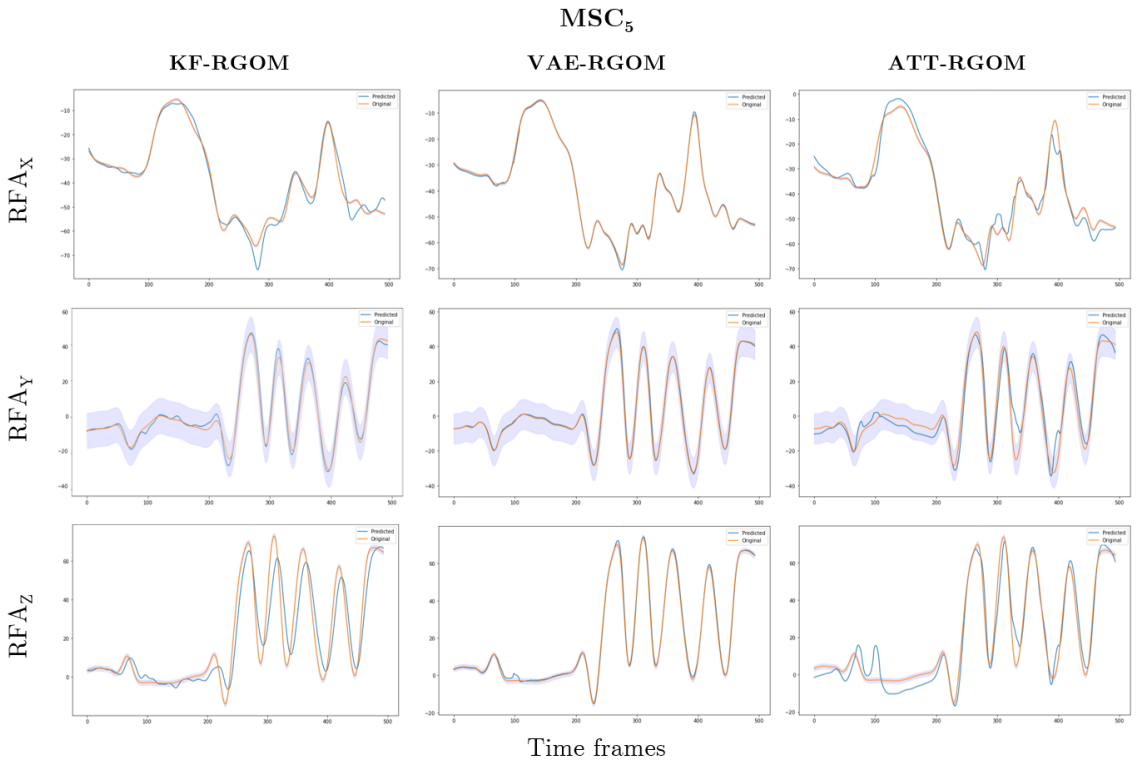


Figure 4.14: Static simulation of RFA for the movement MSC_5. The movement consists of the mastic farmer continuously collecting mastic from the outer bark of the tree using a razor.

Table 4.10: Quantitative comparison of the models for MSC_5.

Model	Joint angle	RMSE	U_1	U_B	U_V	U_C
KF-RGOM	$SP1_X$	2.012	0.023	0.127	0.082	0.791
	$SP1_Y$	4.721	0.050	0.089	0.049	0.862
	$SP1_Z$	2.420	0.058	0.280	0.080	0.640
VAE-RGOM	$SP1_X$	**1.027**	**0.011**	**0.053**	**0.004**	**0.943**
	$SP1_Y$	**0.533**	**0.010**	**0.012**	**0.009**	**0.979**
	$SP1_Z$	**1.264**	**0.024**	**0.030**	**0.009**	**0.961**
ATT-RGOM	$SP1_X$	5.677	0.067	0.223	0.004	0.773
	$SP1_Y$	2.946	0.076	0.194	0.010	0.796
	$SP1_Z$	4.892	0.097	0.358	0.003	0.639

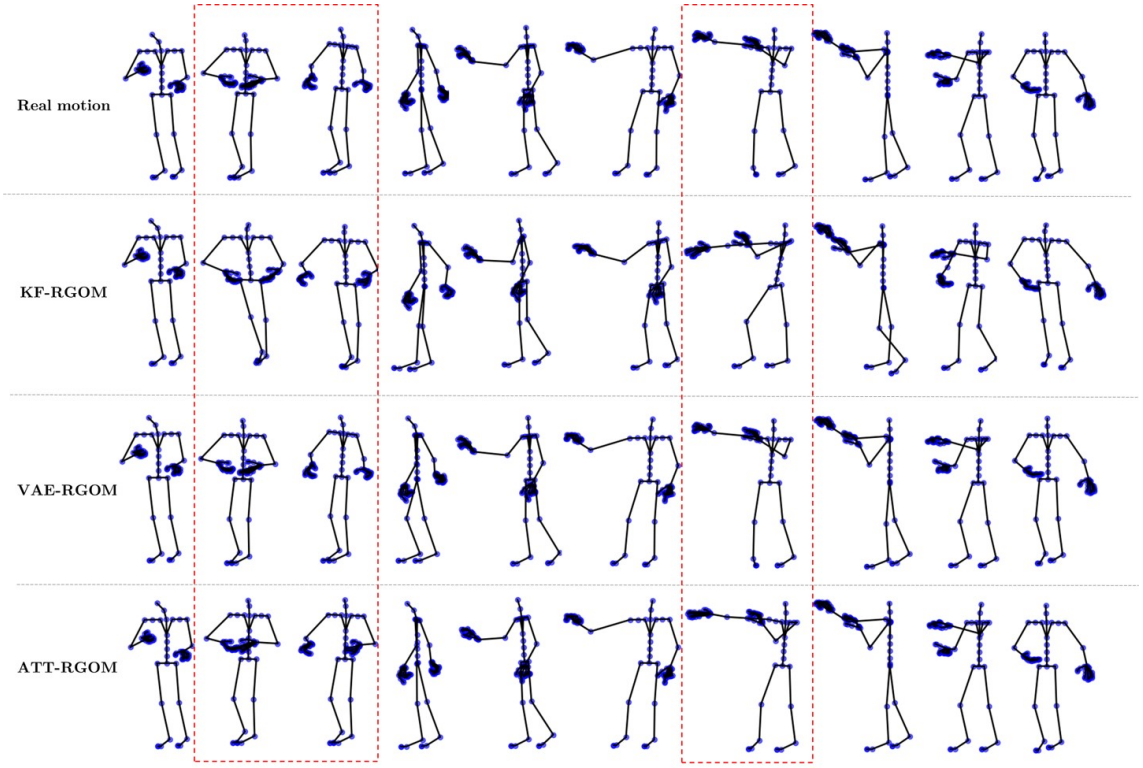


Figure 4.15: Visual comparison of generated posture sequences for TVA_1 and its ground-truth. The operator takes a circuit board from a container (the recording of the operator is shown in Figure 3.1a)

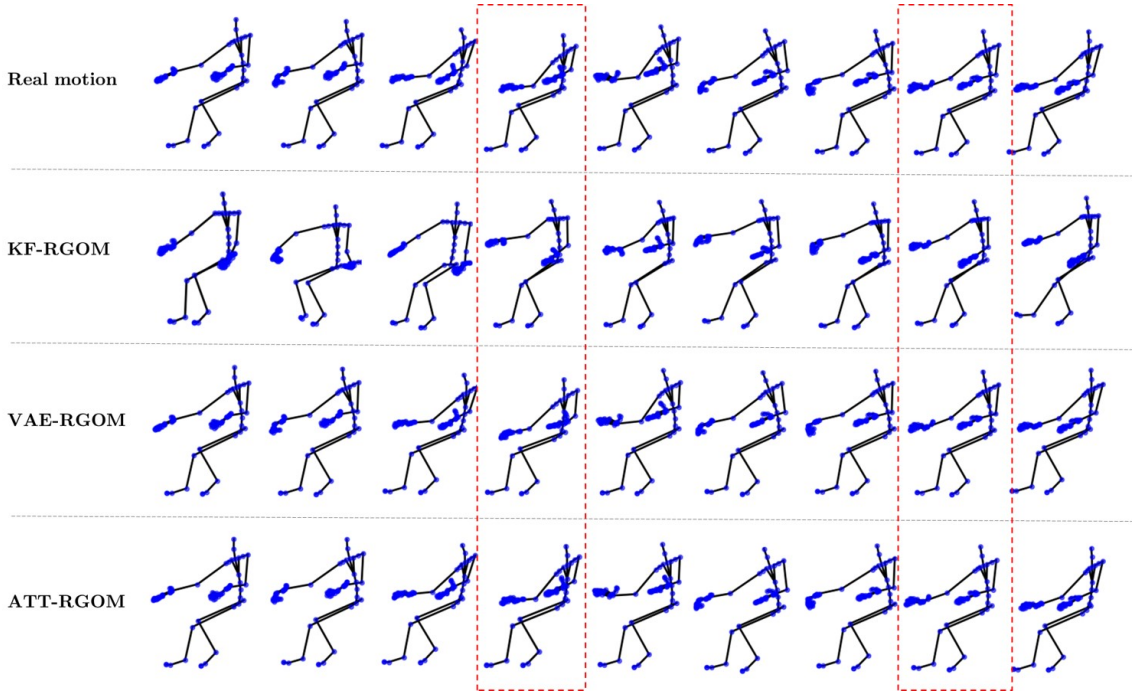


Figure 4.16: Visual comparison of generated posture sequences for GLB_4 and its ground-truth. The glassblower rotates the blowpipe with the left hand while shaping the glass with the right (the recording of the glassblower is shown in Figure 3.5a).

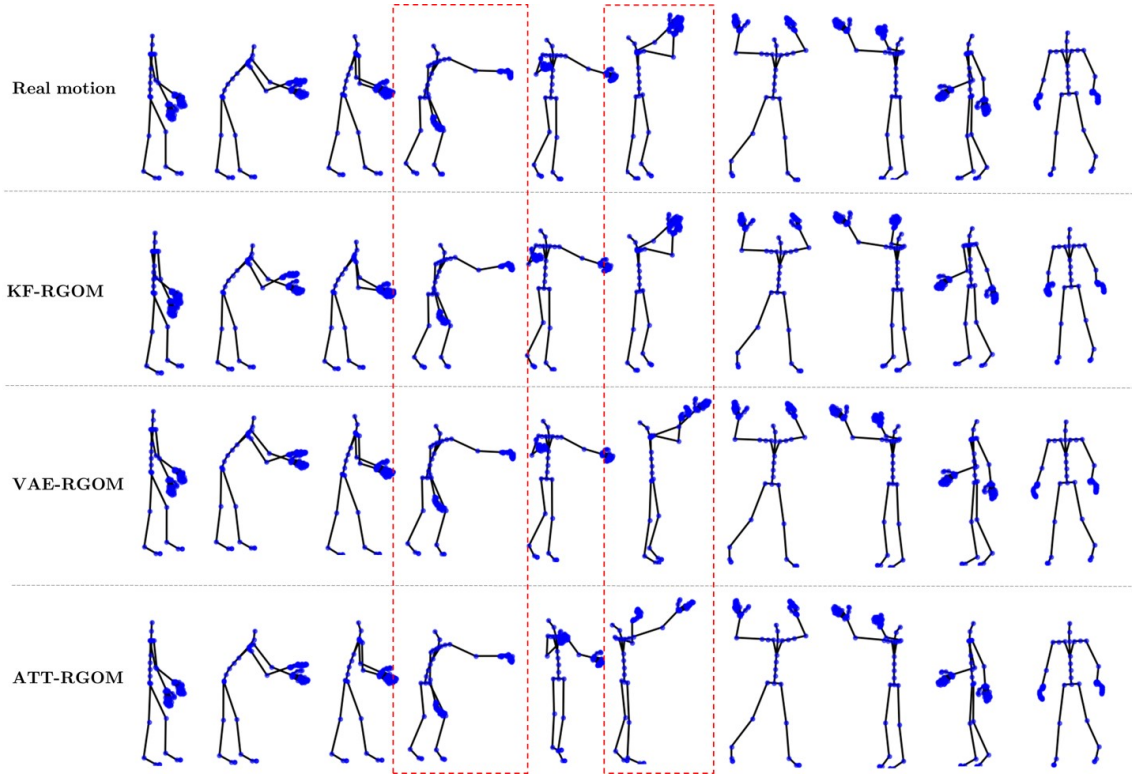


Figure 4.17: Visual comparison of generated posture sequences for TVP_8 and its ground-truth. The operator places a television on the third level of a pallet (picture of the recording in Figure 3.1d).

The potential of KF-RGOM, VAE-RGOM, and ATT-RGOM to dynamically simulate human movements was also examined. In a dynamic simulation, as stated previously, the model's predictions are used to predict the subsequent ones. Thus, only the first two samples of the endogenous data are real motion data values. Then the predictions are used to simulate the rest of the movement using real exogenous data. As prediction errors increase throughout the simulation, the accuracy decreases significantly compared to a static simulation in which each time step is predicted using actual motion data. Figures 4.18 and 4.19 depict the dynamic simulation of the MSC_5 and MSC_{11} movements. Here, endogenous data corresponds to the joint angles of RFA_X and RFA_Y, whereas exogenous data comprises the remaining full-body motion data. The results demonstrate that the accuracy of the simulation is, in fact, lower than that of static simulation but that the models can initially simulate the patterns adequately (up to 2-4 seconds). However, due to the accumulation of errors, the accuracy decreases at the end of each movement, especially for ATT-GOM and KF-RGOM. For the movement MSC_{11}, ATT-RGOM is unable to simulate it in its entirety since the errors are increasing exponentially from the start of the simulation. Because of this, Figure 4.19 only shows the dynamic simulations of VAE-RGOM and KF-RGOM.

The sensitivity analysis was conducted on the three approaches. Next are illustrated the results with movement APA_3. The movement consists of the operator bending to hold a bucking bar during the riveting of an airplane float. The shock of 80% was applied in the

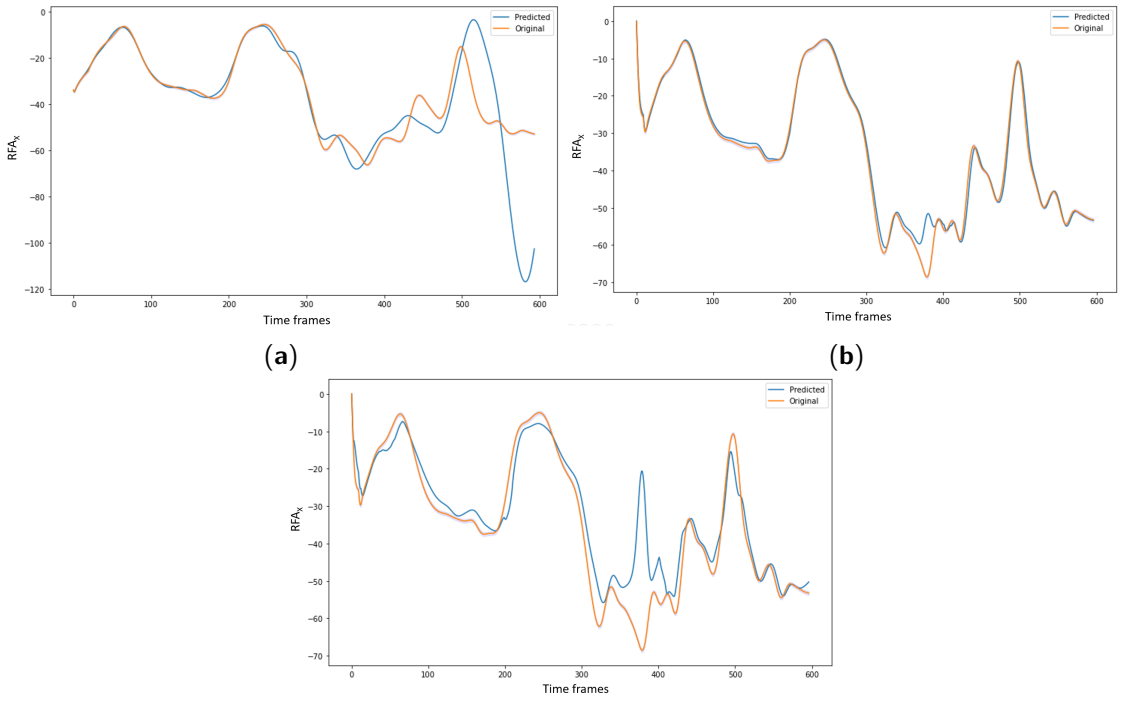


Figure 4.18: Dynamic simulation of RFA$_X$ for the movement MSC$_5$. (a) KF-RGOM; (b) VAE-RGOM.(c) ATT-RGOM.

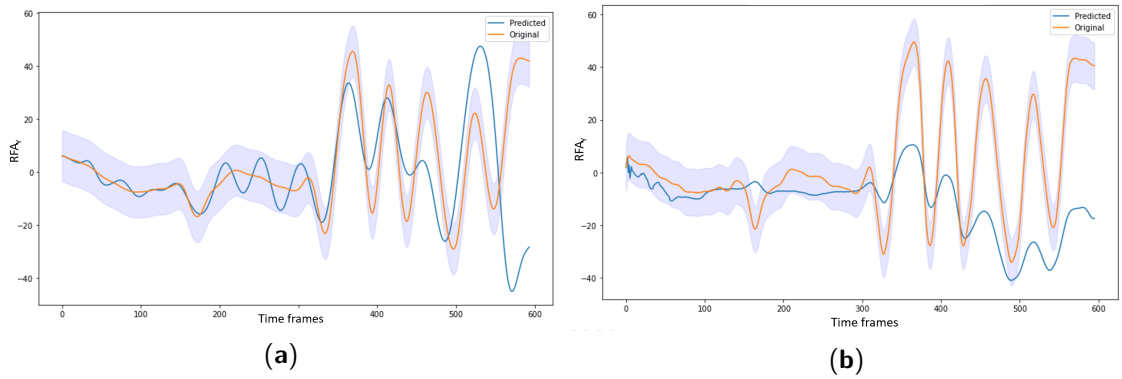


Figure 4.19: Dynamic simulation of RFA$_Y$ for the movement MSC$_{11}$. (a) KF-RGOM; (b) VAE-RGOM.

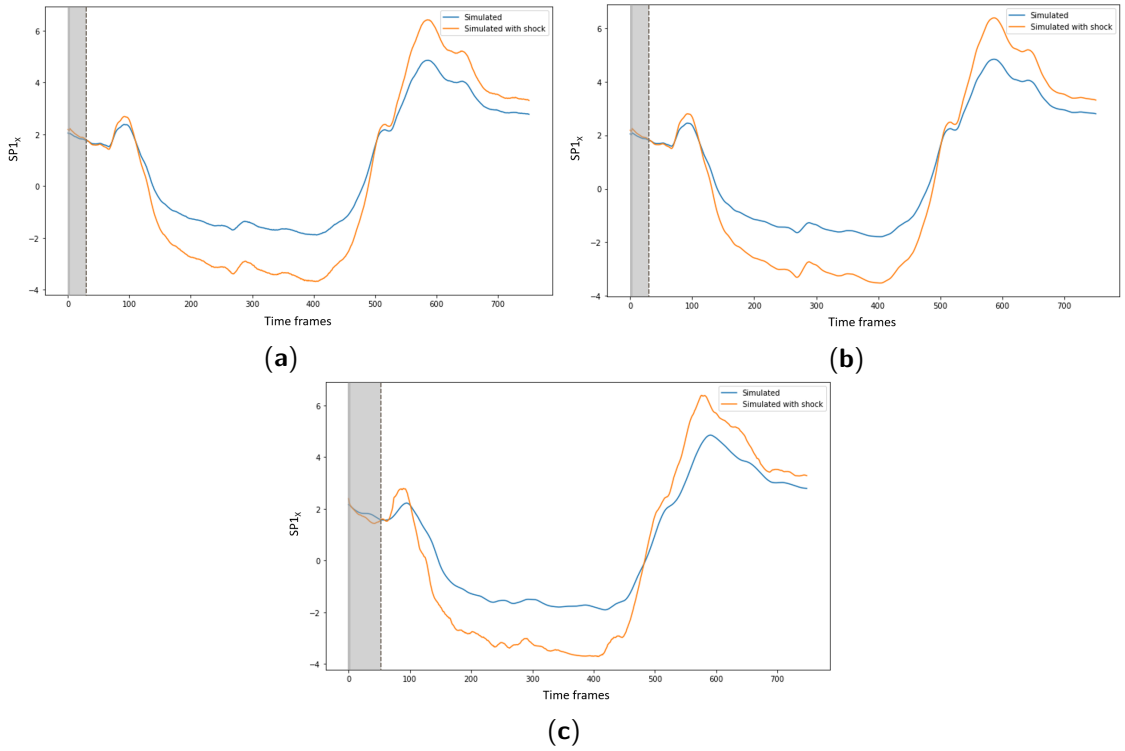


Figure 4.20: Simulated joint angle $SP1_X$ without disturbance (blue line) and with disturbance of 80% on the two initial time frames (orange line). (a) VAE-RGOM; (b) ATT-RGOM; (c) KF-RGOM.

two initial XYZ joint angles frames corresponding to the hips motion, H. Figure 4.20 displays the simulations of $SP1_X$ following the shock. The VAE-RGOM and ATT-RGOM simulations stabilize between the 20 and 30 frames (less than half a second). KF-RGOM, on the other hand, becomes stable after 50 frames (around half a second). This example demonstrates that the three approaches stabilize faster than KF-GOM after an external disturbance. As the influence of motion descriptors with disturbance is not consistent over the entire time series (coefficients change over time), the use of time-varying coefficients may make the models more resilient to disturbances. Yet, they exhibit a small distortion in the simulation following the shock.

4.6 Discussion

The experiments suggest that by solving the simultaneous equations that compose the GOM, it is possible to accurately simulate diverse human movements using Euler joint angles as motion descriptors. Overall, constant and time-varying GOM representations are tolerant of slight variations in human movements and offsets between movements of the same class produced by varying recording conditions (different subjects or different recording days).

Observing simply the simulations generated by each method reveals that VAE-RGOM and ATT-RGOM outperform KF-GOM and KF-RGOM. However, this is expected as motion representations from KF-GOM and KF-RGOM were trained using one-shot training. This implies

that only one movement template was used for computing the model's parameters, unlike VAE-RGOM and ATT-RGOM, which utilized data from all datasets. The variability between the reference and simulated movements may account for the errors exhibited by both KF-GOM and KF-RGOM. Accordingly, it can be inferred that the quality of their simulations depends on the recorded person's ability to replicate their movements while repeatedly performing the same activity. In the case of the datasets used, the majority of recorded subjects were experts in their respective fields, carrying out each task with great precision. Consequently, KF-GOM and KF-RGOM could capture the patterns formed in each motion template and simulate the test movements within the confidence bounds.

Implementing time-varying coefficients increased the modeling performance of KF-GOM, especially for movements with greater variance and longer duration, such as those conducted during glassblowing. This is due to the fact that coefficients were adapted to the change in mediations between the dependent variables and their assumptions throughout the whole time series. In addition, the findings of the sensitivity analysis suggested that this kind of representation was more tolerant of external disturbances. This tolerance can be helpful whenever the MoCap data contains artifacts, as it mitigates the error caused by these artifacts in motion trajectory predictions.

The errors in KF-RGOM depicted in Figure 4.12 may have been caused by the fact that the reference movement and the simulated movement were executed on different looms (reference on a large loom and simulated on a medium-size loom). Therefore, variations in pedal and position may have contributed to the errors in the simulation of the movement. Similarly, in the motion simulation depicted in Figure 4.13. The skilled glassblower progressively adjusted his posture, even for the same repetitive activity, in order to appropriately shape the molten glass. As a result, the training movements for each class of the GLB dataset did not adequately represent all movements from the same class (high intraclass variance), leading to a decrease in simulation performance.

Across all seven datasets, the time-varying parameter models estimated by data-driven approaches performed the best. Two arguments were deduced as to why this improvement in performance. First, both deep SSMs used motion data from all seven datasets for training, which would have allowed them to map diverse relationships between assumptions and dependent variables and accurately estimate the optimal coefficients for one-step prediction. Secondly, the temporal encoder-decoder structure of VAE-RGOM and ATT-RGOM enables these models to learn a low-dimensional (latent space) manifold of the data. Ideally, this manifold untangles variation factors across distinct movements, clusters related motion descriptors, and aids in identifying joint dynamics across sequences. In the case of VAE-RGOM, it disentangles the dynamics and postures in terms of the ELBO. For ATT-RGOM, this latent space allows the attention mechanism to interpret the hidden mechanisms and connections underlying the motion descriptors sequences.

According to the presented metrics, VAE-RGOM gave the most accurate movement simulations. VAE-RGOM may outperform ATT-RGOM since it models a probability distribution over future postures rather than making point estimates. Particularly, VAE-RGOM yields the high-

est simulation performance for movements from the datasets ERGD and TVA. These datasets are the largest ones and correspond to the simplest movements with low intraclass variability. In these, the movements were performed in a more controlled setting. For instance, in ERDG, the subjects performed diverse movements in a laboratory, receiving constant instructions on how to perform them. In the case of TVA, the operators were recorded in a production cell performing the same tasks repeatedly for several hours with little variation in between repeats. In addition, the movements in TVA primarily involved manipulating objects with their hands. In contrast to the movements performed, for instance, by the craftsmen and farmers, who had to employ their entire bodies to perform their work properly.

The most challenging movements to replicate were those associated with mastic cultivation. The reason behind this could be a bias in the training data, as MSC was the smallest dataset and involved movements where the farmer most of the time moved while kneeling. In the other six datasets, the subjects were mostly standing while performing their tasks. This may have prevented the networks from fully learning the dynamics of the legs when they are flexed. Because when the farmer moved to reach the tree or objects, he usually repositioned the legs while kneeling to improve balance.

In the dynamic simulation, a general observation for all models is that the error in long-term predictions increases, but they are able to reproduce motion patterns accurately for two seconds using only two time frames from the endogenous variable. In order to improve this performance, it would be necessary to construct a new loss function that takes into account long-term predictions, as opposed to the current loss function, which is only applicable to evaluating short-term predictions.

In determining the optimal approach for modeling human movements, there is a trade-off to be considered between the accuracy of the modeling and the computing cost of the training procedure. For example, VAE-RGOM and ATT-RGOM are able to simulate human movements more accurately and can be scaled to provide a greater variety of human movements. In addition, these approaches generate the representations of all full-body motion descriptors simultaneously, unlike KF-GOM and KF-RGOM, which require modeling one motion descriptor at a time, meaning training separately 57 models (one per descriptor) for simulating full-body movements. Nevertheless, the training of VAE-RGOM and ATT-RGOM is data-intensive, necessitating a large volume of data depending on the architecture of the neural network and a significant amount of computing power. KF-GOM and KF-RGOM, on the other hand, are sufficiently accurate to generate specific human movements using one-shot training. This training strategy enables users to specify the human movement and motion descriptors to be analyzed. Then construct their mathematical representation according to GOM using straightforward procedures that demand less computational power than data-driven methods. Table 4.11 summarizes the aforementioned advantages and disadvantages of each approach.

For applications requiring the analysis of multiple descriptors or human movements, it would be preferable to use VAE-RGOM and ATT-RGOM. These methods could also be utilized to augment data in deep learning applications. KF-GOM and KF-RGOM would be preferred for analyses when only small sets of human movements are available, as well as for applications

Table 4.11: Summary of each method's advantages and disadvantages.

Method	Advantages	Disadvantages
KF-GOM	▪ Training utilizing a single reference movement. ▪ Appropriate for assessing single human movements or specific motion descriptors. ▪ Analyzing constant representations is simpler than analyzing time-varying representations. ▪ Less computationally intensive than data-driven approaches (only CPU).	▪ Models are trained separately per motion descriptor (requires training 57 models for full-body simulations).
KF-RGOM	▪ Training utilizing a single reference movement. ▪ Superior simulation performance compared to KF-GOM. ▪ Suitable for assessing single human movements and specific motion descriptors. ▪ Less computationally intensive than data-driven approaches (only CPU).	▪ Unpractical while modeling high-dimensional motion descriptors.
VAE-RGOM	▪ High simulation accuracy. ▪ Practical for modeling large datasets. ▪ Represent and simulate a variety of movements using the same trained network (more robust than the statistical approaches).	▪ Demand more computational power (GPU).
ATT-RGOM	▪ Can model the whole body movement simultaneously (57 joint angles). ▪ High-dimensional motion descriptions are easier to process. ▪ Greater tolerance for external disruptions.	▪ Require a bigger dataset.

where modeling only a few motion descriptors is necessary. This is suitable for applications where the movements of two people are compared. For instance, while instructing a craft, the template motion could be based on the teacher's movements. KF-RGOM can learn its representation, and then the motion descriptors of the students can be fed into the trained model. If the simulations go outside the confidence bounds, the students can receive feedback to improve their performance.

4.7 Conclusion of the chapter

The work presented in this chapter has demonstrated, from a modeling perspective, that temporal architectures, whether statistical or data-driven, combined with GOM interpretable representations, are effective for learning the motion dynamics of varied professional activities and producing simulations of good quality. Three novel approaches were proposed, which estimate the parameters of state-space models that generate motion data. Experimental results confirmed that using time-varying representations improves the robustness of the models for accurately simulating a range of human movements. The deep state-space models were able to deal with various data distributions utilizing non-linear network parameterization and

offer interpretable forecasts by introducing exogenous variable data through the GOM representation. It can be concluded that VAE-RGOM and ATT-RGOM are the best approaches for estimating accurate models of human movement. However, as stated previously, the use of statistical or data-driven approaches would depend on the application requiring human movement analysis and the computational power limitations.

The most closely related works are the deep SSMs proposed by Fraccaro et al. [Fraccaro, 2016], Li et al. [Li, 2019], and the DeepAR [Salinas, 2017], as the observations and transition models are too non-linear. In contrast, the proposed approaches not only can predict human motion trajectories but can also produce comprehensible motion representations that may be used to explain how these predictions are made. In addition, ATT-RGOM uses attention to interpret the hidden dynamics of human movement, thus providing an explanation for the mechanics underlying the movement performance.

Next, in Chapter 5, the capability of the models to highlight mediations between joints and their interpretability are examined. The trained models in this chapter are used to gain insight into how specific human movements are performed.

Chapter 5

Body dexterity analysis of expert professionals

"We especially need imagination in science. It is not all mathematics, nor all logic, but it is somewhat beauty and poetry."

— Maria Mitchell

Contents

5.1 Introduction

In the preceding chapter, GOM motion representations were learned by incorporating assumptions about the spatial and temporal dynamics of human movement into the equations. These consisted of mediations between joints and dependencies on their precedent values. Identifying and capturing the interdependence between the motion of various joints with these models not only allows for realistic human motion simulations but also allows for the study of how varied and complex full-body human movements are accomplished. This chapter presents the analysis done over the learned GOM representations from Chapter 4 regarding their capability to explain inter-joint coordination through their mathematical assumptions. Analytical models such as these can be utilized to understand better the neurophysiological mechanisms underlying dexterity and motor learning based on the observed joint movement. Dexterity can be defined as the skill to perform a given movement or task using the hands or other body parts.

The notion is to use the trained motion models to observe and quantify the manifestation of skill in industrial operators and expert artisans. The parameters of the train models can give information about how a person moves in order to achieve a specific goal, such as assembling a TV or making a specific piece of glass. In the future, multidisciplinary frameworks might be built to study how people learn and get better at industrial or craft tasks by looking at the trained analytical models of experts and beginners. Furthermore, GOM could be used to investigate the biomechanical risk factors that lead to work-related musculoskeletal disorders by comparing motion representations from safe and hazardous movements.

5.2 Use of analytical models for human movement analysis

Initially, a statistical analysis is performed on the learned GOM representations to determine the significance of the models' assumptions in relation to the professional movement. The significant assumptions (motion descriptors) and their learned coefficients are then used to describe the cooperation of the joints to perform the movement. The goal here is to provide evidence for the thesis's second hypothesis: *The cooperation of body joints and their contribution during the performance of a human movement can be learned and represented through interpretable models*.

Next, by analyzing the p-values of each assumption, the most important motion descriptors for modeling and recognizing human movements from a professional task are found. In many applications of human movement analysis, it is neither feasible nor practical to use full-body MoCap suits. Therefore, to enable the adoption of less intrusive technologies, such as smartphones and smartwatches, a procedure for finding the minimal set of motion descriptors to measure using GOM is also detailed in this chapter.

Finally, the computation of tolerance intervals is provided as another use of the learned motion representations. These intervals consist of predefined motion ranges that can be used to evaluate a user's ability to replicate a particular movement. These intervals have the potential to be utilized by learning and skill acquisition systems that compare the movements of experts

and apprentices.

The following subsections provide an overview of how dexterity is analyzed using GOM, as well as examples of models learned using each of the approaches described in Chapter 4. Next, Section 5.4 outlines how the most important motion descriptors for each of the seven datasets were determined. The selections are then validated and discussed based on their capacity to enhance the performance of a gesture recognition problem. After, the procedure for calculating the tolerance intervals of joint motions is described in Section 5.5, followed by the chapter's conclusions in Section 5.6.

5.3 Analysis of experts' movements

5.3.1 Full-body dexterity analysis according to GOM

As mentioned in Chapter 2, GOM is an analytical model that learns human movements by means of a set of mathematical equations. The equations are designed based on four assumptions, depicted one by one in Figure 5.1. The concept of GOM is to explain body dexterity through these four assumptions:

- → **H1**: Velocity of the movement.
- ↔ **H2**: Movement of the body joint across the 3D space.
- ↔ **H3**: Cooperation between body limbs.
- ↔ **H4.1** and ↔ **H4.2**: Cooperation between serially and non-serially linked joints of the body.

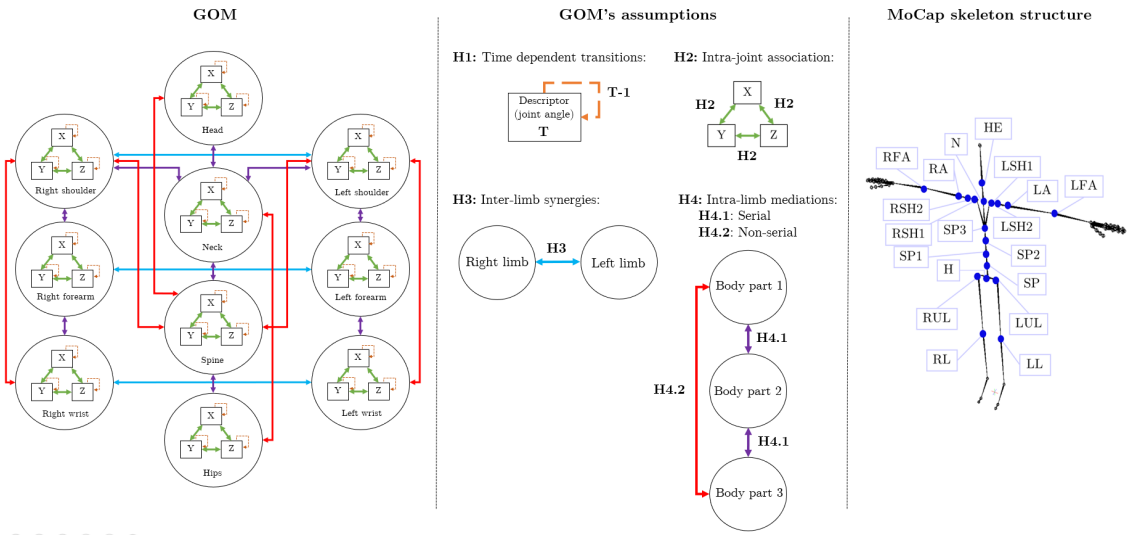


Figure 5.1: The Gesture Operational Model and its assumptions. The mathematical representation of GOM is utilized to model the movements of every joint of the MoCap skeleton. Then, the full-body movement is explained based on each joint motion model's coefficients and their statistical significance.

Each assumption consists of a specific set of variables (in our case, joint angles) that are parametrized (either with a constant or time-varying coefficient) and depict a particular relationship between body joints or a temporal dependency. By examining the generated coefficients and statistical significance of each variable, it can be gleaned how relevant these are according to the movement modeled and the predicted joint angles.

5.3.1.1 Extensive description of each assumption in GOM

The first assumption concerns transitioning or time dependency (H1), illustrated as $\rightarrow$ in Figure 5.1. This assumption is intended to explain how the modeled movement (and motion descriptors) is affected by the velocity with which it is conducted. This is reflected by the statistical significance and values of the coefficients assigned to the two previous endogenous variables included in each representation. For instance, if both preceding values of the predicted joint angle are statistically significant, it can be regarded as slow motion because the prediction is dependent on its predecessor. In contrast, a rapid motion would indicate a low dependence on prior values.

The second assumption corresponds to the *intra-joint association* (H2), depicted as $\leftrightarrow$ in Figure 5.1. This assumption describes how strongly are related the descriptors from which a body joint motion is decomposed. In the case of joint angles, are the connection between the angles X, Y, and Z from the same axis. As an example, if a joint only moves in a single anatomical plane, two joint angles would be closely associated (statistically significant in their motion representation), but their connection with the third angle could be relatively weak. The third angle would not greatly influence the prediction of the other two joint angles.

The third assumption concerns the *inter-limb synergies* (H3), depicted as $\leftrightarrow$ in Figure 5.1. This represents the bidirectional connection between the left and right side limbs. Consider the case where a movement involves manipulating a tool or object. In this situation, these assumptions reveal whether both hands are cooperating to operate the tool properly. This is evidenced by the significance and large values of the coefficients estimated for the right-hand variable in the left-hand's motion model and vice versa.

The fourth and final assumption involves *serial and non-serial intra-limb mediations* (H4), each shown as $\leftrightarrow$ and $\leftrightarrow$, respectively. This assumption describes the potential mediations between directly or indirectly connected joints. This assumption would disclose, for instance, how much the shoulder's motion influences the arm's motion (serial mediation) or the forearm's motion (non-serial mediation).

The following section illustrates how the motion representations learned in Chapter 4 describe the dexterity of the recorded subjects, given the above assumptions.

5.3.2 Statistical analysis of human motion representations

Statistical analysis by applying a t-test over the model's learned coefficients is conducted in order to investigate the relevance of their assumptions in the modeling of a human movement. The coefficients and statistical significance of the assumptions are utilized to interpret how the

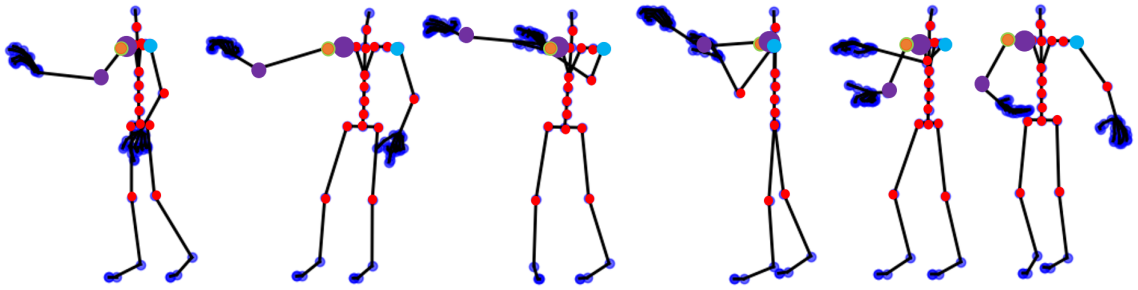


Figure 5.2: Illustration of the movement performed in TVA_1, where the operator grabs from a container a circuit board. The color annotations are based on the assumptions of Equation 5.1, where a larger circle implies an important variable based on coefficients and p-values. The picture of the recording can also be visualized in Figure 3.1a.

hands and other body parts accomplish the modeled movement.

The statistical analysis of four motion representations estimated by KF-GOM is provided next. Here are visualized the coefficients and p-values of the several assumptions that comprise the model of a joint angle, wherein some variables must remain dynamic and others static (coefficients close to zero). In addition, the posture sequence of the modeled movement is provided for each example, along with color annotations to highlight the equations' assumptions. The first example illustrates the equation for the joint angle sequence RAy_t (right arm on the Y-axis) when performing the movement TVA_1 (grab a circuit board from a container, shown in Figure 5.2):

$$RAy_t = \underbrace{(1.010)RAy_{t-1}}_{p = 0.001} + \underbrace{(-0.076)RAy_{t-2}}_{p = 0.188} + \underbrace{(0.720)RAx_{t-1}}_{p = 0.003} + \underbrace{(1.214)RAz_{t-1}}_{p < 0.001} + \underbrace{(-0.324)LAy_{t-1}}_{p < 0.001} + \underbrace{(6.123)RSH1y_{t-1}}_{p < 0.001} + \cdots + \underbrace{(0.555)RFAy_{t-1}}_{p = 0.009} \quad (5.1)$$

The p-values < 0.05 suggest a dependency between the prior value of the dependent variable but not between the value two time steps before. This implies that the movement is carried out at a moderate speed. If both previous values (assumption H1) are significant, this indicates a slow speed motion if neither is a faster one. The movement of the joint RA exhibits an intra-joint association along the X, Y, and Z axes. Inter-limb synergy with LAy (left arm) indicates that the operator performed the motion with both arms moving in synchrony. The movement on $RSH1y$ (right shoulder) and $RFAy$ (right forearm) result in a serial intra-limb mediation. This outcome makes sense, given that most of this arm movement primarily depends on shoulder motions (raising the arm). In addition, if viewing Figure 3.1a in Chapter 3, the operator must lift the shoulder and bend the forearm to reach the circuit board from the container. The bending of the forearm may explain the statistical significance of $RFAy$.

The second example is the equation for the joint angle of the neck on the X-axis (Nx_t)

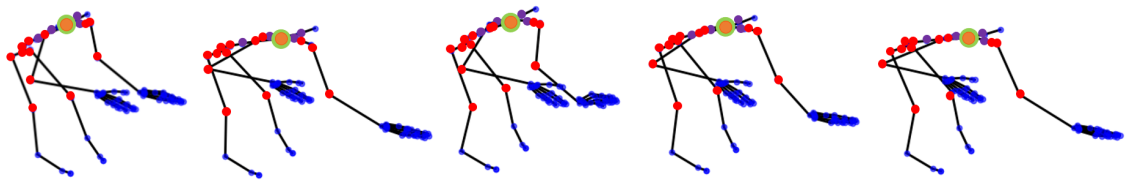


Figure 5.3: Illustration of the movement performed in APA_3, where the operator places the bucking bar to counteract the incoming rivet. The color annotations are based on the assumptions of Equation 5.2, where a larger circle implies an important variable based on coefficients and p-values. The picture of the recording can also be visualized in Figure 3.2c.

while performing APA_3 (hold the bucking bar, shown in Figure 5.3):

$$\begin{aligned} Nx_t = &\underbrace{(1.02)\, Nx_{t-1}}_{p < 0.001} + \underbrace{(-1.2)\, Ny_{t-1}}_{p < 0.001} + \underbrace{(-0.47)\, Nz_{t-1}}_{p < 0.001} + \cdots + \\ &\underbrace{(-0.01)\, SP2x_{t-1}}_{p = 0.002} + \underbrace{(-0.01)\, SP3x_{t-1}}_{p < 0.001} + \underbrace{(0.01)\, Hx_{t-1}}_{p = 0.84} \end{aligned} \tag{5.2}$$

Equation 5.2 reveals an intra-joint association with Ny and Nz, as well as a serial intra-limb mediation with SP3 (upper spine). SP2 (middle spine) exhibits non-serial intra-limb mediation, but H (hips) does not. Holding a bucking bar to counteract a rivet requires bending forward and slightly twisting the torso (as illustrated in figures 5.3 and 3.2c), moving along the X-axis and Y-axis of the spine. This movement is reflected in Equation 5.2, as the joint angles from SP2 and SP3 on the X and Y axes are statistically significant and relevant to the movement on Nx. In addition, the subject had to rotate the neck to see where to position the bucking bar; thus, this is consistent with the intra-joint association indicated by the p-values of Ny and Nz. At last, the movement is performed at a low pace as both transition assumptions are significant (only one is illustrated to show other joint mediations).

The following example is an equation trained with the movement GLB_4 (shape the decanter curves with a block, as depicted in Figure 5.4), and represents the joint angle on the X-axis of the left shoulder ($LSH2x_t$). More precisely, this equation simulates the motion of the left clavicle:

$$\begin{aligned} LSH2x_t = &\underbrace{(1.877) LSH2x_{t-1}}_{p < 0.001} + \underbrace{(-0.913) LSH2x_{t-2}}_{p < 0.001} + \underbrace{(0.292) LSH2y_{t-1}}_{p = 0.002} + \underbrace{(0.252) LSH2z_{t-1}}_{p = 0.004} + \\ &\underbrace{(0.145) RSH2x_{t-1}}_{p = 0.014} + \underbrace{(0.36) LAx_{t-1}}_{0.004} + \cdots + \underbrace{(0.016) LFAx_{t-1}}_{p = 0.030} + \underbrace{(-0.543) SP3x_{t-1}}_{p = 0.049} \end{aligned} \tag{5.3}$$

The statistical analysis of Equation 5.3 reveals a temporal dependence (slow motion); intra-joint association ($LSH2y$ and $LSH2z$); inter-limb synergy with the right shoulder; serial intra-limb mediation with the left arm (LAx), and non-serial mediation with the left forearm ($LFAx$). SP3 is considered marginally significant, as this study uses a p-value threshold of 0.05 to determine significance.

To shape the decanter correctly, both arms must work together during this movement. This

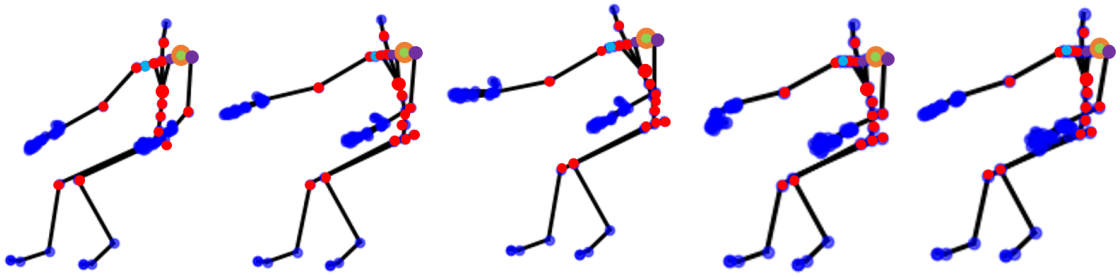


Figure 5.4: Illustration of the movement performed in GLB_4, where the expert glassblower shapes the decanter curve with a block and simultaneously rotates the blowpipe back and forward. The color annotations are based on the assumptions of Equation 5.3, where larger circles imply an important variable based on coefficients and p-values. The picture of the recording is shown in Figure 3.5a.

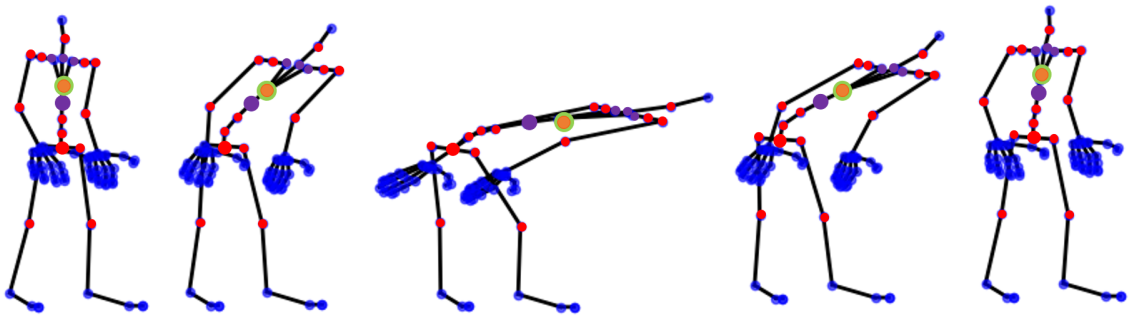


Figure 5.5: Illustration of the movement performed in $ERGD_7$, where the subject bends forward more than 60° for six seconds. The color annotations are based on the assumptions of Equation 5.4, where larger circles imply an important variable based on coefficients and p-values.

is evident by the presence of an inter-limb synergy in Equation 5.3. Accordingly, the joint angles of the right shoulder contribute to the response of the left shoulder, as the glassblower forms the decanter's curves with the right arm while rolling the blowpipe with the left. Furthermore, the expert mostly maintains the torso straight during this movement, as seen in figures 4.16 and 3.5a. Yet, when he rotates the blowpipe forward, there is a slight tilt of the torso to maintain grip on the blowpipe; this could indicate a high p-value for SP3, but not as high to not be significant for the left shoulder motion.

The Equation 5.4 represents the joint angle SP3 on the Y-axis ($SP3y_t$), when performing $ERGD_7$ (shown in Figure 5.5). During this movement, the subject bent forward more than 60°.

$$\begin{aligned} SP3y_t = &\underbrace{(2.13)\, SP3x_{t-1}}_{p = 0.007} + \underbrace{(-0.17)\, SP3z_{t-1}}_{p < 0.001} + \underbrace{(-0.91)\, Hx_{t-1}}_{p = 0.012} + \cdots + \\ &\underbrace{(0.42)\, SP1y_{t-1}}_{p < 0.001} + \underbrace{(-3.24)\, SP2x_{t-1}}_{p < 0.001} + \underbrace{(-0.06)\, HEx_{t-1}}_{p = 0.061} \end{aligned} \tag{5.4}$$

The p-values indicate that the intra-joint association assumptions are significant. The joint angles on the X-axis measured by the sensors SP3, H, and SP2 have the highest coefficient values and are statistically significant. Since the spine moves along the X-axis in order to

bend forward, these estimations are expected. However, the movement of the head (HEx_t) has little impact on the upper spine motion and is not significant. In addition, there is serial and non-serial intra-limb mediation with the spine angles on the Y-axis, most likely because the subjects do not bend fully on the X-axis.

The following are two examples of models with time-varying coefficients. These are summarized in tables, displaying the various representations provided by KF-RGOM, VAE-RGOM, and ATT-RGOM for the same movement. The equation is presented first with their corresponding assumptions and time-varying coefficients $\alpha_{i,t}$, where i is the number of coefficients. Next, a summary of their coefficients and p-values calculated for each time step is presented, along with a figure of graphs containing the time-varying coefficient calculated by a method and the predicted trajectory using these coefficients. The mean and standard deviation of the coefficients are provided in the tables, together with the range of p-values. The range indicates the highest and lowest p-values that were calculated over all time steps. Lastly, the posture sequence of the motion modeled in each equation is also given in a figure, highlighting with colors the equation's assumptions.

The first time-varying model is for the movement TVA_1, which consists of an operator grabbing a circuit board from a container. The joint angle on the Y-axis of the right arm (RAy) is modeled and represented in Equation 5.5. Tables 5.1, 5.2, and 5.3 present the summaries of coefficients and p-values estimated by each approach. Figure 5.6a depicts the predicted trajectory based on the coefficients calculated by ATT-RGOM, whereas Figure 5.6b illustrates the posture sequence highlighting the assumptions shown in Equation 5.5.

$$\begin{aligned} RAy_t = \alpha_{1,t} RAy_{t-1} + \alpha_{2,t} RAy_{t-2} + \alpha_{3,t} RAx_{t-1} + \alpha_{4,t} RAz_{t-1} + \\ \alpha_{5,t} LAy_{t-1} + \alpha_{6,t} RSH1y_{t-1} + \cdots + \alpha_{7,t} RFAy_{t-1} \end{aligned} \tag{5.5}$$

According to each table, the time-varying coefficients show a time dependence, being a low-speed motion; however, at certain periods of the time series, the values two time steps prior to the prediction were not significant in the KF-RGOM representation. This is consistent with the constant representation provided in Equation 5.1. All estimated representations exhibit an intra-joint association with the X and Z axes of RA (RAx and RAz). There is also an inter-limb synergy, but according to the coefficients provided by VAE-RGOM and ATT-RGOM, it was not significant for some periods of the time series. This indicates that there was not always a relationship between the movements of the right arm and the left arm, which could be the case if the operator was simply holding the card with the right hand, as seen in Figure 5.6b. Lastly, there is a serial intra-limb mediation with RSH1 and RFA for KF-RGOM's and VAE-RGOM's representation. For ATT-RGOM, though, the mediation with RSH1 and RFA was not present throughout the entire time series. Specifically, RSH1 and RFA were not significant on transitions where the operator is walking toward the container or just standing and holding the circuit board for a moment.

The second example is a time-varying model of GLB_4, which corresponds to the movement of shaping the glass decanter curves with a wooden block while turning the blowpipe with the

Table 5.1: KF-RGOM estimation for TVA_1.

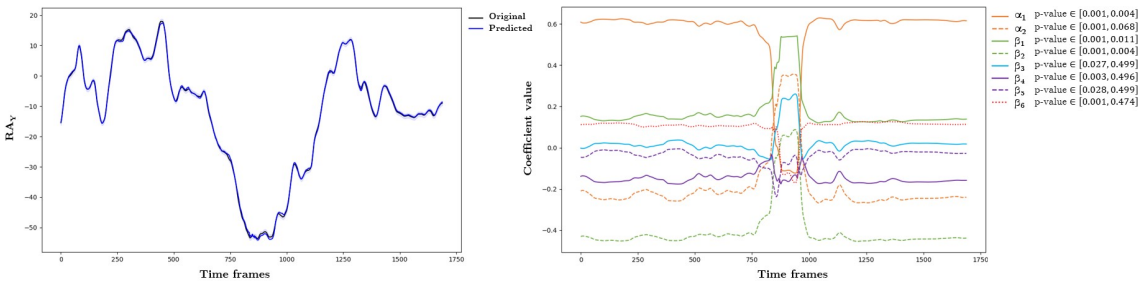
Variable	Coefficients $\mu_\alpha(\sigma_\alpha)$	P-values [min, max]
$\alpha_{1,t}$	0.486 (0.014)	[0.001, 0.004]
$\alpha_{2,t}$	-0.096 (0.009)	[0.078, 0.220]
$\alpha_{3,t}$	-1.236 (0.007)	[0.001, 0.010]
$\alpha_{4,t}$	-0.273 (0.002)	[0.001, 0.005]
$\alpha_{5,t}$	-0.005 (0.03)	[0.001, 0.040]
$\alpha_{6,t}$	-4.711 (0.010)	[0.001, 0.015]
$\alpha_{7,t}$	0.156 (0.020)	[0.026, 0.040]

Table 5.2: VAE-RGOM estimation for TVA_1.

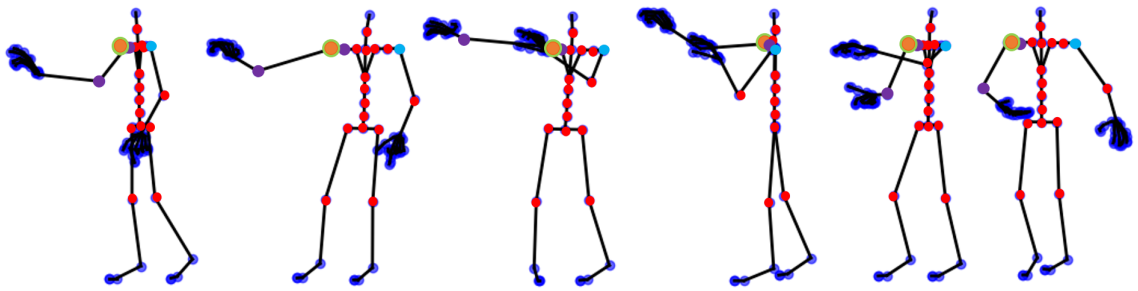
Variable	Coefficients $\mu_\alpha(\sigma_\alpha)$	P-values [min, max]
$\alpha_{1,t}$	1.089 (0.012)	[0.001, 0.005]
$\alpha_{2,t}$	0.051 (0.006)	[0.001, 0.010]
$\alpha_{3,t}$	-0.068 (0.003)	[0.001, 0.014]
$\alpha_{4,t}$	0.108 (0.002)	[0.001, 0.004]
$\alpha_{5,t}$	-0.001 (0.001)	[0.001, 0.474]
$\alpha_{6,t}$	-0.024 (0.002)	[0.001, 0.004]
$\alpha_{7,t}$	-0.003 (0.003)	[0.001, 0.020]

Table 5.3: ATT-RGOM estimation for TVA_1.

Variable	Coefficients $\mu_\alpha(\sigma_\alpha)$	P-values [min, max]
$\alpha_{1,t}$	0.688 (0.012)	[0.001, 0.004]
$\alpha_{2,t}$	0.309 (0.030)	[0.001, 0.008]
$\alpha_{3,t}$	-0.506 (0.012)	[0.001, 0.011]
$\alpha_{4,t}$	-0.100 (0.006)	[0.001, 0.004]
$\alpha_{5,t}$	-0.007 (0.015)	[0.157, 0.499]
$\alpha_{6,t}$	0.009 (0.012)	[0.003, 0.496]
$\alpha_{7,t}$	-0.001 (0.003)	[0.028, 0.499]



(a)



(b)

Figure 5.6: Generation of angle trajectory of *RAy* for the assembly movement TVA_1: (a) shows the predicted angles using Equation 5.5, which computed time-varying coefficients are visualized on the second plot; (b) illustrates the posture sequence with color annotations of the angles included as assumptions, where larger circles imply an important variable based on coefficients and p-values. The picture of the recording can also be visualized in Figure 3.1a.

Table 5.4: KF-RGOM estimation for GLB_5.

Variable	Coefficients $\mu_\alpha(\sigma_\alpha)$	P-values [min, max]
$\alpha_{8,t}$	0.543 (0.001)	[0.001, 0.005]
$\alpha_{9,t}$	0.456 (0.009)	[0.001, 0.002]
$\alpha_{10,t}$	0.223 (0.017)	[0.001, 0.020]
$\alpha_{11,t}$	0.192 (0.026)	[0.001, 0.010]
$\alpha_{12,t}$	-0.087 (0.037)	[0.018, 0.045]
$\alpha_{13,t}$	0.011 (0.002)	[0.001, 0.003]
$\alpha_{14,t}$	-0.002 (0.001)	[0.053, 0.060]
$\alpha_{15,t}$	0.135 (0.001)	[0.045, 0.087]

Table 5.5: VAE-RGOM estimation for GLB_5.

Variable	Coefficients $\mu_\alpha(\sigma_\alpha)$	P-values [min, max]
$\alpha_{8,t}$	0.483 (0.003)	[0.001, 0.007]
$\alpha_{9,t}$	0.343 (0.001)	[0.001, 0.014]
$\alpha_{10,t}$	0.060 (0.001)	[0.001, 0.033]
$\alpha_{11,t}$	-0.094 (0.001)	[0.001, 0.002]
$\alpha_{12,t}$	0.016 (0.004)	[0.001, 0.009]
$\alpha_{13,t}$	0.066 (0.009)	[0.001, 0.005]
$\alpha_{14,t}$	-0.042 (0.004)	[0.001, 0.011]
$\alpha_{15,t}$	0.006 (0.001)	[0.007, 0.015]

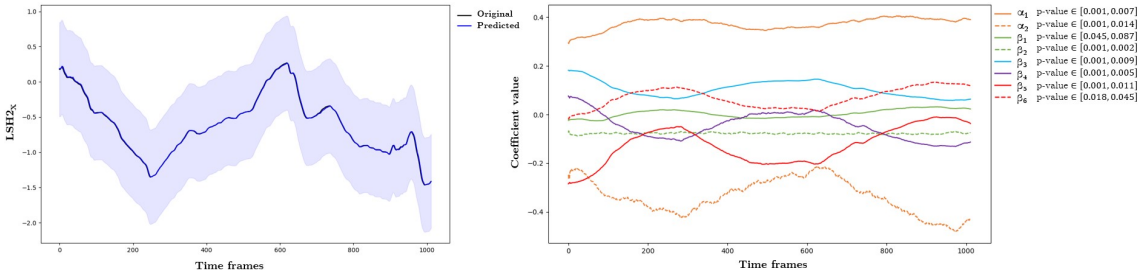
Table 5.6: ATT-RGOM estimation for GLB_5.

Variable	Coefficients $\mu_\alpha(\sigma_\alpha)$	P-values [min, max]
$\alpha_{8,t}$	0.444 (0.008)	[0.001, 0.006]
$\alpha_{9,t}$	0.348 (0.011)	[0.002, 0.018]
$\alpha_{10,t}$	0.022 (0.034)	[0.134, 0.161]
$\alpha_{11,t}$	-0.083 (0.010)	[0.001, 0.005]
$\alpha_{12,t}$	0.014 (0.003)	[0.002, 0.023]
$\alpha_{13,t}$	0.048 (0.003)	[0.001, 0.002]
$\alpha_{14,t}$	-0.013 (0.005)	[0.021, 0.144]
$\alpha_{15,t}$	-0.024 (0.017)	[0.029, 0.043]

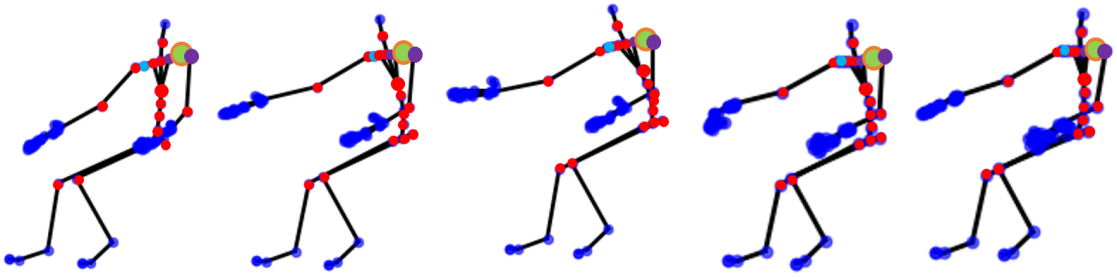
right hand. The Equation 5.6 represents the left shoulder's motion along the X-axis ($LSH2x_t$). Tables 5.4, 5.5, and 5.6 summarize each approach's estimated coefficients and their respective p-values. The time-varying coefficients calculated by VAE-RGOM for GLB_4 and their predicted joint angle trajectory are shown in Figure 5.7a. The posture sequence is illustrated in Figure 5.7b.

$$\begin{aligned} LSH2x_t = \alpha_{8,t} LSH2x_{t-1} + \alpha_{9,t} LSH2x_{t-2} + \alpha_{10,t} LSH2y_{t-1} + \alpha_{11,t} LSH2z_{t-1} + \\ \alpha_{12,t} RSH2x_{t-1} + \alpha_{13,t} LAx_{t-1} + \cdots + \alpha_{14,t} LFAx_{t-1} + \alpha_{15,t} SP3x_{t-1} \end{aligned} \tag{5.6}$$

The statistical analysis of each representation reveals a temporal dependency in GLB_4, indicating that the movement is slow. According to KF-RGOM and VAE-RGOM, there is an intra-joint association for all time series with $LSH2y$ and $LSH2z$, except for ATT-RGOM with respect to $LSH2y$. The movement presents an inter-limb synergy with $RSH2y$, indicated by KF-RGOM, VAE-RGOM, and ATT-RGOM, as it was as well for KF-GOM. This suggests a collaboration between both arms. As visualized in Figure 5.7b, the glassblower manipulated the molten glass with one arm while rotating it with the other. This action requires synchronization between both arms, which is reflected in the representations of all four approaches. Again, all methods considered a non-serial intra-limb mediation with LFA. There is also a mediation with SP3 in all the representations. Similarly to the representation from KF-GOM, the p-



(**a**)



(**b**)

Figure 5.7: Generation of angle trajectory of $LSH2x$ for the glassblowing movement GLB_4: (a) shows the predicted angles using Equation 5.6, which computed time-varying coefficients are visualized on the second plot; (b) illustrates the posture sequence with color annotations of the angles included as assumptions, where larger circles imply an important variable based on coefficients and p-values. The picture of the recording is shown in Figure 3.5a.

value of SP3 was near the threshold in the representation of KF-RGOM and ATT-RGOM. Accordingly, the movement of the upper part of the spine is crucial when performing this particular movement in glassblowing. This may also be observed in Figure 5.7b, where the glassblower moves his arms while bending his torso back and forth.

The preceding examples demonstrated how trained analytical models can be utilized to explain the physical dexterity of operators, craftsmen, and laboratory subjects that cannot be observed directly. The models emphasized the key motion descriptors associated with and contributing to complex whole-body movement. This information can later be utilized to test skill acquisition strategies. A novice can learn to make precise movements by minimizing the variability of their motion representations compared to those of professional artisans or operators.

5.4 Selection of the most significant sensors to maximize recognition accuracy

This section explains how the best motion descriptors for modeling and recognizing a set of human movements from each dataset are determined according to the learned GOM representations. After performing the statistical analysis, the number of times a motion descriptor (assumption) is statistically significant for all equations that comprise GOM is counted. As an example, Tables 5.7 - 5.10 show the top ten variables that were more frequently significant

Table 5.7: KF-GOM - TVA dataset.

p-Value < 0.05					
Spine		**Arms**		**Legs**	
Variable	*Count*	*Variable*	*Count*	*Variable*	*Count*
$SP1_Z$	49	LA_X	56	RUL_Y	32
$SP2_Z$	47	$RSH1_X$	55	LUL_Z	32
H_Y	46	$RSH2_Y$	55	LUL_Y	31
H_Z	45	$RSH1_Z$	54	RL_Y	31
N_Y	44	$RSH2_Z$	53	LUL_X	29
$SP1_X$	43	$RSH2_X$	53	LL_X	29
H_Z	42	RA_Y	49	RL_X	29
N_X	42	LFA_Z	48	H_X	29
$SP1_Y$	41	$LSH1_X$	46	RUL_X	29
$SP3_X$	41	LFA_X	42	RUL_Z	29

Table 5.8: KF-GOM - APA dataset.

p-Value < 0.05					
Spine		**Arms**		**Legs**	
Variable	*Count*	*Variable*	*Count*	*Variable*	*Count*
$SP3_X$	209	$LSH2_X$	243	LUL_Z	39
$SP3_Y$	205	$LSH1_X$	236	RUL_X	39
$SP2_X$	202	RA_Z	230	H_X	38
H_Z	202	LFA_X	229	LL_X	38
H_X	201	RFA_Y	227	LUL_Y	38
SP_X	201	LA_Y	224	LL_Y	37
$SP1_Y$	197	LA_Z	217	LL_Z	37
$SP1_Z$	193	$RSH1_X$	217	RL_X	36
$SP3_Z$	193	LFA_Y	216	RL_Y	36
N_Y	193	LFA_Z	212	RL_Z	36

Table 5.9: KF-GOM - GLB dataset.

p-Value < 0.05					
Spine		**Arms**		**Legs**	
Variable	*Count*	*Variable*	*Count*	*Variable*	*Count*
$SP3_X$	155	$LSH2_Y$	99	H_Y	65
$SP3_Y$	155	RA_X	92	LL_Z	63
$SP3_Z$	149	RFA_Z	90	LL_Y	62
$SP2_X$	118	$LSH2_X$	89	RL_X	60
$SP2_Z$	116	$RSH1_Z$	88	RL_Y	60
$SP2_Y$	110	$LSH1_Z$	86	H_Z	59
$SP1_Y$	105	$RSH1_Y$	85	LL_X	59
$SP1_X$	102	$RSH2_X$	85	RUL_Y	58
$SP1_Z$	93	LA_Y	84	RUL_Z	58
N_X	89	$LSH2_Z$	84	LUL_Y	57

Table 5.10: KF-GOM - ERGD dataset.

p-Value < 0.05					
Spine		**Arms**		**Legs**	
Variable	*Count*	*Variable*	*Count*	*Variable*	*Count*
$SP3_Z$	332	$LSH1_X$	534	RUL_Z	474
$SP2_Y$	330	LA_X	533	RUL_Y	473
$SP2_Z$	330	$RSH1_X$	523	LUL_Y	472
$SP3_X$	326	$LSH1_Y$	520	RL_X	468
$SP3_Y$	316	LFA_X	520	LL_X	465
$SP2_X$	311	$RSH2_X$	518	LUL_X	461
HE_Z	279	$RSH1_Y$	516	LUL_Z	457
SP_Z	264	RA_X	514	RUL_X	456
H_Y	261	$RSH1_Z$	508	LL_Z	455
N_Z	258	LA_Y	507	RL_Y	455

to the movements of the datasets TVA, APA, GLB, and ERGD, based on the representations learned by KF-GOM. These tables are arranged based on the body regions of the spine, arms, and legs. Also, the variables are sorted in descending order according to their incidence. For time-varying representations, counting was performed for each time step of the modeled movement.

Then for the selection, different combinations of descriptors considered most frequently significant were utilized for training in an all-shots approach. Because a single inertial sensor gives three joint angles, all of the sensor's joint angles were used for recognition if at least one was among the joint angles that were more often significant in all movements of a dataset.

The first combination to be tested consisted of a minimal sensor configuration: the best sensor for measuring the motion of the spine, another for the motion of the arms, and a third for the motion of the legs. If the recognition performance was poor, motion data from another relevant sensor was added to improve it. If not, the first tested sensors were swapped with one of the top three sensors measuring a similar body region (spine, arms, or legs).

The sensor configurations that obtained the best recognition results with each dataset's movements are presented in the following section, along with details on the techniques and metrics used to validate the selected sensors.

5.4.1 Validation and discussion of the selected joints

For the recognition of human movements utilizing different sensor combinations, HMMs were trained. In order to properly train the HMMs, a gesture vocabulary containing the movements with the most iterations was specified for each dataset. The total number of motion classes for TVA, APA, and ERGD were four, three, and 28, respectively. The TVP, GLB, and MSC gesture vocabularies contained only movements with at least seven repetitions. Therefore, their respective gesture vocabularies included five, seven, and six classes of movements. Regarding SLW, the gesture vocabulary consisted of only three classes of silk weaving on a loom. Despite the differences in loom size, the movements used to weave on a small, medium, and large loom are similar. Therefore, they were combined into three classes for the gesture recognition problem.

The HMM ergodic and left-to-right topologies, along with a different number of hidden states, were evaluated to determine the best HMM settings for the gesture vocabulary defined in each dataset. The performance metrics utilized were accuracy and F1-score, the last being the harmonic mean of precision and recall:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \qquad F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{5.7}$$

Note that TP corresponds to true positives, TN is true negatives, FP is false positives, and FN is false negatives.

As an illustration of the process for finding the best sensor configuration for each gesture vocabulary, the experiments done for the ERGD dataset with KF-GOM representations are described next. According to Table 5.10, which provides an ordered list of meaningful descriptors for ERGD motions, the sensors SP3 and SP2 provided the most significant joint angles for the modeling of the spine motion; for the arms motion, the sensors LSH1, LA, and RSH1; and for the legs motion, RUL and LUL. Different configurations were then used to train 28 HMMs for gesture recognition in order to determine the optimal configuration. To find the best number of states for the 28 HMMs, 10-fold cross-validation was performed. Initially, the best sensor for each body region was utilized for recognition, followed by the best two, and then the best three. Since the three configurations resulted in high false negatives and false positives in the recognition of movements that only varied in the forearms' posture, the most significant sensor placed on a forearm was added to the configuration. In this case, the left forearm sensor (LFA) was the one with more times being significant in the GOM representations. Table 5.11 shows the F1-score achieved with each configuration of sensors tested and with a different number of states in HMMs. The configuration comprised of the sensors SP2, LSH1, RSH1, RUL, and LFA, and with seven states in the HMMs, yielded the best performance of 0.917.

All gesture vocabularies were subjected to the selection process described previously with each approach's representations. Then, the sensor configurations that achieved the best performance were compared to the recognition performance obtained by utilizing all motion data from all sensors. Additionally, the recognition performance using a minimal set of two sensors was also computed for comparison. This minimal set consisted of two hand-picked sensors

Table 5.11: F1-scores achieved with each configuration of sensors and number of states, tested for ERGD using KF-GOM representations.

Sensors	Number of states in HMMs									
	3 states	4 states	5 states	6 states	7 states	8 states	9 states	10 states	11 states	12 states
SP3, LSH1, RUL	0.683	0.731	0.775	0.841	0.850	0.857	0.811	0.820	**0.860**	0.825
SP2, SP3, LSH1, LA, LUL, RUL	0.799	0.878	0.868	0.867	0.866	0.872	**0.896**	0.879	0.853	0.850
SP2, SP3, LSH1, LA, RSH1, LUL, RUL	0.812	0.849	0.860	0.889	**0.900**	0.849	0.880	0.870	0.848	0.848
SP3, LSH1, RUL, LUL	0.749	0.827	0.856	0.863	0.870	0.861	0.830	0.804	**0.884**	0.855
SP2, SP3, LSH1, LA, LUL, RUL, LFA	0.855	0.887	0.871	0.876	0.899	**0.906**	0.842	0.868	0.833	0.834
SP2, SP3, LSH1, LA, RSH1, LUL, RUL, LFA	0.860	0.885	0.876	**0.900**	0.892	0.879	0.869	0.885	0.867	0.859
SP2, LSH1, RUL, LFA	0.755	0.799	0.839	0.821	**0.894**	0.881	0.884	0.830	0.873	0.849
SP2, LSH1, LA, RUL, LFA	0.784	0.827	0.854	0.889	0.902	0.874	**0.904**	0.885	0.840	0.860
SP2, RSH1, LA, RUL, LFA	0.766	0.862	0.865	0.876	**0.917**	0.893	0.886	0.873	0.869	0.879

Table 5.12: Selected sensors for each dataset.

Motion representation	**Dataset**						
	TVA	*TVP*	*APA*	*GLB*	*MSC*	*SLW*	*ERGD*
KF-GOM	LA, SP1, RUL	RSH1, LFA, SP2	RA, LSH1, LSH2, SP3, SP2, LUL, RUL	LSH2, RFA H, SP3	LSH1, SP3, LUL, LL	RSH1, LSH1, HE, LUL, RL	LA, RSH1, LFA, SP2, RUL
KF-RGOM	LA, RFA, SP	LSH1, RFA, SP2	RA, LSH1, SP3, LUL	RSH1, RFA, SP1, RUL	LSH1, SP3, LL, LUL	LA, RA, SP, LUL, RL	RA, LSH1, RFA, SP1, LL
VAE-RGOM	LA, SP, LL	LA, RFA, SP	LA, SP2, LL	LA, LSH2, SP3, LL	LA, LSH2, SP3, LL	LA, LSH1, RFA, SP3, LL	LA, LSH1, RFA, SP3, LL
ATT-RGOM	LA, LSH1, H	LA, RFA, SP1	LA, LSH1, H	LSH1, LSH2, RFA, LUL	LSH2, RFA, LUL, LL	RA, LSH1, RFA, SP1, LL	RA, LSH1, RFA, SP1, LL

that provided the Euler joint angles of the right forearm (RFA) and hips (H). The sensor positioned on the right forearm was chosen since the majority of individuals in all datasets were right-handed, and the sensor placed on the hips because all spinal movement originates from the hips. The purpose of these comparisons is to assess the method's capability to select the set of sensors that achieves superior recognition performance over configurations that include all 52 inertial sensors or a manually picked set of sensors.

Left-to-right HMM topology produced the best results for all recognition problems. Concerning the number of hidden states, it was defined for the HMMs of TVA and ERGD with seven states, TVP with six states, APA, GLB, and MSW with eight states, and SLW with three states. Table 5.12 illustrates the sensors selected for each dataset based on the motion representations generated by each approach. Then, Figures 5.8 and 5.9 display the calculated metrics for each sensor set per dataset.

The relevance of the sensors selected based on each approach for each dataset was demonstrated by the superior or similar recognition performance attained compared to using all sensor data. By observing Figures 5.8 and 5.9, the best minimal set for TVA and SLW was determined by using the representations estimated by ATT-RGOM. These sensors achieved comparable results to using all sensors' data, having the ATT-RGOM set a mean accuracy of 0.952 and F1-score of 0.949 for TVA, whereas all sensors set had a mean accuracy of 0.967 and an F1-score of 0.966. Then, for SLW, the ATT-RGOM sensor set performed at least 1% better than all other sensor sets in terms of mean accuracy and F1-score.

VAE-RGOM representations provided the best sensor set for TVP, with a mean accuracy

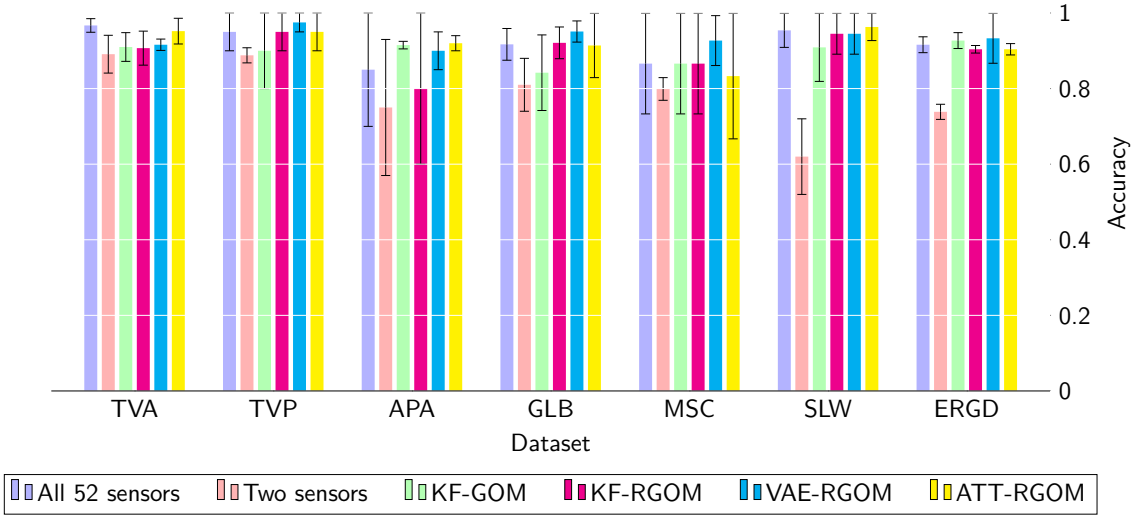


Figure 5.8: Accuracy of the recognition according to each selected set of sensors.

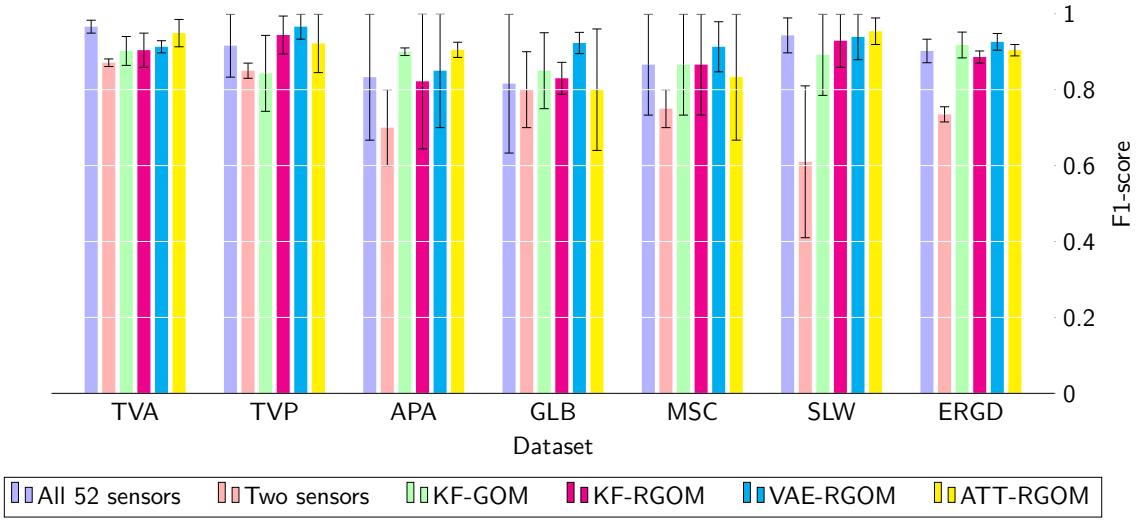


Figure 5.9: F1-score of the recognition according to each selected set of sensors.

of 0.975 and an F1-score of 0.966. When comparing the sensors from the VAE-RGOM and ATT-RGOM sets, it is observed that the VAE-RGOM set included a sensor from the lower spine (SP). This sensor enhanced the ability to distinguish between the movements of placing a box on the first and second levels.

The recognition of movements from APA was better using the data provided by the sensor sets of KF-RGOM and ATT-RGOM, although the ATT-RGOM sensor set contained fewer sensors. The set selected using ATT-RGOM representations attained a mean accuracy of 0.920 and an F1-score of 0.905, outperforming the set containing all sensor data. The APA motions were the most challenging to recognize. This may be because the movements in this vocabulary are more complex and prolonged. The most problematic movement to model and recognize was APA_2, which was expected given that its execution varied the most among the three classes (high intra-class variance). The operator did not prepare the material identically for each repetition. In certain iterations, the operator was slower than in others because he needed more time to adjust the pneumatic hammer or prepare more rivets. In addition, there is a substantial intra-class variance due to the fact that just one airplane structure was constructed for this dataset. There were no repetitions in which the pneumatic hammer was positioned in the same location more than once.

The glassblowing movements performed in GLB were better recognized using the data from the sensor sets estimated using VAE-RGOM representations, reaching a mean accuracy of 0.951 and an F1-score of 0.923. According to all GOM representations, the shoulders contribute the most to the execution of glassblowing movements, which is why the two-sensors configuration performed the worst. In addition, using the motion data measured from the left calf, which VAE-RGOM included, the recognition performance improved by at least 5% in the F1-score compared to all other sets.

The sensors picked using VAE-RGOM representations were also the most effective at discriminating movements from the MSC and ERGD datasets. These were the gesture vocabularies with the greatest number of classes. Figures 3.6 and 3.3 show that between the movements of these two gesture vocabularies, subjects assumed similar postures. Because of this, the selected sensors for each recognition problem are similar. The VAE-RGOM sensor set attained a mean accuracy of 0.927 and an F1-score of 0.913 for MSC. Then, for ERGD, the VAE-RGOM sensor set achieved a mean accuracy of 0.933 and an F1-score of 0.926. The poor performance of the two-sensor configuration for MSC and ERGD may have been caused by its inability to differentiate between movements that differ only in the posture of the legs, as the motion data from the hips was insufficient.

5.5 Computation of tolerance intervals for analyzing movement similarity

Some applications of human movement analysis involve evaluating a subject's performance by analyzing the similarity between two movements. For instance, to examine gait [Ezati, 2019] or to instruct proper golf or tai-chi postures [Liao, 2021; Kamel, 2019]. Another application

of measuring movement similarity is for communicating with a human-computer interface [Caramiaux, 2015; Santos, 2019]. The motion representations presented in this thesis can be used to compare distinct movements. One method is to compare their GOM mathematical representations directly; alternatively, tolerance intervals of the joints' motion can be calculated if the application calls for assessing how well a user can replicate a movement.

These tolerance intervals vary throughout the time series and indicate the range of motion acceptable for properly executing a specific movement. In order to calculate the tolerance intervals, all repetitions of a movement are first aligned in time using dynamic time warping and a template movement. Then, their time-varying GOM representation is estimated so that their aligned coefficients can be extracted. The tolerance intervals can then be defined using the standard distribution of the coefficients ($\sigma_{n,t}$) for each time step t:

$$\mu_{n,t} = \frac{\sum_{i=1}^{R} \alpha_{i,n,t}}{R} \qquad \sigma_{n,t} = \sqrt{\frac{\sum_{i=1}^{R} (\alpha_{i,n,t} - \mu_{n,t})^2}{R}} \tag{5.8}$$

where R is the number of repetitions of a movement and n the number of coefficients.

One or two standard deviations can be defined for estimating the tolerance intervals for the correct execution of a movement. Figure 5.10 illustrates three examples of tolerance intervals defined as two standard deviations. The first example consists of the movement of bending forward more than 60° ($ERGD_7$) for six seconds, the second is the embroidery of a mastic tree (MSC_5), and the third is the movement of moving the shuttle while weaving a silk textile ($SLW_{4,2,1}$). In Figure 5.10a, it can be observed that the tolerance interval is wider at the moment the subjects bend, as the subjects need to readjust their posture after bending to maintain balance and prevent falling forward. The second example illustrates that there is greater variation in the first cuts of the tree, which may be due to the fact that the harvester does not begin cutting at the same location. However, at the end of the movement, there is more precision in the cutting. The third example shows the process of rapidly moving the right forearm to move the loom's shuttle. There is a wide interval at the beginning of the action, presumably due to the different initial positions of his hand in each repetition. The wide interval segment at the second curve may thus represent the variability in how broadly the weaver moved his forearm to move the shuttle. Note that the tolerance intervals in the first example for $ERGD_7$ are wider than in the other two, as they are generated using representations from several subjects. In contrast, the tolerance intervals for MSC_5 and $SLW_{4,2,1}$ are calculated using motion models from a single farmer and a single skilled craftsman, which performed movements with higher precision.

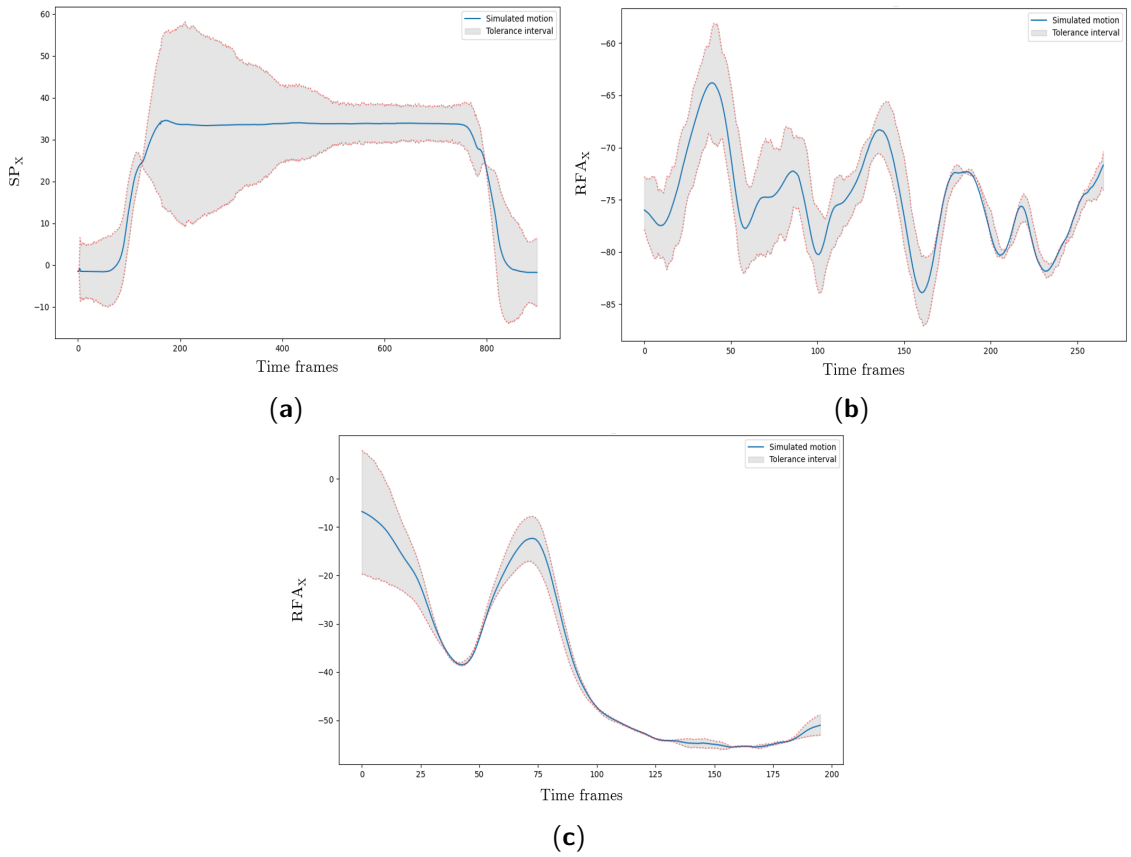


Figure 5.10: Examples of tolerance intervals. (a) $ERGD_7$; (b) MSC_5.(c) $SLW_{4,2,1}$.

5.6 Conclusion of the chapter

This chapter presented the application of GOM representations for dexterity analysis. Also, a methodology for identifying the best motion descriptors for modeling and recognizing movements from a gesture vocabulary.

The GOM representations generated by KF-GOM, KF-RGOM, VAE-RGOM, and ATT-RGOM, were statistically analyzed to examine how statistical methods or data-driven approaches depicted human movements through interpretable mathematical equations. The movement of each body joint was described by the coefficients and p-values of its respective model. These revealed the joints that contributed most to the prediction of the modeled joint motion and the significance of potential joint associations.

The results demonstrated the ability of the proposed approaches to mathematically describe human movements and explain how a movement is conducted in accordance with GOM's assumptions. As potential future work, the proposed analytical models can be integrated with neurophysiological techniques that, for example, account for muscle activity and motor cortex activity. This combination would allow for a more thorough approach that could provide a neurophysiological roadmap of complex body dexterity. However, because of the inherent complexity and the sheer amount of data it requires, such a complete study that considers all these neurophysiological factors hasn't been done to this point.

The most relevant sensors for a set of movements were found and selected using the estimated p-values of each assumption that composes the GOM representations. Per each GOM estimation approach (KF-GOM, KF-RGOM, VAE-RGOM, and ATT-RGOM) was selected a set of sensors using their motion representations. To validate the selection of sensors with different gesture vocabularies, only the motion data of the selected sensors was utilized for gesture recognition. The recognition performance using the motion data from the selected sensors was compared to that obtained using all sensors data from the MoCap suit and data from a minimal configuration of two hand-picked sensors. The results showed that in most datasets, each approach's selected sensors outperformed or matched the recognition performance of the set containing all sensor data. Thus, it was possible to identify the motion descriptors that best solved each recognition problem with the proposed methodology. Overall, the representations given by VAE-RGOM and ATT-RGOM allowed for the selection of the sensors with the greatest discrimination between motion classes.

As stated before, it is neither feasible nor practicable to employ a full-body MoCap suit in many human movement analysis applications. Determining the minimal motion descriptors to measure allows for the adoption of less invasive technologies, such as smartphones and smartwatches, that could also measure these motion descriptors. The following chapter illustrates an example in which a minimal set of sensors determined using GOM representations are used to perform automatic ergonomic evaluations of industrial tasks.

Last but not least, this work contributes to the literature by providing a procedure for calculating tolerance (or expert) intervals of joints' motion. These represent the acceptable range of motion for reproducing a specific movement from a professional task, based on the

recorded movements of skilled operators and artisans. When practicing sports, dancing, or playing an instrument, these tolerance intervals could allow an application to provide effective feedback. Users could learn sophisticated motor skills even in the absence of an instructor by comparing their MoCap data to that of an expert during the teaching process.

Chapter 6

Computational ergonomics for task delegation in Human-Robot Collaboration

"AI is the new electricity - it will transform every business and industry."

— Andrew Ng

Contents

6.1 Introduction

Industry 4.0 has resulted in a rise in research in the field of Human-Robot Collaboration. As a result, robotic agents are being integrated into the work routine, not to take the position of human operators but to assist them in accomplishing complicated and physically demanding tasks. By properly incorporating collaborative robots, operators may be able to avoid developing Work-related Musculoskeletal Disorders (WMSDs). WMSDs are a significant concern in the industry, constituting the majority of work-related health problems in Europe [Jan de Kok, 2019]. These are caused by the repeated performance of complex and repetitive operations that frequently demand operators to push themselves beyond their normal physical limitations.

In designing HRC frameworks, task delegation must be optimized while considering ergonomic aspects to maximize operators' comfort and production efficiency in industrial co-production cells. With this in mind, this chapter presents a methodology for ergonomically effective task delegation to design optimal HRC frameworks. The hypothesis formulated is that by utilizing a reduced amount of MoCap data, operators' movements can be accurately measured, allowing for a more thorough ergonomic analysis of their actions, and facilitating task delegation when implementing HRC frameworks.

This chapter is divided into six main sections. First, Section 6.2 discusses the current state of ergonomic analysis frameworks. Following that, Section 6.3 describes the methodology utilized to test the formulated hypothesis. Section 6.4 explains the development of the automatic ergonomic evaluation system. Then, the ergonomic evaluation of real professional tasks and the outcomes obtained are detailed in Section 6.5. The implementation and evaluation of the optimized HRC framework for TV assembly is later described in Section 6.6. Finally, Section 6.7 gives the conclusion and suggestions for future work.

6.2 Current ergonomic analysis methods

As mentioned earlier, the activities performed by manual laborers in the industrial sector are becoming more challenging and complex in order to meet market demands within certain time limits, job specifications, and budget constraints. Operators must go beyond their natural physical limitations to undertake repetitive jobs for long periods of time in order to complete the tasks required of them. Being subjected to such constant physical strain leads to WMSDs. Ergonomists have developed a variety of methods for evaluating work-related tasks. The methods based their analysis on theoretical knowledge of human physical limitations and abilities indicated by known standards (e.g., ISO 11226:2000 and EN 1005-4). Some of the most popular methods are RULA [McAtamney, 1993], EAWS [Schaub, 2013], and Ovako Working Posture Analysing System (OWAS) [Karhu, 1977]. To implement these methodologies, the ergonomist observes an operator executing the task under evaluation and annotates various body part postures on a worksheet, such as the one illustrated in Figure 6.1. The ergonomic score of the task is then calculated using these annotations. This method of scoring determines which tasks should be changed for better ergonomics. However, because these approaches rely

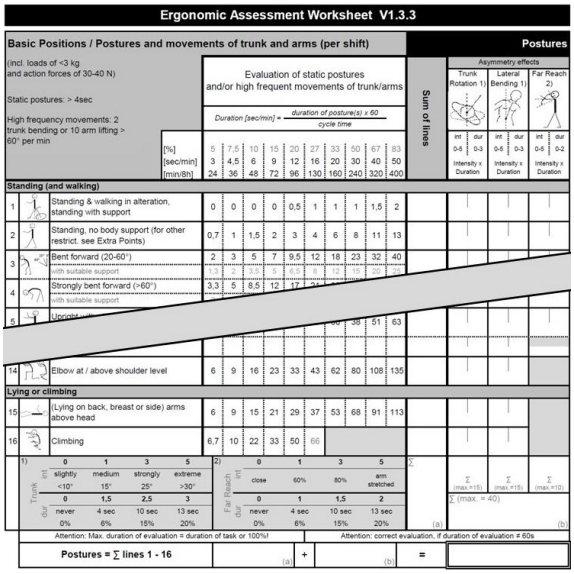


Figure 6.1: EAWS postural assessment section. The ergonomist completes the worksheet to estimate the overall ergonomic score of the task based on the observed posture.

on the ergonomist's perception and experience, scoring can be subjective and have a lot of inter-variability. Alternative sensor-based ergonomic evaluation approaches are now being developed by researchers. Optical and inertial-based MoCap systems have frequently been used to extract upper body posture for ergonomic evaluation [Manghisi, 2017; Shafti, 2019]. By using IMUs, Vignais et al. created a real-time upper-body ergonomic assessment based on RULA. Similarly, Yan et al. [Yan, 2017] used inertial sensors to track the torso inclination of construction workers for ergonomic monitoring.

For designing ergonomic HRC scenarios, previous studies used biomechanical simulations to compute ergonomic metrics (posture, physical effort, and energy spent during the task) [Kim, 2018; Marin, 2018]. However, the main downside of these approaches is that they are hard to incorporate into industrial applications that demand rapid reconfigurability. This is because the human ergonomic analysis in these studies is done offline and in a laboratory. Subjects are asked to simulate the tasks, and then an offline biomechanical analysis is done to compute the ergonomic metrics for the workstation redesign. For an accurate performance evaluation, optical MoCap technology (only available in specialized laboratories) or multiple inertial sensors distributed throughout the body are typically used to measure the movement of subjects. Laboratory recordings might lead to inaccurate measurements since they lack authenticity and are not based on actual workplace scenarios. Moreover, using multiple sensors for tracking operators' movement is impractical to implement in the industry. Consequently, there is still a need for methodologies that employ technologies that are simple to implement in real-world settings and can rapidly and accurately estimate the full-body ergonomic risk level of any representative set of manipulation actions performed in the industry.

6.3 Methodology for ergonomically optimizing HRC in TV assembly

The ability to record accurate measurements for ergonomic analysis is essential as it provides quantitative measures of operators' performance. Firstly, a pipeline for automatic recognition of four postural risk factors and computation of ergonomic scores is designed based on the evaluation protocol of the EAWS. The first risk factor is the posture of the legs (F1), which includes three possible motion patterns: standing, sitting, and kneeling. The second factor focuses on torso inclination (F2), consisting of two patterns: bending forward or upright torso. The third risk factor is lateral flexion and rotation of the torso (F3). Lastly, the posture of the arms is the fourth risk factor (F4). Depending on the detected factors in the movement evaluated, an EAWS-related score is assigned on a scale ranging from 0.5 to 26.5, with higher values being attributed to the more risky postures. HMMs were trained for risk factor recognition using motion primitives that presented various combinations of the four ergonomic risk factors. This dataset consists of ERGD, presented in Chapter 3.

For evaluating professional tasks, these are initially segmented into small motion primitives. These are then processed by the automatic ergonomic evaluation pipeline, which estimates ergonomic scores. The collaborative robot is then assigned the tasks with the most hazardous movements, while the operators are given the ergonomically safe or supervisory control tasks.

In order to implement the proposed methodology in an industrial environment, it is necessary to use less intrusive technologies and minimize the number of sensors placed on the human body. Therefore to overcome this limitation, the performance of the pipeline is evaluated using the motion data of the selected set of sensors identified in Chapter 5. These are the ones estimated using KF-GOM motion representations for the recognition of the ERGD motions (Section 5.4.1). In addition, a smaller set of two sensors was examined for comparison in order to investigate the feasibility of utilizing IMUs of smartphones or smartwatches, which is a more realistic attempt for wide industrial implementation.

Figure 6.2 depicts the methodology for task delegation in HRC, which is evaluated by analyzing actual professional tasks carried out on a television production line. This approach for task delegation is part of the two-step methodology presented in [Olivas-Padilla, 2023], for improving HRC in manufacturing using computational ergonomics. In this context, the resulting task configuration for the television production scenario described in this chapter was then utilized to develop an optimized HRC framework. This HRC framework was proposed by Papanagiotou et al.[Papanagiotou, 2021], and was designed to enhance as well the ergonomics and safety of the production cell by incorporating gesture recognition and pose estimation. The gesture recognition was used to create contactless communication between a robot and a human operator, ensuring the robot's temporal adaptation. Alternatively, the robot adjusted its movements to the anthropometric characteristics of each operator through pose estimation, augmenting its perception and enabling its spatial adaptation in the HRC.

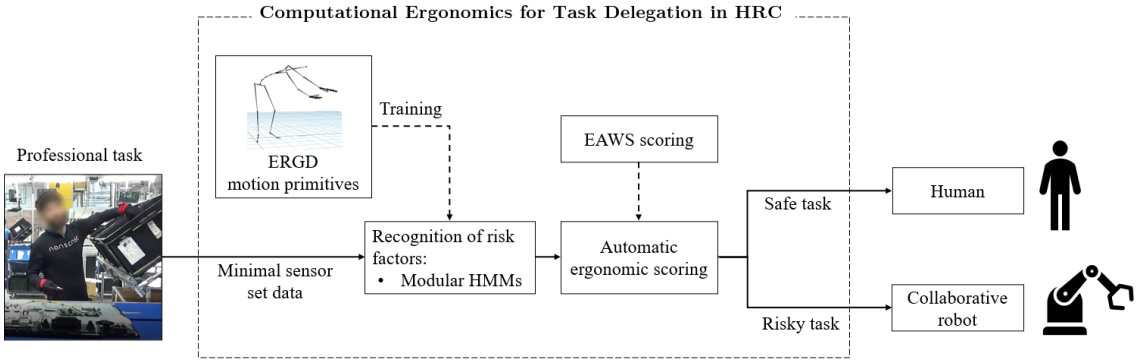


Figure 6.2: Pipeline for ergonomically optimizing industrial co-production cells with HRC.

6.4 Automatic computation of an EAWS-related ergonomic score

Four sets of HMMs were utilized for the recognition of postural risk factors. Figure 6.3 illustrates the modular scheme for recognizing the four factors using the selected sensors for the ERGD dataset: SP2, RSH1, LA, RUL, and LFA.

Three HMMs were trained to recognize F1 using only the joint angles from the RUL (right upper leg). Each HMMs represented one of the three possible leg postures (standing, sitting, and kneeling). The HMMs with the highest likelihood indicated the identified posture. If HMMs F1.1 has the highest likelihood, for instance, the detected posture is standing. Two HMMs were trained to recognize F2, using only the data from the sensor located on the spine (SP2). One HMM modeled the movements when subjects were standing and the other when they were bending forward. The data from the arms and spine (SP2, RSH1, and LA) were used to train two additional HMM for the recognition of F3, as subjects moved both body parts to perform the movements involving the risk factor F3. One HMM modeled the movements where subjects rotated and lateral bent their torsos, and the other the movements where they did not. The recognition of F4 was performed with another HMM, trained using the data from the arms and shoulders. One HMM modeled the movements where the subject's arms

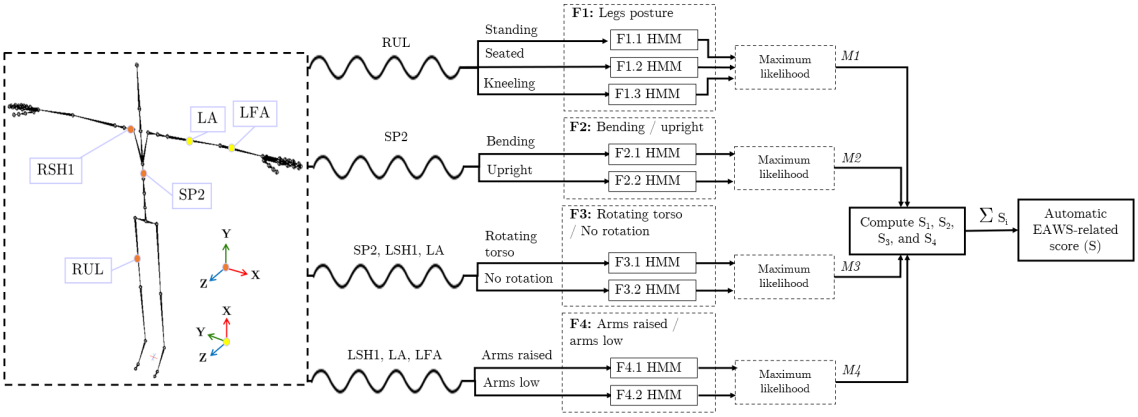


Figure 6.3: The pipeline for the motion modeling using inertial data and the computation of the EAWS-related score.

were raised above shoulder level, while the second HMM modeled the movements where the subject's arms remained below shoulder level. All HMMs followed a left-to-right topology and learned the hidden states using the Baum-Welch algorithm, given each motion's observation sequence (joint angles). A stratified 10-fold cross-validation was utilized to select the number of states for each model.

For the computation of an EAWS-related score, five equations were designed based on the tables provided by the EAWS worksheet, in the working posture assessment section for postures assumed for approximately six seconds [Schaub, 2013]. The automatic EAWS-related score is defined as $S \in \left[0.5, 26.5\right]$. The higher the ergonomic risk score, the greater the risk. The final score S consists of the sum of the scores S_1, S_2, S_3, and S_4 as illustrated in Equation 6.1.

$$S = S_1 + S_2 + S_3 + S_4 \tag{6.1}$$

S_1 is computed as follows:

$$S_1 = L_{M1}, \quad L = \begin{bmatrix} 1.5 \\ 0.5 \\ 7 \end{bmatrix} \tag{6.2}$$

where $M1$ is used as the index of vector L, which is composed of the constants defined by EAWS for standing, sitting, and kneeling. For instance, when the subject is seated, a value of 0.5 is assigned, while when the subject is kneeling, a value of seven is assigned. The second score S_2 is calculated using the following formula:

$$S_2 = (M2 - 1)B_{M1}, \quad B = \begin{bmatrix} 7 \\ 1 \\ 3 \end{bmatrix} \tag{6.3}$$

In Equation 6.3, $M2$ is two if the subjects are bending and one if they are not; B is the vector of constants for forward bending, where a constant is selected based on $M1$. If the subjects are bending forward, $(M2-1)$ is one, and a constant is obtained from the vector B. However, if the subject is standing, the subtraction $(M2 - 1)$ is zero as S_2. The next score, S_3, is calculated as follows:

$$S_3 = 7.5(M3 - 1) \tag{6.4}$$

$M3$ is two if the subject's torso is rotating and one if it is not. Thus, If there is torso rotation, S_3 is equal to 7.5, if not equal to zero. S_4 is finally calculated using the equations 6.5 and 6.6.

$$S_4 = (M4 - 1)(2 - M2)A_{M1} + 5(M4 - 1)(M2 - 1) \tag{6.5}$$

$$A = \begin{bmatrix} 7 \\ 6.5 \\ 9 \end{bmatrix} \tag{6.6}$$

Note that $M4$ is two if the subject's arms are detected to be raised and one otherwise. If the arms are raised, S_4 value would depend on whether the subjects are also bending forward and

Table 6.1: Recognition performance with each configuration of sensors for F1, F2, F3, and F4. Note that All sensors: Configuration with all the sensors data; H and RF: Configuration using only two sensors data.

Risk factor	Sensors	F1-scores
F1	All sensors (19)	0.950
	SP2, RSH1, LA, RUL, and LFA	0.856
	H	0.793
F2	All sensors (19)	0.946
	SP2, RSH1, LA, RUL, and LFA	0.938
	H	0.859
F3	All sensors (19)	0.916
	SP2, RSH1, LA, RUL, and LFA	0.926
	H and RFA	0.927
F4	All sensors (19)	0.928
	SP2, RSH1, LA, RUL, and LFA	0.925
	H and RFA	0.945

whether they are standing, seated, or kneeling. For instance, if the subject is not bending, a constant for having raised arms is obtained from the vector A. This constant differs based on whether the subject is standing, seated, or kneeling, as indicated by the index $M1$. If it is detected that the subject is also bending forward (indicated by the index $M2$), S_4 is equal to five directly.

As mentioned earlier, a configuration of two sensors was additionally evaluated to determine the feasibility of implementing the proposed pipeline using IMUs from smartphones and smartwatches. The sensors utilized in this configuration were the sensor on the right forearm, which represented the inertial sensor of a smartwatch, and the sensor on the hips, which represented the sensor of a smartphone. The right forearm was chosen because most subjects were right-handed, and the hip sensor was selected as the movement for bending forward and rotating the torso originates from the hips.

6.4.1 Experimental results and discussion

A stratified cross-validation procedure with ten iterations was utilized for the evaluation. The data set was randomly divided into ten equal-sized parts. Nine of them were used to train HMMs, while the remainder were used for testing. The process was repeated for all ten parts. Since the data set contained fewer motions where subjects were kneeling or raising their arms, a stratified cross-validation was chosen to maintain the same proportion of movements with different factors across iterations. Consequently, only 180 movements per class were utilized for F1 (standing, sitting, and kneeling), 90 per class for F2 (upright and bending), 90 per class for F3 (no torso rotation and torso rotation), and 90 per class for F4 (arms low and arm raised). The recognition performance with each configuration of sensors for F1, F2, F3, and F4 after the ten cross-validation iterations is shown in Table 6.1, using as a metric the F1-score.

The overall F1-score achieved with the selected set of sensors was 0.911, 0.881 with the two sensors, and 0.935 with the all sensors configuration. The factor that was the most challenging for the two-sensor set was F1, as there is only one sensor on the hips, which was insufficient to

discriminate between the three different legs postures. The two-sensor set is thus preferred for upper body monitoring. These results indicate that it is possible to accurately compute EAWS-related scores of human movements using a minimal set of sensors. This could enable the daily use of smartwatches and smartphones for ergonomic assessment in the workplace.

6.5 Evaluation of television assembly movements

This section describes the task delegation approach in which real professional tasks are automatically evaluated using the pipeline proposed in the preceding section, using a minimal set of sensors. The professional tasks performed in a television assembly line are evaluated, which consist of the movements captured in the dataset TVA, presented in Chapter 3. The entire assembly procedure can be broken down into four main tasks, each illustrated in Figure 6.4. The first task is grabbing a circuit board from a container (T_1); the second is taking a wire from a second container (T_2); the third involves connecting the board and wire and placing them on the TV chassis (T_3); the fourth task corresponds to drilling the circuit boards on the TV chassis (T_4).

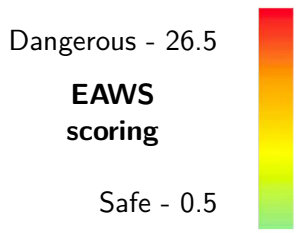
For the ergonomic evaluation, the tasks are segmented into short windows of similar duration to the motion primitives in ERGD (less than 4 seconds) and then provided to the automatic evaluation system. All ergonomic scores estimated per window for each task are annotated and used to calculate each task's mode, standard deviation, and mean ergonomic score. The statistics are then used to identify which tasks expose operators to a higher ergonomic risk. These are proposed to be delegated to a collaborative robot, leaving only the safer tasks to be performed by human operators.

6.5.1 Results of the ergonomic evaluation

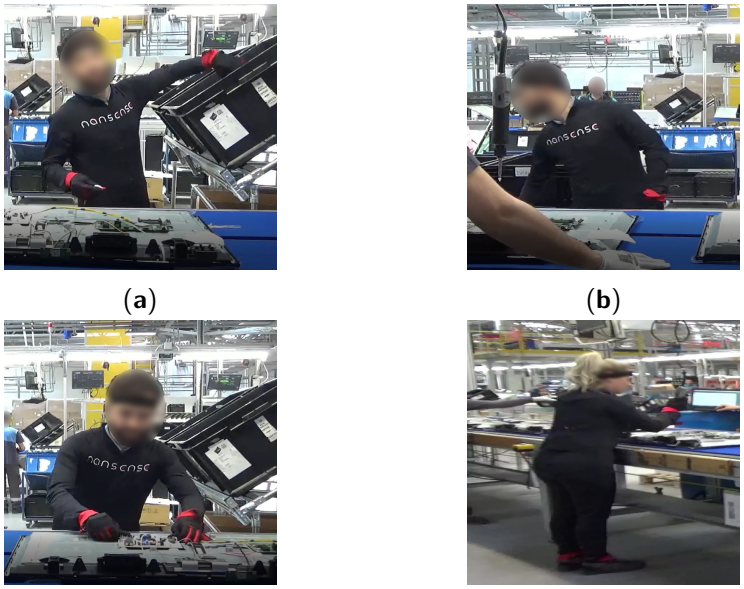
The estimated EAWS scores for each professional task are summarized in Table 6.2. Table 6.2 contains the mean, standard deviation, and mode of the ergonomic scores calculated per task.

According to these results, the majority of iterations of tasks T_1 and T_2 can be classified as medium-risk movements, while iterations of tasks T_3 and T_4 are classified as low-risk movements. The most prevalent risk factors for T_1 were movements in which the elbows were raised above shoulder level while the torso was flexed laterally. These results are expected based on the movements performed in T_1, as operators must rotate and laterally bend their torsos to retrieve a circuit board from a container. In addition, operators must raise their arms

Table 6.2: Summary statistics of the EAWS scores calculated for each task.



Tasks	EAWS scores		
	Mean	STD	Mode
T_1	16.02	2.65	17.50
T_2	15.02	3.43	16.00
T_3	10.76	3.68	8.50
T_4	11.50	3.19	12.50



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

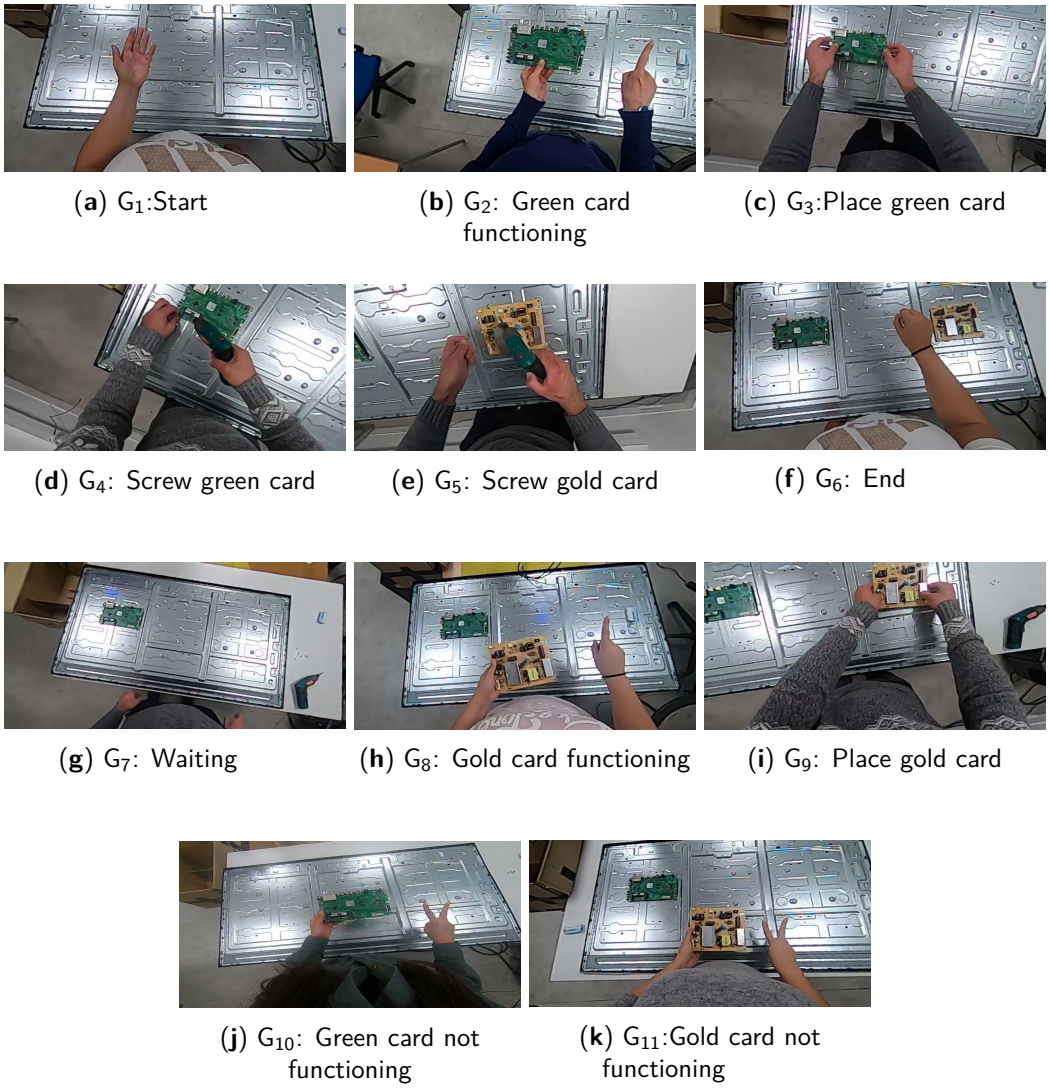
Figure 6.4: Professional tasks for TV assembly. (**a**) T_1: Grab the circuit board from a container; (**b**) T_2: Take a wire from a container; (**c**) T_3: Connect the circuit board and wire and place them on the TV chassis; (**d**) T_4: Drilling circuit boards to the TV chassis.

above shoulder level due to the container's location, as illustrated in Figure 6.4a. The risk factors detected for T_2 corresponded to movements where there is both bending and rotation of the torso and stretching of the arms. These results match T_2, where the operators bend to retrieve a wire from a container and then connect it to the circuit board. Figures 6.4c and 6.4d depict the torso rotations that were identified as risk factors for T_3 and T_4. In these tasks, operators were only required to slightly rotate their torsos due to the constant movement of the TV chassis caused by the conveyor belt, but they did not need to raise their arms highly or strongly bend forward to reach the TV chassis.

Since T_1 and T_2 involve assuming awkward postures such as rotating the torso while bending forward or raising the arms above shoulder level, they represent a greater ergonomic risk than T_3 and T_4. EAWS recommends that tasks with moderate risk be redesigned whenever possible; otherwise, the risk must be controlled through other means. Therefore, it is proposed to delegate T_1 and T_2 to a collaborative robot and leave T_3 and T_4 to the operators.

6.6 Optimization of the work-space scenario

According to the results of the task delegation, the assembling routine was first divided into sub-tasks performed by the robot or operators. 3D Convolutional Neural Networks (3DCNNs), a type of deep Learning architecture, are used for the gesture recognition module that concerns the communication between the operator and the robot. These networks were trained using a dataset consisting of command gestures adapted for TV assembly and recorded with an RGB camera (GoPro Black) in an egocentric view. The recording was made at an 848x480



(**a**) G_1:Start (**b**) G_2: Green card functioning (**c**) G_3:Place green card

(**d**) G_4: Screw green card (**e**) G_5: Screw gold card (**f**) G_6: End

(**g**) G_7: Waiting (**h**) G_8: Gold card functioning (**i**) G_9: Place gold card

(**j**) G_{10}: Green card not functioning (**k**) G_{11}:Gold card not functioning

Figure 6.5: Command gestures of the gesture recognition module.

resolution and 20 frames per second. A group of 14 operators, four females and ten males, were recorded performing six gestures and five postures. Thus, there are 11 classes in total, each corresponding to a unique command for the collaborative robot[1]. These command gestures are illustrated in Figure 6.5.

In the new routine, the operator performs the start gesture (G_1) to notify the robot that the assembly routine starts. Next, the robot approaches the card container, retrieves the initial green card, and hands it to the operator in the defined handover position. By pressing the force sensor[2] on the robot, the operator releases the card and verifies its functionality. If the card is functional, the operator performs G_2 and places it on the TV chassis (G_3), while the robot moves toward the card box to retrieve the gold card. If the green card is not functional, the operator executes G_{10} to notify the robot, which then brings a replacement green card.

[1]UR3 robotic arm from Universal Robots: https://www.universal-robots.com/products/ur3-robot/
[2]Force torque sensor FT-300-S: https://robotiq.com/products/2f85-140-adaptive-robot-gripper

When the robotic arm delivers a functional card, the operator performs G_2 and then places and screws the card on the TV chassis (G_3 and G_4). This procedure is repeated until both the green and gold cards are placed on the TV chassis. Then, the human operator executes G_{11} to signal the routine's completion.

The introduction of the robotic arm and a gesture recognition module requires only minor torso rotations from the human operator. To facilitate natural collaboration and assist the operator in performing only ergonomically safe movements, a posture estimation module is added that enables the robot to spatially adapt to the operator. Papanagiotou et al. [Papanagiotou, 2021] described in full this posture estimation module and spatial adaptation procedure. The spatial adaptation refers to the fact that the robotic arm does not place the cards in a fixed position but rather adapts to the operator's anthropometric characteristics, thereby improving the operator's posture. The installation complied with all applicable safety regulations for collaborative robotics (ISO 10218 and TS 15066).

6.6.1 Evaluation and validation of the proposed HRC framework

6.6.1.1 Experiments and key performance indicators

In order to evaluate the HRC scenario in terms of collaboration and operator performance, 14 operators, who did not participate in the creation of the training dataset, were recorded performing the proposed routine in three separate experiments. The 14 subjects consisted of seven males and seven females, either right- or left-handed, ranging in height from 1.60 to 1.90 meters. Three experiments were conducted to determine whether the proposed HRC framework enhances ergonomics by assessing the adaptation of the robotic arm and the operator's movement during each. In the first experiment, gesture recognition and spatial adaptation were disabled. Hence, operators were required to interrupt their routine and inform the robotic arm of their current action by pressing its force-torque sensor. The gesture recognition module was enabled for the second experiment, but not the spatial adaptation, so the operators received the circuit cards from a predefined handover position. Finally, in the third experiment, gesture recognition and spatial adaptation were enabled and continuously provided information about the operators' actions to the robotic arm.

The percentage of robot spatial adaptation (SA) and reduction in operator's movement ($RiOM$) [Papanagiotou, 2021; Olivas-Padilla, 2023] were used as key performance indicators (KPI). The KPI of robot spatial adaptation represents the ratio of the distance covered by the robot without spatial adaptation to the distance covered when the robot adjusts to the operator-specified position. The following formula is used to determine this KPI:

$$SA(\%) = \frac{\|AHP - WP\| - \|PHP - WP\|}{\|PHP - WP\|} \tag{6.7}$$

where SA is spatial adaptation, AHP the adapted handover position, WP is the waiting point and PHP the particular handover position. Centimeters are used to measure distances. The higher the rate of adaptation, the more effort the operator had to put in during the HRC

scenario without the spatial adaptation of the robot.

RiOM quantifies the difference in operators' movement before and after introducing gesture recognition. This KPI is calculated as follows:

$$RiOM(\%) = \frac{\|MwoGR\| - \|MwGR\|}{\|MwoGR\|} \tag{6.8}$$

MwoGR corresponds to the movement without gesture recognition, and *MwGR* is the movement with gesture recognition. This KPI measures the amount of effort reduced by the operator as a result of gesture recognition.

SA is primarily used to compare the first experiment (physical interaction) with the third experiment (spatial adaptation with pose estimation), whereas *RiOM* is used to compare the first experiment with the second experiment (temporal adaptation with gesture recognition).

6.6.1.2 Results and discussion

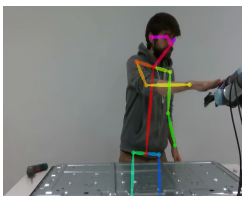
The 3DCNNs were trained with the command gesture dataset for the gesture recognition module and demonstrated 98.50% accuracy in recognizing the 11 gestures. All the operators completed the collaboration procedure successfully, indicating that the accuracy of the recognition algorithm is sufficient even for users that were not part of the training dataset.

The calculated KPIs for each operator are shown in Table 6.3. Note that the greater the percentage of *SA*, the more difficult it was for the operator to receive the cards without spatial adaptation enabled. This is because the predefined handover position was not close to where the operator would prefer to receive the cards. For *RiOM*, however, the larger the percentage, the better, as it indicates a greater reduction in operator's movement when gesture recognition is utilized. The average rate of spatial adaptation and reduction in operator's movement were 29.37% and 28.37%, respectively, among the 14 subjects. Compared to the configuration without gesture recognition, the optimized HRC configuration significantly reduced the operators' movement. This is also demonstrated by seeing Figure 6.6a, where the operator was required to rotate his torso in order to touch the robot's sensor, which is located outside the TV chassis. In Figure 6.6b, the robot recognizes when the operator has completed his task and can proceed to the next action in the work routine.

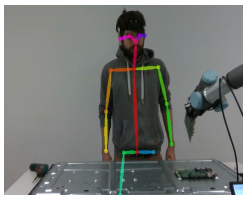
As demonstrated by the KPIs, the proposed HRC scenario enhances ergonomics and efficiency. This is accomplished by first assigning the hazardous tasks from the original scenario to the collaborative robot. Then, integrating gesture recognition and spatial adaptation to prevent operators from performing unnecessary movements that could cause physical discomfort.

Table 6.3: Measured KPIs for each operator.

KPI	Operator													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
SA(%)	39.10	33.30	21.10	27.50	30.40	31.90	27.10	31.80	13.40	33.90	43.50	32.10	18.70	27.40
RiOM(%)	31.40	33.10	24.40	27.10	32.10	27.30	24.50	26.80	37.40	20.60	45.90	20.80	21.30	24.50



(**a**) MwoGR: Press robot to start



(**b**) MwGR: The robot recognizes G_7 to start

Figure 6.6: TV assembly with and without gesture recognition.

6.7 Conclusion of the chapter

This chapter presents a methodology for designing a human-robot collaboration framework that maximizes ergonomics and production efficiency in a television co-production cell. The first step was creating a system for identifying four postural risk factors. According to the risk factors detected, an ergonomic risk score is calculated based on EAWS. This system was trained using the 28 motion primitives from the ERGD dataset. The trained system successfully recognized the four risk factors using only data from a minimal set of sensors, selected in Chapter 5. Following the training of the automatic ergonomic scoring system, real professional tasks performed on a television production line were evaluated for task delegation. Two of the four evaluated tasks were assigned to a collaborative robot. Hence, the suggested HRC framework for this scenario consists of the robot grabbing the circuit board and wire from their respective containers and handing them to the human operator. The human operator then connects the wire to the circuit board, positions the board and wire on the television chassis, then drills the board into the chassis.

This analysis can be applied to other professional tasks for rapid reconfigurability, just as it was done with TV assembly tasks. First, the professional tasks to be evaluated need to be recorded with the designated inertial sensors and placed according to the standards of the ISB. Next, calculate the Euler angles of each body part measured by each sensor. Then, segment the data of the tasks into four-second windows, ideally with a two-second overlap to cover the entire task. Finally, apply the automatic postural evaluation to the segmented tasks, which would indicate the motion primitives detected and the estimated EAWS score. Depending on the nature of the identified high-risk tasks, it is determined whether they should be delegated to a collaborative robot or the production cell should be redesigned to prevent the performance of hazardous movements.

Thus, HRC frameworks can be ergonomically improved by implementing the task delegation process presented in this chapter and the HRC design proposed by Papanagiotou et al. [Papanagiotou, 2021; Olivas-Padilla, 2023]. First, it is identified which risky tasks should be assigned to the collaborative robot. Then, by applying gesture recognition and spatial adaptation, the robot can assist operators in reducing their range of motion so that the operators perform only safe and convenient movements. As a result, less physical effort is required from them to fulfill their professional duties.

Lastly, it is worth noting that wearables that measure working postures have the potential to decrease the incidence of WMSDs. High-frequency and easily-accessible monitoring technology can provide feedback to managers and operators on how to address exposures to ergonomic risks. In this regard, Appendix B describes automatic ergonomic evaluation applications designed to utilize MoCap data for ergonomically analyzing human movements. Future plans for this platform include implementing the proposed automatic ergonomic evaluation in this chapter and the human movement analysis methods presented so that users can apply them to analyze their own recorded movements or learn from those in Chapter 3.

Conclusions

7.1 Summary

This dissertation was primarily focused on creating methodologies for training interpretable human motion models utilizing practical and portable MoCap technologies, such as IMUs. These models could be used to generate accurate human movements and gain insight into the dynamics of human movement during the execution of a movement. The applicability of state-space models for developing a generalized motion understanding framework was investigated. Consequently, three approaches that adhere to the structure of the Gesture Operational Model were proposed.

The fourth chapter presented ideas for combining state-space models and data-driven approaches to train interpretable GOM representations of human movements that enable the simulation of accurate 3D human postures. The proposed methods were able to parameterize the conditional distributions specified in the state-space models. The generated models exhibited their potential to learn human movements in a general and scalable way, as they were able to fit data distributions from reduced data sets and recorded with different subjects in different scenarios. The difference in performance between the three approaches may be influenced by their structure and number of parameters. Additional research is required to get a complete understanding of how these two factors interact. The use of either a statistical or data-driven approach for human motion representation would depend on the nature of the motion-based application. For instance, whether a single or several human movements are examined, as well as the processing power constraints. The proposed approaches can be simply implemented utilizing existing statistical and deep learning libraries. Additionally, they can be upgraded with new advancements in deep learning or incorporated into more sophisticated architectures, such as Generative Adversarial Networks (GANs) [Yoon, 2019], that compute the gradients of any differentiable architecture using automatic differentiation techniques.

As described in Chapter 5, the trained models allowed the body dexterity analysis of industrial operators and skilled craftsmen. This analysis described how body joints collaborate to accomplish specified motion trajectories. With the motion representations, it was also possible to perform a selection of meaningful motion descriptors for modeling a set of human movements. This selection method could be utilized for a broad range of applications requiring the modeling of a specific set of movements using a minimal sensor configuration.

Chapter 6 presented an application of the selection of meaningful motion descriptors while creating a methodology for automatic ergonomic analysis and task delegation in HRC frameworks. Professional tasks were assessed by first recording them with a minimal set of selected sensors. Afterward, an EAWS-based ergonomic score was automatically calculated based on the detected motion patterns. The task delegation then consisted of assigning the tasks with the most dangerous movements to the collaborative robot, while the operators were allocated the ergonomically safe or supervisory control tasks.

The main scientific and technological contributions of this thesis are listed next.

7.1.1 Scientific contributions

Three methods for learning of interpretable human motion models

The methods generate interpretable time-varying models of human movements to study body dexterity and create realistic motion simulations. Human movements are represented using GOM, which consists of a set of autoregressive models, each modeling a different joint motion descriptor (joint angle). The first method estimates the model's parameters using Kalman filters (KF-RGOM) with one-shot training. The second and third methods correspond to deep state-space models. The second method utilizes a stochastic autoencoder (VAE-RGOM) to train multiple interpretable human motion representations, while the third method utilizes an autoencoder with the Luong attention mechanism (ATT-RGOM). KF-RGOM can be utilized for analyses where only small sets of human movements are available, as well as for applications that require only a few motion descriptors to be modeled. VAE-RGOM and ATT-RGOM are proposed for applications that need the analysis of multiple motion descriptors and human movements. Additionally, these methods can be used to augment data in deep learning applications.

Analysis of full-body dexterity in industrial operators and expert artisans

A methodology based on trained motion representations is provided for analyzing the body dexterity of industrial operators and skilled artisans. This comprises a statistical analysis of the trained GOM representations to determine the significance of their assumptions in modeling a specific human movement. The results highlighted the key motion descriptors associated with and contributing to the whole-body movement, providing insights into how body joints collaborate to accomplish the predicted motion trajectories. In addition, identifying the most significant motion descriptors of all human movements associated with a professional task reveals the optimal set of motion descriptors for modeling and recognizing them. Finally, a procedure for computing tolerance intervals based on experts' motion representations is provided to supplement the dexterity analysis. These tolerance intervals, which specify the acceptable range of motion for replicating a specific movement, are calculated using the parameters of multiple motion representations trained with different repetitions of the same movement.

7.1.2 Industrial and technological contributions

Motion capture benchmark of industrial tasks and European historic crafts

A motion capture benchmark featuring datasets containing the full-body movements of real industrial operators and skilled craftsmen was developed. Currently, the most used datasets consist of common daily human movements. Therefore, seven new MoCap datasets of actual professional tasks performed in industry and crafts were created using an inertial based full-body MoCap suit.

Methodology for improving HRC through computational ergonomics

A methodology for task delegation is developed in order to design the optimal HRC framework that maximizes ergonomics and production efficiency in a television co-production cell. Professional tasks done by human operators on the television production line are initially evaluated by recognizing postural risk factors using HMMs. According to the European Assembly Worksheet, an ergonomic risk score is calculated based on the detected risk factors. The optimal HRC configuration is then defined by delegating to the collaborative robot the hazardous tasks. Finally, the HRC scenario is enhanced by applying gesture recognition and spatial adaptation, which allow the human operator to collaborate with the robot using gestures while avoiding unnecessary movements that could cause physical strain.

7.2 Open questions and perspectives

This section concludes the dissertation by discussing the open questions and ideas regarding how the presented work could be expanded. Although the results obtained in this thesis already look promising, there is still room for improvements in the proposed frameworks and follow-up work on their implementation in real-world scenarios.

Human motion representation with GOM

The GOM's representation of human movement and the proposed estimation approaches can be further optimized. One of the advantages of GOM is its ability to easily incorporate new assumptions into its representations of human movement. In this dissertation, the mathematical model of human movements was based on kinematic measures of the body joints. Thus, it remains an open question if the accuracy of motion modeling can be improved by incorporating other types of measures, such as kinetic motion descriptors (joint torques or external forces), into the representation. Diverse applications, particularly in ergonomics, could benefit from the ability to get insight into how kinetic and kinematic measures interact to accomplish a specific movement. Previous studies have calculated joint torques for identifying the joints that accumulate the most strain during a variety of tasks [Menychtas, 2020]. Consequently, GOM representations with kinetic and kinematic measures could be utilized to create novel ergonomic monitoring systems for recognizing potential posture risks. For instance, the sensitivity analysis performed in Section 4.5 demonstrates the potential for using the estimated motion representations

to analyze anomalous motion descriptor behavior. When performing an ergonomic analysis, it may be helpful to examine how the models react to shocks applied to various joint motion descriptors in order to later identify any physical strains (such as on the shoulders or lower back) or loads that may be affecting the workers' performance during their shift. The professional tasks can then be modified to reduce the danger of injury.

Parametrization of state-space models

The selection of the best architecture for parameterizing the GOM representations utilized in this thesis was not a trivial task, and there may be better default settings than those offered in this thesis. For example, the autoregressive order of the GOM models, the number of layers, units, or activation functions. In this dissertation, standard optimization algorithms were applied to determine the optimal settings for each architecture based on the dataset utilized. However, in order to avoid prolonged run times, the search for the best hyperparameter values was restricted to a specific range. As always, when using data-driven approaches, training tricks can make a huge difference in terms of the final performances of a model. Some were used in the training of VAE-RGOM and ATT-RGOM, but it would be interesting to find even more effective ones based on a deeper theoretical understanding of the learning process.

Inertial sensors for human motion capturing

Working with inertial-based MoCap data requires awareness of the limitations of using inertial sensors in real-world work environments. Inertial sensors can offer precise and reliable measurements to study human movement; however, the degree of this precision and reliability depends on the site, movements, and tools handled during the performance. In the recording for datasets GLB and APA, for example, subjects used plastic gloves or did not wear the gloves that come with the inertial suit to prevent measurement disturbances. Therefore, for implementing motion-based applications with inertial sensors, it is necessary to account for the possibility of magnetic disturbances during the recording of new datasets and to apply post-processing techniques to eliminate drifts in the measures that may influence the results of the proposed approaches.

Explainable AI

This thesis is a step toward the development of explainable AI (XAI) for human motion modeling [Hagras, 2018]. Humans can easily comprehend and analyze the actions in XAI. As discussed in Chapter 2, conventional data-driven approaches and other supervised methods, such as linear or logistic regressions, can be difficult to interpret for high-dimensional motion data. These approaches do not permit the interpretation of human movements, nor do they explain the logic behind the trajectory predictions of joint motion descriptors. The work done in this thesis offers a first approach for generating interpretable human motion representations that can be used for dexterity analysis and other applications that require describing different human movements using only kinematic descriptors. Implementing human-computer interfaces that automatically apply statistical analysis to learned motion models and highlight significant motion descriptors

in an intuitive and meaningful way could be the focus of future work. The intention would be to ease the user's implementation of the proposed approaches on their own recorded movements or their understanding of movements contained in the benchmark presented in Chapter 3. Additionally, through the computation of the tolerance intervals outlined in Section 5.5, applications that provide users with real-time feedback about their capability to replicate particular movements could be designed. These interactive applications have the potential to enable novices to learn gross and fine motor skills even in the absence of an instructor by comparing their motion data to that of an expert during the instruction process.

Simulating the movement of multiple individuals

Another potential application of the proposed motion representations can be their implementation in more sophisticated methods that simulate the movements of multiple individuals. Simulating human movements with the proposed motion representations implies fully utilizing the information in the MoCap data to predict future postures. Consequently, these learned representations might serve as the foundation for simulation. For example, to simulate the movement of several individuals, methods often use complex hypothetical decision rules, which frequently fail to produce realistic movements [Patterson, 2008; Rudenko, 2020]. These might be replaced by simulating from the proposed representations to generate the expected spatial distribution of the population.

Appendix A

General simulation results with all seven datasets

Tables A.1, A.2, and A.3 present the average Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Theil Inequality Coefficient (U_1), respectively, achieved with each dataset and approach. All movements were generated with the respective motion representation of their class, then the MAE, RMSE, and U1 were calculated between the generated movement and the original. Tables A.4 to A.8 illustrate the simulation performance based on MAE for each movement within the datasets.

Table A.1: Average MAE for each dataset.

Dataset	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
TVA	16.830 (σ: 15.379)	6.938 (σ: 0.459)	0.093 (σ: 0.016)	0.191 (σ: 0.024)
TVP	7.947 (σ: 5.278)	9.867 (σ: 5.592)	0.213 (σ: 0.058)	0.398 (σ: 0.085)
APA	3.312 (σ: 2.816)	10.946 (σ: 0.664)	0.091 (σ: 0.019)	0.203 (σ: 0.048)
GLB	19.211 (σ: 9.981)	12.916 (σ: 3.665)	0.119 (σ: 0.044)	0.220 (σ: 0.066)
SLW	14.267 (σ: 7.739)	9.207 (σ: 3.307)	0.115 (σ: 0.041)	0.246 (σ: 0.078)
MSC	23.313 (σ: 14.514)	16.002 (σ: 5.887)	0.247 (σ: 0.101)	0.457 (σ: 0.126)
ERGD	13.461 (σ: 8.699)	13.569 (σ: 4.931)	0.095 (σ: 0.052)	0.198 (σ: 0.085)

Table A.2: Average RMSE for each dataset.

Dataset	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
TVA	32.438 (σ: 27.247)	14.903 (σ: 1.574)	0.962 (σ: 0.430)	1.126 (σ: 0.410)
TVP	15.978 (σ: 5.614)	20.937 (σ: 2.391)	3.231 (σ: 1.402)	3.339 (σ: 1.389)
APA	17.814 (σ: 17.652)	19.127 (σ: 10.266)	0.885 (σ: 0.147)	1.034 (σ: 0.146)
GLB	42.918 (σ: 22.873)	29.097 (σ: 10.373)	2.049 (σ: 1.384)	2.204 (σ: 1.388)
SLW	28.787 (σ: 13.616)	22.868 (σ: 10.798)	0.467 (σ: 0.287)	0.721 (σ: 0.328)
MSC	51.455 (σ: 19.791)	36.828 (σ: 22.618)	3.103 (σ: 2.043)	3.311 (σ: 1.980)
ERGD	21.732 (σ: 13.926)	15.126 (σ: 11.006)	1.134 (σ: 0.758)	1.279 (σ: 0.782)

Table A.3: Average U_1 for each dataset.

Dataset	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
TVA	0.427 (σ: 0.368)	0.384 (σ: 0.057)	0.015 (σ: 0.005)	0.023 (σ: 0.005)
TVP	0.195 (σ: 0.068)	0.125 (σ: 0.073)	0.019 (σ: 0.004)	0.025 (σ: 0.003)
APA	0.310 (σ: 0.192)	0.210 (σ: 0.101)	0.009 (σ: 0.003)	0.016 (σ: 0.003)
GLB	0.540 (σ: 0.221)	0.292 (σ: 0.123)	0.026 (σ: 0.015)	0.028 (σ: 0.015)
SLW	0.390 (σ: 0.341)	0.201 (σ: 0.102)	0.043 (σ: 0.016)	0.048 (σ: 0.013)
MSC	0.586 (σ: 0.262)	0.361 (σ: 0.099)	0.024 (σ: 0.010)	0.030 (σ: 0.009)
ERGD	0.394 (σ: 0.281)	0.2742 (σ: 0.056)	0.010 (σ: 0.003)	0.015 (σ: 0.003)

Table A.4: Mean absolute angle errors for TVA and TVP.

Dataset	Motion	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
TVA	TVA_1	25.610	7.226	**0.106**	0.218
	TVA_2	37.600	6.271	**0.103**	0.211
	TVA_3	2.742	6.804	**0.066**	0.160
	TVA_4	1.368	7.450	**0.096**	0.176
TVP	TVP_1	2.050	3.241	**0.130**	0.310
	TVP_2	4.779	5.746	**0.201**	0.309
	TVP_3	5.545	5.164	**0.179**	0.375
	TVP_4	2.267	9.657	**0.125**	0.256
	TVP_5	12.305	13.669	**0.239**	0.446
	TVP_6	15.816	15.192	**0.233**	0.480
	TVP_7	16.452	21.746	**0.301**	0.508
	TVP_8	4.132	6.898	**0.222**	0.480
	TVP_9	8.178	7.491	**0.288**	0.419

Table A.5: Mean absolute angle errors for APA and MSC.

Dataset	Motion	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
APA	APA_1	1.550	10.531	**0.089**	0.186
	APA_2	7.286	10.423	**0.114**	0.269
	APA_3	1.100	11.884	**0.069**	0.155
MSC	MSC_1	2.143	7.880	**0.139**	0.341
	MSC_2	12.283	15.033	**0.156**	0.342
	MSC_3	34.171	28.742	**0.323**	0.557
	MSC_4	24.116	12.710	**0.183**	0.370
	MSC_5	25.538	16.444	**0.204**	0.389
	MSC_6	31.718	10.472	**0.230**	0.458
	MSC_7	2.197	8.836	**0.132**	0.314
	MSC_8	20.070	15.076	**0.273**	0.455
	MSC_9	26.040	16.920	**0.432**	0.677
	MSC_{10}	47.524	20.863	**0.288**	0.489
	MSC_{11}	48.210	23.696	**0.453**	0.732
	MSC_{12}	5.860	20.343	**0.160**	0.343
	MSC_{13}	23.202	11.009	**0.236**	0.472

Table A.6: Mean absolute angle errors for SLW.

Dataset	Motion	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
SLW	SLW_1	8.841	10.820	**0.097**	0.121
	$SLW_{2,1}$	7.411	8.097	**0.069**	0.164
	$SLW_{2,2}$	7.104	8.985	**0.068**	0.139
	$SLW_{2,3}$	10.217	11.100	**0.099**	0.246
	$SLW_{2,4}$	3.097	5.117	**0.067**	0.151
	$SLW_{2,5}$	3.120	5.886	**0.118**	0.303
	SLW_3	10.666	18.936	**0.092**	0.270
	$SLW_{4,1,1}$	15.427	10.960	**0.217**	0.406
	$SLW_{4,1,2}$	13.489	8.516	**0.185**	0.344
	$SLW_{4,1,3}$	30.356	12.912	**0.130**	0.268
	$SLW_{4,2,1}$	15.635	10.108	**0.154**	0.339
	$SLW_{4,2,2}$	25.050	5.884	**0.098**	0.206
	$SLW_{4,2,3}$	18.471	6.064	**0.133**	0.272
	$SLW_{4,3,1}$	14.747	7.718	**0.123**	0.276
	$SLW_{4,3,2}$	18.740	8.947	**0.081**	0.184
	$SLW_{4,3,3}$	25.895	7.264	**0.111**	0.246

Table A.7: Mean absolute angle errors for GLB.

Dataset	Motion	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
GLB	GLB_1	36.340	18.491	**0.128**	0.231
	GLB_2	2.929	10.239	**0.088**	0.170
	GLB_3	21.410	19.363	**0.135**	0.228
	GLB_4	11.502	9.898	**0.073**	0.146
	GLB_5	24.251	11.521	**0.081**	0.155
	GLB_6	17.904	9.500	**0.108**	0.209
	GLB_7	5.213	10.483	**0.092**	0.187
	GLB_8	28.045	16.302	**0.091**	0.177
	GLB_9	21.055	16.196	**0.088**	0.180
	GLB_{10}	25.070	15.138	**0.126**	0.245
	GLB_{11}	32.240	15.081	**0.163**	0.276
	GLB_{12}	7.859	11.642	**0.167**	0.265
	GLB_{13}	18.714	15.717	**0.248**	0.407
	GLB_{14}	33.678	12.708	**0.100**	0.193
	GLB_{15}	9.203	8.919	**0.176**	0.333
	GLB_{16}	7.300	6.998	**0.072**	0.146
	GLB_{17}	15.606	7.684	**0.097**	0.191
	GLB_{18}	27.487	16.608	**0.113**	0.221

Table A.8: Mean absolute angle errors for ERGD.

Dataset	Motion	KF-GOM	KF-RGOM	VAE-RGOM	ATT-RGOM
ERGD	$ERGD_1$	6.516	2.825	**0.020**	0.039
	$ERGD_2$	4.378	7.014	**0.034**	0.077
	$ERGD_3$	2.633	6.187	**0.025**	0.057
	$ERGD_4$	3.332	6.329	**0.027**	0.065
	$ERGD_5$	2.759	8.543	**0.038**	0.084
	$ERGD_6$	4.750	14.041	**0.062**	0.162
	$ERGD_7$	3.657	10.544	**0.048**	0.109
	$ERGD_8$	3.090	12.626	**0.053**	0.124
	$ERGD_9$	3.063	13.229	**0.073**	0.190
	$ERGD_{10}$	18.234	15.835	**0.087**	0.192
	$ERGD_{11}$	22.209	20.697	**0.124**	0.256
	$ERGD_{12}$	6.586	13.920	**0.073**	0.181
	$ERGD_{13}$	4.952	11.037	**0.082**	0.200
	$ERGD_{14}$	16.796	13.568	**0.091**	0.221
	$ERGD_{15}$	17.635	6.105	**0.067**	0.167
	$ERGD_{16}$	8.923	14.033	**0.093**	0.219
	$ERGD_{17}$	10.190	14.444	**0.112**	0.227
	$ERGD_{18}$	18.354	15.565	**0.123**	0.261
	$ERGD_{19}$	22.967	13.936	**0.137**	0.308
	$ERGD_{20}$	18.294	15.073	**0.148**	0.329
	$ERGD_{21}$	15.333	17.560	**0.077**	0.155
	$ERGD_{22}$	16.639	12.833	**0.255**	0.262
	$ERGD_{23}$	15.077	12.473	**0.103**	0.207
	$ERGD_{24}$	23.038	19.642	**0.114**	0.241
	$ERGD_{25}$	27.285	19.341	**0.143**	0.283
	$ERGD_{26}$	29.911	20.268	**0.151**	0.308
	$ERGD_{27}$	23.421	19.252	**0.123**	0.259
	$ERGD_{28}$	26.899	23.024	**0.180**	0.362

Appendix B

Web-based and Android-based applications for automated ergonomic evaluation

B.1 Introduction

Manual laborers in the industry sector are often subject to critical physical strain that leads to work-related musculoskeletal disorders. Lifting, poor posture, and repetitive movements are among the causes of these disorders. In order to prevent them, several rules and methods have been established to identify ergonomic risks that workers might be exposed to during their activities. However, the ergonomic assessment through these methods is not a trivial task, and a relevant degree of theoretical knowledge on the part of the analyst is necessary. Therefore, this appendix presents a web-based and an android-based application for automatic ergonomic evaluation using MoCap data. The proposed applications use segment rotations (or joint angles) acquired from IMUs for the assessment and provide as feedback RULA scores, color visualizations, and limb angles in a simple, intuitive and meaningful way. RULA is one of the most commonly used observational methods for assessing occupational risk factors for upper-extremity musculoskeletal disorders. By automatizing RULA, an interesting perspective for extracting posture analytics for ergonomic assessment is opened, as well as the inclusion of new features that may complement it.

B.2 Automatic ergonomic evaluation module

Both applications use a module that automatically computes RULA scores based on a skeleton constructed using MoCap data. In this first version, the MoCap data is retrieved from BVH files generated either by the Notch Interfaces[1] or Nansense MoCap systems. The module's design is divided into three steps. The first step is to extract the segment rotations per

[1]Notch Interfaces Inc. website: https://wearnotch.com/

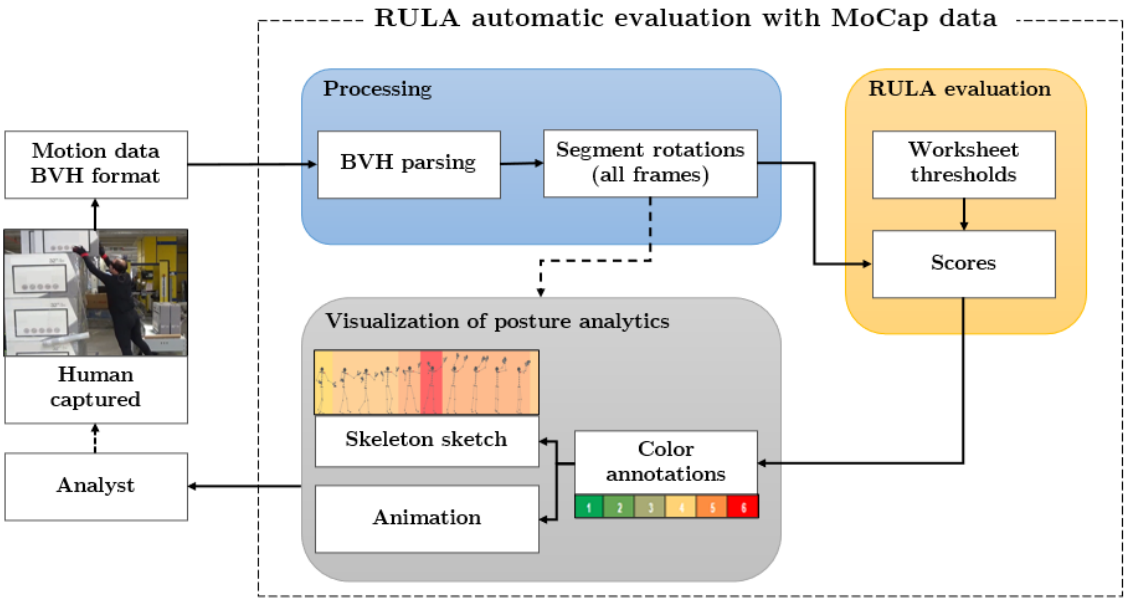


Figure B.1: General scheme of the proposed RULA evaluation module.

frame from the BVH file. Next, the RULA score is computed by applying thresholds to the joint rotations, followed by the generation of color maps depending on the obtained scores. Lastly, visual feedback is produced for ergonomic analysis. The main visual feedback for each application consists of three sections. The first section is composed of colored annotations based on the scores and color maps computed in the previous step. The second section is the Skeleton Sketch, which displays color annotations and skeleton drawings of various frames. The third section is comprised of the animation of the human motion data. Figure B.1 depicts the overall structure of the created module, whose primary components are described in the following subsections.

B.2.1 RULA computation

RULA is used to evaluate workers' risk of developing upper extremity WMSDs. The evaluation considers posture, muscle use, and force applied during a task. According to RULA, the upper human body is divided into eight segments. Those segments are the trunk, the neck, two upper arms, two forearms, and two wrists. A score is assigned to each segment posture, as well as a score for exerted force and muscle activation [McAtamney, 1993]. As an example, Figure B.2 shows the thresholds defined in RULA for scoring the upper arm posture. The scores for the upper arms (S_{UPA}), neck (S_N), and trunk (S_T) can be from 1 to 6, the lower arms (S_{LA}) and wrists position (S_{WP}) from 1 to 4, and the legs (S_L) and wrist twist (S_{WT}) from 1 to 2. RULA has two other scores, Force score (S_F) and Muscle use score (S_M), where it considers external forces that the human might be exposed to and if the work posture is sustained for a long period or intermittently. After calculating all previous scores, the final RULA risk score (S_{RULA}) is computed from RULA's Table C using the scores defined as score A (S_A) and score B (S_B) [McAtamney, 1993]. SA and SB can vary from 1 to 13. The S_A is

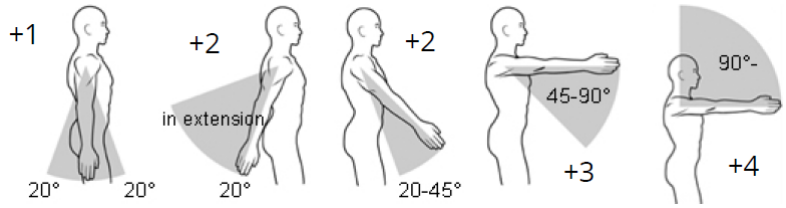


Figure B.2: RULA scoring for the upper arm posture.

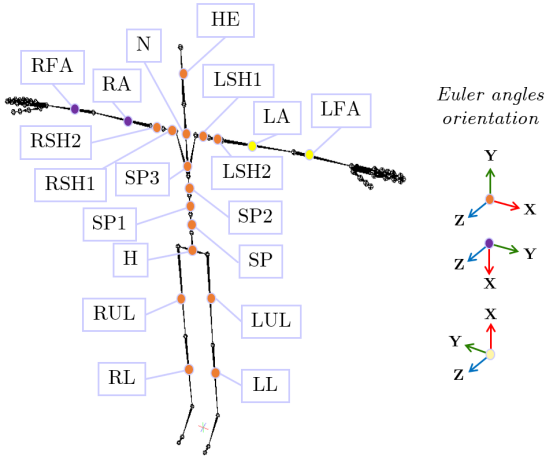


Figure B.3: Location and Euler orientation of the joint angles provided by the Nansense system.

obtained by the RULA's Table A according to the scores S_{UPA}, S_{LA}, S_{WP}, S_{WT}, S_F, and S_M. S_B is calculated from Table B using S_N, S_T, S_L, S_F, and S_M [McAtamney, 1993]. S_{RULA} can vary from 1 to 7. The highest score indicates a severe ergonomic risk, implying that the work posture must be changed immediately, and the lowest score a low risk, meaning that the work posture is acceptable and no change is needed.

For automatically computing the RULA scores using MoCap data, joint angles sequences are extracted from BVH files. This initial version of the module utilizes BVH files generated from data recorded by IMUs. The BVH file format is split into two sections. The first section describes the skeleton's hierarchy and initial posture. This section also lists the degrees of freedom and Euler orientation for each body part. The second section describes the channel data for each frame, which corresponds to the local joint angle sequences. For illustration purposes, the computation of the RULA scores using MoCap data recorded from the Nansense system is described next. Figure B.3 illustrates the joints measured and their Euler orientation in BVH files generated by the Nansense system.

Next is explained the RULA score calculation that is applied to each motion data frame, with each frame representing a posture. Note that the ergonomist typically chooses one arm posture (left or right) in order to calculate the overall RULA score. Nonetheless, the module provides both scores, one based on the posture of the left arm and the other on the right arm. Also, due to the fact that RULA does not provide predefined thresholds for some scores, such as the wrist twist score, where the analyst must subjectively determine if the worker's wrist is twisted or not, additional thresholds were introduced to achieve proper RULA scoring. These

are detailed next, which consists of two score adjustments of the scores: S_{UPA}, S_N, and S_T, and the threshold used to determine if the lower arms are twisted. The remaining thresholds are the same as the predefined by RULA [McAtamney, 1993].

S_{UPA} is evaluated according to the rotation on the X-axis of the shoulders' joints. In order to determine if the shoulder is raised, the rotation of the collars' joints on the Z-axis (Z_{SH1}) are examined. Since RULA does not define any angle threshold to determine if a shoulder is raised, the thresholds specified in Equation B.1 are used. Through these thresholds, S_{UPA} is modified, generating its adjusted version S'_{UPA}.

$$S'_{UPA} = \begin{cases} S_{UPA} + 1 & \text{if } Z_{SH1} \geq 10^\circ \\ S_{UPA} & \text{if } Z_{SH1} < 10^\circ \end{cases} \tag{B.1}$$

The upper arm is indicated as abducted according to Equation B.2. The angle on the X-axis of the shoulder joint is defined as X_{SH2}, the angle on the X-axis of the corresponding elbow (LFA or RFA) is X_{FA}, and the final adjusted score is S''_{UPA}.

$$S''_{UPA} = \begin{cases} S'_{UPA} + 1 & \text{if } X_{SH2} \geq 90^\circ \ \& \ X_{FA} \geq 5^\circ \\ S'_{UPA} & \text{otherwise} \end{cases} \tag{B.2}$$

The angle on the X-axis of the elbow is used to calculate S_{LA}, and the angle on the Z-axis of the same joint to identify whether the arm is moving across the body's midline or outside. For S_{WP}, the angles on the X and Z axes of the respective wrist joint are utilized to determine if the wrist is bent in a way that crosses the midline. S_{WT} is calculated using the Y-axis angle of the corresponding elbow (Y_{FA}). Since RULA does not establish thresholds for this situation, this score is defined as follows:

$$S_{WT} = \begin{cases} 1 & \text{if } -45^\circ < Y_{FA} < 45^\circ \\ 2 & \text{otherwise} \end{cases} \tag{B.3}$$

The neck flexion/extension, which corresponds to S_N, is assessed by using the angle on the X-axis of the neck joint. To determine if the neck is twisted, the angle on the Y-axis (Y_N) is used, where S_N is adjusted (S'_N) according to Equation B.4.

$$S'_N = \begin{cases} S_N + 1 & \text{if } Y_N \geq 20^\circ || Y_N \leq -20^\circ \\ S_N & \text{otherwise} \end{cases} \tag{B.4}$$

The angle on the Z-axis of the neck joint (Z_N) is utilized to establish whether or not the neck is flexed to a side; the threshold defined is the following:

$$S''_N = \begin{cases} S'_N + 1 & \text{if } Z_N \geq 20^\circ || Z_N \leq -20^\circ \\ S'_N & \text{otherwise} \end{cases} \tag{B.5}$$

where S''_N is the final RULA score of the neck region. S_T is computed by analyzing the angle of the middle spine (SP2). The angle on the X-axis is used for measuring the bending, the Y-axis

(Y_{SP2}) to determine if the trunk is twisted, and the Z-axis (Z_{SP2}) if there is a side bending. Adjustments to S_T for a twisted trunk are defined by Equation B.6 and for side bending by Equation B.7.

$$S'_T = \begin{cases} S_T + 1 & \text{if } Y_{SP2} \geq 20^\circ || Y_{SP2} \leq -20^\circ \\ S_T & \text{otherwise} \end{cases} \tag{B.6}$$

$$S''_T = \begin{cases} S'_T + 1 & \text{if } Z_{SP2} \geq 20^\circ || Z_{SP2} \leq -20^\circ \\ S'_T & \text{otherwise} \end{cases} \tag{B.7}$$

where S'_T represents the first score adjustment and S''_T is the final score for the trunk region. The final adjusted versions of S''_{UPA}, S''_N, and S''_T are the scores used for these body regions in the visual feedback and final RULA score calculation.

B.2.2 Visual feedback for posture analytics

B.2.2.1 Color mapping of RULA scores

Following the computation of all scores, a color mapping is performed. First, a color map is created for use in the mapping process. The number of possible values for a score is specified and denoted as n. For example, the score for the upper arm can vary from 1 to 6, so $n = 6$ for this score. The color maps are obtained as follows:

$$C = [ch_1, ch_2, ch_3] \tag{B.8}$$

$$M_n = [C_1, C_2, .., C_n] \tag{B.9}$$

ch is between $[0, 255]$ and M_n is an n-dimensional vector of RGB colors. The lower indexes in M_n are represented by green tones, the intermediate indexes by yellow tones, and the higher indexes by red tones. Next, the color mapping is done by using the following equation:

$$C_R = M_n[S_R] \tag{B.10}$$

In Equation B.10, S_R is a score of the RULA evaluation (S''_{UPA}, S_{LA}, S_{WP}, S_{WT}, S''_N, S''_T, S_L, S_B, S_A, or S_{RULA}), and C_R is the color corresponding to S_R.

B.2.2.2 Graphical user interface

The main interfaces of the web-based and Android applications are shown in Figures B.4 and B.5, respectively. The applications were created using HTML, CSS, JavaScript, PHP, and Java. The Android app is compatible with operating systems 11 and up. The interface comprises four parts: the menu, the skeleton sketch, the human animation, and the score list. In the menu, it is possible to select the scores to display, configure the settings for score computation, and manipulate the skeleton sketch's display. After the MoCap data (BVH file) has been uploaded, the evaluation is performed by computing the RULA scores and showing them in the Score List area. As seen in Figure B.4, various RULA scores may be added and

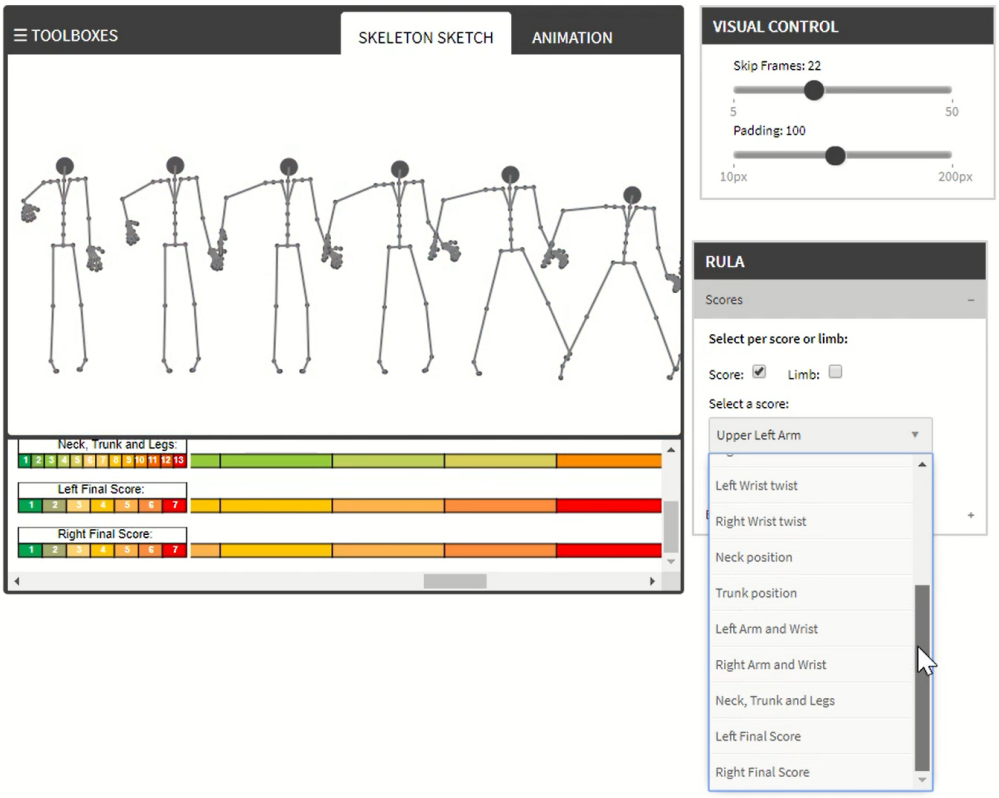


Figure B.4: User interface of the web-based application.

displayed in the scores list area. A score can be selected by either selecting it on the score dropdown list or selecting the joint of the body region that is wished to assess.

There is a settings menu for selecting how scores should be computed. For instance, if external forces are present or indicate the movement's repetition. In the External Factors section of the web-based application, illustrated in Figure B.6a, the default values set for the computing of the Legs, Muscle Use, and Force scores can be modified. First, it must be indicated if the human has any support on the legs and feet while doing the movement under analysis. Then, for the computation of the Muscle use and Force scores, it is necessary to indicate if the work posture is static, repeated in certain periods, or intermittent. If it is a static posture, the module will request the duration the human spends in that posture. If the posture is intermittent or repeated over a period of time, the module will request the number of repetitions that the human performs in one minute. In addition, the load the human is subjected to during the movement can also be specified. The effect of these manually set parameters on the overall RULA score is indicated in [McAtamney, 1993]. These parameters can be similarly modified in the Android application in the section RULA computation, illustrated in Figure B.6b.

The skeleton sketch and animation are organized in tabs. For these visualizations, the skeleton posture of each frame was obtained by using the segmented and initial posture offsets provided by the BVH file. The skeleton sketch illustrates the worker's posture on different frames according to the parameters set in the visual control section. The postures shown on

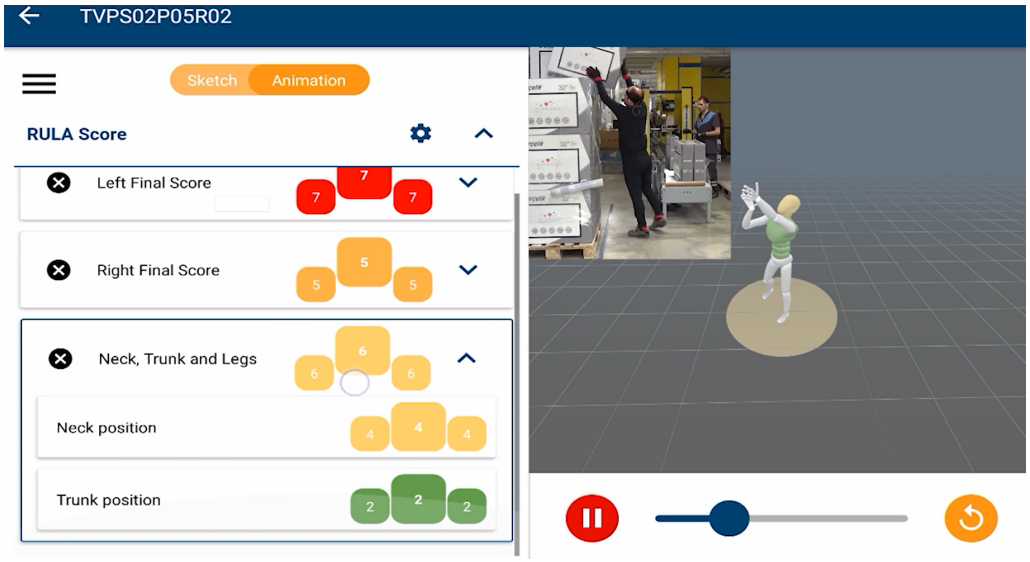
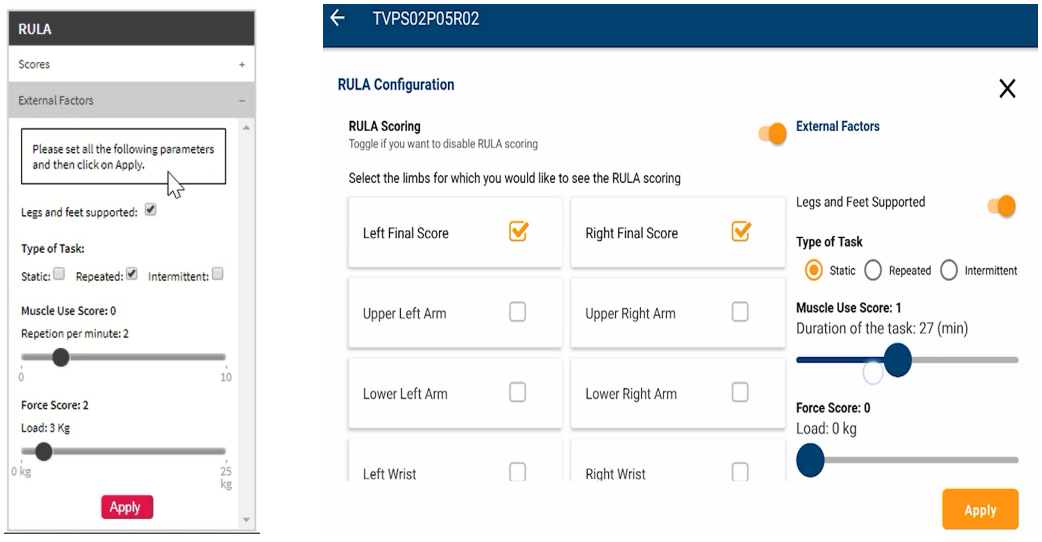


Figure B.5: User interface of the android application.



(a) (b)

Figure B.6: Settings menu for adjusting manual parameters. (a) Web-based application; (b) Android application.

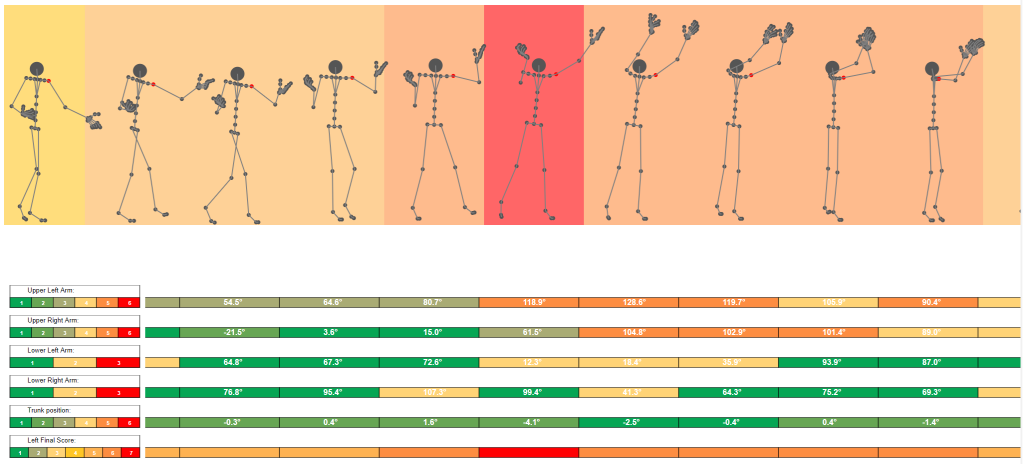


Figure B.7: Skeleton sketch with color-coded scores.

the skeleton sketch match the color-coded scores, which are aligned with the timeline of the recording. The animation displays the video of the uploaded motion data as well as the buttons for pausing and playing it.

When the cursor is positioned over a score annotation in the score list, its colors are placed on the background of the skeleton sketch with the end to better visualize the scoring for each selected frame. In addition, as shown in Figure B.7, the most critical joint for this score annotation is emphasized in red. When the cursor is put over the Upper Left Arm annotation, for instance, the left shoulder joint is highlighted because the angle ranges for this score assignment are with respect to this joint. In addition to the score and colors, the angles assessed during the RULA evaluation are also presented on the score annotation (note the angles displayed in the color bars in Figure B.7). Figure B.7 illustrates the increase in ergonomic risk as the task progresses, as indicated by the Left Final RULA score. This visualization demonstrates that when the arm is raised, the ergonomic risk of the posture increases. This event is clearly denoted by the transition from the color yellow to red and the increase in angle depicted on its color annotation (Upper Left Arm).

B.3 Conclusion and future work

The developed applications are intended to be a useful tool for analysts, facilitating the evaluation of workers' exposure to ergonomic risk factors related to WMSDs[2]. The proposed solution for automatic ergonomic evaluation utilizing MoCap data offers a number of advantages and presents interesting perspectives for ergonomists, factory production directors, workers, and anyone else interested in movement analysis. These applications could permit recording and storing analysis results from different executions of the same movement. From the ergonomic perspective, the applications allow comparing results and monitoring workers' performance to detect any progress or regress. If progress can be observed in this intrapersonal performance study, the analyst may attempt to uncover aspects that bring the worker to

[2]Note to reviewers: The applications have not yet been released as they are still under testing.

improvement.

In order to enrich the application, its compatibility with motion data recorded with other acquisition systems could be implemented. This by taking into consideration the diverse issues that might be faced (occlusions, noise, inaccurate data, etc.). In addition to the ergonomic evaluation, other human movement analyses, such as the dexterity analysis presented in Chapter 5, could be implemented in the applications. The motion representations could be learned using cloud services connected to the application.

List of publications

Journal papers

Brenda Elizabeth Olivas-Padilla, Sotiris Manitsaris, Dimitrios Menychtas, and Alina Glushkova. "*Stochastic-biomechanic modeling and recognition of human movement primitives, in industry, using wearables*". Sensors, 2021.
DOI: https://doi.org/10.3390/s21072497

Brenda Elizabeth Olivas-Padilla, Sotiris Manitsaris, and Alina Glushkova. "*Motion Capture Benchmark of Real Industrial Tasks and Traditional Crafts for Human Movement Analysis*". IEEE Access, 2023.
DOI: https://doi.org/10.1109/ACCESS.2023.3269581

Articles in progress

Brenda Elizabeth Olivas-Padilla, Dimitris Papanagiotou, Gavriela Senteri, Sotiris Manitsaris, and Alina Glushkova. "*Improving Human-Robot Collaboration in TV assembly through computational ergonomics: effective task delegation and robot adaptation*". The IEEE International Conference on Systems, Man, and Cybernetics (SMC), Hawaii, USA, 2023 - Under review.

Brenda Elizabeth Olivas-Padilla and Sotiris Manitsaris. "*Deep state-space modeling for explainable representation, analysis, and generation of professional human movements*". Pending.

International conferences

Agnès Aubert, **Brenda Elizabeth Olivas-Padilla**, Vasileios Syrris, and Sotiris Manitsaris. "*Deep learning architectures applied for recognizing human motion primitives from the Ergonomic Assessment Worksheet*". ICRA workshop - Unlocking the potential of human-robot collaboration for industrial applications, Xi'an, China, 2021.

Brenda Elizabeth Olivas-Padilla, Dimitrios Menychtas, Alina Glushkova, and Sotiris Manitsaris. "*Hidden Markov modelling and recognition of Euler-based motion patterns for automatically detecting risks factors from the European assembly worksheet*". The 27th IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020.
DOI: https://doi.org/10.1109/ICIP40778.2020.9190756

Brenda Elizabeth Olivas-Padilla, Alina Glushkova, Dimitrios Menychtas, and Sotiris Manitsaris. "*Designing a web-based Automatic Ergonomic Assessment using Motion Data*". Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA), Rhodes, Greece, 2019.
DOI: https://doi.org/10.1145/3316782.3322758

Brenda Elizabeth Olivas-Padilla, Alina Glushkova, and Sotiris Manitsaris. "*Motion analysis for identification of overused body segments: the packaging task in industry 4.0*". 17th IFIP TC 13 International Conference of Human-Computer Interaction (INTERACT), Paphos, Cyprus, 2019.
DOI: https://doi.org/10.18573/book3.as

Résumé en français

Chapitre 1

L'introduction décrit le contexte de cette thèse en donnant un aperçu général de la recherche sur l'analyse du mouvement humain ainsi que de ses défis actuels. Les hypothèses et objectifs formulés sont présentés, ainsi qu'un résumé des contributions et la structure de la thèse. Cette thèse se concentre sur la modélisation analytique de la dynamique du mouvement humain. Elle explore l'estimation des paramètres du mouvement dans le but de développer une méthode généralisée de compréhension du mouvement. L'analyse est effectuée sur des patrons de mouvement globaux (corps entier) plutôt que seulement sur des patrons locaux tels que les gestes de la main ou les expressions faciales. Pour une simulation réaliste du mouvement humain, le modèle de transition approprié a été recherché, qui devrait également décrire le phénomène ou les liens entre les suppositions spatiales et temporelles. Les suppositions spatiales doivent tenir compte des interdépendances potentielles entre les articulations de la structure squelettique articulée. D'autre part, les suppositions temporelles impliquent le principe fondamental que la plupart des séries temporelles, telles que les mouvements humains, présentent intrinsèquement, qui est la dépendance entre les observations adjacentes. Par conséquent, le modèle proposé doit être une représentation suffisamment précise du système dynamique qu'est le mouvement humain pour atteindre les objectifs précédents (simulation précise du mouvement humain et compréhension de la réalisation du mouvement). Les deux hypothèses suivantes ont été formulées pour guider cette recherche :

Hypothèse 1 La dynamique du mouvement humain peut être modélisée de manière analytique en tenant compte de la stochastique du mouvement et de la structure physique du corps humain.

Hypothèse 2 L'association des articulations du corps et leur contribution pendant l'exécution d'un mouvement humain peuvent être apprises et représentées par des modèles interprétables.

Pour prouver chacune des hypothèses formulées, des objectifs spécifiques ont été définis. Pour la première hypothèse :

1. **Étudier les approches statistiques et d'apprentissage profond pour paramétrer les représentations mathématiques du mouvement humain :** Le modèle analytique

qui serait entraîné avec des mouvements du corps entier est d'abord désigné. Il devrait simplifier le système musculo-squelettique pour une interprétation facile, tout en incorporant des suppositions pertinentes concernant la dynamique du corps humain. Ensuite, des approches statistiques et d'apprentissage profond seront explorées pour capturer les patrons de mouvement du corps entier et estimer les paramètres du modèle.

2. **Évaluer la capacité de la méthode proposée à modéliser et à simuler des mouvements humains spécifiques de manière statique et dynamique :** La capacité du modèle analytique à simuler les mouvements humains de manière statique ou dynamique est évaluée par une série de tests. La simulation statique implique que toutes les entrées du modèle sont des données réelles, alors que, dans la simulation dynamique, l'entrée correspond à des prédictions antérieures.

Deuxième hypothèse :

1. **Estimer la signification statistique des associations dynamiques à partir des représentations de mouvement générées :** Les modèles de mouvement humain entraînés sont soumis à une analyse statistique pour découvrir les associations significatives entre les descripteurs de mouvement des articulations (dynamique spatiale) et leur dépendance aux transitions précédentes (dynamique temporelle). Ensuite, à partir des suppositions significatives, il est déduit comment les articulations du corps collaborent pour réaliser un mouvement spécifique.

2. **Développer une méthode pour sélectionner les meilleures articulations à mesurer pour l'analyse discriminante et la reconnaissance d'un groupe spécifique de mouvements humains :** Une méthodologie est développée à partir des modèles de mouvements humains entraînés afin de déterminer le set optimal de capteurs pour la reconnaissance précise d'un ensemble de mouvements humains. L'objectif est de valider la possibilité d'identifier un set minimal de capteurs qui pourrait être plus pratique pour les applications quotidiennes liées au mouvement.

Enfin, les contributions de cette thèse peuvent être résumées de la manière suivante :

Modèles interprétatifs pour l'analyse et la simulation du mouvement humain

Cette thèse propose de nouvelles méthodes pour créer des représentations explicables du mouvement humain. Les méthodes actuelles d'apprentissage profond ont été entraînées efficacement pour produire des simulations réalistes de mouvements humains. Cependant, il y a un manque de méthodes qui peuvent aussi bien expliquer le raisonnement derrière leurs prédictions d'une manière qu'un humain peut interpréter. Par conséquent, trois approches ont été développées, qui suivent une représentation d'espace d'état et intègrent des suppositions sur la stochasticité du mouvement humain et les médiations des articulations du corps. Pour la paramétrisation des modèles d'espace d'état, la première méthode utilise l'estimation du maximum de vraisemblance en utilisant des filtres de Kalman. Les deux autres utilisent des approches d'apprentissage profond avec des architectures d'encodeur-décodeur. L'une encode les séries temporelles dans un espace latent

avec une forme de distribution gaussienne pour la prédiction probabiliste, et l'autre utilise un mécanisme d'attention pour capturer la dynamique de l'état. D'autres contributions sont présentées et discutées au chapitre 4.

Analyse de la dextérité des opérateurs industriels et des artisans spécialisés Cette thèse analyse les mouvements humains effectués dans les secteurs industriels et artisanaux en interprétant les représentations de mouvement proposées. La méthode consiste à examiner les paramètres appris des suppositions temporelles et spatiales incorporées dans les représentations de mouvement afin de mieux comprendre comment les experts exécutent un mouvement. En plus, une méthode pour identifier les descripteurs de mouvement articulaire les plus significatifs pour la modélisation et la reconnaissance d'un ensemble de mouvements humains est proposée. Ces informations peuvent ensuite être utilisées pour déterminer la configuration idéale des capteurs pour les problèmes de reconnaissance des mouvements humains. Le chapitre 5 présente et discute d'autres contributions, telles que la création d'intervalles de tolérance.

Ergonomie computationnelle pour la délégation de tâches
La prévalence élevée des troubles musculo-squelettiques (TMS) liés au travail pourrait être atténuée en optimisant les structures de collaboration homme-robot (HRC en anglais) pour les applications de la manufacture. Lors de la conception de ces structures HRC , la délégation des tâches doit être optimisée et les facteurs ergonomiques doivent être pris en compte pour améliorer le confort des opérateurs et l'efficacité de la production. Les ergonomes ont créé de nombreuses méthodes pour évaluer les tâches liées au travail. Cependant, comme ces méthodes se basent sur la perception et l'expérience de l'ergonome, la mesure est subjective et présente une forte inter-variabilité. Par conséquent, cette thèse présente une méthodologie pour une délégation efficace des tâches pendant l'intégration des systèmes de collaboration homme-robot dans une cellule de manufacture. La délégation de tâches est basée sur l'analyse ergonomique automatique des tâches professionnelles, où les scores ergonomiques des tâches sont estimés en fonction des facteurs de risque détectés. Afin de pouvoir être intégré dans des applications industrielles réelles, l'algorithme proposé peut calculer avec précision les scores ergonomiques des mouvements humains en utilisant un set minimal de capteurs. D'autres contributions sont présentées et discutées dans le chapitre 6 et l'annexe B.

Bases de données de tâches industrielles et de métiers historiques européens Cette thèse présente un motion capture benchmark composé de sept bases de données. Ceux-ci comprennent des mouvements exécutés par des opérateurs industriels réels et des artisans qualifiés. Selon les bases de données trouvées, la plupart sont composées de mouvements exécutés lors d'activités quotidiennes, de sports ou de danses et sont capturés dans un laboratoire. Par conséquent, ces sept bases de données sont parmi les seules qui contiennent des enregistrements de mouvements professionnels capturés sur des lieux de travail réels. Le chapitre 3 contient plus d'informations sur le développement du benchmark.

Chapitre 2

Le chapitre 2 présente le contexte nécessaire pour le reste de cette thèse. Tout d'abord, les recherches en modélisation du mouvement humain sont abordées, en commençant par les technologies de capture de mouvement et les descripteurs de mouvement. Ensuite, un aperçu des différentes méthodologies de modélisation du mouvement humain et de leurs principales applications est présenté. Ces méthodes peuvent être divisées en quatre catégories : modélisation biomécanique, modélisation stochastique, modèles hybrides stochastiques-biomécaniques et approches d'apprentissage profond pour la modélisation du mouvement humain. Les modèles stochastiques-biomécaniques ont prouvé leur capacité à simuler les mouvements humains et à produire des représentations mathématiques interprétables. Cependant, ils n'ont pas la robustesse et l'évolutivité des approches d'apprentissage profond pour les applications nécessitant l'analyse et la modélisation de plusieurs mouvements humains. Pour presque tous les problèmes de modélisation des mouvements humains, les approches d'apprentissage profond ont supplanté les méthodes traditionnelles telles que les méthodes biomécaniques ou stochastiques, où la sélection des caractéristiques est cruciale. Néanmoins, les travaux sur les approches d'apprentissage profond qui simulent avec précision les mouvements humains et sont explicables en termes de modèles et de résultats sont encore peu nombreux.

Chapitre 3

Le chapitre 3 présente les bases de données utilisées dans les expérimentations rapportées dans cette thèse. Ces bases de données ont été collectées en utilisant des capteurs inertiels et sont composées de mouvements effectués par des opérateurs industriels et des artisans spécialisés. Les mouvements professionnels ont été collectés dans l'intention d'être utilisés pour la recherche sur l'analyse et la modélisation du mouvement humain. Les procédures d'enregistrement et de traitement sont décrites en détail, ainsi que les mouvements capturés.

Chapitre 4

Le chapitre 4 décrit trois nouvelles approches de modélisation des mouvements humains à travers des représentations mathématiques interprétables. La première approche utilise la modélisation statistique (KF-RGOM) pour estimer les paramètres de mouvement d'un système d'espace d'état, tandis que les deuxième et troisième approches utilisent l'apprentissage profond (VAE-RGOM et ATT-RGOM). Les représentations du mouvement suivent la structure hybride stochastique-biomécanique de GOM pour modéliser la dynamique des mouvements humains. Les trois approches estiment des représentations de mouvements qui varient dans le temps et qui peuvent être utilisées pour simuler des mouvements humains de manière statique ou dynamique. De plus, les représentations de mouvement peuvent être utilisées pour obtenir des informations sur la relation dynamique entre les articulations du corps pendant l'exécution d'un mouvement. Les expérimentations ont prouvé l'hypothèse 1 et révélé que l'utilisation de

représentations qui varient dans le temps augmente la robustesse des modèles pour simuler avec précision divers mouvements humains. En outre, selon une analyse de sensibilité des modèles générés, les modèles ont montré une tolérance aux perturbations externes.

Chapitre 5

Le chapitre 5 détaille l'application des représentations de mouvement estimées au chapitre 4 pour l'analyse de la dextérité afin de démontrer l'hypothèse 2. Les modèles sont analysés statistiquement, et les résultats sont utilisés pour évaluer l'importance des suppositions de chaque modèle concernant les associations de parties du corps spécifiées dans la représentation du mouvement. Le calcul de la significativité statistique des descripteurs de mouvement articulaire a permis d'identifier les plus significatifs pour le mouvement humain modélisé. Un groupe de capteurs a été sélectionné en utilisant la représentation du mouvement estimée par chaque approche (KF-GOM, KF-RGOM, VAE-RGOM et ATT-RGOM). Les capteurs sélectionnés ont été validés en fonction de leur capacité à améliorer les performances de reconnaissance des vocabulaires de gestes extraits de sept bases de données distinctes. La performance de reconnaissance utilisant les données des capteurs sélectionnés a été comparée à celle obtenue en utilisant les données de tous les capteurs et les données d'une configuration minimale de deux capteurs. Les représentations du mouvement ont prouvé leur capacité à capturer et à décrire les mouvements humains en fonction de leurs suppositions. De plus, les performances de reconnaissance des capteurs sélectionnés par chaque approche ont dépassé ou égalé celles de tous les capteurs. Ainsi, il a été possible de déterminer les descripteurs de mouvement qui résolvaient le plus efficacement chaque problème de reconnaissance.

Chapitre 6

Le chapitre 6 propose une méthodologie de délégation de tâches pour concevoir des structures HRC qui améliorent l'ergonomie dans les applications de manufacture. L'hypothèse formulée est que les mouvements des opérateurs peuvent être correctement évalués en utilisant les données de mouvement capturées avec un minimum de capteurs, permettant une analyse ergonomique plus approfondie de leurs activités et facilitant la délégation des tâches lors de la réalisation de structures HRC . Tout d'abord, un système de détection de quatre facteurs de risque postural a été créé. Le système d'évaluation automatique de la posture est composé de modèles de Markov cachés qui ont appris à reconnaître les patrons de mouvement causés par l'exposition aux facteurs de risque. A partir des facteurs de risque détectés, un pointage du risque ergonomique est calculé selon l'EAWS. La méthodologie est évaluée en examinant les tâches professionnelles effectuées dans un processus de fabrication de téléviseurs. Pour identifier les tâches les plus dangereuses, des scores ergonomiques ont été calculés sur la base des facteurs de risque détectés dans chaque tâche. Un HRC optimisé a ensuite été proposé dans lequel les tâches potentiellement dangereuses ont été déléguées à un robot collaboratif.

Conclusion et perspectives

Finalement, le dernier chapitre présente un retour global sur le travail effectué dans cette thèse et suggère quelques directions de recherche futures. Bien que les résultats obtenus dans cette thèse soient déjà encourageants, il est encore possible d'améliorer les modèles proposés et de poursuivre le travail sur leur implémentation dans des scénarios réels. La représentation du mouvement humain par GOM et les approches d'estimation proposées peuvent être optimisées davantage. Par ailleurs, la création d'interfaces homme-machine qui appliquent automatiquement une analyse statistique aux modèles de mouvement entraînés et qui soulignent les descripteurs de mouvement significatifs d'une manière intuitive et utile pourrait être au centre des travaux futurs. L'objectif serait de faciliter aux utilisateurs l'application des approches proposées sur leurs propres mouvements enregistrés, ou leur compréhension des mouvements inclus dans les bases de données présentées.

Appendix A

L'annexe A présente les applications web et Android développées pour l'évaluation ergonomique automatisée à partir des données MoCap. Les applications proposées évaluent les rotations des segments du corps captées par les IMUs et fournissent un retour simple, intuitif et significatif sous la forme de scores ergonomiques, d'annotations en couleur et d'angles des membres. Les scores sont basés sur l'approche RULA, l'une des méthodes d'observation les plus utilisées pour mesurer les facteurs de risque liés aux troubles musculo-squelettiques des membres supérieurs. L'automatisation de RULA ouvre une perspective intéressante pour l'extraction de l'analytique de la posture pour l'évaluation ergonomique et l'incorporation de caractéristiques complémentaires. Les travaux futurs consistent à implémenter d'autres analyses du mouvement humain, comme l'analyse de la dextérité décrite au chapitre 5, et à rendre les applications compatibles avec les données des systèmes optique de capture du mouvement.

Bibliography

[Agarwal, 2004] Ankur Agarwal and Bill Triggs. "Tracking articulated motion using a mixture of autoregressive models". *Computer Vision - ECCV 2004*. 2004, pp. 54–65 (cit. on p. 22).

[Ahlrichs, 2016] Claas Ahlrichs, Albert Samà, Michael Lawo, Joan Cabestany, Daniel Rodríguez-Martín, Carlos Pérez-López, et al. "Detecting freezing of gait with a tri-axial accelerometer in Parkinson's disease patients". *Medical & Biological Engineering & Computing* 54.1 (2016), pp. 223–233 (cit. on p. 19).

[Aicher, 2020] Christopher Aicher, Nicholas J. Foti, and Emily B. Fox. "Adaptively Truncating Backpropagation Through Time to Control Gradient Bias". *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 115. Proceedings of Machine Learning Research. PMLR, 2020, pp. 799–808 (cit. on p. 34).

[Alatise, 2017] Mary B. Alatise and Gerhard P. Hancke. "Pose estimation of a mobile robot based on fusion of IMU data and vision data using an extended kalman filter". *Sensors (Switzerland)* 17.10 (2017), pp. 1–22 (cit. on p. 15).

[Aliakbarian, 2021] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. "Contextually Plausible and Diverse 3D Human Motion Prediction". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 11313–11322. arXiv: 1912.08521 (cit. on p. 44).

[Alo, 2020] Uzoma Rita Alo, Henry Friday Nweke, Ying Wah Teh, and Ghulam Murtaza. "Smartphone Motion Sensor-Based Complex Human Activity Identification Using Deep Stacked Autoencoder Algorithm for Enhanced Smart Healthcare System". *Sensors* 20.21 (2020), p. 6300 (cit. on p. 20).

[Álvarez, 2016] Diego Álvarez, Juan C. Alvarez, Rafael C. González, and Antonio M. López. "Upper limb joint angle measurement in occupational health". *Computer Methods in Biomechanics and Biomedical Engineering* 19.2 (2016), pp. 159–170 (cit. on p. 17).

[Bahdanau, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". *n 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 2014, pp. 1–15. arXiv: 1409.0473 (cit. on p. 40).

[Baldi, 2021] Pierre Baldi. "Autoencoders". *Deep Learning in Science*. Cambridge University Press, 2021, pp. 71–98 (cit. on p. 38).

[Bancroft, 2011] Jared B. Bancroft and Gérard Lachapelle. "Data fusion algorithms for multiple inertial measurement units". *Sensors* 11.7 (2011), pp. 6771–6798 (cit. on p. 15).

[Bank, 2020] Dor Bank, Noam Koenigstein, and Raja Giryes. "Autoencoders". *ArXiv* abs/2003.05991 (2020) (cit. on p. 38).

[Barbič, 2004] Jernej Barbič, Alla Safonova, Jia-Yu Pan, Christos Faloutsos, Jessica K. Hodgins, and Nancy S. Pollard. "Segmenting Motion Capture Data into Distinct Behaviors". *Proceedings of Graphics Interface 2004*. GI '04. London, Ontario, Canada: Canadian Human-Computer Communications Society, 2004, pp. 185–194 (cit. on p. 19).

[Barth, 2008] Alexander Barth and Uwe Franke. "Where will the oncoming vehicle be the next second?" *2008 IEEE Intelligent Vehicles Symposium*. 2008, pp. 1068–1073 (cit. on pp. 22, 23).

[Ben-Shabat, 2021] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, et al. "The IKEA ASM Dataset: Understanding People Assembling Furniture through Actions, Objects and Pose". *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 846–858 (cit. on p. 49).

[Binelli, 2005] E. Binelli, A. Broggi, A. Fascioli, S. Ghidoni, P. Grisleri, T. Graf, et al. "A modular tracking system for far infrared pedestrian recognition". *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.* 2005, pp. 759–764 (cit. on pp. 22, 23).

[Bologna, 2020] E. Bologna, N. Lopomo, G. Marchiori, and M. Zingales. "A non-linear stochastic approach of ligaments and tendons fractional-order hereditariness". *Probabilistic Engineering Mechanics* 60 (2020), p. 103034 (cit. on p. 29).

[Bottou, 2012] Léon Bottou. "Stochastic Gradient Descent Tricks." *Neural Networks: Tricks of the Trade (2nd ed.)* Ed. by Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller. Vol. 7700. Lecture Notes in Computer Science. Springer, 2012, pp. 421–436 (cit. on p. 32).

[Busch, 2017] Baptiste Busch, Guilherme Maeda, Yoan Mollard, Marie Demangeat, and Manuel Lopes. "Postural optimization for an ergonomic human-robot interaction". *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 2778–2785 (cit. on p. 13).

[Cai, 2020] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, et al. "Learning Progressive Joint Propagation for Human Motion Prediction". *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Cham: Springer International Publishing, 2020, pp. 226–242 (cit. on p. 44).

[Calinon, 2011] S. Calinon, A. Pistillo, and D. G. Caldwell. "Encoding the time and space constraints of a task in explicit-duration Hidden Markov Model". *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 3413–3418 (cit. on p. 26).

[Cao, 2019] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) (cit. on p. 13).

[Caputo, 2019] Francesco Caputo, Alessandro Greco, Egidio D'Amato, Immacolata Notaro, and Stefania Spada. "IMU-Based Motion Capture Wearable System for Ergonomic Assessment in Industrial Environment". *Structural Health Monitoring on composite components*. Vol. 795. August 2019. Springer International Publishing, 2019, pp. 325–334 (cit. on p. 14).

[Caramia, 2018] Carlotta Caramia, Diego Torricelli, Maurizio Schmid, Adriana Munoz-Gonzalez, Jose Gonzalez-Vargas, Francisco Grandas, et al. "IMU-Based Classification of Parkinson's Disease from Gait: A Sensitivity Analysis on Sensor Location and Feature Selection". *IEEE Journal of Biomedical and Health Informatics* 22.6 (2018), pp. 1765–1774 (cit. on p. 13).

[Caramiaux, 2015] Baptiste Caramiaux, Nicola Montecchio, Atau Tanaka, and Frédéric Bevilacqua. "Adaptive Gesture Recognition with Variation Estimation for Interactive Systems". *ACM Transactions on Interactive Intelligent Systems* 4.4 (2015), pp. 1–34 (cit. on pp. 22, 23, 113).

[Carnegie Mellon University,] Carnegie Mellon University. *CMU Graphics Lab Motion Capture Database* (cit. on p. 48).

[Challis, 1995] J.H. Challis. "An examination of procedures for determining body segment attitude and position from noisy biomechanical data". *Medical Engineering & Physics* 17.2 (1995), pp. 83–90 (cit. on p. 15).

[Chang, 2020] Michael Chang, Nicholas O'Dwyer, Roger Adams, Stephen Cobley, Kwee-Yum Lee, and Mark Halaki. "Whole-body kinematics and coordination in a complex dance sequence: Differences across skill levels". *Human Movement Science* 69 (2020), p. 102564 (cit. on p. 19).

[Cho, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2014, pp. 1724–1734. arXiv: 1406.1078 (cit. on p. 34).

[Chung, 2014] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". *NIPS 2014 Deep Learning and Representation Learning Workshop*. 2014, pp. 1–9. arXiv: 1412.3555 (cit. on p. 34).

[Chung, 2015] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. "A Recurrent Latent Variable Model for Sequential Data". *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 2980–2988 (cit. on p. 44).

[Davidson, 2004] Peter L. Davidson, David J. Chalmers†, and Barry D. Wilson‡. "Stochastic-rheological Simulation of Free-fall Arm Impact in Children: Application to Playground Injuries". *Computer Methods in Biomechanics and Biomedical Engineering* 7.2 (2004), pp. 63–71 (cit. on p. 29).

[Devanne, 2017] Maxime Devanne, Stefano Berretti, Pietro Pala, Hazem Wannous, Mohamed Daoudi, and Alberto Del Bimbo. "Motion segment decomposition of RGB-D sequences for human behavior understanding". *Pattern Recognition* 61 (2017), pp. 222–233 (cit. on p. 23).

[Dewancker, 2016] Ian Dewancker, Michael McCourt, and Scott Clark. "Bayesian Optimization for Machine Learning : A Practical Guidebook" (2016). arXiv: 1612.04858 (cit. on p. 74).

[Donnell, 2014] Drew Michael S. Donnell, Jessica L. Seidelman, Christopher L. Mendias, Bruce S. Miller, James E. Carpenter, and Richard E. Hughes. "A stochastic structural reliability model explains rotator cuff repair retears". *International Biomechanics* 1.1 (2014), pp. 29–35 (cit. on p. 29).

[Drotar, 2015] Peter Drotar, Jiri Mekyska, Irena Rektorova, Lucia Masarova, Zdenek Smekal, and Marcos Faundez-Zanuy. "Decision Support Framework for Parkinson's Disease Based on Novel Handwriting Markers". *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23.3 (2015), pp. 508–516 (cit. on p. 19).

[Du, 2016] Han Du, Martin Manns, Erik Herrmann, and Klaus Fischer. "Joint Angle Data Representation for Data Driven Human Motion Synthesis". *Procedia CIRP* 41 (2016), pp. 746–751 (cit. on p. 16).

[Duchi, 2011] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". *Journal of Machine Learning Research* 12.61 (2011), pp. 2121–2159 (cit. on p. 32).

[Dumas, 2007] R. Dumas, L. Chèze, and J. P. Verriest. "Adjustments to McConville et al. and Young et al. body segment inertial parameters". *Journal of Biomechanics* 40.3 (2007), pp. 543–553 (cit. on pp. 15, 18).

[Ezati, 2019] Mahdokht Ezati, Borna Ghannadi, and John McPhee. "A review of simulation methods for human movement dynamics with emphasis on gait". *Multibody System Dynamics* 47.3 (2019), pp. 265–292 (cit. on p. 112).

[Faber, 2016] G. S. Faber, C. C. Chang, I. Kingma, J. T. Dennerlein, and J. H. van Dieën. "Estimating 3D L5/S1 moments and ground reaction forces during trunk bending using a full-body ambulatory inertial motion capture system". *Journal of Biomechanics* 49.6 (2016), pp. 904–912 (cit. on pp. 17, 21).

[Feldmann, 2019] Felix Feldmann, Robin Seitz, Veronika Kretschmer, Nicole Bednorz, and Michael Ten Hompel. "Ergonomic Evaluation of Body Postures in Order Picking Systems Using Motion Capturing". *2019 Prognostics and System Health Management Conference (PHM-Paris)*. IEEE, 2019, pp. 204–209 (cit. on p. 13).

[Ferrari, 2002] Robert Ferrari, Tom Bohr, and Asa Wilbourn. "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion". *Clinical Spine Surgery* 15.4 (2002), p. 334 (cit. on p. 15).

[Fraccaro, 2016] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. "Sequential Neural Models with Stochastic Layers". *Advances in Neural Information Processing Systems* (2016), pp. 2207–2215. arXiv: 1605.07571 (cit. on pp. 40, 44, 96).

[Françoise, 2015] Jules Françoise. "Motion-Sound Mapping by Demonstration". PhD thesis. Université Pierre et Marie Curie, 2015, p. 254 (cit. on p. 28).

[Ghorbani, 2021] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, et al. "MoVi: A large multi-purpose human motion and video dataset". *PLoS ONE* 16.6 June (2021) (cit. on p. 49).

[Glushkova, 2018] Alina Glushkova and Sotiris Manitsaris. "Gesture recognition and sensorimotor learning-by-doing of motor skills in manual professions: A case study in the wheel-throwing art of pottery". *Journal of Computer Assisted Learning* 34.1 (2018), pp. 20–31 (cit. on pp. 22, 26).

[Grood, 1983] E. S. Grood and W. J. Suntay. "A Joint Coordinate System for the Clinical Description of Three-Dimensional Motions: Application to the Knee". *Journal of Biomechanical Engineering* 105.2 (1983), pp. 136–144 (cit. on p. 15).

[Hagras, 2018] Hani Hagras. "Toward Human-Understandable, Explainable AI". *Computer* 51.9 (2018), pp. 28–36 (cit. on p. 134).

[Haid, 2019] Thomas H. Haid, Matteo Zago, Arunee Promsri, Aude-Clémence M. Doix, and Peter A. Federolf. "PManalyzer: A Software Facilitating the Study of Sensorimotor Control of Whole-Body Movements". *Frontiers in Neuroinformatics* 13 (2019) (cit. on p. 19).

[Halilaj, 2018] Eni Halilaj, Apoorva Rajagopal, Madalina Fiterau, Jennifer L. Hicks, Trevor J. Hastie, and Scott L. Delp. "Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities". *Journal of Biomechanics* 81 (2018), pp. 1–11 (cit. on p. 19).

[Hasan, 2020] S. M. Shafiul Hasan, Masudur R. Siddiquee, and Ou Bai. "Asynchronous Prediction of Human Gait Intention in a Pseudo Online Paradigm Using Wavelet Transform". *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.7 (2020), pp. 1623–1635 (cit. on p. 19).

[Hernández, 2019] Fabio Hernández, Luis F. Suárez, Javier Villamizar, and Miguel Altuve. "Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network". *2019 22nd Symposium on Image, Signal Processing and Artificial Vision, STSIVA 2019 - Conference Proceedings*. 2019, pp. 1–5 (cit. on p. 36).

[Hsu, 2018] Yu-Liang Hsu, Shih-Chin Yang, Hsing-Cheng Chang, and Hung-Che Lai. "Human Daily and Sport Activity Recognition Using a Wearable Inertial Sensor Network". *IEEE Access* 6 (2018), pp. 31715–31728 (cit. on p. 17).

[Ionescu, 2014] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339 (cit. on p. 48).

[Jan de Kok, 2019] Jan de Kok, Paul. Vroonhof, Jacqueline. Snijders, Georgios. Roullis, Martin. Clarke, Kees. Peereboom, et al. *Work-related musculoskeletal disorders : prevalence, costs and demographics in the EU*. 2019, p. 215 (cit. on p. 118).

[Jogin, 2018] Manjunath Jogin, Mohana, M S Madhulika, G D Divya, R K Meghana, and S Apoorva. "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning". *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. 2018, pp. 2319–2323 (cit. on p. 20).

[Jun, 2020] Kooksung Jun, Deok-Won Lee, Kyoobin Lee, Sanghyub Lee, and Mun Sang Kim. "Feature Extraction Using an RNN Autoencoder for Skeleton-Based Abnormal Gait Recognition". *IEEE Access* 8 (2020), pp. 19196–19207 (cit. on p. 20).

[Kalman, 1960] Rudolph Emil Kalman et al. "A new approach to linear filtering and prediction problems". *Journal of basic Engineering* 82.1 (1960), pp. 35–45 (cit. on p. 23).

[Kamel, 2019] Aouaidjia Kamel, Bowen Liu, Ping Li, and Bin Sheng. "An Investigation of 3D Human Pose Estimation for Learning Tai Chi: A Human Factor Perspective". *International Journal of Human–Computer Interaction* 35.4-5 (2019), pp. 427–439 (cit. on p. 112).

[Karhu, 1977] Osmo Karhu, Pekka Kansi, and Iikka Kuorinka. "Correcting working postures in industry: A practical method for analysis". *Applied Ergonomics* 8.4 (1977), pp. 199–201 (cit. on p. 118).

[Kim, 2018] W. Kim, J. Lee, L. Peternel, N. Tsagarakis, and A. Ajoudani. "Anticipatory Robot Assistance for the Prevention of Human Static Joint Overloading in Human–Robot Collaboration". *IEEE Robotics and Automation Letters* 3.1 (2018), pp. 68–75 (cit. on p. 119).

[Kingma, 2013] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114* (2013) (cit. on p. 37).

[Kingma, 2014] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". *arXiv e-prints* (2014). arXiv: 1412.6980 (cit. on pp. 32, 75).

[Kitzig, 2018] Andreas Kitzig, Julia Demmer, Tobias Bolten, Edwin Naroska, Gudrun Stockmanns, Reinhard Viga, et al. "An HMM-based averaging approach for creating mean motion data from a full-body Motion Capture system to support the development of a biomechanical model". *Current Directions in Biomedical Engineering* 4.1 (2018), pp. 389–393 (cit. on p. 26).

[Kok, 2014] Manon Kok, Jeroen D. Hol, and Thomas B. Schön. "An optimization-based approach to human body motion capture using inertial sensors". *IFAC Proceedings Volumes* 47.3 (2014), pp. 79–85 (cit. on p. 13).

[Koller, 2009] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009 (cit. on p. 23).

[Krishnan, 2017] Rahul Krishnan, Uri Shalit, and David Sontag. "Structured Inference Networks for Nonlinear State Space Models". *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (2017) (cit. on p. 44).

[Kulsoom, 2022] Farzana Kulsoom, Sanam Narejo, Zahid Mehmood, Hassan Nazeer Chaudhry, Ayesha Butt, and Ali Kashif Bashir. "A review of machine learning-based human activity recognition for diverse applications". *Neural Computing and Applications* 34.21 (2022), pp. 18289–18324 (cit. on p. 42).

[Kurbiel, 2017] Thomas Kurbiel and Shahrzad Khaleghian. "Training of Deep Neural Networks based on Distance Measures using RMSProp". *arXiv e-prints* (2017). arXiv: 1708.01911 (cit. on p. 32).

[Langenderfer, 2006] Joseph E. Langenderfer, James E. Carpenter, Marjorie E. Johnson, Kai-nan An, and Richard E. Hughes. "A Probabilistic Model of Glenohumeral External Rotation Strength for Healthy Normals and Rotator Cuff Tear Cases". *Annals of Biomedical Engineering* 34.3 (2006), pp. 465–476 (cit. on p. 29).

[Larsen, 2020] Frederik Greve Larsen, Frederik Petri Svenningsen, Michael Skipper Andersen, Mark de Zee, and Sebastian Skals. "Estimation of Spinal Loading During Manual Materials Handling Using Inertial Motion Capture". *Annals of Biomedical Engineering* 48.2 (2020), pp. 805–821 (cit. on p. 21).

[Lee, 2020] Hoonyong Lee, Kanghyeok Yang, Namgyun Kim, and Changbum R. Ahn. "Detecting excessive load-carrying tasks using a deep learning network with a Gramian Angular Field". *Automation in Construction* 120 (2020), p. 103390 (cit. on p. 49).

[Lee, 2017] Wonil Lee, Edmund Seto, Ken Yu Lin, and Giovanni C. Migliaccio. "An evaluation of wearable sensors and their placements for analyzing construction worker's trunk posture in laboratory conditions". *Applied Ergonomics* 65.2017 (2017), pp. 424–436 (cit. on p. 17).

[Leva, 1996] Paolo de Leva. "Adjustments to Zatsiorsky-Seeluyanov's segment inertia parameters". *Journal of Biomechanics* 29.9 (1996), pp. 1223–1230 (cit. on pp. 15, 18).

[Li, 2019] Longyuan Li, Junchi Yan, Xiaokang Yang, and Yaohui Jin. "Learning Interpretable Deep State Space Model for Probabilistic Time Series Forecasting". *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Vol. 2019-Augus. California: International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 2901–2908. arXiv: 2102.00397 (cit. on pp. 40, 44, 96).

[Liao, 2021] Chen-Chieh Liao, Dong-Hyun Hwang, and Hideki Koike. "How Can I Swing Like Pro?: Golf Swing Analysis Tool for Self Training". *SIGGRAPH Asia 2021 Posters*. SA '21 Posters. Tokyo, Japan: Association for Computing Machinery, 2021 (cit. on p. 112).

[Lin, 2009] Cheng Feng Lin, Michael Gross, Chuanshu Ji, Darin Padua, Paul Weinhold, William E. Garrett, et al. "A stochastic biomechanical model for risk and risk factors of non-contact anterior cruciate ligament injuries". *Journal of Biomechanics* 42.4 (2009), pp. 418–423 (cit. on p. 29).

[Lin, 2012] Cheng-Feng Lin, Hui Liu, Michael T. Gros, Paul Weinhold, William E. Garrett, and Bing Yu. "Biomechanical risk factors of non-contact ACL injuries: A stochastic biomechanical modeling study". *Journal of Sport and Health Science* 1.1 (2012), pp. 36–42 (cit. on p. 29).

[Liu, 2020a] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.10 (2020), pp. 2684–2701 (cit. on p. 49).

[Liu, 2017] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. "A survey of deep neural network architectures and their applications". *Neurocomputing* 234 (2017), pp. 11–26 (cit. on p. 42).

[Liu, 2020b] Xiaoli Liu, Jianqin Yin, Huaping Liu, and Jun Liu. "DeepSSM: Deep State-Space Model for 3D Human Motion Prediction". *arXiv e-prints* (2020). arXiv: 2005.12155 (cit. on p. 44).

[Liu, 2019] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, et al. "Towards Natural and Accurate Future Motion Prediction of Humans and Animals". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on p. 43).

[Lopez Pinaya, 2020] Walter Hugo Lopez Pinaya, Sandra Vieira, Rafael Garcia-Dias, and Andrea Mechelli. "Chapter 11 - Autoencoders". *Machine Learning*. Ed. by Andrea Mechelli and Sandra Vieira. Academic Press, 2020, pp. 193–208 (cit. on p. 37).

[Lu, 2012] Tung Wu Lu and Chu Fen Chang. "Biomechanics of human movement and its clinical applications". *Kaohsiung Journal of Medical Sciences* 28.2 SUPPL. (2012), S13–S25 (cit. on p. 20).

[Luinge, 2005] H. J. Luinge and Peter H. Veltink. "Measuring orientation of human body segments using miniature gyroscopes and accelerometers". *Medical and Biological Engineering and Computing* 43.2 (2005), pp. 273–282 (cit. on p. 15).

[Luong, 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation". *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 1412–1421 (cit. on pp. 40, 42, 76).

[Luttinen, 2014] Jaakko Luttinen, Tapani Raiko, and Alexander Ilin. "Linear State-Space Model with Time-Varying Dynamics". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8725 LNAI. PART 2. 2014, pp. 338–353. arXiv: 1410.0555 (cit. on p. 68).

[Mahmood, 2019] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. "AMASS: Archive of Motion Capture as Surface Shapes". *International Conference on Computer Vision*. 2019, pp. 5442–5451 (cit. on p. 48).

[Mäkela, 2021] Satu-Marja Mäkela, Arttu Lämsä, Janne S. Keränen, Jussi Liikka, Jussi Ronkainen, Johannes Peltola, et al. "Introducing VTT-ConIot: A Realistic Dataset for Activity Recognition of Construction Workers Using IMU Devices". *Sustainability* 14.1 (2021), p. 220 (cit. on p. 49).

[Malaisé, 2018] Adrien Malaisé, Pauline Maurice, Francis Colas, François Charpillet, Adrien Malaisé, Pauline Maurice, et al. "Activity Recognition With Multiple Wearable Sensors for Industrial Applications". *ACHI 2018 - Eleventh International Conference on Advances in Computer-Human Interactions*. 2018, pp. 1–7 (cit. on pp. 22, 26).

[Malaisé, 2019] Adrien Malaisé, Pauline Maurice, Francis Colas, and Serena Ivaldi. "Activity Recognition for Ergonomics Assessment of Industrial Tasks with Automatic Feature Selection". *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1132–1139 (cit. on p. 17).

[Manghisi, 2017] Vito Modesto Manghisi, Antonio Emmanuele Uva, Michele Fiorentino, Vitoantonio Bevilacqua, Gianpaolo Francesco Trotta, and Giuseppe Monno. "Real time RULA assessment using Kinect v2 sensor". *Applied Ergonomics* 65.2017 (2017), pp. 481–491 (cit. on pp. 13, 119).

[Manitsaris, 2014] S. Manitsaris, A. Glushkova, F. Bevilacqua, and F. Moutarde. "Capture, modeling, and recognition of expert technical gestures in wheel-throwing art of pottery". *Journal on Computing and Cultural Heritage* 7.2 (2014), pp. 1–15 (cit. on pp. 22, 23, 26, 69).

[Manitsaris, 2020] Sotiris Manitsaris, Gavriela Senteri, Dimitrios Makrygiannis, and Alina Glushkova. "Human movement representation on multivariate time series for recognition of professional gestures and forecasting their trajectories". *Frontiers in Robotics and AI* 7 (2020), pp. 1–20 (cit. on pp. 22, 23, 26, 29, 30, 66).

[Mao, 2021a] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. "Generating Smooth Pose Sequences for Diverse Human Motion Prediction". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 13289–13298 (cit. on p. 44).

[Mao, 2019] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. "Learning Trajectory Dependencies for Human Motion Prediction". *ICCV*. 2019. arXiv: 1908.05436 (cit. on p. 43).

[Mao, 2021b] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. "Multi-level Motion Attention for Human Motion Prediction". *International Journal of Computer Vision* 129.9 (2021), pp. 2513–2535 (cit. on p. 44).

[Marin, 2018] A. G. Marin, M. S. Shourijeh, P. E. Galibarov, M. Damsgaard, L. Fritzsch, and F. Stulp. "Optimizing Contextual Ergonomics Models in Human-Robot Interaction". *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018, pp. 1–9 (cit. on p. 119).

[Martinez, 2017] Julieta Martinez, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks". *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-Janua (2017), pp. 4674–4683. arXiv: 1705.02445 (cit. on p. 36).

[Martinez-Hernandez, 2018] Uriel Martinez-Hernandez, Imran Mahmood, and Abbas A. Dehghani-Sanij. "Simultaneous Bayesian Recognition of Locomotion and Gait Phases with Wearable Sensors". *IEEE Sensors Journal* 18.3 (2018), pp. 1282–1290 (cit. on p. 13).

[Maurice, 2019] Pauline Maurice, Adrien Malaisé, Clélie Amiot, Nicolas Paris, Guy Junior Richard, Olivier Rochel, et al. "Human movement and ergonomics: An industry-oriented dataset for collaborative robotics". *International Journal of Robotics Research* 38.14 (2019), pp. 1529–1537 (cit. on p. 49).

[McAtamney, 1993] Lynn McAtamney and Nigel Corlett. "RULA: A survey method for the investigation of work-related upper limb disorders". *Applied Ergonomics* 24.2 (1993), pp. 91–99 (cit. on pp. 118, 142–144, 146).

[Mcculloch, 1943] Warren Mcculloch and Walter Pitts. "A Logical Calculus of Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biophysics* 5 (1943), pp. 127–147 (cit. on p. 32).

[Menychtas, 2020] Dimitrios Menychtas, Alina Glushkova, and Sotiris Manitsaris. "Analyzing the kinematic and kinetic contributions of the human upper body's joints for ergonomics assessment". *Journal of Ambient Intelligence and Humanized Computing* (2020), pp. 1–23 (cit. on pp. 21, 133).

[Mohammadzade, 2021] Hoda Mohammadzade, Soheil Hosseini, Mohammad Reza Rezaei-Dastjerdehei, and Mohsen Tabejamaat. "Dynamic Time Warping-Based Features With Class-Specific Joint Importance Maps for Action Recognition Using Kinect Depth Sensor". *IEEE Sensors Journal* 21.7 (2021), pp. 9300–9313 (cit. on p. 20).

[Mozer, 1989] M. C. Mozer. "A focused backpropagation algorithm for temporal pattern recognition". *Complex Systems* 3 (1989), pp. 349–381 (cit. on p. 34).

[Muller, 2020] A. Muller, C. Pontonnier, X. Robert-Lachaine, G. Dumont, and A. Plamondon. "Motion-based prediction of external forces and moments and back loading during manual material handling tasks". *Applied Ergonomics* 82.August 2019 (2020), p. 102935 (cit. on pp. 17, 20).

[Nath, 2017] Nipun D. Nath, Reza Akhavian, and Amir H. Behzadan. "Ergonomic analysis of construction worker's body postures using wearable mobile sensors". *Applied Ergonomics* 62.2017 (2017), pp. 107–117 (cit. on p. 14).

[Nath, 2018] Nipun D. Nath, Theodora Chaspari, and Amir H. Behzadan. "Automated ergonomic risk monitoring using body-mounted sensors and machine learning". *Advanced Engineering Informatics* 38.August (2018), pp. 514–526 (cit. on p. 17).

[Nielsen, 2011] Johnny L G Nielsen, S Holmgaard, Ning Jiang, K B Englehart, D Farina, and P A Parker. "Simultaneous and Proportional Force Estimation for Multifunction Myoelectric Prostheses Using Mirrored Bilateral Training". *IEEE Transactions on Biomedical Engineering* 58.3 (2011), pp. 681–688 (cit. on p. 19).

[Olivas-Padilla, 2019] Brenda E. Olivas-Padilla, Alina Glushkova, and Sotiris Manitsaris. "Motion analysis for identification of overused body segments : the packaging task in industry 4 . 0". *Proceedings of the Human-Computer Interaction- INTERACT 2019: IFIP TC 17 International Conference*. Paphos, Cyprus: Springer, 2019, pp. 2–5 (cit. on p. 19).

[Olivas-Padilla, 2021] Brenda Elizabeth Olivas-Padilla, Sotiris Manitsaris, Dimitrios Menychtas, and Alina Glushkova. "Stochastic-biomechanic modeling and recognition of human movement primitives, in industry, using wearables". *Sensors* 21.7 (2021) (cit. on p. 66).

[Olivas-Padilla, 2023] Brenda Elizabeth Olivas-Padilla, Dimitris Papanagiotou, Gavriela Senteri, Sotiris Manitsaris, and Alina Glushkova. "Computational ergonomics for task delegation in Human-Robot Collaboration: spatiotemporal adaptation of the robot to the human through contactless gesture recognition". *arXiv e-prints* (2023). arXiv: 2203.11007 (cit. on pp. 120, 127, 129).

[Olsson, 2007] Rasmus Kongsgaard Olsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. "State-Space Models: From the EM Algorithm to a Gradient Approach". *Neural Computation* 19.4 (2007), pp. 1097–1111 (cit. on p. 26).

[Otten, 2015] Paul Otten, Jonghyun Kim, and Sang Hyuk Son. "A framework to automate assessment of upper-limb motor function impairment: A feasibility study". *Sensors (Switzerland)* 15.8 (2015), pp. 20097–20114 (cit. on p. 13).

[Papanagiotou, 2021] Dimitris Papanagiotou, Gavriela Senteri, and Sotiris Manitsaris. "Egocentric Gesture Recognition Using 3D Convolutional Neural Networks for the Spatiotemporal Adaptation of Collaborative Robots". *Frontiers in Neurorobotics* 15.November (2021) (cit. on pp. 120, 127, 129).

[Pascanu, 2012] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training Recurrent Neural Networks". *arXiv e-prints* (2012). arXiv: 1211.5063 (cit. on p. 34).

[Patterson, 2008] Toby Patterson, Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos. "State–space models of individual animal movement". *Trends in Ecology & Evolution* 23.2 (2008), pp. 87–94 (cit. on pp. 23, 135).

[Pavllo, 2020] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. "Modeling Human Motion with Quaternion-Based Neural Networks". *International Journal of Computer Vision* 128.4 (2020), pp. 855–872. arXiv: 1901.07677 (cit. on p. 16).

[Pavllo, 2018] Dario Pavllo, David Grangier, and Michael Auli. "QuaterNet: A quaternion-based recurrent model for human motion". *British Machine Vision Conference 2018, BMVC 2018*. 2018, pp. 1–14. arXiv: 1805.06485 (cit. on p. 43).

[Peng, 2022] Zhe Peng, Qing Xu, Runlin Zhang, Klaus Schoeffmann, and Simon Parkinson. "Eye movement based information system indicates human behavior in virtual driving" (2022) (cit. on p. 19).

[Petrovich, 2021] Mathis Petrovich, Michael J. Black, and Gül Varol. "Action-Conditioned 3D Human Motion Synthesis with Transformer VAE". *arXiv e-prints* (2021), arXiv:2104.05670. arXiv: 2104.05670 (cit. on p. 44).

[Portnova-Fahreeva, 2020] Alexandra A. Portnova-Fahreeva, Fabio Rizzoglio, Ilana Nisky, Maura Casadio, Ferdinando A. Mussa-Ivaldi, and Eric Rombokas. "Linear and Non-linear Dimensionality-Reduction Techniques on Full Hand Kinematics". *Frontiers in Bioengineering and Biotechnology* 8 (2020) (cit. on p. 19).

[Rabiner, 1989] Lawrance R. Rabiner. "A Tutorial on Hidden Markov Models and Seected Applications in Speech Recognition". *Proceedings of the IEEE* 77.2 (1989), pp. 257–286 (cit. on pp. 22, 26, 28).

[Rangapuram, 2018] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. "Deep State Space Models for Time Series Forecasting". *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018 (cit. on p. 44).

[Ribeiro, 2020] Antonio H. Ribeiro, Koen Tiels, Luis A. Aguirre, and Thomas Schön. "Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness". *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2370–2380 (cit. on p. 34).

[Roetenberg, 2009] Daniel Roetenberg, Henk Luinge, and Per Slycke. *Xsens MVNN: full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV*. Tech. rep. 2009 (cit. on p. 15).

[Rudenko, 2020] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M. Kitani, Dariu M. Gavrila, and Kai O. Arras. "Human motion trajectory prediction: a survey". *International Journal of Robotics Research* 39.8 (2020), pp. 895–935 (cit. on pp. 42, 135).

[Rumelhart, 1986] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning Internal Representations by Error Propagation". *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. Ed. by David E. Rumelhart and James L. Mcclelland. Cambridge, MA: MIT Press, 1986, pp. 318–362 (cit. on p. 36).

[Salinas, 2017] David Salinas, Valentin Flunkert, and Jan Gasthaus. "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks" (2017). arXiv: 1704.04110 (cit. on pp. 44, 96).

[Samadani, 2020] Ali Samadani, Rob Gorbet, and Dana Kulic. "Affective Movement Generation using Laban Effort and Shape and Hidden Markov Models". *arXiv e-prints* (2020). arXiv: 2006.06071 (cit. on p. 26).

[Santos, 2019] Olga C. Santos. "Artificial Intelligence in Psychomotor Learning: Modeling Human Motion from Inertial Sensor Data". *International Journal on Artificial Intelligence Tools* 28.04 (2019), p. 1940006 (cit. on p. 113).

[Santos, 2004] V.J. Santos and F.J. Valero-Cuevas. "A Bayesian approach to biomechanical modeling to optimize over large parameter spaces while considering anatomical variability". *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 4. IEEE, 2004, pp. 4626–4629 (cit. on p. 29).

[Saraygord Afshari, 2022] Sajad Saraygord Afshari, Fatemeh Enayatollahi, Xiangyang Xu, and Xihui Liang. "Machine learning-based methods in structural reliability analysis: A review". *Reliability Engineering & System Safety* 219 (2022), p. 108223 (cit. on p. 30).

[Sato, 2019] Taiki Sato and Yuko Osana. "Automatic Generation of Dance and Facial Expressions Linked to Music using HMM". *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3999–4006 (cit. on p. 26).

[Schaub, 2013] K. Schaub, G. Caragnano, B. Britzke, and R. Bruder. "The European Assembly Worksheet". *Theoretical Issues in Ergonomics Science* 14.6 (2013), pp. 616–639 (cit. on pp. 48, 118, 122).

[Shafti, 2019] A Shafti, A Ataka, B Urbistondo Lazpita, A Shiva, H A Wurdemann, and K Althoefer. "Real-time Robot-assisted Ergonomics*". *International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, 2019, pp. 1975–1981 (cit. on p. 119).

[Shaheen, 2016] Fatma Shaheen, Brijesh Verma, and Md. Asafuddoula. "Impact of Automatic Feature Extraction in Deep Learning Architecture". *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2016, pp. 1–8 (cit. on p. 20).

[Sharma, 2019] Shubham Sharma, Shubhankar Verma, Mohit Kumar, and Lavanya Sharma. "Use of Motion Capture in 3D Animation: Motion Capture Systems, Challenges, and Recent Trends". *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. 2019, pp. 289–294 (cit. on p. 13).

[Shi, 2003] Pengcheng Shi and Huafeng Liu. "Stochastic finite element framework for simultaneous estimation of cardiac kinematic functions and material parameters". *Medical Image Analysis* 7.4 (2003), pp. 445–464 (cit. on p. 29).

[Shojaei, 2016] Iman Shojaei, Milad Vazirian, Emily Croft, Maury A. Nussbaum, and Babak Bazrgari. "Age related differences in mechanical demands imposed on the lower back by manual material handling tasks". *Journal of Biomechanics* 49.6 (2016), pp. 896–903 (cit. on pp. 17, 21).

[Shu, 2022] Xiangbo Shu, Liyan Zhang, Guo-Jun Qi, Wei Liu, and Jinhui Tang. "Spatiotemporal Co-Attention Recurrent Neural Networks for Human-Skeleton Motion Prediction". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.6 (2022), pp. 3300–3315 (cit. on p. 43).

[Sigal, 2010] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion". *International Journal of Computer Vision* 87.1-2 (2010), pp. 4–27 (cit. on p. 49).

[Slaton, 2020] Trevor Slaton, Carlos Hernandez, and Reza Akhavian. "Construction activity recognition with convolutional recurrent networks". *Automation in Construction* 113 (2020), p. 103138 (cit. on p. 36).

[Snoek, 2012] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012 (cit. on p. 74).

[Sousa Lima, 2019] Wesllen Sousa Lima, Eduardo Souto, Khalil El-Khatib, Roozbeh Jalali, and Joao Gama. "Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview". *Sensors* 19.14 (2019), p. 3213 (cit. on p. 17).

[Stergiou, 2018] Nicholas Stergiou. *Nonlinear Analysis for Human Movement Variability*. Ed. by Nicholas Stergiou. Boca Raton : Taylor & Francis, Taylor & Francis, a CRC title, part of the: CRC Press, 2018 (cit. on pp. 20, 23).

[Sun, 2017] Li Sun, Zhi Yan, Sergi Molina Mellado, Marc Hanheide, and Tom Duckett. "3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data". *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2017, pp. 5942–5948 (cit. on pp. 36, 43).

[Switonski, 2019] Adam Switonski, Henryk Josinski, and Konrad Wojciechowski. "Dynamic time warping in classification and selection of motion capture data". *Multidimensional Systems and Signal Processing* 30.3 (2019), pp. 1437–1468 (cit. on p. 20).

[Takeda, 2009] R. Takeda, S. Tadano, A. Natorigawa, M. Todoh, and S. Yoshinari. "Gait posture estimation by wearable acceleration and gyro sensor". *IFMBE Proceedings* 25.9 (2009), pp. 111–114 (cit. on p. 13).

[Tang, 2018] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. "Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamic". *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, 2018, pp. 935–941 (cit. on p. 43).

[Van Der Aa, 2011] N. P. Van Der Aa, X. Luo, G. J. Giezeman, R. T. Tan, and R. C. Veltkamp. "UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction". *Proceedings of the IEEE International Conference on Computer Vision* (2011), pp. 1264–1269 (cit. on p. 49).

[Vignais, 2013] Nicolas Vignais, Markus Miezal, Gabriele Bleser, Katharina Mura, Dominic Gorecky, and Frédéric Marin. "Innovative system for real-time ergonomic feedback in industrial manufacturing". *Applied Ergonomics* 44.4 (2013), pp. 566–574 (cit. on pp. 14, 17).

[Vince, 2011] John Vince. *Quaternions for Computer Graphics*. London: Springer London, 2011 (cit. on p. 16).

[Von Marcard, 2016] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. "Human Pose Estimation from Video and IMUs". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2016), pp. 1533–1547 (cit. on p. 13).

[Wang, 2010] Xiaoyue Wang, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. "Experimental Comparison of Representation Methods and Distance Measures for Time Series Data" (2010). arXiv: 1012.2789 (cit. on pp. 20, 70, 78).

[Wang, 2013] Zhikun Wang, Katharina Mülling, Marc Peter Deisenroth, Heni Ben Amor, David Vogt, Bernhard Schölkopf, et al. "Probabilistic movement modeling for intention inference in human-robot interaction". *International Journal of Robotics Research* 32.7 (2013), pp. 841–858 (cit. on p. 22).

[Woltring, 1994] Herman J. Woltring. "3-D attitude representation of human joints: A standardization proposal". *Journal of Biomechanics* 27.12 (1994), pp. 1399–1414 (cit. on p. 15).

[Wu, 2002] Ge Wu, Sorin Siegler, Paul Allard, Chris Kirtley, Alberto Leardini, Dieter Rosenbaum, et al. "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine". *Journal of Biomechanics* 35.4 (2002), pp. 543–548 (cit. on p. 15).

[Wu, 2005] Ge Wu, Frans C.T. Van Der Helm, H. E.J. Veeger, Mohsen Makhsous, Peter Van Roy, Carolyn Anglin, et al. "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion - Part II: Shoulder, elbow, wrist and hand". *Journal of Biomechanics* 38.5 (2005), pp. 981–992 (cit. on p. 15).

[Xia, 2012] L. Xia, C.C. Chen, and JK Aggarwal. "View invariant human action recognition using histograms of 3D joints". *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE. 2012, pp. 20–27 (cit. on p. 49).

[Xu, 2015] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 2048–2057 (cit. on pp. 41, 42).

[Xue, 2018] Hao Xue, Du Q. Huynh, and Mark Reynolds. "SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction". *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 1186–1194 (cit. on p. 43).

[Yan, 2018] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, et al. "MT-VAE: Learning Motion Transformations to Generate Multimodal Human Dynamics". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11209 LNCS (2018), pp. 276–293. arXiv: 1808.04545 (cit. on pp. 36, 44).

[Yan, 2017] Xuzhong Yan, Heng Li, Angus R. Li, and Hong Zhang. "Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention". *Automation in Construction* 74.2017 (2017), pp. 2–11 (cit. on pp. 17, 119).

[Yoon, 2019] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. "Time-series Generative Adversarial Networks". *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 131).

[Yoshida, 2022] Kaya Yoshida, Drew Commandeur, Sandra Hundza, and Marc Klimstra. "Detecting Differences in Gait Initiation between Older Adult Fallers and Non-Fallers Through Time-Series Principal Component Analysis (PCA)". *SSRN Electronic Journal* (2022) (cit. on p. 19).

[Zaremba, 2014] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. "Recurrent Neural Network Regularization". *arXiv e-prints* (2014). arXiv: 1409.2329 (cit. on p. 34).

[Zhang, 2013] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. "From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding". *2013 IEEE International Conference on Computer Vision*. 2013, pp. 2248–2255 (cit. on p. 49).

[Zhou, 2019] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. "On the Continuity of Rotation Representations in Neural Networks". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on p. 16).

[Zhu, 2015] Chun Zhu, Weihua Sheng, and Meiqin Liu. "Wearable Sensor-Based Behavioral Anomaly Detection in Smart Assisted Living Systems". *IEEE Transactions on Automation Science and Engineering* 12.4 (2015), pp. 1225–1234 (cit. on p. 13).

[Zhu, 2020] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. "S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation". *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), pp. 6537–6546. arXiv: 2005.11437 (cit. on p. 36).

RÉSUMÉ

L'analyse des mouvements humains a été étudiée de manière approfondie dans le passé en raison de sa grande variété d'applications pratiques, telles que l'interaction homme-robot, les applications d'apprentissage humain, le diagnostic clinique et la surveillance des activités humaines. Néanmoins, l'état de l'art reste confronté à des défis scientifiques lors de la modélisation des mouvements humains. D'abord, pour modéliser la dynamique spatiale et temporelle des mouvements humains et prédire avec précision l'évolution des descripteurs de mouvement, il faut considérer la stochasticité des mouvements humains et la structure physique du corps. Ensuite, l'explicabilité des algorithmes d'apprentissage profond existants concernant leurs prédictions doit encore être améliorée car ils manquent pour la plupart de représentations du mouvement humain compréhensibles par les humains. Par conséquent, cette thèse étudie et présente des méthodes d'apprentissage machine pour l'analyse automatique et la représentation du mouvement humain. Le mouvement humain est formulé comme un modèle d'espace d'état d'un système dynamique dont les paramètres sont estimés à partir d'algorithmes d'apprentissage profond et de statistiques. Les modèles adhèrent à la structure du Gesture Operational Model (GOM), qui intègre des hypothèses sur la dynamique spatiale et temporelle dans la représentation mathématique du mouvement humain. Deux nouveaux modèles profonds d'espace d'état sont présentés, qui modélisent une variété de mouvements humains à partir d'une paramétrisation non linéaire et fournissent des prédictions interprétables en utilisant la représentation GOM. La troisième méthode estime les représentations GOM en utilisant l'estimation par maximum de vraisemblance avec des filtres de Kalman. Contrairement aux modèles d'état profond, l'approche statistique est suffisamment précise pour produire des mouvements humains précis en utilisant des procédures d'entraînement simples qui nécessitent moins de puissance de traitement que les méthodes d'apprentissage profond.Enfin, deux applications des modèles créés sont décrites. La première est destinée à l'analyse de la dextérité des mouvements professionnels, où les associations dynamiques entre les articulations du corps et les descripteurs de mouvement significatifs sont identifiées. La seconde application concerne la réalisation d'une méthodologie de délégation de tâches pour optimiser l'ergonomie des structures de collaboration humain-robot.

MOTS CLÉS

Analyse automatique du mouvement; Modélisation du mouvement humain; Modèles d'espace d'état; Apprentissage profond; Capture du mouvement.

ABSTRACT

The analysis of human movements has been extensively studied in the past due to its wide variety of practical applications, such as human-robot interaction, human learning applications, clinical diagnosis, and monitoring of human activities. Nevertheless, the state-of-the-art still faces scientific challenges while modeling human movements. Firstly, to model the spatial and temporal dynamics of human movement and accurately predict the evolution of motion descriptors over time, the stochasticity of human movement and the physical body structure must be considered. Second, the explainability of existing deep learning algorithms regarding their predictions still needs to be improved as they lack human-comprehensible representations of human movement. This thesis studies and introduces machine learning approaches for the automatic analysis and representation of human movement. Human movement is formulated as a state-space model of a dynamic system whose parameters are estimated using deep learning and statistical algorithms. The models adhere to the structure of the Gesture Operational Model (GOM), which incorporates spatial and temporal dynamics assumptions in the mathematical representation of human movement. Two novel deep state-space models are presented that model a variety of human movements using nonlinear network parameterization and provide interpretable predictions using the GOM representation. The encoder-decoder structure of the models not only allows them to simulate full-body human movement, but also to disentangle variation factors across distinct movements, cluster related motion descriptors, and identify joint dynamics across sequences. The third method estimates GOM representations using Maximum Likelihood Estimation via Kalman Filters. In contrast to the deep state models, the statistical approach is sufficiently accurate to generate specific human movements utilizing one-shot training. This training strategy enables users to model single human movements and estimate their mathematical representation using simple procedures that require less computational power than data-driven methods. Finally, two applications of the generated models are described. The first is for dexterity analysis of professional movements, where dynamic associations between body joints and meaningful motion descriptors are identified. The second application is for implementing an ergonomically effective task delegation methodology to optimize human-robot collaboration frameworks.

KEYWORDS

Automatic movement analysis; Human motion modeling; State-space modeling; Data-driven learning; Motion capture.