Contributions to the in-situ biomechanical and physical ergonomic analysis of workstations using machine learning and deep learning techniques

Hasnaa Ouadoudi Belabzioui

COLLEGE DOCTORAL BRETAGNE

MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

école normale supérieure rennes

THÈSE DE DOCTORAT DE

L'ÉCOLE NORMALE
SUPÉRIEURE DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, télécommunications, informatique, signal, systèmes, électronique
Spécialité : *Informatique*

Par
Hasnaa OUADOUDI BELABZIOUI

Contributions to the In-Situ Biomechanical and Physical Ergonomic Analysis of Workstations Using Machine Learning and Deep Learning Techniques.

Thèse présentée et soutenue à INRIA Rennes, le 20/12/2024
Unité de recherche : IRISA - UMR 6074

Rapporteurs avant soutenance :

Laetitia FRADET — Professeure des universités à l'université de Bretagne sud
Nasser REZZOUG — Maître de conférences, HDR, à l'université de Poitiers

Composition du Jury :

Président·e :	Elisa FROMONT	Professeure des universités à l'université de Rennes
Examinateurs :	Laetitia FRADET	Professeure des universités à l'université de Bretagne sud
	Elisa FROMONT	Professeure des universités à l'université de Rennes
	Nasser REZZOUG	Maître de conférences, HDR, à l'université de Poitiers
	Jonathan SAVIN	Responsable d'études (PhD) à l'INRS
Dir. de thèse :	Franck MULTON	Directeur de recherche à l'Inria Rennes
Co-dir. de thèse :	Charles PONTONNIER	Maître de conférences, HDR, à l'École normale supérieure de Rennes

Acknowledgement

Reflecting on the acknowledgements feels like stepping back into a journey filled with invaluable connections and experiences. As I look back on these past three years, **from January 7, 2022, to December 20, 2024**, I am reminded of the countless interactions and support that have guided me to the successful completion of this thesis. This journey has been enriched by the remarkable individuals who have accompanied me every step of the way.

First and foremost, I would like to express my sincere gratitude to Nasser Rezzoug and Laetitia Fradet for agreeing, without hesitation, to serve as my thesis reviewers. Their insightful feedback has provided me with valuable perspectives on this work, and I am deeply thankful for their constructive and thoughtful remarks. I also extend my gratitude to Elisa Fromont and Jonathan Savin for kindly agreeing to be part of this jury.

I am also profoundly grateful to my supervisors, Charles Pontonnier, Franck Multon, and Georges Dumont (listed by first name). Over the past three years, I have had the privilege of being their PhD student. They are three individuals of great humanity, remarkable intelligence, exceptional teaching skills, and unwavering honesty. Above all, they share a burning passion for Science. As Charles often emphasizes, *Il faut faire avancer la Science*, a philosophy that has been a driving force throughout this research project and will continue in my post-doc. The quote Georges shared with me still resonates in my mind, urging me to read and tirelessly revise my thesis manuscript: *Hâtez-vous lentement et sans perdre courage, vingt fois sur le métier remettez votre ouvrage : polissez-le sans cesse et le repolissez ; ajoutez quelquefois, et souvent effacez* (L. Boileau). Honestly, my thesis supervisors will be missed, and it is thanks to them that this experience will remain special and unforgettable. I want to say thank you to them!

Naturally, I must also express my gratitude to the company Moovency, which made this thesis possible. First and foremost, I would like to thank Pierre Plantard, who co-supervised me during this thesis, and François Morin, for his kindness and support. It was truly a pleasure to be part of this company, which granted me the freedom to pursue my research with autonomy and creativity.

I would like to extend a special thank you to Aurélien for all his help. It has been a pleasure to finish my thesis with such a kind person. I'll be waiting for you in Boston so we can run a half-marathon together!

I would also like to express my gratitude to my friends from Math&Maroc, with whom I've shared countless wonderful memories—through both the highs and lows. A special thanks goes to Omar, the President of Math&Maroc, a remarkably kind individual who

has helped me move many things forward, especially regarding the post-thesis period. Looking forward to working together on new projects for the association. To Mohammed Ali, working alongside someone as determined, honest, and hardworking as you has been a true pleasure. You've taught me the real meaning of fully committing to our beliefs and never giving up. I would also like to specifically thank Chaimaa for attending my thesis defense; it will always remain engraved in my heart. I can't forget Abdelhakim, my friends from 1337 (Mounia, Khalid, Abderahmane, Jawaher, Hamza, Mouad), and Ahmed (1337); it's always a real joy to work with all of you.

I would like to thank the remarkable staff of the Inria cafeteria: Laurence, Maeva, and Anne-Claire. Over the course of these three years of my PhD, it has been a real pleasure seeing you at the cafeteria. I will miss you, and I will also miss the atmosphere of the cafeteria. I would also like to thank Tassadit for mentoring me during these three years of my PhD. It has been a real pleasure to converse with you and to exchange thoughts on the ups and downs of the PhD journey and what comes after. Thank you so much, Tassadit!

I would also like to extend my heartfelt thanks to Pauline, Claire, Louise, and Aurélie for their support and help. It has been a true pleasure to have crossed paths with such interesting and kind people. I also want to extend my gratitude to all the people—far too many to name individually—with whom I've shared these past three years. A special thanks to the members of the Mimetic team at Inria, as well as everyone in the M2S group and the residents of the Rocard building at ENS Rennes. I appreciate the coffee breaks and meals we shared. I also want to express my heartfelt thanks to Théo, Ewen, Valentin, Nolwenn, Sony, Pierre, Julian, Victor, Lucas, Rebecca, Benjamin, Shubhendu, Tangui, Thomas, Suzan, Salomé, João, Amine, Nathalie, and Mohamed for their friendship and support. Additionally, I am grateful to Anne-Josiane, a PhD student I met at a summer school in Germany, for the interesting discussions we had together.

I dedicate this thesis to the person I am most grateful to, my mother Zahra. Mama, I want you to know that I love you with all my heart. You will always be the best mother in the world, and I will strive to be the best daughter. I love you, and I look forward to living by your side and sharing everything that life has in store for us. Together, we can overcome anything. I also want to thank my only sister Ghizlane, the one who always makes me laugh. I love you, my sis, and I am so happy to see you growing in life. Know that I will always be there for you, through the highs and lows. I love you, Kbida and Sis.

I also want to express my gratitude to my uncle Mohamed, my aunt Rachida (I miss you so much Khalti), my cousin Imad, my aunt Kabira, my cousins Sabah and Kamal, and my favorite little cousins, Ritaj, Mohamed, Abderahmane and Yaakoub. You are the ones with whom I share the ups and downs of my life, and you have always been there to support me. I feel incredibly fortunate to have you in my life. I love you all!

TABLE OF CONTENTS

List of Figures

List of Tables

INTRODUCTION

In the context of globalization, the rapid evolution of manufacturing processes is having a significant impact on industrial organization and working conditions. Modern production environments are characterized by a multitude of factors that directly influence operators' activity. These factors include work organization, managerial practices, production rate, as well as tools and the working environment. Each of these elements plays a crucial role in defining the tasks and conditions with which workers are confronted on a daily basis. However, this increased complexity and intensification of the pace of work can lead to significant health risks for workers. Among the most common physical disorders are Musculoskeletal Disorders (MSDs), which account for an alarming share of occupational illnesses. Indeed, according to statistics published by (Ameli, 2019), MSDs make up over 80% of reported occupational illnesses and rank as the second leading cause of sick leave in France. A detailed analysis of the 25 leading causes of "years lived with disability" highlights that work-related MSDs are significant contributors (Cieza et al., 2020). Furthermore, data from the Global Burden of Disease (GBD) study further reveals that an estimated 1.71 billion individuals globally are affected by MSDs. These disorders result from an interaction between the factors mentioned above and physical risk factors, and their prevalence highlights the importance of rethinking work practices and ergonomic conditions in order to improve workers' health and well-being.

To effectively address these constraints, it is essential to monitor the physical activity of operators in their work environment. This monitoring enables real-time assessment of the biomechanical constraints to which workers are subjected. By integrating this data into a more global approach to activity analysis, it becomes possible for ergonomists to design or modify workstations in such a way as to reduce the risk of MSDs. Such an approach involves not only the adjustment of equipment and procedures, but also in-depth consideration of work organization, in order to create optimal ergonomic conditions and promote workers' health and comfort.

In the context of monitoring operators' physical activity, Moovency (Moovency, 2024) has developed an objective methodology for measuring biomechanical constraints in the workplace using minimally invasive systems. Currently, Moovency offers systems like KIMEA to monitor operator movements and evaluate posture during work activities. The collected data is essential for conducting accurate ergonomic assessments; however, this postural analysis provides only a partial view of the biomechanical constraints experienced by operators. Indeed, it does not provide detailed information on specific physical solicitations, such as the forces exerted or generated by the operator. This dimension

is crucial for identifying the physical risk factors associated with particular efforts, but its assessment requires more sophisticated calculations and often relies on more invasive measurements, sometimes difficult to carry out on site. Moovency seeks to improve the existing method and aims, through this thesis, to develop a methodology for objectively measuring biomechanical quantities, such as joint torques. The goal is to quantify these quantities using on-site data collection without disrupting the operator's workflow. The central question of this research is: *How can we accurately estimate these quantities from limited data sources, such as depth cameras or RGB videos, which are constrained in terms of available degrees of freedom, have a lower sampling frequency, and may experience occlusions that introduce measurement noise, while avoiding reliance on additional effort measurements or oversimplified assumptions that could compromise biomechanical accuracy?* More specifically, *how can we preserve the complexity and richness of biomechanical models while estimating these quantities from limited, low-frequency data, where occlusion could occur*, and without the need to estimate external forces?

Traditional methods for estimating joint torques, include Inverse Dynamics(ID). Inverse method requires precise, high-frequency, low-noise data, which is typically obtained from opto-electronic systems. These systems involve multiple infrared cameras, skin markers, calibration, and extensive data processing. However, collecting such data in real industrial settings is challenging due to time constraints, space limitations, and the lengthy setup and processing times involved. Recent advances in affordable, markerless, and calibration-free sensors, such as Microsoft Kinect or RGB cameras, can not only provide real-time feedback to workers, but also offer an alternative to traditional motion capture systems. However, studies have shown that joint angles can be poorly estimated in certain situations, particularly in environments with occlusions or improper sensor placement (Plantard, Shum, & Multon, 2017). Furthermore, these methods often rely on simplifying assumptions (Plantard, Muller, et al., 2017), such as limiting the degrees of freedom in the biomechanical model or requiring floor detection to estimate ground reaction forces. In this thesis, we aim to overcome these limitations with Machine Learning (ML) and Deep Learning (DL)-based approaches that enable the use of a richer biomechanical model while relying on less explicit data.

To address the research questions, we first positioned our work scientifically by reviewing recent studies, as outlined in chapter 1. The second chapter 2 compares the accuracy and robustness of computer vision-based measurement systems for RULA assessment including those with one or more cameras, based on RGB or depth images, and using only vision information or coupled with a few wearable sensors (hybrid systems). The third chapter 3 evaluates different learning architectures designed to emulate the inverse dynamics step in motion analysis. In chapter 4, we assess the generalizability of DL-based tools, such as OpenCap (Uhlrich et al., 2023), in bimanual manipulation and picking tasks through fine-tuning—a widely used technique in DL for adapting models to new datasets.

Additionally, Chapter 5 provides a summary of the key contributions, limitations and perspectives to this work.

Chapter 1

Related work

1.1 General Context

This chapter reviews recent studies in the literature to highlight the gaps in current physical ergonomic and biomechanical assessments in industrial settings, ultimately addressing the key research questions of this thesis. In the first section of our literature review, we present the factors that contribute to the onset of musculoskeletal disorders. Then, we present the methods used to evaluate workstation ergonomics and posture measurement systems (part 1.1.1). In the second section, we discuss laboratory-based methods for estimating joint torques (part 1.1.2). In the third section, we explored a variety of machine learning and deep learning algorithms (part 1.1.3). Finally, in the last section, we present various approaches from the literature that aim to estimate these quantities in industrial environments (part 1.1.4).

1.1.1 Physical Ergonomics Analysis

MSD risk factors

The risk factors that contribute to the development of MSDs include professional factors related to external constraints induced by work, as well as non-professional factors, which encompass both intrinsic factors (such as age, height, and medical history) and extrinsic factors (such as lifestyle and life events) of individuals. These risks can manifest in various forms, including environmental, psychosocial, organizational, or even factors specific to each individual. With this distinction in mind, it is crucial to delve deeper into the analysis of non-professional risk factors and professional risk factors to better understand their respective contributions to the development of MSDs:

- **Non-professional risk factors** are distinguished between intrinsic and extrinsic factors. The former are intrinsic (such as age or gender (Yamalik, 2007)), which cannot be modified but must be considered in ergonomic evaluation. The latter are related to lifestyle factors outside of professional activity. The direct relationship between these factors and MSDs is not fully clarified (Bernard & Putz-Anderson, 1997).
- **Professional risk factors** refer to external constraints related to the operator's

activity. These are mainly divided into psychosocial and organizational factors, which include the work pace, production mode, as well as the worker's perception of their work. Moreover, the work environment also plays a role in the occurrence of MSDs, including elements such as thermal conditions, impacts, or mechanical pressures resulting from contact with objects. Several biomechanical factors play a critical role in the development of musculoskeletal disorders (MSDs). This thesis focuses specifically on biomechanical risk factors, which are primarily categorized into four key elements: **Repetitiveness**, **Force**, **Posture**, and **Recovery Periods**. **Repetitiveness** refers to the frequent, repeated use of the same body tissues, either through repetitive movements or by holding postures for extended periods. A task is considered repetitive if it lasts for at least one hour and involves cycles that are similar and relatively short (Silverstein et al., 1987). **Force** represents the effort required to perform an action or maintain a position. It can be external (applied force) or internal (force produced by the muscle) (Colombini et al., 2001). External force is measured using tools such as the manipulated weight or dynamometers, while internal force is estimated through methods such as electromyography (EMG), inverse dynamics, or the Borg scale. **Posture** refers to the configuration of body segments adopted by the operator during a task. Joint angles are generally used to assess the risk of MSDs, although there is no consensus on the anatomical zones to consider or the angular limits not to be exceeded (Hagberg, 1995). Finally, **Recovery Periods** represent the times during which the muscles stressed during a task, are at rest. They allow muscle tissues to return to their initial state after an effort (Colombini et al., 2001).

Ergonomic Analysis

To evaluate the ergonomics of workstations, various methods have been developed, falling into three main groups: **self-assessment**, **observation**, and **direct methods**.

* **Self-assessment** allow individuals to report their perceived effort and stress at work, usually via interviews or questionnaires. While they are easy and inexpensive, they may lack reliability and can be misinterpreted (Delépine et al., 2011). Scales like the Borg (Borg, 1998)(RPE (Ratings Perceived Exertion) (Borg, 1990) and CR10 (Category Ratio 10)) and Body Part Discomfort Scale (BPD) help quantify and evaluate discomfort and task-related difficulties.
* **Observation methods** involve trained observers assessing workstations and are practical and adaptable. Techniques like RULA (McAtamney & Corlett, 1993), REBA (Rapid Entire Body Assessment) (Hignett & McAtamney, 2000), and LUBA (Postural Loading on the Upper Body assessment) (Kee & Karwowski, 2001) evaluate postural risks, while OCRA (Occupational Repetitive Actions) (Occhipinti, 1998) and the NIOSH lifting equation (Valero et al., 2016) address repetitiveness

and lifting risks. The Strain Index method assesses wrist, hand, and forearm/elbow risks.

* **Direct methods** use devices such as goniometers, accelerometers, EMG, and dynamometers to collect data on posture, muscle activity, and strength. These methods are effective for research but can cause discomfort and disrupt work processes.

Posture Measurement Systems

They are specialized tools used to assess and quantify body posture. In this paragraph, we will present various categories of motion capture systems, including marker-based systems, sensor-based systems, RGB/RGB-D systems, and hybrid systems. Each of these categories offers unique advantages and challenges in terms of accuracy, ease of use, and suitability for different environments, making them valuable tools for both research and practical applications in posture analysis.

* **Marker-based systems** use opto-electronic cameras with infrared LEDs to track reflective markers on a subject. The cameras capture the reflected infrared light and use preprocessing to generate coordinates of the markers. Calibration involves intrinsic parameters (focal length, optical center, distortion) and extrinsic attributes (spatial arrangement). Commercial software like Vicon Nexus (Vicon, 2024) and Qualisys Track Manager (QTM) (Qualisys, 2024) are often used. Despite their advantages, marker-based systems are prone to artifacts caused by soft tissue movements, particularly at the thigh, which can lead to significant errors in joint angle predictions. Tissue artifacts can results in marker displacements of up to 2.5 cm, causing errors of up to 3° in knee joint angles (Benoit et al., 2015). Therefore, the placement of markers by different operators introduces variability ranging from 13 to 25 mm, potentially leading to errors in joint angle estimation of up to 10° (Gorton III et al., 2009). Errors in explicitly calculated joint positions using marker-based methods can reach up to 5 cm, contributing to joint angle errors of up to 3° in the lower limb (Leboeuf et al., 2019).
* **Inertial Measurement Units systems** IMUs are wearable sensors composed of an accelerometer (measuring linear acceleration), a gyroscope (measuring rotational speed), and a magnetometer (measuring the Earth's magnetic field orientation) to provide 3D orientation data. When combined with a skeletal model, they can be used to infer posture. Among the advantages of IMU systems, they are more affordable than marker-based systems and easier to use, as they do not require complex setups, calibration, or controlled environments. They can also function outside of controlled environments and operate in real time. Additionally, they require minimal storage and avoid issues like self-occlusion or gear occlusion, which are common in camera systems. However, IMU systems have certain limitations. They are susceptible to drift over time, necessitating frequent recalibration

(Brodie et al., 2008; S. Kim & Nussbaum, 2013; Lebel et al., 2013; Oliveira et al., 2022; Plamondon et al., 2007). They are sensitive to ferromagnetic disturbances and have limited accuracy in rotational movements outside flexion/extension, with errors exceeding 5° for most motionsAzure (Rekant et al., 2022). IMU systems, such as APDM (APDM, 2024) and Xsens (Xsens, 2024), are integrated into OpenSim software (Al Borno et al., 2022).

* **RGB/RGB-D systems** RGB-D systems, such as the Kinect Azure (SDK, 2024), use infrared light to provide depth information, delivering comprehensive 3D data. While earlier depth sensors faced significant challenges in direct sunlight (Bhoi, 2019), struggled with distances beyond 5 meters, and often operated at lower frame rates (Pagliari & Pinto, 2015), advancements in hardware have significantly mitigated these limitations. Modern systems, such as Intel RealSense cameras, demonstrate improved performance in bright environments (Corporation, 2019). Monocular RGB systems are valued for their simplicity and cost-effectiveness, but they often fall short in depth perception and spatial coverage, which can lead to missed crucial details (Masoumian et al., 2022). They are sensitive to changes in lighting and may struggle with occlusions, making accurate pose estimation difficult in dynamic scenes (X. Zhang et al., 2023). In contrast, multi-RGB systems capture a wider range of joint angles and leverage advanced deep learning algorithms, providing a more comprehensive and accurate analysis of human movement (Ahn et al., 2023). These systems require Human Pose Estimation (HPE) algorithms to process the visual data and infer human poses (Khan et al., 2020; Liang et al., 2023). Some widely used HPE algorithms include traditional methods such as OpenPose (Cao et al., 2017) and recent deep learning-based models like HRNet (High-Resolution Network) (Sun et al., 2019) and AlphaPose (Fang et al., 2022). OpenPose, known for its efficiency and real-time capabilities, uses a bottom-up approach, which means its computational cost does not increase with more people detected. In this approach, all joint keypoints are found first, and then matched to the correct person. This is different from the top-down approach, where the system first finds the people in bounding boxes and then looks for the joint keypoints inside them. However, it is limited by its reliance on 2D keypoints and struggles with occlusions and extreme poses (Khirodkar et al., 2021). HRNet offers higher accuracy, particularly in complex poses, but can be computationally expensive. AlphaPose achieves high accuracy and robust performance in crowded scenes but is sensitive to image quality and lighting conditions. Despite the strengths of these algorithms, challenges remain in terms of processing speed, handling occlusions, and achieving consistent performance under varying environmental conditions. Since RGB/RGB-D systems rely on pose estimation algorithms designed for specific tasks and populations, the challenge of applying them to new scenarios has yet to be addressed.

* **Hybrid systems**, such as KIMEA (Moovency, 2024) and VIMU (Vision Inertial Measurement Unit) (Adjel et al., 2023), combine the strengths of computer vision and inertial measurement unit (IMU) data to leverage the complementary advantages of both measurement modalities. The KIMEA system integrates IMUs into gloves, enabling more accurate and reliable tracking of wrist joint angles compared to relying solely on computer vision. This is especially crucial as computer vision, while powerful, often struggles to precisely track small body parts, particularly those that are frequently occluded during dynamic movements in a worker's activity. By fusing these data sources, hybrid systems can overcome such limitations and provide more robust and dependable results in real-world applications. However, since these systems also rely on HPE algorithms that were trained on specific tasks and populations, the issue of their ability to generalize to new scenarios remains unresolved.

1.1.2 Estimation of Biomechanical Quantities in the Laboratory

In the processing pipeline of musculoskeletal analysis through inverse methods, as illustrated in the figure 1.1, the process begins with the acquisition of experimental measurements, including motion capture data and external forces.The motion data capture system relies on an optoelectronic reference system. To reconstruct these positions and orientations, markers are attached to specific anatomical points on the participant's limbs, and their positions are recorded. Each segment is associated with at least three markers placed on the corresponding limb. Using the positions of these markers, an anatomical reference frame unique to the limb can be constructed (Wu et al., 2002; Wu et al., 2005). The position and orientation of the limb are tracked by calculating the position and orientation of the associated anatomical reference frame. These reflective markers are detected using an array of infrared cameras. External forces, such as ground reaction forces (GRF&M) and load contact forces (LCF&M), are recorded using force plates or sensors. To describe the mechanical behavior of the human body, it is represented by a musculoskeletal model composed of a set of rigid bodies driven by muscles action. To study the kinematics and dynamics of rigid segments, an osteoarticular model is used to represent the influence of muscles on joint movements. Each rigid segment is associated with a reference frame and is connected to other segments through joints. The movement of these segments is described using joint coordinates, which vary in number depending on the total number of segments and the modeling approach applied to each joint.

In the figure 1.1, we distinguish four important phases: the experimental motion capture data enables obtaining joint angles (kinematic variables) following IK (using an osteoarticular model). Subsequently, these joint angles, combined with measurements of external forces, allow obtaining joint torques (dynamic variables) following ID (using an

osteoarticular model and an inertial model). In this thesis, we are primarily focused on the estimation of joint torques.

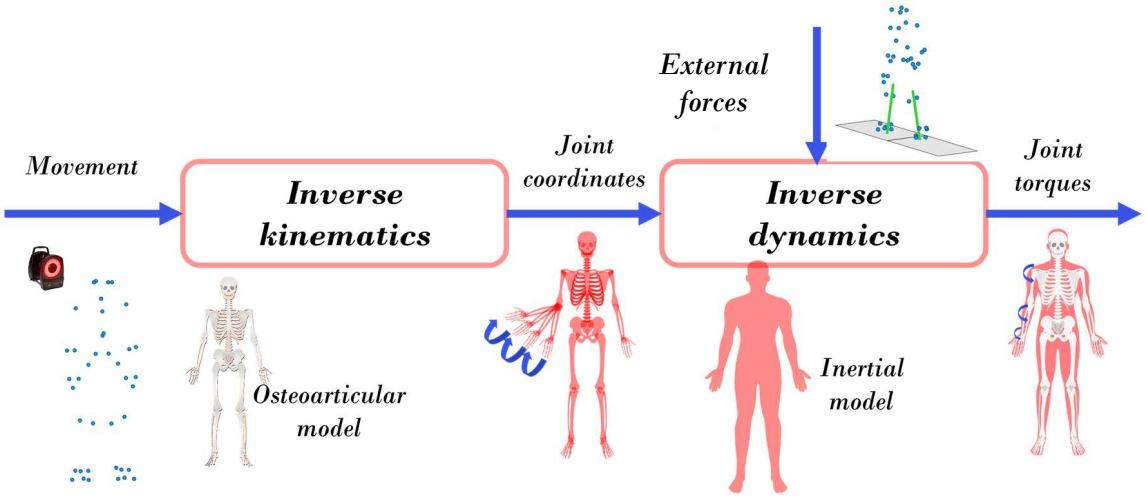


Figure 1.1 – Movement analysis processing pipeline using inverse dynamics (inspired by Claire Livet's thesis (Livet, 2022)).

Inverse Methods and Biomechanical Variables

Inverse kinematics involves determining a model's joint configuration from its position in Cartesian space, specifically identifying the vector of joint angles corresponding to the experimental marker positions at any given moment (Begon et al., 2018). Joint angles represent the angular positions of various joints, which characterize the posture and movement of the limbs and body. In contrast, inverse dynamics aims to estimate the joint torques associated with movement, by determining external forces and motion parameters during a subject's movement. Joint torques quantify the total moments acting on a joint, resulting from the contributions of muscles and ligament.

Building on the demonstration from (Featherstone, 2014), the general formulation of inverse dynamics for a biomechanical model can be expressed as follows:

$$\tau = M(\mathbf{q})\ddot{\mathbf{q}} + C(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) \tag{1.1}$$

where $M(\mathbf{q})$ denotes the joint-space inertia matrix, $C(\mathbf{q}, \dot{\mathbf{q}})$ represents the joint-space Coriolis and centrifugal matrix, $\mathbf{g}(\mathbf{q})$ is the joint-space gravitational vector, and τ denotes the unknown joint torques.

Several software programs specialized in motion analysis, such as Anybody (Damsgaard et al., 2006), OpenSim (Delp et al., 2007), and CusToM (Muller, Pontonnier, Puchaud, et al., 2019), use the inverse dynamics method. Joint torques τ can be computed either using the recursive Newton-Euler algorithm (Featherstone, 2014). The Newton-

Euler algorithm (Featherstone, 2014) is broken down into three steps: calculating joint velocities and accelerations, determining the acceleration vector, and calculating the transmitted forces by isolating each segment along the kinematic chain. However, this method has the drawback of propagating errors along the chain and is unable to handle structures with closed loops, such as the shoulder.

Why Movement Dynamics Are Essential to Physical Ergonomics Analysis?

As mentioned in paragraph 1.1.1, the prevention of musculoskeletal disorders requires considering numerous factors, particularly biomechanical constraints. Monitoring the operator's physical activity in their work environment is a means to evaluate these biomechanical constraints and integrate them into a comprehensive activity analysis approach aimed at designing or adapting work situations. Analyses focusing on levels of postural demands are particularly useful for assessing the operator's actual activity in real work conditions. However, such analyses offer a limited view of biomechanical constraints, as they do not assess the physical demands—namely the forces experienced or exerted by the operator. Evaluating these demands is essential for gaining a deeper understanding of the physical risk factors associated with exertion. By leveraging precise kinematic data and external force measurements, inverse dynamics enables ergonomists to estimate joint torques (Plantard, Muller, et al., 2017). In this thesis, we focus on estimating joint torques, which are considered crucial complement to physical ergonomic analyses.

1.1.3 Learning-Based Approaches: Machine Learning (ML), Deep Learning (DL), and Fine-tuning

Learning-based approaches involve training algorithms on datasets to identify patterns, approximate functions, or make predictions (Alpaydin, 2020; Goodfellow et al., 2016; Mitchell & Mitchell, 1997). The algorithms are trained on datasets to optimize performance for specific tasks—often without explicit programming for those tasks (Hardt & Recht, 2021). Depending on the nature of the input data and the desired outcomes, learning-based approaches can be trained using supervised, unsupervised, semi-supervised, or reinforcement learning techniques (Sindhu Meena & Suriya, 2020). To better understand the principles underlying ML and DL, we will first present their core algorithms, focusing on their applications and essential mechanisms, along with their limitations. Next, we will discuss the concept of fine-tuning, a key technique for adapting pre-trained models to specific tasks.

* **Machine Learning (ML)** involves learning algorithms that identify meaningful patterns and features from training data to make predictions. In ML, we distinguish between the following training modes: **Supervised Learning**, where models are trained on labeled data (e.g., regression, classification); **Unsupervised Learning**,

where models uncover patterns in unlabeled data (e.g., clustering, dimensionality reduction) (Russell & Norvig, 2016); and **Reinforcement Learning**, where agents learn decision-making strategies by interacting with an environment and receiving feedback in the form of rewards (Sutton, 2018). Among the key ML models, we can mention the following:

- **Decision Trees** are non-linear models that recursively split data based on feature values to construct a tree structure. The splitting criteria create decision rules that are easy to interpret. While decision trees are interpretable and computationally efficient, they are prone to overfitting when the tree becomes too deep (Breiman, 2017; Quinlan, 1986).
- Random Forests build upon decision trees by creating an ensemble of them, using bootstrapped data samples and random subsets of features at each split. This ensemble method reduces overfitting by averaging predictions across multiple trees, improving accuracy and robustness. However, while random forests mitigate overfitting, they are less interpretable than individual decision trees and can become computationally expensive when working with large datasets (Breiman, 2001; Liaw, 2002).
- **Linear Regression** is used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of squared differences between the observed and predicted values. While simple and interpretable, it is sensitive to issues like multicollinearity, outliers, and overfitting (Hastie, 2009; Hoerl & Kennard, 1970; Seber & Lee, 2012). Ridge regression, addresses some of these issues by adding a regularization term, which shrinks the coefficients, preventing overfitting, particularly in high-dimensional spaces. However, both methods are limited to capturing linear relationships.

- **Deep Learning (DL)** algorithms excel at processing unstructured data, including images, text, and audio (Dixon et al., 2019; Sapoval et al., 2022). Despite their remarkable performance across various domains, DL models face significant challenges, particularly their reliance on large amounts of labeled data and substantial computational resources during training (Sarker, 2021). They are often criticized for being "black boxes"—a term underscoring their lack of transparency and interpretability compared to traditional machine learning approaches (Qamar & Bawany, 2023). It is well-known that, in addition to requiring training data that is representative in terms of probability density (Goswami, 2020), the architecture of the network itself plays a crucial role in shaping the quality of the learned deep representations (Sarker, 2021). In this thesis, we primarily focus on the extraction of spatio-temporal features from training data. In the following section, we

will present the various existing architectures used for capturing spatio-temporal features.

* **Learning Spatial Features** Spatial features refer to the patterns in data that are related to the position, structure, and arrangement of elements within a spatial domain (Mourot et al., 2022b). In images, these features can include edges, textures, shapes, or object parts. The spatial arrangement of these features (i.e., where they appear in the image) helps define what the image represents. In videos, spatial features might be objects or motion patterns occurring in specific areas within a sequence of frames. In skeleton data (which represents human pose or keypoints corresponding to body joints), spatial features describe the relationships between joints and their positions in space, including their relative positions and arrangements over time (Mourot et al., 2022b). When it comes to learning spatial features, Convolutional Neural Networks (CNNs) play a key role (Krizhevsky et al., 2012; LeCun et al., 1998), they are widely used in computer vision tasks such as image classification (Litjens et al., 2017; G. Xu et al., 2024), object detection, semantic segmentation (Hatamizadeh, 2020), and face recognition, as well as in video analysis and medical imaging. A CNN consists of multiple layers, each performing different operations to extract relevant features. The convolutional layer is the core of the network, where a convolution operation is applied to the input data using kernels that slide across the data to generate feature maps. The pooling layer reduces the spatial dimensions (height and width) of the feature maps, while preserving important information. Despite their impressive performance in the computer vision domain, CNNs have inherent limitations. They require vast amounts of labeled data to train effectively and often demand significant computational power, especially when dealing with high-dimensional data like images and videos (Dagès et al., 2023). Moreover, like other deep learning models, CNNs are often referred to as "black-box" models, as it is challenging to interpret the learned features and understand how the network makes decisions. This lack of transparency remains a significant challenge (Simonyan & Zisserman, 2014).
* **Learning Temporal Features** Temporal features refer to patterns of data that are related to the time dimension. They are used in models analyzing sequential or time-series data, where the value of a feature at a particular time step is dependent on previous of future values (Mourot et al., 2022b). These features are essential for capturing the dynamics and trends over time in various applications such as forecasting, classification, natural language processing (NLP) or speech recognition (Ismail Fawaz et al., 2019). To capture these temporal features, certain specific architectures are used: **Long Short-Term Memory (LSTM)** are suited for tasks where past events influence future ones,

such as predicting the next word in a sentence or recognizing a pattern in a time-series dataset. However, LSTMs can struggle with very long sequences and may require a lot of computational resources to train effectively, especially as the length of the sequence grows (Hochreiter, 1997). **Bidirectional LSTM (BiLSTM)** improves upon standard LSTM by processing the data in both forward and backward directions. This allows the model to capture context from both past and future events, which is particularly useful in NLP tasks like machine translation or speech recognition. The major limitation of BiLSTMs is the increased computational burden, as the model needs to process the data in two directions (Graves & Schmidhuber, 2005). **Attention mechanisms** often combined with LSTMs or other models like Transformers, allow the model to focus on different parts of the input sequence, giving it a way to selectively weight the importance of certain temporal features.

* **Fine-tuning** is part of the Learn From Model (LFM) paradigm, which focuses on adjusting the parameters of pre-trained models for adaptation to downstream tasks (H. Zheng et al., 2023)— the tasks that the pre-trained model will be adapted or fine-tuned for, after it has already been trained on some initial data. These tasks could include things like classification, regression, question answering, sentiment analysis, etc., depending on the model and the problem at hand. In contrast, to retraining a model entirely on new task-specific data (Learn From Data), the primary advantage of model tuning through transfer learning is its ability to mitigate the risks associated with limited datasets and high training costs (Hinton, 2015; Pan & Yang, 2009). This approach leverages the general knowledge embedded in the pre-trained model to initialize parameters for new tasks. Successful fine-tuning requires a deep understanding of the internal structure and behavior of the pre-trained model, including how it encodes input data and which components have the most influence on predicted outcomes. In the case of *weight tuning*—refers to the process of adjusting the weights (parameters) of a pre-trained model to improve its performance on a specific downstream task (Donahue et al., 2014), fine-tuning is a specific transfer learning technique that applies knowledge from pre-trained neural networks to solve new, relevant tasks. For example, in computer vision, fine-tuning is often used to adapt pre-trained CNNs for tasks such as image classification, object detection, and facial recognition (He et al., 2016; Sharif Razavian et al., 2014). This process involves adjusting the parameters of a pre-trained model (e.g., ResNet (He et al., 2016), VGG (Simonyan & Zisserman, 2014)) by continuing training on a smaller, task-specific dataset. By leveraging the insights gained from large, diverse datasets, fine-tuning enables the model to recognize new patterns and improve its performance on downstream tasks. One common strategy in fine-tuning is to freeze the weights of earlier layers while adjusting the weights

of the later layers (Donahue et al., 2014). The rationale behind this approach is that early layers often learn basic, transferable features—such as edges in images or basic linguistic structures in text—that are useful across many tasks (Yosinski et al., 2014). In contrast, later layers tend to capture more task-specific features, making their adjustment more critical for successful fine-tuning. However, despite its many advantages, fine-tuning has limitations, for instance, if the fine-tuning dataset does not accurately represent the target domain, there is a risk of overfitting (Caruana, 1997). Additionally, fine-tuning can lead to catastrophic forgetting, where the model loses the general knowledge it gained during pre-training (Z. Li & Hoiem, 2017). This issue becomes more pronounced when too many layers are frozen, limiting the model's ability to retain valuable information from the original large dataset. Finally, since pre-trained models may inherit biases from the data they were trained on, fine-tuning can propagate these biases if the new task is not sufficiently diverse, potentially leading to biased predictions or decisions (Bolukbasi et al., 2016).

1.1.4 Estimation of Biomechanical Quantities in an Industrial Environment

In the state of the art, to the best of our knowledge, we observe two main trends. Some studies focus on the information extracted from a RGB-D/RGB systems by correcting/enriching the extracted skeleton in order to estimate joint torques in real-time. The following study (Plantard, Muller, et al., 2017) emphasizes that the Kinect (RGB-D system), when combined with an adapted inverse dynamics method, provides reliable joint torque estimates that can be used for practical, on-site ergonomic assessments. RGB-D and RGB systems use a variety of deep learning and computer vision algorithms (as discussed in paragraph 1.1.1) to process data and predict human body posture. However, these algorithms are often trained on specific datasets, which can limit their generalizability to other measurements, tasks, or populations beyond the scope of the training data. Furthermore, the resulting 3D keypoints are typically sparse, failing to comprehensively capture the movements of all body segments. This limitation may lead to an incomplete representation of the body's full range of translations and rotations, raising concerns about the expressiveness and accuracy of these keypoints for detailed movement analysis. To address these two limitations, in this study (Uhlrich et al., 2023), two long short-term memory (LSTM) networks were trained (for more details about the architecture, see the paragraph 1.1.3) to map joint centers to precise anatomical markers. The set of anatomical markers corresponds to those typically used in marker-based motion capture, which are robust for determining 3D joint kinematics (Falisse, Uhlrich, et al., 2024). The tool developed is called OpenCap. It is an open-source web tool designed to compute kinemat-

ics and dynamics from two or more smartphone videos. It uses HPE algorithms, such as OpenPose and HRNet, to detect 2D positions of body keypoints in video frames. These 2D keypoints are then triangulated to estimate their corresponding 3D positions. Once the 3D pose is determined, two learning algorithms predict the locations of anatomical markers based on the 3D keypoints from the video. These predicted markers are subsequently used as inputs in a musculoskeletal analysis pipeline using inverse methods. The learning algorithms were mainly trained and tested on walking and running. However, these learning algorithms rely on specific training datasets, which restricts its ability to generalize across a broader range of movements. Although the development of a more extensive dataset by (Falisse, Uhlrich, et al., 2024) seeks to address some of these limitations, a critical gap remains in capturing the nuances of motions beyond the scope of the training data. In the same context, combining markerless motion capture techniques (e.g., pose detection) with scaled musculoskeletal models, we find the studies by (Pagnon et al., 2021, 2022a, 2022b) that introduced Pose2Sim, a markerless kinematics workflow designed to capture the movements of elite athletes in real-world environments. This workflow focuses on three physical activities: walking, running, and cycling. It uses four camera perspectives and OpenPose to initially detect 2D video keypoints. After camera calibration and subject identification, the 2D keypoints are triangulated to obtain 3D keypoints. These 3D keypoints are then filtered, and inverse kinematics (IK) are applied to estimate joint angles using a full-body OpenSim model (Delp et al., 2007; Seth et al., 2018). Pose2Sim has been tested in real-world scenarios, and its inverse kinematics results have demonstrated high accuracy, with mean absolute errors typically under 4° when compared to traditional marker-based methods. These approaches still rely on solving an inverse kinematics problem, using keypoints from pose detection algorithms rather than optical markers.

Other studies aim to address or emulate inverse dynamics using learning algorithms. (Aghazadeh et al., 2020) applied a neural network to predict moments at the L5-S1 joint during static load-handling tasks. The training dataset included individuals performing both symmetric and asymmetric load-handling tasks with varying weights and load positions. The algorithm tested was a neural network (for further details on ML and deep learning, see paragraph 1.1.3). The neural network predicted the lumbosacral (L5-S1) moment based on hand-load magnitude, 3D position, body height, and body weight. After outlier removal and input data normalization (Mohseni et al., 2022), the Root Mean Square Error (RMSE) decreased by 29%. However, their study mainly focused on a limited set of variables and specific tasks, raising concerns about the generalizability of their model to different dynamic movements, tasks and populations. In their work, (Zell & Rosenhahn, 2017) developed a learning-based algorithm to solve the inverse dynamics problem in human motion. Their method used Random Forest regression to predict joint torques and ground reaction forces (GRFs) from motion data, specifically joint coordinates

and velocities. The approach extends incomplete force plate data by using Random Forest to predict force vectors and then estimates joint torques through a modified, physics-based predictive dynamics model. Their results were evaluated by comparing them to state-of-the-art methods and measured force plate data, demonstrating robustness against noisy and incomplete inputs. The approach was tested on a dataset containing walking and running motions. However, the study's limitation lies in its reliance on high-quality training data that is generated with specific contact property constraints, which limits its flexibility. Subsequently, (Zell & Rosenhahn, 2020) compared several learning-based approaches—including neural networks, Random Forests, and ridge regression (for more details about the architectures, see the paragraph 1.1.3)—to predict the same outputs from the same input variables used in the previous study. While their results were promising, they were limited to controlled activities, which restricts the model's ability to generalize to more complex and unstructured movements. In a later study, (Zell et al., 2020) introduced a new approach that reduced the need for supervision by incorporating forward dynamics optimization and inverse dynamics optimization into their joint torque and GRF prediction pipeline.

1.1.5 A Critical Review of Current Solutions

In the previous sections (see paragraphs 1.1.2 and 1.1.4), we discussed the methods commonly used to estimate joint torques in both laboratory and industrial settings. While these approaches provide valuable insights, they also present several limitations that we aim to address in this work.

First, the literature highlights several studies that have validated the application of motion capture systems in the field of ergonomics. These systems have been used to evaluate workplace postures and assess ergonomic risk factors, in both controlled and field settings. For example, (Menolotto et al., 2020) conducted a systematic review of motion capture technology, emphasizing its industrial applications and potential to improve ergonomic assessments, while (Humadi et al., 2021) compared the effectiveness of Inertial Measurement Units and Kinect V2 for in-field ergonomic risk assessments, focusing on their precision and adaptability to real-world conditions. However, the majority of these studies have focused on earlier technologies, particularly systems like the now-obsolete Kinect V2 depth camera (Manghisi et al., 2017; Plantard, H. Shum, et al., 2017). While recent research has begun to explore and validate the use of monocular (L. Li et al., 2020; McKinnon et al., 2022; Nayak & Kim, 2021; Yuan & Zhou, 2023) or multi-camera RGB systems (W. Kim et al., 2021) for postural assessment, these studies have largely evaluated these systems in isolated conditions. To the best of our knowledge, no study has compared different motion capture systems—such as monocular depth systems, RGB systems, and hybrid systems that integrate IMUs with monocular and RGB systems,

like KIMEA (Moovency, 2024) and VIMU (Adjel et al., 2023)—under simulated working conditions, limiting their practical relevance for industrial applications.

Recent advancements in human pose estimation (HPE) algorithms, driven by deep learning (DL), have significantly enhanced the accuracy of extracting skeletal data from images and videos (see paragraph 1.1.1), allowing for more detailed analyses of human movement and posture (C. Zheng et al., 2023). However, many current studies on joint torque estimation still rely on datasets that do not include skeletal data (Aghazadeh et al., 2020; Mohseni et al., 2022; Zell & Rosenhahn, 2017; Zell et al., 2020), despite the fact that such data can be easily obtained using depth and RGB cameras. Furthermore, these studies primarily focus on human locomotion and static manipulation movements, leaving dynamic activities such as lifting tasks insufficiently explored. This underscores the need to evaluate the integration of skeletal data as inputs for learning algorithms to estimate joint torques during both the static and dynamic phases of lifting tasks.

Third, most current deep learning-based approaches for estimating kinematics and dynamics from videos primarily concentrate on human locomotion movements, such as walking and running (Pagnon et al., 2021; Uhlrich et al., 2023). Furthermore, OpenCap, although effective to some extent, is constrained by the types of movements included in its training data. Consequently, its application to real-world tasks—especially those involving complex bimanual activities, such as industrial work—remains limited.

1.2 Chapter Conclusion and Present Contributions

Given the limitations highlighted in section 1.1.5, this thesis contributes to the in-situ biomechanical and ergonomic analysis of workstations by evaluating learning algorithms for estimating joint torques from limited and low-frequency data collected in industrial environments.

The current chapter (**Chapter 1**) presented a review of risk assessment methods from the literature, highlighting both laboratory-based techniques and those used in industrial settings. This overview establishes the context for the subsequent chapters, illustrating the current state of knowledge and identifying gaps that this thesis aims to address (see the section 1.1.5).

Chapter 2 compares the accuracy and robustness of computer vision-based measurement systems for Rapid Upper Limb Assessment (RULA), particularly concerning the needs of on-site physical ergonomic evaluation. The study was evaluated on simulated work conditions. It focuses on evaluating the performance of different types of computer vision-based systems, including those with one or more cameras, based on RGB or depth images, and using only vision information or coupled with a few wearable sensors (hybrid systems).

Chapter 3 benchmarks various learning algorithms aimed at estimating upper limb

joint torque during both static and dynamic phases of lifting tasks. It introduces a generalized torque estimation learning model that enhances the accuracy and adaptability of torque predictions by accounting for variability across different subjects and tasks. This contributes to the understanding of how learning algorithms can be used to emulate inverse dynamics step in motion analysis by learning, without the use of external effort measurements.

In **Chapter 4**, we focus on evaluating and enhancing the generalizability of OpenCap's learning algorithms in the context of bimanual manipulation and picking tasks. To achieve this, we evaluate various fine-tuning strategies to adapt the pre-trained learning models to new markersets and motion. Through this approach, we aim to enhance the learning model's generalizability in industrial settings scenarios, making these algorithms more robust and adaptable to a wider range of markersets and motion.

Chapter 2

Comparison of Computer Vision-Based Motion Capture Systems for Ergonomic Postural Assessment in Work Conditions

2.1 Introduction

As seen in the section 1.1.1, considering a worker's posture and movement is important for assessing the risks of development of musculoskeletal injuries in the workplace (see section). Risk factors can directly contribute to the onset of musculoskeletal disorders, act as triggers, or create conditions conducive to the progression of the pathology. These risk factors are categorized into professional origins, representing external constraints induced by the worker's activity, and non-professional origins, defining the individual's capacity to respond to these constraints. Both types of risk factors are closely interconnected (Bernard, 1997). Within professional risk factors, several biomechanical constraints play a crucial role in the development of musculoskeletal disorders. Exposure to biomechanical risk factors, including force, posture, and repetition, along with individual factors affecting the worker, increases the risk of work-related musculoskeletal disorders (WMSDs) (as highlighted in the section 1.1.1). Self-assessment, direct measurement, and observational techniques (David, 2005; G. Li & Buckle, 1999) are common methods to assess the risk of WMSDs (as explained in section 1.1.1).

Self-assessment methods can take many forms, such as scales, questionnaires or interviews. This type of method focuses on assessing physical workload, perceived discomfort, or work-related stress, which are difficult to measure objectively. Therefore, although this type of method is easy to use, it is not sufficiently reliable and can lead to erroneous interpretations (Burdorf & Laan, 1991; Wiktorin et al., 1993). Observational methods, such as the RULA method (McAtamney & Corlett, 1993), involve directly evaluating the performance of the worker at the workstation. The accuracy and validity of results obtained using these observational methods depend directly on the input information (Fagarasanu & Kumar, 2002). However, data collection is generally achieved by subjective observation

or simple estimation of angles projected in videos/photos, leading to low accuracy and high inter- and intra-observer variability (Burdorf et al., 1992; Lowe, 2004b). Indeed, when using RULA assessment grid, approximately one out of three assessments conducted by practitioners in actual work situations does not adequately evaluate the level of potential WMSD (Diego-Mas et al., 2017). In a previous study (Robertson et al., 2009), the RULA grand score leaded to only "fair" inter-rater reliability (ICC<0.5) among four trained raters. Moreover, (Dockrell et al., 2012) showed that intra-rater reliability was stronger than inter-rater reliability, suggesting that assessments should ideally be conducted by the same person. This can be challenging for companies with production sites located in different geographical areas. To mitigate this problem, direct methods consist in using sensors to estimate the human body poses. The recent commercial solutions could be divided into two main families: either based on wearable sensors (sensor-based), or based on cameras (computer vision-based). Sensor-based systems generally use goniometric, magnetic, and inertial sensors. Inertial sensors have become very popular, and proposed in many commercial solutions. They are the most cost-effective and can be used in wide spaces with on-board recording or wide wireless communication. Nowadays, inertial measurement units (IMUs) are the most popular systems, as they fuse information from three different sensors (accelerometers, gyroscopes, and magnetometers) to estimate more robust and reliable joint angles. However, these wearable sensors are difficult to implement in a real work situation (G. Li & Buckle, 1999; Shiao et al., 2024) due to many practical factors, including discomfort and the fact that they may influence the posture (David, 2005; Sibson et al., 2024; Zhao et al., 2021).

The workers may have to stop their activity to put and calibrate the sensors. Incorrect placement of sensors on the worker can also lead to important measurement errors (Caputo et al., 2019; Niswander et al., 2020). The calibration process is also a sensitive step that can lead to an increase in measurement errors (Poddar et al., 2016). Moreover, it may not always be compatible with professional equipment or could potentially interfere with the execution of professional tasks, particularly if the subject is wearing clothes under the sensor (Plamondon et al., 2007). Inertial sensors may also be subject to drift, especially when exposed to magnetic field disturbances (Yunus et al., 2021) during long tasks (Robert-Lachaine et al., 2017). The magnetic disturbance error can be reduced with advanced signal processing methods (Roetenberg et al., 2005) or with additional sensors such as a potentiometer. However, recent advancements in machine learning algorithms have shown potential in predicting both kinematics and dynamics from partial or noisy sensor data, thereby reducing the dependency on perfect sensor placement and calibration (Lawson et al., 2024; Long et al., 2024; Moghadam et al., 2023).

Computer vision-based methods cover all technologies that use images captured by one or more color (RGB) or infrared cameras to estimate movements using computer vision methods. Historically, marker-based methods (mostly based on passe reflective markers or

active LED) can be used to help the system to track anatomical landmarks are considered as the most accurate motion capture system, especially when using infrared cameras. However, these systems are difficult to use in a work environment: they are very expensive and require a set of markers to be placed on the body, a large number of cameras to be placed in the space, the capture space to be calibrated, exposure to natural light to be avoided, etc. With the recent rapid development of machine learning and deep learning, markerless systems have gained interest. Depth sensors based on RGB-D cameras, such as Microsoft Kinect, associated with machine learning algorithm can track the human skeleton without the need of placing markers on the body (Shotton et al., 2011). This approach has been recently improved using deep learning, to enhance accuracy with a single-view depth camera (Orbbec, 2024), and algorithms to handle the transition from 2D to 3D posture (SDK, 2024). Deep learning (Cao et al., 2017) has boosted computer vision performance, enabling the tracking of human poses with a single camera, without using markers (Pavllo et al., 2019). The THEIA system (THEIA, 2024) applied this approach to multiple calibrated RGB cameras to enhance the accuracy of markerless human pose estimation.

Hybrid systems, such as the KIMEA (Moovency, 2024) or Visual-Inertial Measurement Units (VIMUs) (Adjel et al., 2023) systems, fuse computer vision and IMU data, to benefit from the advantages of both measurements. The KIMEA system is a hybrid setup consisting of a depth sensor and IMU devices positioned on the hands, while the KIMEA Cloud system integrates a single RGB camera with IMUs embedded in gloves. The KIMEA or KIMEA Cloud systems propose to place IMUs in gloves to obtain more reliable joint angles for the wrist than using computer vision techniques only. Indeed, computer vision generally fails to accurately track small body parts which are generally occluded during the worker activity.

The main advantage of computer vision-based (or eventually hybrid-based) methods remains their ease of use and the low perturbations for the worker and the production line (Needham et al., 2021). Nevertheless, the accuracy and robustness of pose estimations is subject to many factors, such as the nature of the learning algorithm, the training datasets, alongside considerations like background dependence, the presence of multiple workers within the scene, and occlusions (Plantard et al., 2015).

As mentioned in section 1.1.5, previous studies have highlighted the potential of motion capture systems in ergonomics, initially concentrating on older technologies such as the Kinect V2 (Manghisi et al., 2017; Plantard, H. Shum, et al., 2017). More recent advancements have investigated the use of monocular and multi-camera RGB systems for posture analysis (W. Kim et al., 2021; L. Li et al., 2020; McKinnon et al., 2022; Nayak & Kim, 2021; Yuan & Zhou, 2023). Although these previous works provide interesting evaluation of isolated systems, up to our knowledge, they have not been compared in the context of realistic work task condition.

In this chapter, we aim to analyze the accuracy and robustness of various computer vision and hybrid measurement systems under simulated working conditions. The systems evaluated include: THEIA (a markerless system using multiple RGB cameras), Microsoft's Kinect Azure DK (a single depth camera), the KIMEA system (a hybrid system combining a depth sensor with IMUs placed on the hands), and the KIMEA Cloud system (a hybrid setup consisting of a single RGB camera and IMUs embedded in gloves). The XSens inertial motion capture system (Roetenberg et al., 2009) is used as the reference standard. The study detailed in this chapter has been submitted to the international journal of industrial ergonomics (See 4.5). Section 2.2 introduces the methodology used to evaluate these systems. Section 2.3 presents the results of this evaluation. These results are discussed in section 2.4, with some perspectives.

2.2 Materials and methods

This section describes the experimental protocol designed to evaluate the accuracy and robustness of different motion capture systems in simulated work conditions. The aim of the experiment is to compare different pose estimation systems using computer vision techniques:

— the THEIA system with multiple calibrated RGB cameras,
— the Microsoft's Kinect Azure DK based on a single depth camera,
— the KIMEA system composed of a single depth camera and IMUs sensors on the hands,
— and the KIMEA Cloud system composed of a single RGB camera and IMUs placed on the hands.

These tests were carried-out in a context of postural assessment in work conditions. The specifications of each tested motion capture system are summarized in Table 2.1.

Motion capture systems	Point of view	Image type	Hybrid/Vision only
THEIA	**multi-cam**	**RGB**	**Vision Only**
Kinect Azure DK	**monocular**	**Depth**	**Vision only**
KIMEA	**monocular**	**Depth**	**Hybrid**
KIMEA Cloud	**monocular**	**RGB**	**Hybrid**

Table 2.1 – Motion capture systems tested in this chapter, including monocular video, multi-cameras and hybrid systems. These systems use RGB or depth (RGB-D) images as input data. Some of the systems are hybrid, combining video and sparse inertial sensors.

In this study, we focus on two main scientific objectives. First, we assess the accuracy and robustness of vision systems, including monocular systems (such as Kinect Azure DK), multi-camera systems, and hybrid systems combining depth and RGB cameras with

IMU sensors. A key aspect of this evaluation is exploring whether adding IMU sensors to areas often obscured in cluttered environments can improve the reliability of wrist joint angle measurements, especially when combined with monocular depth or RGB cameras, compared to using vision systems alone. Second, we investigate the robustness of vision and hybrid systems under different camera placements. In real-world settings, it is challenging to place cameras freely due to cluttered environments. To address this, we simulate such conditions in a controlled laboratory environment, where reference systems are available. This setup allows us to quantify joint angle measurement errors and assess the corresponding RULA scores, which are commonly used in ergonomics evaluations.

2.2.1 Joint angles estimation and RULA scores computation

The RULA assessment grid requires relevant joint angles computed on standardized 3D points selected on a skeleton description. Hence, a posture can be defined as $p = \{(x_j, y_j, z_j) \mid j = 1, 2, \ldots, N\}$, where N stands for the number of joints in the posture, and (x_j, y_j, z_j) stand for the 3D Cartesian coordinates of the jth joint. Joint angles are computed using the ISB recommendation (Wu et al., 2005), based on p. However, the joint positions delivered by most of computer vision systems are limited to sparse anatomical landmarks, generally joint centers (Hsiao et al., 2022; Needham et al., 2021; T. Xu et al., 2021), too limited to actually use this standard joint angle computation (Wade et al., 2022; T. Xu et al., 2021).The RULA method requires the computation of joint angles from sparse skeleton data. A correction approach for such data has been proposed in (Plantard, Shum, Le Pierres, et al., 2017). While the ISB recommendation (Wu et al., 2005) outlines a standard for calculating joint angles based on estimated joint positions, sparse skeleton data lacks the necessary anatomical landmarks for full compatibility. The referenced study adapted joint angle definitions to align with the joints available in various motion capture systems. In this study, we build on that approach by incorporating wrist (flexion/extension, radial/ulnar deviation, pronosupination) and neck angles.

Hence, for all the evaluated measurement systems, the aim is to provide the similar set of joint angles, used to compute the RULA scores: *back flexion*, *back side bend*, *back twist*, *neck flexion*, *neck side bend*, *neck twist*, and for left and right upper limbs, the *shoulder flexion*, *shoulder abduction*, *shoulder raise*, *elbow flexion*, *wrist flexion*, *wrist deviation* and *wrist twist*.

To compute the RULA score, each joint angle is assigned a value according to a range of predefined angles (McAtamney & Corlett, 1993). For example, the arm score varies from 1 to 4 if shoulder flexion is between [-20°; 20°], < -20° or between [20°; 45°], between [45°; 90°], or > 90° respectively. The same types of threshold are applied to the other joint angles.

The scores for each joint are grouped into the *A Scores*, for arms, forearms and wrists,

and the *B Score*, for neck, trunk and legs. An *A score* is calculated for the joint of the left upper limb and the right limb respectively. Other elements such as *Muscle use* and *Force score* are included in these *A and B scores* to give *C Scores* (for the left and right upper limbs) and a *D Score* (for neck, trunk and legs). The *Muscle use* and *Force score*, assess the repetitiveness and external loads imposed on the worker during his or her task. These additional items are entered manually, as they cannot be deduced automatically from the video or the sensors. Each *C Score* is combined with the *D Score* to provide the *Final Score* for the left and right parts of the body, ranging from 1 to 7. These left and right *Final Scores* lead to a *RULA Action Level Score* summarized in four levels of intervention (from "acceptable posture" to "workstation requiring immediate changes").

2.2.2 Experimental procedure under simulated work conditions

In this section, we present the experimental protocol used to assess the accuracy of the various motion capture systems in simulated work conditions. To this end, we set up an experimental protocol with 12 participants, 3 women and 9 men (age: 32.6 ± 10 years, height: 1.73 ± 0.079 m, mass: 76 ± 16 kg). This study was approved by the Operational Committee for the Evaluation of Legal and Ethical Risks (COERLE) No. 2021-32.

The participants were equipped with the XSens inertial motion capture system (Roetenberg et al., 2009), considered as the reference system for our experiment (W. Kim et al., 2021; Robert-Lachaine et al., 2017). An anthropometric measurement and system calibration phase were carried out for each participant as recommended by the supplier.

Three Orbbec depth cameras were installed around the participant, and the resulting depth images were used to run the Kinect and KIMEA systems, with different viewpoints. The 3 depth cameras used in this protocol allowed us to evaluate the impact of camera position relatively to the participant. The depth cameras were placed in front, to the side and behind the participant.

Six RGB cameras were also placed around the subject to provide RGB images required for the THEIA and KIMEA Cloud systems. The THEIA system used data from the 6 camera viewpoints to assess the participant's movement. Only one RGB camera was needed for the KIMEA Cloud system. Hence, KIMEA Cloud has been tested with 6 different points of view, using one single camera information at a time. Details about the depth and RGB camera placement is given in Table 2.2, and depicted in figure 2.1.

As stated in introduction, KIMEA and KIMEA Cloud systems combined images and four IMUs integrated into specific gloves. One sensor was placed to the midpoint between the styloids, and the other one on the dorsal surface of the hand at level of the third metacarpal bone, for each arm.

Unlike the KIMEA and KIMEA Cloud systems, the Kinect Azure DK and THEIA systems estimated the wrist movements based on visual information only.

ID	Placement	Position XYZ [m] - X-left, Y-up, Z-front
KIMEA Cloud / THEIA		
A1	**Left side**	[4.5, 2.5, 2.1]
A2	**Right side**	[-3.4, 2.4, 1.5]
A3	**Facing left**	[3.0, 2.4, -3.3]
A4	**Facing right**	[-0.5, 2.4, -3.3]
A5	**Back left**	[2.5, 1.3, 5.7]
A6	**Back right**	[-2.6, 1.3, 5.6]
Kinect Azure DK / KIMEA		
B1	**Front**	[0.7, 1.0, -2.1]
B2	**On the side**	[-1.7, 1.0, 0.4]
B3	**Back**	[0.6, 1.0, 2.5]

Table 2.2 – The depth camera placements (B1, B2 and B3) for Kinect Azure DK and KIMEA, and RGB cameras (A1 to A6) used for KIMEA Cloud and THEIA during the experiment. With X Y and Z the position of the camera relative to the subject (X-left, Y-up, Z-front).

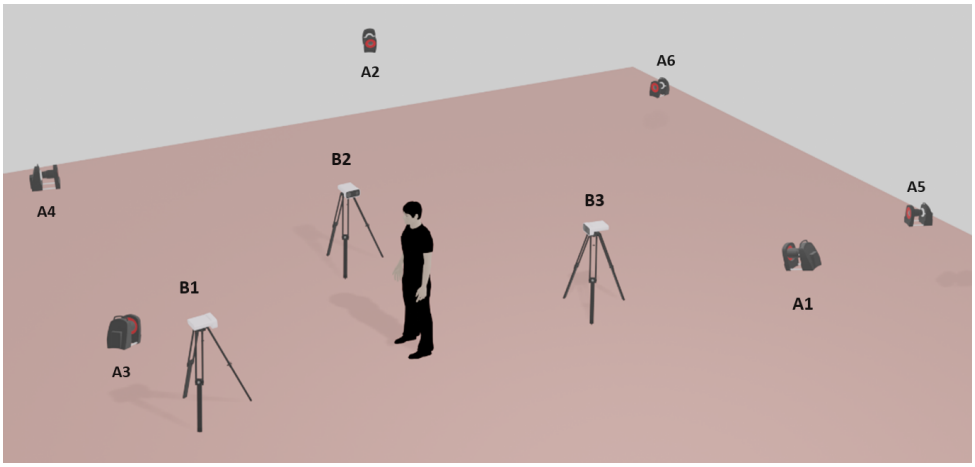
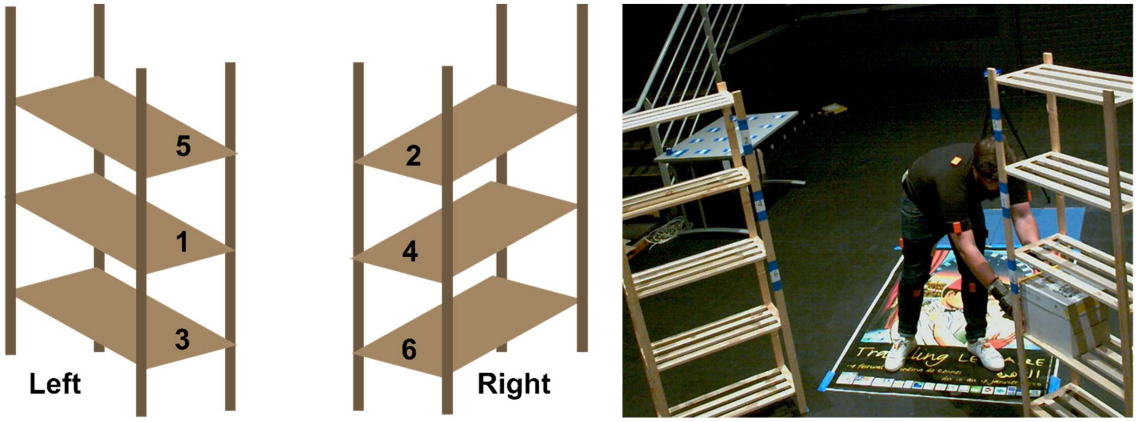


Figure 2.1 – Illustration of camera placement in relation to the subject.

Bimanual handling tasks are frequently used in industry, involving repetitive motions with various masses. Thus, we decided to analyze this work task for this comparison. The task involved removing an empty cardboard box (size: 39x29.5x19cm, weight: 250g) from a three-tier shelf and transferring it to another one. This task was repeated five times consecutively with no waiting period in-between. The two shelves were positioned at 45° to the subject, with three different heights: 51 cm, 89 cm and 127 cm from the floor. The average task duration was about 25 seconds. The order of shelves was predetermined and remained consistent across subjects. Pick and place order for bimanual handling is illustrated in figure 2.2.

We chose this task because it involved external occlusion (with the box) and self-occlusion according to the different depth and RGB camera viewpoints. Indeed, one of the main problems in computer vision systems is the robustness to occlusions. In addition, this task involved all the degrees of freedom taken into account in the RULA method.



(a). Illustration of pick and place orders for evaluated work task.

(b). Illustration of the work task performed by the subjects, the bimanual handling task.

Figure 2.2 – The **(a)** sub-figure on the left shows the pick-and-place orders for the evaluated work task, while the sub-figure **(b)** on the right depicts the work task performed by the subjects.

A reference pose (corresponding to 0 value for each degree of freedom) was preliminary performed for each subject. It enabled us to estimate the angular offsets for each pose, compared to this reference pose. Hence, it provided us with comparable angular values for the different biomechanical models of each motion capture system, as proposed in (W. Kim et al., 2021) and (Antico et al., 2021). Synchronization between the systems was performed using a clap performed at the beginning and the end of each trial. It enabled us to synchronize the different measurements and to resample all the collected data to 30Hz.

2.2.3 Statistics

We compared the joint angles and corresponding RULA scores (*RULA Action Level Score*, *left and right Final Score*, *left and right C Scores* and *D Score*) calculated with each evaluated system and the reference one (obtained with the XSens system), for each subject.

The Root Mean Square Error ($RMSE$) was calculated to evaluate the performance of each system with the reference one for the joint angles and global RULA scores. Following (Yuan & Zhou, 2023), the Mean Absolute Error (MAE) was used to evaluate the accuracy of each measurement system. While $RMSE$ emphasizes larger errors due to the

squaring of differences—making it more sensitive to outliers—MAE reflects the average magnitude of all errors without disproportionately weighting larger deviations. Specifically, MAE represents the average size of the absolute difference between the evaluated system and the reference system. MAE_j^{eval} is the absolute value of the error between the results x^{eval} of the evaluated system e for the joint angle j, and the results x_j^{ref} of the reference system ref for the joint angle j. The MAE_j^{eval} is calculated as follows:

$$MAE_j^{eval}(x_j^{eval}, x_j^{ref}) = \frac{\sum_{i=1}^{n} |x_j^{eval}(i) - x_j^{ref}(i)|}{n}$$

We computed the normalised mean absolute error ($nMAE$) to facilitate comparison of the joint angles with different level of ranges of motion. We normalized the MAE^{eval} of each joint j angles by the range of motion of the reference system (ROM) as follows:

$$nMAE_j^{eval} = \frac{MAE^{eval}(x_j^{eval}, x_j^{ref})}{max(x_j^{ref}) - min(x_j^{ref})}$$

The correlation between each evaluated system and the reference one was also calculated for the joint angles. A Kolmogorov-Smirnov test was used to verify the normality of the error distribution for these analyses. Since the distributions did not follow a normal distribution for this experiment, Spearman's correlation coefficient (ρ) was selected.

We compared the sensitivity of the different systems with the reference one, by computing the number of times the RULA scores change during the task. Moreover, we analyzed the Proportion agreement index (Po) of the RULA score (no difference between the RULA score obtained with the reference system, and the one based on the tested systems), for each system and camera placement.

2.3 Results

Table 2.3 shows the $RMSE$ in degrees for the 4 tested systems, for depth and RGB cameras placed in front of the subject, as generally recommended. These results show that the RMSE of the various calculated joint angles were close to 10°, expect for Kinect Azure DK ($RMSE$: 17.2°). The RMSE of the shoulder and elbow joints were lower for the THEIA system compare to the other evaluated systems.

	KIMEA B1	**THEIA**	**Kinect Azure DK B1**	**KIMEA Cloud A3**	**KIMEA Cloud A4**
Back flexion	**3.7 ± 1.0**	4.3 ± 1.4	**3.7 ± 1.0**	5.1 ± 1.7	6.1 ± 1.1
Neck flexion	**5.8 ± 2.0**	9.7 ± 2.8	5.9 ± 2.0	6.8 ± 1.4	**6.3 ± 1.7**
Left shoulder flexion	15.2 ± 4.8	**14.2 ± 5.4**	15.2 ± 4.8	16.5 ± 3.4	16.6 ± 3.2
Right shoulder flexion	16.0 ± 5.1	**14.2 ± 5.6**	15.9 ± 5.1	18.4 ± 3.0	15.5 ± 2.6
Left elbow flexion	11.9 ± 2.3	**9.7 ± 3.3**	14.7 ± 4.3	15.6 ± 2.7	15.6 ± 4.1
Right elbow flexion	11.2 ± 2.7	**9.4 ± 2.7**	15.7 ± 5.4	16.7 ± 2.7	14.1 ± 1.7
Left wrist flexion	**3.8 ± 2.7**	14.1 ± 3.7	38.5 ± 19.2	4.3 ± 3.1	4.3 ± 3.1
Right wrist flexion	4.0 ± 2.4	13.0 ± 3.9	28.3 ± 5.2	4.0 ± 2.2	**3.9 ± 2.2**
Overall	**8.9 ± 2.9**	11.1 ± 3.6	17.2 ± 4.6	10.9 ± 2.5	10.3 ± 2.5

Table 2.3 – $RMSE$ ± standard deviation expressed in degrees [°] for the work task performed during experimentation with frontal camera placement and for the main joint angles required for RULA. Results in bold highlight the smallest errors for each angle evaluated.

Table 2.4 shows the MAE in degrees and $nMAE$ in percent for the 4 tested systems, for depth and RGB cameras placed in front of the subject. These results show that the errors were the lowest for the KIMEA, THEIA and KIMEA Cloud systems, with a maximum MAE error of 13.3°, 11.9° and 13.3° respectively. We also reported a mean MAE error of 7.4°, 9.0° and 8.2°respectively. For these systems, the largest MAE values occurred mainly for joints with large movements, such as the shoulders or elbows joints, leading to a lower percentage of error ($nMAE$). The results show that the Kinect Azure DK system suffers from even greater error for the wrist, with a maximum error of 32.83° and an average error of 27.71° for this joint.

	KIMEA B1	THEIA	Kinect Azure DK B1	KIMEA Cloud A3	KIMEA Cloud A4
Back flexion	**3.0 ± 0.9 (5.3)**	3.6 ± 1.3 (6.4)	**3.0 ± 0.9 (5.3)**	3.8 ± 1.2 (6.8)	4.8 ± 0.9 (8.5)
Neck flexion	**4.7 ± 1.7 (26.0)**	8.0 ± 2.5 (44.5)	**4.7 ± 1.7 (26.2)**	4.8 ± 0.8 (26.8)	**4.7 ± 1.2 (26.4)**
Left shoulder flexion	13.1 ± 4.7 (12.9)	**12.0 ± 5.2 (11.8)**	13.2 ± 4.7 (12.9)	13.0 ± 2.7 (12.7)	13.0 ± 2.7 (12.7)
Right shoulder flexion	13.4 ± 5.3 (13.1)	**11.7 ± 5.5 (11.6)**	13.3 ± 5.2 (13.2)	14.6 ± 2.5 (14.3)	12.1 ± 2.4 (12.0)
Left elbow flexion	9.6 ± 2.2 (9.1)	**7.9 ± 3.0 (7.5)**	12.1 ± 3.8 (11.6)	12.3 ± 2.1 (11.6)	12.2 ± 4.0 (11.1)
Right elbow flexion	9.3 ± 2.3 (8.7)	**7.8 ± 2.4 (7.3)**	12.9 ± 5.2 (12.1)	12.4 ± 2.1 (11.7)	10.8 ± 1.3 (10.2)
Left wrist flexion	**3.1 ± 2.4 (5.8)**	10.9 ± 3.5 (20.6)	32.8 ± 19.9 (61.9)	3.3 ± 2.4 (6.3)	3.3 ± 2.4 (6.3)
Right wrist flexion	**3.1 ± 1.9 (6.5)**	10.1 ± 3.3 (21.0)	22.6 ± 4.9 (47.1)	**3.1 ± 1.7 (6.5)**	**3.1 ± 1.7 (6.4)**
Overall	**7.4 ± 2.7 (10.9)**	9.0 ± 3.3 (16.3)	14.3 ± 4.5 (23.8)	8.4 ± 1.9 (12.1)	8.0 ± 2.1 (11.8)

Table 2.4 – MAE ± standard deviation expressed in degrees [°] ($nMAE$ expressed in percent [%]) for the work task performed during experimentation with frontal camera placement and for the main joint angles required for RULA. Results in bold highlight the smallest errors for each angle evaluated.

Please refer to Table 2.5 for more details about the joint range of motions used to calculate the $nMAE$ error. This table shows the average joint range of motion (ROM), measured in degrees, for various body parts based on reference data.

	Mean joint range of motion [°]
Back flexion	55,96
Neck flexion	17,97
Left shoulder flexion	101,18
Right soulder flexion	100,93
Left elbow flexion	105,06
Right elbow flexion	105,76
Left wrist flexion	53,03
Right wrist flexion	47,76

Table 2.5 – Mean joint range of motion [°] calculated from reference data.

Table 2.6 shows the correlations between joint angles obtained with the tested systems compared to those obtained with the reference one, for the main joints. These results support the hypothesis that wrist angles are difficult to estimate with a computer vision approaches. For example, the Kinect Azure DK B1 exhibits very low correlation for the two wrist angles (0.14 and 0.18 for the left and right wrist flexion respectively). However, hybrid systems benefit from additional information for the wrists, leading to correlation

greater than 0.88. The neck flexion seems also difficult to estimate with computer vision methods, especially for KIMEA Cloud system, with correlation between 0.34 and 0.39.

	KIMEA B1	**THEIA**	**Kinect Azure DK B1**	**KIMEA Cloud A3**	**KIMEA Cloud A4**
Back flexion	0.92 ± 0.03	**0.96 ± 0.02**	0.92 ± 0.03	0.87 ± 0.05	0.77 ± 0.07
Neck flexion	**0.59 ± 0.20**	0.51 ± 0.21	**0.59 ± 0.20**	0.34 ± 0.22	0.39 ± 0.21
Left shoulder flexion	0.95 ± 0.02	**0.96 ± 0.02**	0.95 ± 0.02	0.85 ± 0.05	0.85 ± 0.06
Right shoulder flexion	0.95 ± 0.03	**0.96 ± 0.02**	0.95 ± 0.02	0.81 ± 0.05	0.86 ± 0.03
Left elbow flexion	0.90 ± 0.04	**0.94 ± 0.03**	0.88 ± 0.05	0.86 ± 0.05	0.88 ± 0.06
Right elbow flexion	0.93 ± 0.03	**0.95 ± 0.02**	0.89 ± 0.05	0.85 ± 0.05	0.89 ± 0.05
Left wrist flexion	**0.88 ± 0.28**	0.55 ± 0.18	0.14 ± 0.14	**0.88 ± 0.23**	**0.88 ± 0.24**
Right wrist flexion	**0.91 ± 0.11**	0.60 ± 0.14	0.18 ± 0.16	0.90 ± 0.14	0.90 ± 0.13
Overall	**0.88 ± 0.09**	0.80 ± 0.08	0.69 ± 0.08	0.80 ± 0.11	0.80 ± 0.11

Table 2.6 – Spearman's correlation coefficient (ρ) for the work task performed during the experiment with frontal camera placement and for the main joint angles required for RULA and used during the tasks.Results in bold highlight the highest correlations for each angle evaluated.

Table 2.7 shows the $nMAE$ of the joint angles depending on the camera placements. The results show that the camera placement have a real impact on the error of the joint angles estimation. We noticed greater errors for camera placement on the back and the side for both KIMEA and Kinect Azure DK (KIMEA back position: 17.4%, side position: 14.0%; Kinect Azure DK back position: 33.8%, side position: 29.4%). The KIMEA Cloud system seams less impacted by the camera placement with $nMAE$ values ranging between 11.8% and 14.6%.

	KIMEA			THEIA	Kinect Azure DK			KIMEA Cloud					
	B1	**B2**	**B3**		**B1**	**B2**	**B3**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
Back flexion	**5.3**	5.6	8.6	6.4	**5.3**	5.6	8.6	8.2	9.7	**6.8**	8.5	12.3	15.2
Neck flexion	**25.9**	37.1	54.3	44.46	**26.3**	38.6	55.6	34.4	38.6	26.8	**26.4**	37.8	38.2
Left shoulder flexion	**12.9**	14.0	18.1	11.8	19.2	**14.1**	18.1	13.1	17.5	12.7	12.7	**11.4**	14.4
Right shoulder flexion	**13.2**	19.8	20.3	11.6	**13.2**	19.6	20.1	15.6	11.5	14.4	12.0	**11.4**	15.1
Left elbow flexion	**9.1**	**9.1**	12.6	7.5	**11.5**	14.2	16.8	11.4	12.7	11.6	11.5	12.0	**10.4**
Right elbow flexion	**8.7**	13.4	13.3	7.3	**12.1**	15.9	18.6	11.1	**8.8**	11.7	10.2	10.9	11.3
Left wrist flexion	**5.8**	6.6	6.1	20.6	**61.9**	64.9	72.2	**6.2**	**6.2**	6.3	6.3	6.3	**6.2**
Right wrist flexion	**6.5**	6.6	6.0	21.0	**47.1**	62.2	59.9	6.5	6.5	6.5	**6.4**	**6.4**	**6.4**
Overall	**10.9**	14.0	17.4	16.3	**23.8**	29.4	33.8	13.3	13.9	12.1	**11.8**	13.6	14.6

Table 2.7 – The $nMAE$ in precent for the main joint angles required for RULA, during the work task for different motion capture systems and camera placements. Results in bold highlight the camera placement with the lowest errors for each angle evaluated.

As the RULA scores are based on joint angles intervals, errors on the joint angle estimation may have smaller impact on the RULA scores. Table 2.8 reports the $RMSE$ values between the RULA scores calculated with the tested and the reference system. Errors for the *C scores (Upper Limb)* were higher for the Kinect Azure DK. Errors for the *D scores (neck, back and legs)* were higher for THEIA.

	KIMEA B1	THEIA	Kinect Azure DK B1	KIMEA Cloud A3	KIMEA Cloud A4
RULA Action Level Score	0.36 ± 0.08	0.40 ± 0.12	0.51 ± 0.06	0.38 ± 0.07	**0.33 ± 0.06**
Left Final Score	**0.49 ± 0.13**	0.62 ± 0.13	0.62 ± 0.12	0.56 ± 0.10	0.51 ± 0.08
Right Final Score	**0.45 ± 0.12**	0.60 ± 0.13	0.58 ± 0.09	0.57 ± 0.10	0.50 ± 0.08
Left C Score (upper limbs)	0.54 ± 0.15	**0.52 ± 0.13**	0.73 ± 0.21	0.58 ± 0.10	0.58 ± 0.10
Right C Score (upper limbs)	0.53 ± 0.13	**0.52 ± 0.14**	0.67 ± 0.12	0.62 ± 0.09	0.55 ± 0.09
D Score (neck, back and legs)	**0.58 ± 0.17**	0.73 ± 0.16	0.59 ± 0.17	0.64 ± 0.11	0.60 ± 0.09

Table 2.8 – $RMSE$ ± standard deviation for global RULA scores on work task performed during experimentation with frontal camera placement. Results in bold highlight the smallest $RMSE$ for each RULA score evaluated. Results in bold highlight the number of score changes closest to the reference, for all the RULA scores evaluated.

Table 2.9 reports the number of times the RULA score changes from one value to

another during the task. The results showed that the number of score changes is close to the reference (XSens) for all evaluated systems, with a mean difference of 1.9 score change. The largest difference is obtained with the THEIA system for the *D Score*. KIMEA Cloud system exhibits the best agreement with the reference number of score changes, except for the *Right C Score (upper limbs)*.

	XSens	**KIMEA B1**	**THEIA**	**Kinect Azure DK B1**	**KIMEA Cloud A3**	**KIMEA Cloud A4**
RULA Action Level Score	15.6 ± 4.1	13.3 ± 3.7	12.8 ± 5.0	11.0 ± 5.0	17.8 ± 3.3	**15.1 ± 3.2**
Left Final Score	22.5 ± 3.1	21.8 ± 5.1	20.2 ± 5.4	24.7 ± 8.7	**22.4 ± 4.3**	20.3 ± 4.4
Right Final Score	22.0 ± 4.2	20.3 ± 4.2	20.6 ± 5.5	24.8 ± 4.9	22.7 ± 5.0	**21.9 ± 3.5**
Left C Score (upper limbs)	38.4 ± 3.5	37.6 ± 5.5	38.5 ± 6.1	40.0 ± 11.1	**38.4 ± 4.8**	39.1 ± 4.1
Right C Score (upper limbs)	38.2 ± 4.9	**38.1 ± 5.1**	39.1 ± 5.0	41.8 ± 6.6	37.4 ± 6.3	40.0 ± 3.0
D Score (neck, back and legs)	19.1 ± 4.9	22.9 ± 7.7	26.0 ± 5.7	23.2 ± 8.0	16.2 ± 5.4	**17.8 ± 5.0**

Table 2.9 – Mean (± standard deviation) number of RULA score changes during work task performed during experimentation with frontal camera placement.

Table 2.10 shows the proportion agreement index (Po) of the global RULA scores obtained with the different tested systems, but also the different camera placements. The agreement of the *RULA Action Level Score* is higher than 0.8 for each evaluated system and camera placement, except for Kinect Azure DK. The results show that the highest levels of agreement were obtained for front camera placement, which is in accordance with previous results. The results were very similar between THEIA Multi-camera system and KIMEA and KIMEA Cloud systems when the camera is placed in front of the subject.

	KIMEA			THEIA	Kinect Azure DK			KIMEA Cloud					
	B1	B2	B3		B1	B2	B3	A1	A2	A3	A4	A5	A6
RULA Action Level Score	**0.87**	0.80	0.81	0.86	**0.74**	0.71	**0.74**	0.80	0.81	0.87	**0.89**	0.85	0.84
Left Final Score	**0.80**	0.76	0.71	0.70	**0.67**	0.60	0.59	0.72	0.65	0.74	**0.76**	0.73	0.69
Right Final Score	**0.82**	0.72	0.71	0.71	**0.72**	0.59	0.61	0.66	0.72	0.72	**0.77**	0.71	0.68
Left C Score (upper limbs)	**0.72**	0.69	0.56	0.73	**0.56**	0.50	0.42	0.66	0.58	0.70	0.71	**0.72**	0.67
Right C Score (upper limbs)	**0.72**	0.60	0.55	0.72	**0.60**	0.45	0.42	0.60	0.71	0.68	**0.73**	0.72	0.66
D Score (neck, back and legs)	**0.74**	0.63	0.51	0.62	**0.73**	0.63	0.51	0.63	0.59	0.69	**0.70**	0.57	0.53

Table 2.10 – Propotion agreement index (Po) for global RULA scores, during the work task for different motion capture systems and camera placements. Results in bold highlight the camera placement with the higher agreement for each RULA scores evaluated.

2.4 Discussion

2.4.1 Main findings and contributions

The aim of this study was to evaluate the forces and weaknesses of computer vision-based motion capture systems to estimate the joint angles and RULA scores in work movement. We tested the Kinect Azure DK system, the THEIA multi-cameras system, the hybrid KIMEA system, and the hybrid KIMEA Cloud system. During our laboratory experiments, we measured the accuracy of these different systems compared to a reference XSens inertial system.

The results for the joint angles were comparable to those found in (W. Kim et al., 2021; Plantard, Shum, Le Pierres, et al., 2017; Yuan & Zhou, 2023), who evaluated the the accuracy of motion capture systems based on depth or RGB cameras. More precisely, (Plantard, Shum, Le Pierres, et al., 2017) evaluated the RMSE of the Kinect V2 depth camera placed in front of the subject handling a box. In this previous study, wrist flexion was not measured and an average RMSE around 10° was found. Excluding the RMSEs for wrist flexions, our results were similar with an average RMSE of 10.6° and 11.8° for KIMEA and Kinect Azure DK, respectively. In a more recent study, (W. Kim et al., 2021) proposes to calculate wrist flexion angles using data from the Kinect V2 depth camera. For all the flexion angles required for RULA, the authors reported an average RMSE of around 18°, which is in agreement with our results for the Kinect Azure DK (RMSE: 17.2°). These previous studies evaluated the former Kinect V2 depth camera, but the overall performance with Kinect Azure DK were similar to former versions (Bertram

et al., 2023).

(Yuan & Zhou, 2023) reported RMSE of 12.9° and a MAE of 9.4°, for all the joint angles measured, when using a system based on a single RGB camera. These results are very close to those found for KIMEA Cloud: a RMSE of 10.6° and an error of 8.2°. (W. Kim et al., 2021) proposed to evaluate the OpenPose 3-cameras system (Cao et al., 2017) for static postures and lifting movements. They reported a RMSE of 8.3°, which is better than the one obtained with the THEIA system in our study (RMSE: 11.1°). This difference may be due to the evaluation of the nature of the static postures used in this previous study. The joint angles errors obtained in our work for the THEIA system are consistent to those reported in previous works for other types of movements (Lahkar et al., 2022).

The joint angles obtained with the tested systems exhibit good correlation ($\rho \geq 0.82$) with the reference system for back, shoulder and elbow flexion angles. For neck joint angles, the correlation was weak to moderate (between 0.34 and 0.59). This is partially explained by a small variation of these angles in the studied movements, resulting in a large normalised error (nMAE between 26% to 44.5%). Nevertheless, KIMEA Cloud, based on a single RGB camera had more difficulties to measure the neck flexion correctly than systems using depth images or multiple cameras.

Computer vision-based only measurement systems, such as THEIA and Kinect Azure DK, suffer from more important errors and lower correlation for the wrist joint angles (THEIA: $MAE = 10.5° \pm 3.4$, $\rho \leq 0.60$; Kinect Azure DK: $MAE = 27.7° \pm 12.4$, $\rho \leq 0.18$), compared to the other systems ($MAE \leq 3.3°$, $\rho \geq 0.88$). This result supports the hypothesis that current computer vision algorithms cannot accurately estimate movements of small body parts with large range of motions, and high risk of occlusion, such as the hands. In working tasks, hands are often used, and errors in estimating their motion may lead to unreliable postural assessments.

The results found for the global RULA score evaluation were globally consistent with those found in previous studies (Abobakr et al., 2019; W. Kim et al., 2021; Manghisi et al., 2017; Plantard, Shum, Le Pierres, et al., 2017; Yuan & Zhou, 2023). We obtained a proportion agreement (Po) for the *RULA Action Level Scores* of 0.83, 0.81, 0.73 and 0.84 for KIMEA, THEIA, Kinect Azure AD and KIMEA Cloud, respectively. Results showed relatively few variations according to camera placement, with Po ranging from 0.80 to 0.87 for KIMEA, 0.71 to 0.74 for Kinect Azure AD, and 0.81 to 0.89 for KIMEA Cloud. These results are comparable to (Yuan & Zhou, 2023), who obtained an average Po of 0.85, with very little difference between front and side camera placement (between 0.84 and 0.87). Our results were slightly better than those found by (W. Kim et al., 2021), which found much higher variation with occlusion and camera placement (between 0.68 and 0.82 for OpenPose). On the contrary, our results showed larger errors than those reported in previous works (Abobakr et al., 2019) (Global RULA Score RMSE=0.49) and (Plantard, Shum, Le Pierres, et al., 2017) (Global RULA Score RMSE=0.43). In our work, we report

Global RULA Score RMSE of 0.58, 0.57, 0.72 and 57 for KIMEA, THEIA, Kinect Azure AD and KIMEA Cloud, respectively. These differences could be due to disparities in the methodology and the design of the experimental protocol. (Plantard, Shum, Le Pierres, et al., 2017) proposed to evaluate a lifting task involving limited movement of the back or neck. This task would cause limited change of the RULA score for the back and neck, and therefore tends to minimize error. (Abobakr et al., 2019) fine-tuned the human pose estimator to enhance the results on the testing dataset, which is not the case in our study.

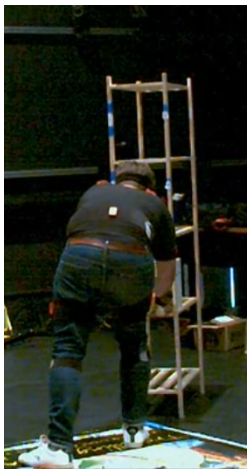


Figure 2.3 – Illustration of auto-occlusions of the left arm with a camera positioned behind the subject for the work task

As expected, our results show that the camera placement affects the performance of computer vision methods. We report lower joint angle errors when the camera was placed in front of the subject (nMAE for KIMEA: 10.9%, Kinect Azure DK: 23.8% and KIMEA Cloud: 12.0%). The impact of the camera placement seems limited for the KIMEA Cloud system, with nMAE ranging from 11.8% to 14.6% for all camera placements. The most significant joint angle errors were found when the camera was placed on the back for the KIMEA (nMAE: 17.4%) and Kinect Azure DK (nMAE: 33.8.4%) systems. The results showed that the neck joint error was over 50% when the depth camera was positioned on the back of the subject. The evaluated tasks involved leaning forward, which caused significant occlusion of the head with this camera placement, as illustrated in figure 2.3. Occlusions can cause significant measurement errors for systems using a depth camera (Jo et al., 2022; Plantard et al., 2015). Therefore, it is important to consider an overhead view to minimize measurement errors during this type of task taken from behind the person.

2.4.2 Limitations

In this work, we considered that the XSens motion capture system was a reference system, as described in previous works (W. Kim et al., 2021). We may have obtained different results with another reference system, such as laboratory optoelectronic systems.

We chose to eliminate optoelectronic reference system because the reflective markers interfere with the depth camera (using infrared sensor), leading to failure in tracking a human body in the resulting images, as reported in previous works (Jo et al., 2022; Naeemabadi et al., 2018; Özsoy et al., 2022). The measurement error of the XSens is about 5° for walking activities (Schepers et al., 2018), or around 2.8° for handling tasks (Robert-Lachaine et al., 2017), but it can reach up to 14.5° (Benjaminse et al., 2020). Additionally, the sensors may shift from their anatomical landmarks, introducing additional error. Finally, these inertial systems may suffer from drift error over long measurement sequences (S. Kim & Nussbaum, 2013; Lebel et al., 2013; Plamondon et al., 2007). However, due to the strong constraints of infrared perturbation, XSens system was a good candidate to deliver reference values. And all the systems were compared frame-per-frame to this reference system, equally.

Even if we designed the protocol to mimic real work conditions, the experiment was carried-out in a laboratory condition. Further works would be necessary to actually evaluate these systems on real working conditions. But this is a very difficult task as it is almost impossible to control the test condition and to ensure that a reference system would deliver actual reliable values.

It is also important to note that various measurement systems rely on custom biomechanical models, which differ from the ISB recommendations (Wu et al., 2005). The placement and number of anatomical landmarks provided by these systems for calculating joint angles may vary. While we used the method described by (W. Kim et al., 2021) to mitigate the impact of these model differences, incorporating a kinematic calibration phase to approximate anatomical reference points (Robert-Lachaine et al., 2017; X. Xu et al., 2017) could further reduce discrepancies between models and improve the validity of comparisons.

2.5 Conclusion

This study evaluated the accuracy of different types of computer vision-based systems for carrying-out postural assessment using the RULA method. The results showed that hybrid systems (coupling Depth or RGB cameras to sparse IMU sensors) correctly scored RULA in more than 80% of cases, regardless of the camera's placement. The system with multiple RGB-camera also achieved a high rate of correct measurement (86%). However, placing and calibrating several cameras on real working environments may be difficult.

Systems based on a single depth camera provided promising results for RULA evaluation (between 71% and 74% of correct measurements), but suffered from higher error in measuring specific joint angles, such as the wrist. They seem to be also very sensitive to the camera placement. Manually entering the wrist joint score in the RULA method would help to mitigate this significant measurement error. However, this joint is the most

difficult one to estimate with the eyes, unlike the shoulder and elbow joints (Lowe, 2004a, 2004b). Hybrid systems, which associate sparse IMU sensors for the hands, clearly help to better reconstruct wrist motion (Manghisi et al., 2017), which is very relevant for postural assessment of tasks involving the upper-limbs.

Although the measurement of RULA scores does not reach 100% accuracy, the results reported in this work are very promising, compared to the traditional hand-made evaluation. Hence, considering the lower inter-subject reliability obtained with the traditional hand-made method (average agreement of only 58.25% for the RULA method (Widyanti, 2020)), the use of these automatic systems is of considerable interest, especially for low experienced raters. In future work, it could be interesting to propose confidence intervals for each risk score instead of discrete scores, which would make it possible to define MSD risk intervals for a task.

To summarize, the different motion capture systems based on computer vision enable a correct evaluation of the risk of WMSDs, when using the RULA method. However, systems based on depth cameras only suffer from high joint angle estimation errors (especially for wrist joint angles), even if the resulting Global RULA score is correctly estimated.

These results open up numerous perspectives in the field of ergonomics, potentially extending beyond the measurement of kinematic data. Indeed, with a more reliable joint angle estimation, this input can be used to estimate internal forces through inverse dynamics frameworks. Although the low sampling frequency of some measurement systems can affect results in the case of dynamic movements, several studies have already addressed this topic with promising results (Plantard, Muller, et al., 2017; Uhlrich et al., 2023). In the next chapter, we will benchmark various learning algorithms aimed at estimating upper limb joint torque during static and dynamic phases of load-carrying tasks. It presents a generalized learning model for torque estimation that considers variability among different subjects and task conditions.

Chapter 3

Estimation of Upper-Limb Joint Torques in Static and Dynamic Phases for Lifting Tasks

3.1 Introduction

Evaluating in-situ physical risk factors generally relies on noisy and/or incomplete motion data captured by videos or depth cameras. However, standard inverse dynamics cannot handle such limited and low-frequency data to compute reliable joint torques. Instead, recent development in machine learning opens new possibilities to estimate these torques in such a difficult condition, robust to noise. However, designing such a machine learning approach requires extensive tests. In this chapter, we propose a study to explore which machine learning approach would best approximate joint torques estimated with standard approach, firstly with noise-free data. This evaluation has been applied to one-handed load carrying tasks.

As mentioned in chapter 1, (Mohseni et al., 2022) used neural networks to estimate L5-S1 coronal and sagittal moments (namely M_x(N.m) and M_y(N.m) along the X and Y axes, respectively) at the L5-S1 joint during static load handling. The trained algorithm maps the relationships between six features (Load location (X, Y, Z) (cm), hand load (Kg), body height (cm) and body weight (Kg)) and two targets (M_x(N.m) and M_y(N.m)) with a Root Mean Square Error (RMSE) of 16.5 N.m and a correlation of $R^2 = 0.97$. After normalizing the inputs and removing the outliers from static trials, the mean torques error decreased from 16.5 N.m to 11.8 N.m.

The estimation of joint torques in dynamic phases using learning algorithms is commonly performed based on the joint coordinates and velocities. (Zell et al., 2020) used Random forest approaches to estimate joint torques from motion parameters (joint coordinates and velocities) and acting forces for human gait. They also presented a weakly-supervised learning approach, aimed at inferring human dynamics (Ground Reaction Forces & Moments, and joint torques). To this end, they used an artificial neural network (NN), which architecture incorporated inverse and forward dynamics layers to minimize a pure motion loss. This motion loss consisted in minimizing the difference between the simulated motion

generated by the model and the observed one, without using any additional information on the ground reaction forces, moments, or joint torques.

Human motion can be described as a combination of spatial and temporal information, where spatial components capture the geometric arrangement of body joints and temporal components represent their evolution over time (as explained in chapter 1.1.3). Recent research has proposed the development of specialized neural network architectures that decouple the processing of spatial and temporal information, enabling more effective estimation of human posture (Mourot et al., 2022a).

Another key point when using machine learning approaches, is to ensure that the inputs have similar range of values. To address this problem, the input data are generally normalized before being processed by the method. This problem has been widely explored when designing action recognition methods (Tang et al., 2022).

Our study aims to estimate joint torques using skeletal data (reference data), subject mass, and load mass. We particularly propose two key contributions:

1. A comparison of state of the art machine learning methods to estimate the back and upper limb joint torques from 3D joint positions, the mass carried and the mass of the subject in static poses;
2. An evaluation of a NN architecture to deal with dynamic motions.

The study has been published in the International Conference on Digital Human Modeling (Belabzioui et al., 2023) (See 4.5).

3.2 Overview

Inverse dynamics aims to determine the joint torques τ associated with external forces and motion quantities for a given motion performed by the subject. The movement is fully defined by the joint coordinates $\mathbf{q}$, velocities $\dot{\mathbf{q}}$, and accelerations $\ddot{\mathbf{q}}$. The general inverse dynamics applied to a biomechanical model of the subject was formulated in Equation 1.1 of Chapter 1.

The software platforms, AnyBody (Damsgaard et al., 2006), OpenSim (Delp et al., 2007), and the open-source CusToM MATLAB toolbox (Muller, Pontonnier, Puchaud, et al., 2019), integrate the inverse dynamics approach. In this study, we utilized the CusToM toolbox to determine joint torques based on a given motion, which served as the reference joint torques for the remainder of the chapter. The joint torques estimated by the learning algorithms were compared to the joint torques computed by CusToM for the same motion. To this end, we devised learning architectures for the static and dynamic phases of a motion. The static phase was defined as a phase in which the subject remain still, whereas the dynamic phase was characterized by non negligible motion quantities. For static mass handling poses, we utilized a regression model to compare four architectures, namely,

linear regression, decision tree, random forest, and neural network. In the dynamic phase, we first extracted spatio-temporal features from the skeleton data and then incorporated the subject's mass and the mass of the load to regress the 16 joint torques of the right upper limb and the trunk.

3.3 Data collection and preparation

In this section, we detail the data used to train and evaluate the machine learning methods used to estimate the joint torques based on static or dynamic input motion data.

3.3.1 Experimental data and biomechanical model

Experimental data: The study used data from a previous experiment (Haj Mahmoud et al., 2021) that involved 11 right-handed subjects lifting and placing objects with varying masses, positions, handling height and timing. Each trial was composed of static phases and dynamic phases when displacing the carried object. An optoelectronic motion capture system Qualisys (23 12-Mpixels cameras, sampled at 200Hz) was used to record the motion of the subjects, and 2 force plates were used to record the ground reaction forces during the tasks. Three different loads were used (0Kg, 1Kg, 3Kg), with five final placing positions (175cm height in front, 175cm height on the right, 75cm height in front, 0cm height in front, and 0cm height on the right). By combining the load and the final placing conditions, it leads to $3\times5 = 15$ configurations that are repeated 9 times each per subject ($15\times9 = 135$ trials per subject). Among the total number of trials ($11\times135 = 1485$), 286 were discarded due to unexpected motions.

In a one-handed load-carrying task, two distinct phases can be identified: the static phase and the dynamic phase. The static phase involves holding the object in a stationary position for a specified duration. In contrast, the dynamic phase encompasses the entire sequence of motion, including the initial static holding phase as well as the subsequent actions of grasping, transporting, and placing the object from one height to another. Static phases lasted 5.38 s in mean per trial, which corresponds to 1076 frames. Hence, a total of $949 \times (1485 - 286) \simeq 1M$ pose samples were available for static phases learning. Trials lasted 11 s in mean, corresponding to 2200 frames. Consequently, $2200 \times (1485 - 286) \simeq 2.6M$ samples were available for the dynamic phase training. As we apply a sliding window with an overlap of 1 and the length of each sequence is 5, the total number of sequences obtained was similar, $2.6M$ 5-frames sequences.

Biomechanical model: The biomechanical model was composed of eleven segments including the head, upper/lower trunks, left/right arms, left/right hands, left/right fore-

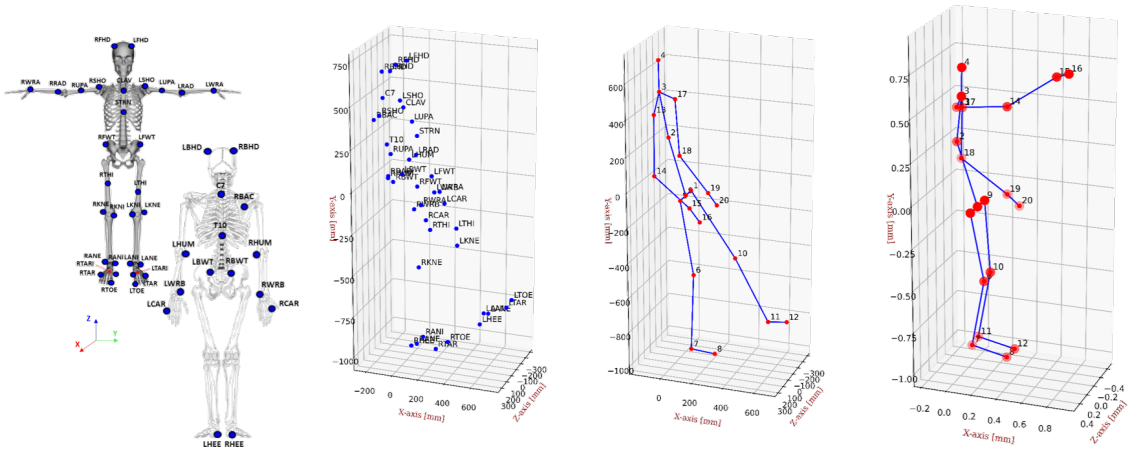
arms, articulated with 38 degrees of freedom: a 6 dofs mobile base, and 31 anatomical joint angles following the International Society of Biomechanics recommendations. The joints included in the torque estimation are described in the table 3.1.

Joint	Corresponding exertion	Joint	Corresponding exertion
1	Lumbar spine flexion/extension	9	Right clavicle axial rotation
2	Lumbar spine lateral flexion/extension	10	Right shoulder plane of elevation
3	Lumbar spine axial rotation	11	Right shoulder depression/elevation
4	Trunk flexion/extension	12	Right Upper arm axial rotation
5	Trunk lateral flexion	13	Right elbow flexion extension
6	Trunk axial rotation	14	Right forearm pronation/supination
7	Right clavicle protraction/retraction	15	Right wrist flexion/extension
8	Right clavicle depression/elevation	16	Right wrist radial/ulnar deviation

Table 3.1 – Joint torques estimated in the study. In particular, the 3 first torques correspond to the classical L5/S1 joint torques.

The CusToM toolbox was used to compute the reference joint torques using the biomechanical model described above.

3.3.2 Joint centers estimation and data normalization



(a). *Anatomical markers in a global Euclidean space* (b). *Anatomical markers in a pelvis reference frame* (c). *Estimated joint centers* (d). *Normalized 3D skeleton*

Figure 3.1 – Marker coordinates were expressed in the pelvis reference frame. Joint centers were estimated using regression equations. The 3D human skeleton was normalized using the AABB (Axis-Aligned Bounding Box) approach.

3D anatomical markers positions in global Euclidean space were expressed in the pelvis reference frame. Joint centers were estimated based on the markers positions using regression equations (Leardini et al., 1999; Reed et al., 1999). The resulting human pose

(See Figure 3.1) was modeled as 20 3D joints $\mathbf{J}_{i=1..20}(t)$ at each time t, that can be gathered in a single vector denoted $\mathbf{J}$. 3D coordinate of each joint was normalized by its range of values, using the Axis-Aligned Bounding Box approach (L. Chen & Qin, 2010). For the $k = [x, y, z]$ component of $\mathbf{J}_i(t)$, denoted $J_i^k(t)$, the normalized value was given by:

$$\hat{J}_i^k(t) = \frac{J_i^k(t) - \min(J_i^k(t))}{\max(J_i^k(t)) - \min(J_i^k(t))}; \qquad \hat{m} = \frac{m - m_{min}}{m_{max} - m_{min}}; \quad \hat{M} = \frac{M}{M_{max}} \tag{3.1}$$

where $min(J_i^k(t))$ and $max(J_i^k(t))$ are the minimal and maximal values of J_i^k across all the subjects and all the trials, m_{min} and m_{max} are the minimum and maximum values of the load mass, and M_{max} is the maximum value of the subject mass respectively.

3.4 Joint torque estimation

In this section, we describe the machine learning approaches developed and evaluated in this chapter to estimate the joint torques based on static or dynamic poses.

3.4.1 Static phases

Estimating the joint torques in a static phase is a particular case of the inverse dynamics problem. Specifically, this problem assumes that the joints velocities $\dot{\mathbf{q}}$ and accelerations $\ddot{\mathbf{q}}$ are negligible.

The objective of this estimation is to learn the function $\tau = f(\hat{\mathbf{J}}, \hat{m}, \hat{M})$ from the set of samples described in the previous section.

Four classical estimators were benchmarked in the current study to solve this issue: linear regression, decision tree, random forest, and neural network. The number of trees in the random forest model was set to 10 to provide a balance between model performance and computational efficiency (Breiman, 2001).

This neural network architecture is designed to balance performance and generalization. It consists of an input layer, a dense layer with 64 units, batch normalization, `ReLU` activation, and a dropout rate of 10%. The *input layer* processes the input data, while the *first dense layer* with 64 units extracts key features. *Batch normalization* helps stabilize training by normalizing activations (Ioffe, 2015), and the *ReLU activation* introduces necessary nonlinearity without causing vanishing gradients. The 10% dropout rate reduces overfitting by randomly deactivating units during training (Srivastava et al., 2014). *A second dense layer with L2 regularization* (weight decay of 0.0001) penalizes large weights to prevent overfitting further (Goodfellow et al., 2016). With 5072 parameters, the model is complex enough to capture important patterns without being overly prone to overfitting.

In order to train our deep network f_θ, we adopted the standard mean squared error (MSE) loss function.

3.4.2 Dynamic phases

The objective of this estimation is to learn the function $\tau = f(G, \hat{m}, \hat{M})$ that maps the input features G (spatiotemporal features, see below), $\hat{m}$ and $\hat{M}$ to the output torques τ at a given time step.

In dynamic phases, velocities and acceleration are not negligible. Thus, we considered a time window of 5 frames prior to the estimation to extract spatiotemporal features G, by applying a one-dimensional convolutional layer to the $\hat{\mathbf{J}}(t_1), \hat{\mathbf{J}}(t_2), \ldots, \hat{\mathbf{J}}(t_5)$ 3D joint positions gathered as a table. The layer had 64 filters with a kernel size of 2 and applied the Rectified Linear Unit (ReLU) activation function to its outputs. We also set the padding to 'valid', which means no padding is added to the input, and the stride to 1, which specifies the step size of the convolutional operation. Next, we applied a max pooling layer with a pool size of 2 and stride of 2 to the output of the convolutional layer (for more details about the key principles of architectures, see the paragraph 1.1.3). This reduced the dimensionality of the output by taking the maximum value in each 2-element segment. Then we applied a dropout rate of 25% to the output tensor from the previous max pooling operation.

We created four distinct neural network architectures to estimate the torques from these features: CNN-LSTM, which consists of a CNN layer followed by an LSTM layer; CNN-LSTM-Attention, which includes a CNN layer, followed by an LSTM layer and an attention mechanism; CNN-BiLSTM, comprising a CNN layer followed by a Bidirectional LSTM (Bi-LSTM) layer; and CNN-BiLSTM-Attention, which consists of a CNN layer followed by a Bi-LSTM layer and an attention mechanism (See the figure 3.2).

To train these networks, we minimized the standard mean squared error (MSE) loss.

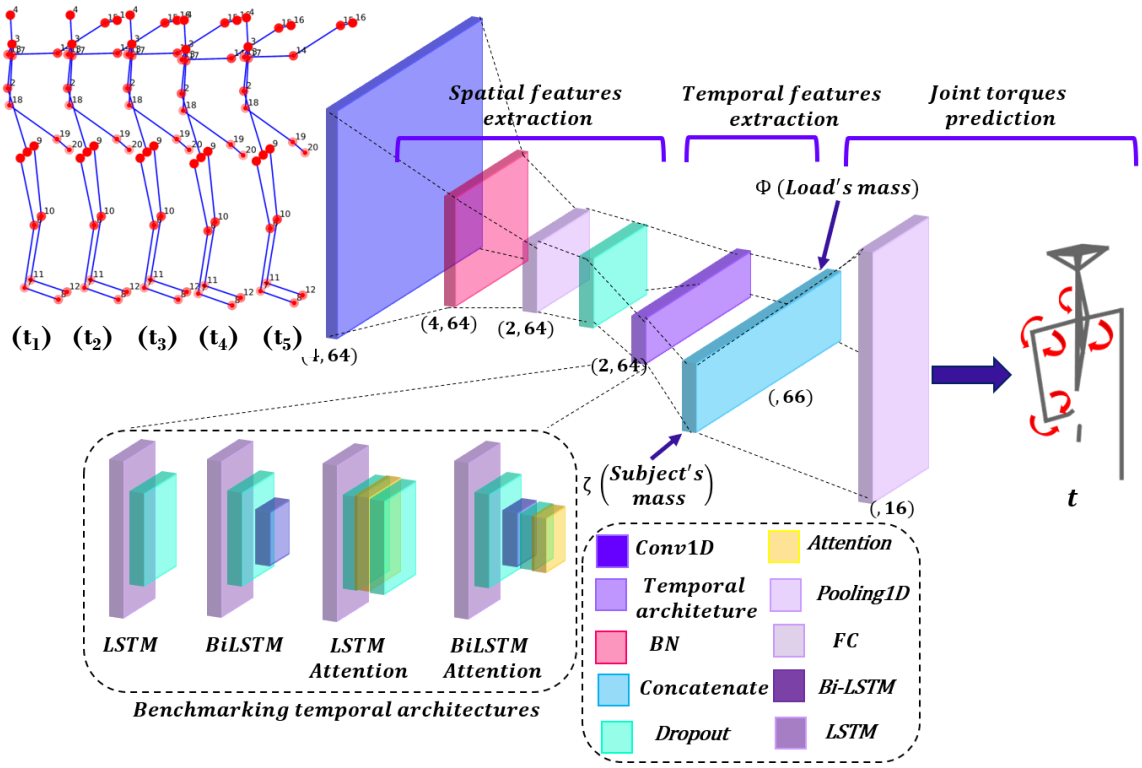


Figure 3.2 – Learning-based algorithm scheme of dynamic phase.

The overall architecture used for extracting spatiotemporal features and regressing the 16 joint torques, had an average of 21232 parameters, which was high relative to the number of samples per parameter (See section 3.3.1). To prevent overfitting, a kernel regularizer with `L2 weight regularization` of 0.01 was added to the `Conv1D` layer, and a `BatchNormalization` layer was added after the `Conv1D` layer to normalize the output. Dropout regularization with a rate of 25% was added after the max pooling and LSTM layers, and kernel and recurrent regularizers with `L2 regularization` of 0.01 were added as arguments to the LSTM layer.

3.4.3 Learning and evaluation

We implemented our learning algorithms in static and dynamic phases using Keras/Tensorflow (Gulli et al., 2019).

Training and validation were executed on an NVidia RTX A3000 GPU. We trained our learning algorithms through stochastic gradient descent. The optimal model was obtained after 4000 iterations using an early stopping technique. The training was terminated when the loss did not decrease with a minimum delta of `1e-4` and a patience value of 50 epochs. We used Adam optimization with a batch size of 64, learning rate $\alpha = 3 \times 10^{-5}$ and hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

We evaluated our learning models by computing the average RMSE and the average nRMSE, resulting from the Leave-One-Out procedure by subject and the Leave-One-

Out procedure by task. The first procedure involves iteratively removing each subject from the training set and testing on that subject. The second procedure involves creating five groups of tasks `[175cm, Right]`, `[75cm, Front]`, `[0cm, Front]`, `[0cm, Right]`, `[175cm, Front]`, with tasks iteratively removed from the training set for testing.

The quantitative evaluation of the regression results was done using the root mean squared error (RMSE) between reference and estimated joint torques, represented by ϵ and the normalized root mean squared error (nRMSE) represented by ϵ_r.

3.5 Results and discussion

Static phase Average RMSEs and standard deviations for the four considered architectures are given in table 3.2. In the inter-subjects scenario, the neural network algorithm achieved the lowest ϵ (5.54 N.m) and the lowest ϵ_r of 0.03%. In the inter-tasks scenario, the linear regression algorithm achieved the lowest ϵ (5.11 N.m) and the lowest ϵ_r of 0.05%. Even if the neural network achieved a better performance from one subject to one other, the results suggest that it may not generalize to other tasks. Such a behavior may be due to overfitting, that is a quite common issue for such a simple problem. Indeed, the static problem is linear and it was well estimated from the linear regression. The figure 3.3 shows the estimations made by the neural network on a test set consisting of trials from a subject that was not included in the training set, in both figures, we observe that the learning model encountered difficulties in accurately estimating the torques. The presence of postural variability in the test data in the `0cm` trials may have further contributed to the model's inaccuracy. Further investigation on the impact of postural variability on the model's performance could be a topic for future research.

	Inter-subjects scenarios		**Inter-tasks scenarios**	
Algorithms	$\epsilon \pm \rho$ **(N.m)**	$\epsilon_r \pm \rho$ **(%)**	$\epsilon \pm \rho$ **(N.m)**	$\epsilon_r \pm \rho$ **(%)**
Decision Tree	8.53 ± 1.62	0.05 ± 0.01	6.80 ± 2.22	0.06 ± 0.02
Linear Regression	6.96± 1.49	0.04 ± 0.01	**5.11 ± 0.87**	**0.05 ± 0.03**
Neural Network	**5.67 ± 0.91**	**0.03 ± 0.01**	5.51 ± 0.67	0.06 ± 0.03
Random Forest	7.48 ± 1.43	0.04 ± 0.01	5.68 ± 1.66	0.05 ± 0.02

Table 3.2 – Static phases Inter-Subjects Scenarios Results and Inter-Tasks Scenarios Results.

Dynamic phase Average RMSEs and standard deviation for the four considered architectures are given in table 3.3. We observed in inter-subjects scenario that all four algorithms have similar mean values of ϵ, with the lowest mean value and standard deviation achieved by the CNN-LSTM-Attention. This suggests that the CNN-LSTM-Attention algorithm was not only more accurate on average, but also more consistent in its performance across multiple runs. In inter-tasks scenarios the CNN-LSTM algorithm had

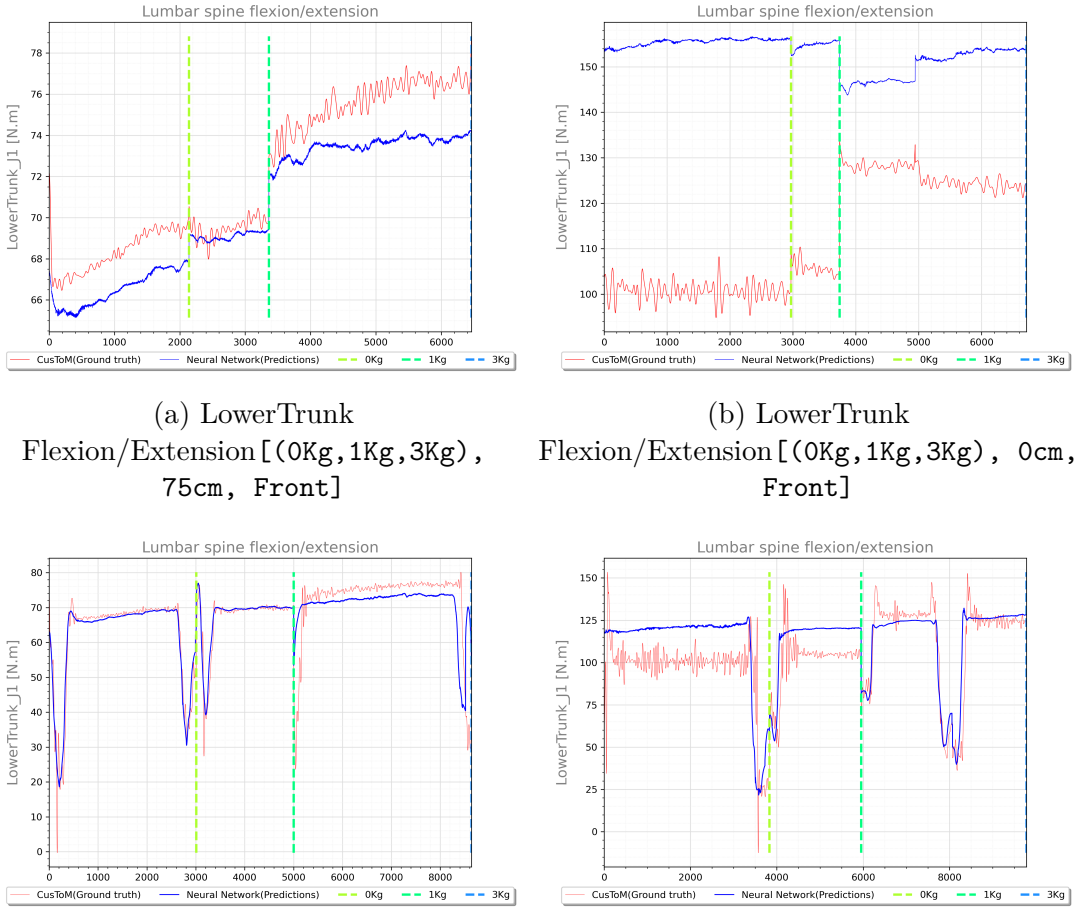
the lowest mean value of ϵ (5.56 N.m). The CNN-BiLSTM-Attention algorithm has a slightly higher mean value of ϵ (5.85 N.m) but still performs reasonably well with a similar standard deviation of 1.72 N.m. On the other hand, the two attention-based models, CNN-LSTM-Attention and CNN-BiLSTM-Attention, have higher mean values of ϵ (5.80 N.m and 5.81 N.m, respectively) and also have slightly higher standard deviations (1.58 N.m and 1.55 N.m, respectively). In general, these results suggest that the addition of attention mechanisms to the CNN-LSTM and CNN-BiLSTM models did not improve their performance on this particular task. It is noteworthy to mention that the variations in the algorithms' performance are comparatively minor, with a maximum difference of only 0.25 N.m between the best and worst performing algorithms. Both figures **3.3** show that the learning model for the dynamic phase predicts the static phases better compared to the model designed for the static phase.

	Inter-subjects scenarios		**Inter-tasks scenarios**	
Algorithms	$\epsilon \pm \rho$ **(N.m)**	$\epsilon_r \pm \rho$ **(%)**	$\epsilon \pm \rho$ **(N.m)**	$\epsilon_r \pm \rho$ **(%)**
CNN-LSTM	6.04 ± 1.72	$0.018 \pm 5 \times 10^{-3}$	**5.56 ± 1.49**	**0.015** $\pm 5 \times 10^{-3}$
CNN-LSTM-Attention	**5.73± 1.41**	**0.017** $\pm 5 \times 10^{-3}$	5.80± 1.58	$0.016 \pm 5 \times 10^{-3}$
CNN-BiLSTM	5.82 ± 1.72	$0.017 \pm 5 \times 10^{-3}$	5.66 ± 1.58	$0.016 \pm 5 \times 10^{-3}$
CNN-BiLSTM-Attention	5.85 ± 1.60	$0.017 \pm 5 \times 10^{-3}$	5.81 ± 1.55	$0.016 \pm 6 \times 10^{-3}$

Table 3.3 – Dynamic phases Inter-Subjects Scenarios Results and Inter-Tasks Scenarios Results.

3.6 Conclusion

This chapter evaluates learning models to estimate joint torques using skeletal data, subject mass, and load mass, as an alternative to standard inverse dynamics methods for future application to noisy and low-frequency data. The resulting mean torque RMSE for the corresponding L5/S1 moments (τ_2 and τ_3, see table 3.1) were equal to 7.29 ± 2.24 N.m and 5.52 ± 1.41 N.m, respectively, for the static phases, and 10.70 ± 2.60 N.m and 5.74 ± 1.41 N.m, respectively, for the dynamic phases. These values are in line with the results of (Mohseni et al., 2022). The results showed a better performance of the estimators used for the dynamic phases than those used for the static phases. Better performance in dynamic phases may be explained by higher variations of torques experienced in this phase. Following this first result, we may potentially significantly enhance the performance of our learning model CNN-LSTM-Attention by replacing the spatiotemporal features extraction component with a pre-trained model and by increasing the number of frames. As the inverse dynamics problem is determined by the equation of motion, we may potentially utilize this prior knowledge to employ Physics Informed Neural Networks (PINNs), which integrate known physical laws and principles into their



(a) LowerTrunk Flexion/Extension `[(0Kg,1Kg,3Kg), 75cm, Front]`

(b) LowerTrunk Flexion/Extension `[(0Kg,1Kg,3Kg), 0cm, Front]`

(c) LowerTrunk Flexion/Extension `[(0Kg,1Kg,3Kg), 75cm, Front]`

(d) LowerTrunk Flexion/Extension `[(0Kg,1Kg,3Kg), 0cm, Front]`

Figure 3.3 – Sample trials. Computed joint torques (red line) and estimated joint torques (blue line) based on the test data using the neural network during **(a-b)** the static phase and CNN-LSTM-Attention during **(c-d)** the dynamic phase are represented.

architecture and training processes (J. Zhang et al., 2022). A limitation of the learning-based approach is the impact of the quantity and quality of the data used for training (Budach et al., 2022; Habehh & Gohel, 2021). We can increase our training set by using simulated data. Our results tend to show that most of the methods tend to generalize to subjects (inter-subjects tests) that were not used during training, which is promising, and should be confirmed by more extensive tests.

At the conclusion of this study, we evaluated learning-based approaches for the direct estimation of joint torques from input data, with the aim of simulating inverse dynamics using a deep learning model. The next chapter will focus on evaluating and improving the generalizability of deep learning-based methods, such as OpenCap (Uhlrich et al., 2023), on unseen data. These methods aim to map the joint centers to precise anatomical landmarks, which can then be used as input to an inverse dynamics method.

Chapter 4

Generalization of Inverse Kinematics Frameworks Based on Deep Learning to New Motor Tasks and Markersets

4.1 Introduction

Unlike the previous chapter, which focused on estimating joint torques directly from a set of input data, this chapter shifts its attention to enhancing the generalizability of OpenCap learning algorithms. As discussed in the introduction and in the related work section (see the part 1), when dealing with biomechanical factors, most assessment methods rely on estimating joint angles. IK aims to compute these quantities according to a predefined and scaled skeleton, aligned with experimental positions of anatomical markers. It is the first step of several assessments, such as completing ergonomic assessment grids or calculating mechanical joint constraints (joint forces and torques). Based on precise, low-noise, high-frequency motion data, inverse kinematics is formulated as a global optimization problem at each frame, with the objective of minimizing the distance between experimental markers and kinematic model markers (Lu & O'connor, 1999). Nevertheless, obtaining precise, low-noise and high-frequency motion capture data in real industrial work conditions is impractical due to several constraints: the time required for installation and subject preparation, the significant space needed for equipment setup, and the extensive processing time involved. Recent advances in computer vision and deep learning offer the possibility to use repeatable posture measurements on site, in industrial context, with a simple RGB camera. For example, (Abobakr et al., 2019) leverages deep learning and vision-based techniques to estimate joint angles directly from single depth images. Other authors (Plantard, Muller, et al., 2017) showed that correcting Kinect data and adapted inverse dynamics approach, enables to correctly estimate internal joint torques, which provides relevant information for ergonomic assessment in real working environment. Several companies and researchers have proposed RGB-based Human Pose Estimation (HPE) as a promising alternative for biomechanical analysis of human movement in industrial

settings. This method allows for movement analysis using just a smartphone, without the need for calibration or markers, and imposes minimal constraints on workers, who can perform their tasks as usual. Despite these promising advancements, a systematic review in (Egeonu & Jia, 2024) highlights that while RGB-based HPE is convenient and minimally intrusive, it typically provides sparse 3D keypoints. This sparse noisy information might be not sufficient to compute well admitted ergonomic assessments grids, or compute physical values. Hence, previous studies (Falisse et al., 2023) have reported an inaccuracy of 5 degrees in the joint angle estimation when using HPE compared to those obtained with optoelectronic systems, but these tests are generally performed in laboratory condition. However, HPE generally returns sparse 3D keypoints information, such as 3D joint centers solely. Inverse kinematics based on this sparse data consequently leads to higher error rates, compared to using 3D positions of a large set of anatomical markers as input (Uhlrich et al., 2023).

Opencap (Uhlrich et al., 2023) has recently proposed to overcome this limitation by augmenting the number of anatomical markers based on the sparse joint positions. It consists in an open-source platform for computing both kinematic (i.e., motion) and dynamic (i.e., forces) variables using videos captured from two or more smartphones. The calibrated videos are used by HPE systems to estimate sparse 3D keypoints trajectories. Then, Opencap proposes a marker augmenter (based on deep learning) algorithm that estimates additional anatomical markers positions based on these few available 3D keypoints. The resulting anatomical markers can be used by standard IK algorithm to estimate joint angles, and apply inverse dynamics. Opencap marker augmenter contains two deep learning (DL) models, namely the **Body Model** and the **Arm Model**. The **Body Model** aims at predicting the 3D positions of the lower-limb and torso anatomical markers. The **Arm Model** aims at predicting the 3D positions of the two arms anatomical markers. These models have been trained and tested on a dataset that contains the following motions:walking, running, squatting, cutting, drop, jumping, and stair ascending and descending. This dataset has also been obtained with a given set of experimental conditions (such as camera intrinsic and extrinsic parameters, 3D keypoints definitions, etc.) and for a given output set of anatomical markers. To the best of our knowledge, the ability of these models to generalize to new experimental conditions and different sets of anatomical markers has not yet been explored.

In Deep Learning, generalization aims at adapting the model: to understand the patterns and relationships within its training data and apply them to previously unseen examples, from within the same distribution as the training set. A more complex problem consists in extending this generalization to unseen examples from within a different distribution, i.e. a set of examples that have never been used for training and testing. Transfer learning consists in using a model trained on one task as the starting point, as a basis for a model addressing a new task, or on data with different distribution (Zhuang et al.,

2020). This is done by transferring the knowledge that the first model has learned about the features of the input and output data to the second model. This is an interesting approach to train a new Opencap marker augmenter model that is able to handle new types of motion and markersets. Hence, fine-tuning or adapting a pre-trained model to a labeled target dataset (Han et al., 2024), represents a prevalent methodology in transfer learning, and is progressively establishing itself as a standard procedure within the computer vision and natural language processing research communities (Shi et al., 2024).

For example, ResNet (He et al., 2016) and EfficientNet (Tan & Le, 2019) architectures, initially trained on the ImageNet dataset (Deng et al., 2009), are extensively fine-tuned for a multitude of computer vision applications. Concurrently, models such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020), which are pre-trained on extensive corpora, exhibit robust performance across a wide spectrum of NLP tasks. There are multiple approaches to implementing fine-tuning of deep networks in practice. A common method is to optimize all the parameters of the deep network using the target training data, after initializing them with the pre-trained model's parameters. However, when the target dataset is small and the network parameters are numerous, this can lead to overfitting (Yosinski et al., 2014). Alternatively, one may fine-tune only the last few layers of the deep network, while keeping the parameters of the initial layers fixed at their pre-trained values (Mao et al., 2023). This approach is motivated by the limited training data in the target task, and the empirical evidence that initial layers learn low-level features that are transferable across various but similar tasks. However, this approach assumes that the input data have the same nature and distribution, which may not be the case if a different HPE or set of keypoints is used as inputs of the Opencap marker augmenter. Moreover, determining the optimal number of initial layers to freeze remains a manual and potentially inefficient process, particularly for networks with hundreds or thousands of layers. We also have to figure out that fine tuning generally has to deal with a small dataset containing the new distribution, which may rapidly lead to overfitting.

One of the main objectives of this chapter is to evaluate the accuracy of the Opencap marker augmenter (Falisse et al., 2023) when dealing with new types of motion, such as those frequently used in industry: bimanual tasks, including asymmetric handling tasks (denoted Lifting Movement), and handling and picking tasks (denoted Picking Movement). These tests also involve different experimental set-up/conditions and different definitions of the anatomical markers. Each company may have its own markerset, HPE with predefined 3D keypoints, and specific motions. Hence, by performing these evaluations, we aim at dealing with similar constraints than these companies may face to adapt the Opencap system. Hence, for each new task, HPE system or specific markerset, the company should be able to collect a small set of motions (concurrently with the HPE and ground truth values) to retrain the system before exploiting it on several workstations and places.

The second main contribution of this chapter is to propose a method to retrain the Opencap marker augmenter to handle such new conditions, with a limited set of examples. To this end, we explored two main fine tuning strategies for the **Body Model** and **Arm Model**. The first strategy consists in retraining all the layers of the DL architecture, assuming that the resulting models could better adapt to the new condition, compared to retraining only part of the network. However, this involves to adapt a huge number of parameters, while the number of examples of the new dataset may be small. Hence, it may lead to overfitting, with difficulties to generalize to new data in the future. The second strategy consists in tuning only the last output layers (to deal with the different output markerset), while freezing the remaining of the network. It leads to a smaller number of parameters to adapt, which may be more appropriate for the available small dataset of new examples. The study presented in this chapter has been submitted to the International Journal of Industrial Ergonomics and is currently in the first revision stage (See 4.5).

4.2 Materials and methods

In this section, we introduce the experimental data and methods used to evaluate and train the Opencap marker augmenter models. We also describe the fine tuning processes used to adapt the models to new upper-limb industrial motions.

4.2.1 Overview

In this study, we evaluated two fine tuning training strategies to adapt the Opencap marker augmenter models to new motions, input and output data. These models aim at estimating a dense set of anatomical markers based on sparse 3D video keypoints computed by HPE methods. Our proposed experimental pipeline consists of two phases. In the initial phase, we fine tuned Opencap's marker augmenter models (the **Body Model** and **Arm Model**) using two different strategies. In the subsequent phase, we applied geometric calibration and inverse kinematics based on the resulting anatomical markers to compute joint angles, as illustrated in Figure 4.1.

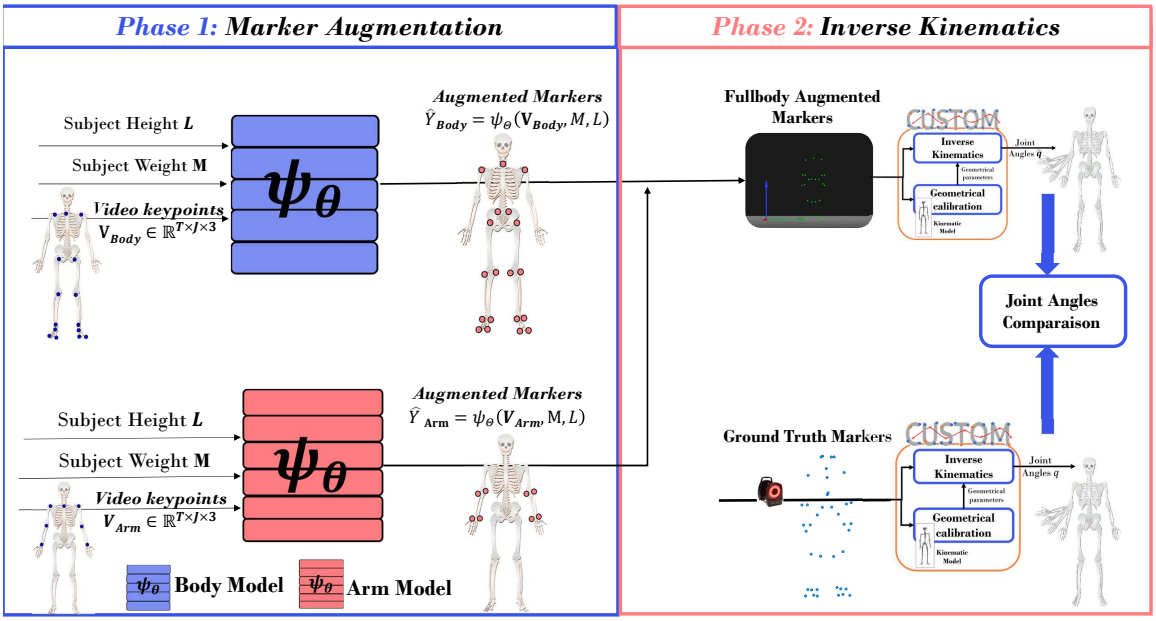


Figure 4.1 – The Proposed Pipeline: Fine tuning the Opencap marker augmenter **Body Model** and **Arm Model** Models to better estimate anatomical markers based on sparse 3D video keypoints, followed by using Custom software to calibrate a biomechanical model and apply inverse kinematics to compute the related joint angles.

In this section, we first recall the Opencap marker augmenter models (see Subsection 4.2.2). Next, we describe the fine tuning process of these models (see Subsection 4.2.3). To evaluate the two fine tuning strategies, we collected a dataset of upper-limb motions (see subsection 4.2.4). We then applied geometric calibration and inverse kinematics to estimate joint angles (see subsection 4.2.5). Finally, we evaluated the resulting anatomical markers and joint angles against ground truth values (see subsection 4.2.6).

4.2.2 Opencap marker augmenter models

The Opencap marker augmenter models (Uhlrich et al., 2023) aim at computing dense anatomical markers position according to sparse 3D video keypoints provided by HPE methods. The 3D video keypoints delivered by the HPE model, and the output anatomical markers are detailed in Supplementary material section .1.1.

As described above, the Opencap marker augmenter is based on two models associated with various body parts. The **Body model** architecture comprised four Long Short-Term Memory (LSTM) layers, each with 128 units, followed by an output layer, as illustrated in figure 4.2. It aims at predicting the 3D positions of 35 body anatomical markers thanks to 15 3D positions of lower-limb and torso 3D video keypoints, along with subjects-specific parameters such as height and weight.

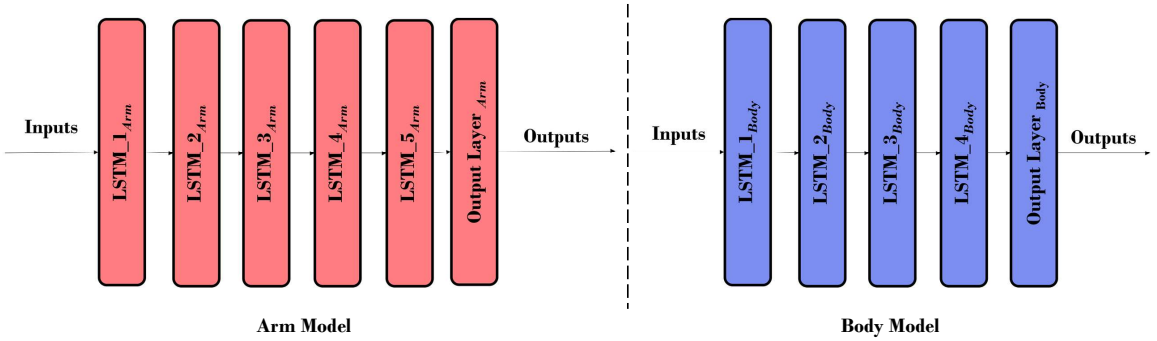


Figure 4.2 – Detailed architecture of Opencap learning models (**Body Model** and **Arm Model**)

The **Arm model** architecture is composed of five stacked Long Short-Term Memory (LSTM) layers, each comprising 128 units, followed by an output layer (as illustrated in figure 4.2). It aims at predicting the 3D positions of 8 arm anatomical markers using 7 3D positions of arm and torso video keypoints, along with subject height and weight.

4.2.3 Fine tuning the Opencap marker augmenter models

In this subsection, we describe how the Opencap marker augmenter is fine tuned to adapt to new anatomical landmark and to a new dataset composed of unseen motion. The same datasets, asymmetric handling tasks (denoted as "Lifting Movement") and handling and picking tasks (denoted as "Picking Movement"), were used to train and test both fine-tuning strategies. We also tested the direct use of the pretrained Opencap augmenter models, denoted ***Inference*** in the remaining of the chapter.

For all the strategies, the objective of the fine tuning process is to learn the mapping function:

$$\psi_{\Theta}(\boldsymbol{V}, \boldsymbol{M}, \boldsymbol{L}) = \hat{Y} \tag{4.1}$$

where $\boldsymbol{V}$ stands for the input features (the sparse 3D keypoints obtained with the HPE at time t), $\boldsymbol{M}$ and $\boldsymbol{L}$ are subject's weight and height respectively. The output of this function is the position of the additional anatomical markers $\hat{Y}$ at a given time step.

Fully strategy

A first strategy consists in retraining all the network, including the input, the intermediate LSTM, and the output layers. Hence, we updated the parameters of all layers in the network based on gradients computed from the new dataset (Fu et al., 2023). We assumed that fine-tuning all layers of the pre-trained model will allow it to better learn features related to the new tasks/motions at all the layers of the network. As the set of output markers is slightly different from the one initially used in Opencap, we need either to adapt and retrain the output layer, or to add a new output layer. As the number and

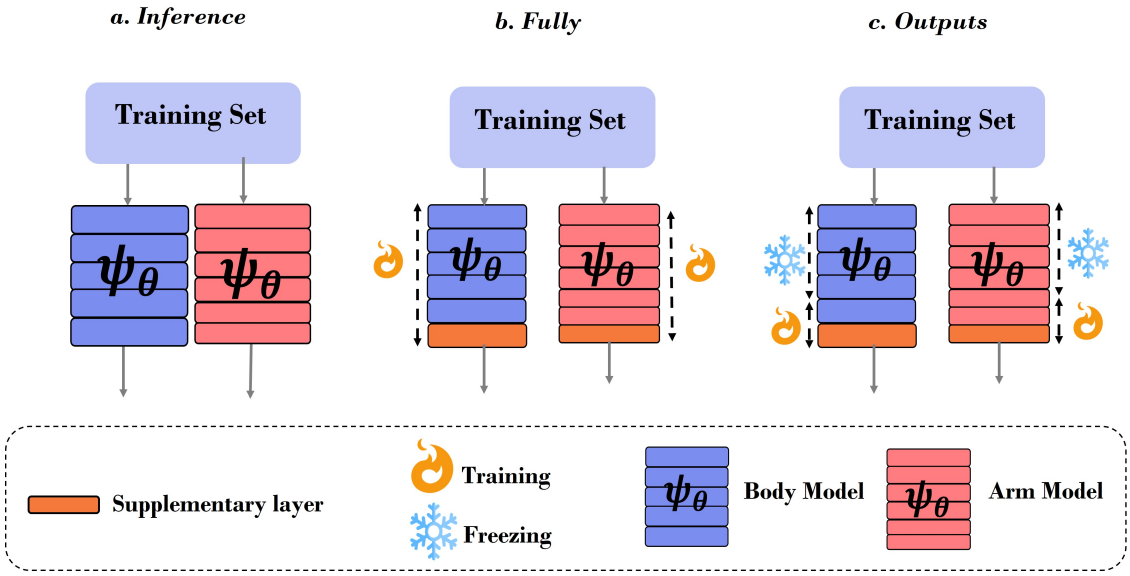


Figure 4.3 – Inference and the two Fine-tuning strategies. a. ***Inference*** consists in directly applying the pretrained marker augmenter models on the new dataset. b. ***Fully*** consists in retraining all the original network + an additional output layer with a subset of the new dataset. c. ***Outputs*** consists in adding an output layer and retraining only the two resulting output layers (the remaining layers are frozen) with a subset of the new dataset.

definition of the output markers may differ, we preferred the latter solution: we added an output layer corresponding to the new set of anatomical markers, as illustrated in figure 4.2. The already existing layers were initialized with the pretrained values of Opencap marker augmenter model to start from a pretrained initial point. As no pretrained value is available for the new output layer, it was initialized using a normal distribution, with a mean of 0 and a standard deviation of 0.022 ($\mathcal{N}(0, 0.02^2)$). During the fine tuning phase, a weight decay of 0.01 was applied to the parameters of the last added layers, excluding biases, in accordance with the methodology outlined by (Barone et al., 2017; D. Li & Zhang, 2021), with the aim to avoid overfitting.

Outputs strategy

The second strategy, denoted ***Outputs*** strategy consists in retraining only the last output layer, while freezing the remaining ones. We hypothesized that features in the first layers were strongly linked to the input data processing, which is supposed to be similar in both the new dataset and those used to train the original Opencap marker augmenter models. However, this hypothesis is not completely true, as the 3D video keypoints may also differ a bit depending on the HPE that is used. Similarly, the last layers are supposed to be linked to the output data, namely the output estimated anatomical markers (Bordes et al., 2022), which are supposed to be different from the ones used to train the initial Opencap marker augmenter. Hence, we propose to freeze all layers except the last one (output layer). As for the ***Fully*** strategy, we also added a supplementary output layer

to handle the new markerset. This method presupposes that the pretrained model has acquired valuable hierarchical features transferable to the new task. By preserving these features and solely adjusting the output layers (both the original output and new inserted layers), the model could swiftly adjust to the new task while mitigating the risk of overfitting, especially when working with a limited dataset. Let $\Theta_{\text{past\_output}}$ represent the parameters of the past output layers (respectively, $\Theta_{\text{past\_output}_{\text{Body}}}$ and $\Theta_{\text{past\_output}_{\text{Arm}}}$), and let $\Theta_{\text{new\_output}}$ represent the parameters of the new output layers. Here, X and Y represent the input and output data for this stage, respectively. The objective function J quantifies the performance of this stage's model.

$$J(\Theta_{\text{past\_output}}, \Theta_{\text{new\_output}}, X, Y) = J_{\text{task}}(\Theta_{\text{past\_output}}, \Theta_{\text{new\_output}}, X, Y) + \lambda R(\Theta_{\text{new\_output}})$$

Where J_{task} denotes the original task loss, which in this case is the mean squared error. The regularization term R is introduced to prevent overfitting; in this implementation, L^2 regularization is applied with a regularization parameter of 0.01. The hyperparameter λ controls the regularization strength, determining the trade-off between fitting the training data and minimizing the complexity of the model.

4.2.4 Datasets

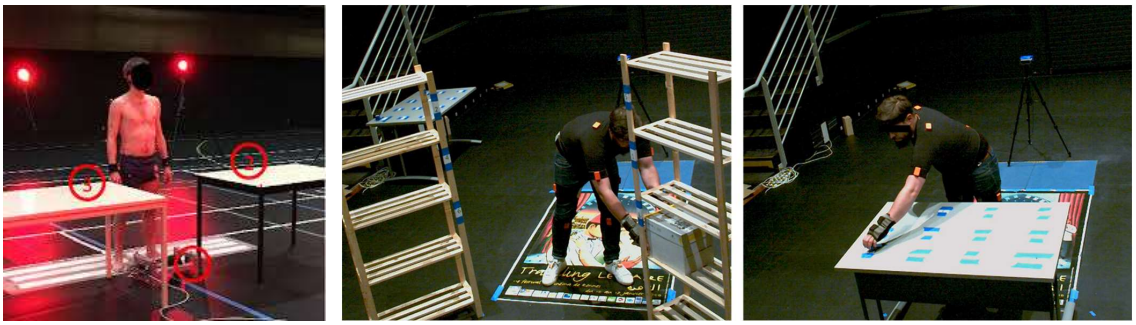
As the Opencap augmenter models were originally mainly trained on lower-limbs motions, such as locomotion, we collected motion capture data associated with upper-limb motions, as mostly used in industry. Hence, we used data collected in two different experiments: asymmetric handling tasks, and handling and picking tasks. Not only the motion are different, but also the markersets, which is an interesting property for testing the fine tuning strategies.

The denoted "Lifting dataset" consists in asymmetric handling tasks (Muller, Pontonnier, & Dumont, 2019). It involves thirteen male participants who had to move a load between three areas, leading to cycles of three displacements: from area 1 to area 2, area 2 to area 3, and area 3 back to area 1. Each participant completed two cycles with a standard load of 6.9 kg and two cycles with an additional 3 kg load. The experimental setup included 47 motion capture markers on standardized anatomical markers, following the recommendations of the International Society of Biomechanics (Wu, Cavanagh, et al., 1995). Motion capture data was recorded at 200 Hz using a 16-camera Vicon motion capture system; considered as the reference system for the experiment, as illustrated in the figure 4.4. The 200Hz resulting data were downsampled to 60Hz, similarly to the video data used to train Opencap. The input 3D keypoints were estimated using the described method in Supplementary material section .1.1. The resulting data (estimated 3D keypoints and ground truth anatomical markers) were used to retrain the models,

and perform the quantitative comparison between predicted and actual joint angles and anatomical landmark positions.

The denoted "Picking dataset" consists in handling and picking tasks. It involves 12 participants (3 women and 9 men, age: 32.6±10 years, height: 1.73±0.079 m, weight: 76± 16 kg). Participants were filmed (to use real HPE system) and equipped with the XSens inertial motion capture system. Once the skeleton model of each subject was calibrated, the XSens software (Roetenberg et al., 2009) simulated skin marker positions, including also the ones following the recommendations of the International Society of Biomechanics. The objective of this experiment was to emulate real work conditions: bimanual handling and picking tasks. Bimanual handling involved picking up a box (dimensions: 39×29.5×19 cm) from a three-tiered shelf and placing it on another shelf, repeating this process 5 times following a specified order on the shelves. Picking task required picking up and replacing a small cubic object (dimensions: 5 × 5 × 5 cm) at 16 different locations arranged on a table in front of the participant, following a specific order. The participants had to perform picking in ascending and then descending order, using their right and left hands, respectively. In total, each participant performed 4 picking actions. In the context of this chapter, these tasks are interesting to challenge the HPE system, as they involve external occlusions (with the box and the table) and self-occlusion depending on different measurement viewpoints. Consequently, it may affect the quality of the HPE outputs before estimating the anatomical markers using the Opencap augmenter models. Whereas the Opencap system required multiple calibrated cameras, we used a single RGB camera, placed facing right during the experiment (see figure 4.4). To process the unique RGB camera, we used the KIMEA Cloud solution developed by Moovency. The video and XSens files were synchronized using a clapping signal at the start and end of each trial. Spatial alignment involved removing translational and orientation information from the resulting 3D pose data, ensuring that each 3D pose captured only the execution of motion, independent of location or viewpoint, as detailed in previous studies (X. Chen & Koskela, 2013; Yasin et al., 2023; Yasin et al., 2020).

Before the training phase, we expressed the 3D positions of anatomical markers relatively to a root marker, specifically the midpoint of the hip keypoints. Additionally, we normalized these 3D positions based on the subject's height. The data was standardized to have zero mean and unit standard deviation, before being used for retrained the Opencap marker augmenter models.



(a). Asymmetric handling tasks **(b). Handling and picking tasks**

Figure 4.4 – The experimental protocol for the asymmetric handling tasks (denoted lifting tasks) is presented in sub-figure **(a)**, as detailed in the study by (Muller, Pontonnier, & Dumont, 2019). The protocol for the handling and picking tasks (denoted picking tasks) is presented in sub-figure **(b)**.

4.2.5 Inverse kinematics

Opencap augmenter models are used to estimate the anatomical markers that are necessary to calibrate a biomechanical model and perform inverse kinematics, to compute the joint angles.

The whole body biomechanical model consisted of eighteen segments: *upper/lower trunk,left/right clavicle left/right arm, left/Right forearm, left/right thigh, left/right shank, left/right foot, and pelvis*. This model was articulated with 42 degrees of freedom, comprising a 6 degrees of freedom (DoFs) mobile base and 43 anatomical joint angles in accordance with recommendations from the International Society of Biomechanics (Wu, Cavanagh, et al., 1995) and summarized in table 4.1. Considering the output markerset delivered by the Opencap marker augmenter models, head and hands segments were removed from the initial model (no available head and hands markers in the markerset).

Joint	Corresponding exertion	Joint	Corresponding exertion
Lumbar Spine F/E	Lumbar Spine Flexion/Extension	**R/L Hip F/E**	R/L Hip Flexion/Extension
Lumbar Spine LF/LE	Lumbar Spine Lateral flexion/extension	**R/L Hip A/A**	R/L Hip Abduction/Adduction
Lumbar Spine I/E	Lumbar Spine Axial Rotation	**R/L Hip I/E**	R/L Hip Internal/External rotation
Thoracic Spine F/E	Thoracic spine Flexion/Extension	**R/L Knee F/E**	R/L Knee Flexion/Extension
Thoracic Spine LF/LE	Thoracic Spine Lateral Flexion/extension	**R/L Ankle F/E**	R/L Ankle Flexion/Extension
Thoracic Spine I/E	Thoracic Spine Axial Rotation	**R/L Ankle I/E**	R/L Subtalar Inversion/Eversion
R/L Clavicle P/R	R/L Clavicle Protraction/Retraction	**R/L Elbow F/E**	R/L Elbow Flexion Extension
R/L Clavicle D/E	R/L Clavicle Depression/Elevation	**R/L Forearm P/S**	R/L Forearm Pronation/Supination
R/L Clavicle I/E	R/L Clavicle Axial Rotation	**R/L Glenohumeral PoE**	R/L Glenohumeral Plane of Elevation
R/L Glenohumeral D/E	R/L Glenohumeral Depression/Elevation	**R/L Glenohumeral nPoE**	Negative Glenohumeral plane of elevation
R/L Glenohumeral I/E	Glenohumeral Internal/External rotation		

Table 4.1 – Biomechanical model depicting joint angles with the following notations: **R/L** indicates Right/Left, **F/E** denotes Flexion/Extension, **LF/LE** represents Lateral Flexion/Lateral Extension, **I/E** stands for Internal/External, **P/R** refers to Protraction/Retraction, **D/E** signifies Depression/Elevation, **PoE** is Plane of Elevation, **nPoE** denotes Negative Plane of Elevation, **A/A** stands for Abduction/Adduction, **I/E** indicates Inversion/Eversion, and **P/S** represents Pronation/Supination.

We used the Custom software (Muller et al., 2017; Muller, Pontonnier, Puchaud, et al., 2019; Puchaud et al., 2020) to perform the geometrical calibration of the model according to the estimated anatomical landmark positions. This calibration was formulated into an optimization problem trying to minimize the distance between the experimental markers and the corresponding anatomical points of the model, by adjusting the segment lengths. This method was applied with ground truth motion capture data, and anatomical markers estimated by the Opencap augmenter models.

Once the calibration was performed, Custom was again used to perform inverse kinematics: a penalty method for constrained multibody kinematics optimization using the Levenberg-Marquardt algorithm (Livet et al., 2023). This method aimed to determine the joint angles $\mathbf{q}$ according to the position of the 3D anatomical markers.

4.2.6 Evaluation methodology

The goal of this work was to evaluate the performance of the Opencap augmenter models to predict the position of dense anatomical markers based on sparse 3D keypoints. As described in subsection 4.2.4, we used two types of datasets:

— Lifting dataset: the asymmetric handling task, composed of ground truth optical motion capture data,

— Picking dataset: the handling and picking tasks, composed of ground truth XSens motion capture data and RGB videos.

Let us consider now the evaluation metrics and implementation details used for this work.

Evaluation metrics

As the markersets are different in all the datasets, compared to the one used by the Opencap pretrained models, we only compared markers with similar definitions. The **Body Model** inputs are composed of 15 video keypoints, and the subject's Height and Weight. The outputs consist in 35 outputs markers, with 19 of them corresponding to anatomical markers that also exist in our two markersets. Thus, we did not use the following markers from the original **Body Model** to carry-out the comparisons: [r_thigh1_study, r_thigh2_study, r_thigh3_study, L_thigh1_study, L_thigh2_study, L_thigh3_study, r_sh1_study, r_sh2_study, r_sh3_study, L_sh1_study, L_sh2_study, L_sh3_study, RHJC, LHJC]. The **Arm Model** input data consist in 7 video keypoints, and the subject's weight and height. Its outputs are composed of 8 anatomical markers, similar to our anatomical markers.

Hence, for each motion clip of the datasets, we can compare the landmark position and joint angles estimated by Opencap augmenter models (using either motion capture or video input data) to ground truth values. For each trial, we can compare the results of direct ***Inference*** of the Opencap data augmenter models, without retraining, to those obtained with the ***Fully*** and ***Outputs*** fine tuning strategies.

For this comparison, we computed the average Root Mean Square Error RMSEm and the corresponding standard deviations (ρm) to quantify the disparities between measured and estimated 3D positions of anatomical markers. Additionally, we estimated the 95% confidence interval (CI) to further assess the precision of the measurements (Simundic et al., 2008).

Similarly, for the resulting joint angles, after IK, the average root mean squared error (RMSE_{jc}) and corresponding standard deviation (ρ_{jc}) were computed, along with the 95% confidence interval (CI) for these measures were computed, to compare joint angles obtained from ground truth marker position and augmented models ones. For **Mean Error (All joint angles)**, we considered all the angles of the biomechanical model. For **Mean Error (OpenCap joint angles)**, we only considered the following angles: [**R/L Hip F/E, R/L Hip A/A, R/L Hip I/E, R/L Knee F/E, R/L Ankle F/E, R/L Ankle I/E, Lumbar Spine F/E, Lumbar Spine LF/LE, Lumbar Spine F/E**]. All the pelvis degrees of freedom (rotation/translation) were removed from the computation as they represent the position and orientation of the pelvis in the global coordinate system and vary depending on the experimental setup.

The evaluation was conducted using a Leave-One-Out procedure by subject. In this method, one subject is systematically removed from the training set, iteratively, and the

model is tested on that removed subject. This procedure was repeated 5 times, leading to 5 subsets of training and testing sets randomly selected among the available subjects. For the lifting task, we had 11 subjects, with an average of 120000 samples in the training set and 12000 samples in the test set. For the picking task, we had 13 subjects, with an average of 160000 samples in the training set and 18000 samples in the test set.

Implementation details

We implemented our learning algorithms using Keras/Tensorflow (Géron, 2022), and used an NVidia RTX A3000 GPU for training and tests. The optimal model was achieved using an early stopping technique: training concluded when the loss failed to decrease with a minimum delta of 1×10^{-4} and a patience value of 10 epochs. The Adam optimization algorithm was used with a batch size of 64 and a learning rate (α) set to 6×10^{-6}. In order to train the learning algorithms, we adopted the standard mean squared error (MSE) loss function, after processing 64 training samples: the model updates its parameters based on the average errors calculated over these 64 samples.

4.3 Results

In this section, we present the performance of the two fine-tuning strategies in comparison to using the pretrained Opencap marker augmenter models. Firstly, we compare the accuracy of the two strategies for predicting the 3D anatomical markers and evaluate the effect of varying training data sizes on model performance by training on smaller datasets (see subsection 4.3.1).

4.3.1 3D anatomical markers positions

The table 4.2 presents the average RMSE (RMSE_m) in millimeters along with their corresponding standard deviations (ρ_m) and 95% confidence interval **(CI)** for ***Inference*** and both fine tuning strategies, namely ***Fully*** and ***Outputs***, for the Lifting task. For the **Body Model**, the results show an important decrease of the prediction error from 39 ± 2 mm down to 15 ± 2 mm and and 16 ± 1 mm for the ***Fully*** and ***Outputs*** fine tuning strategies respectively. More important error decreases were observed for the **Arm Model**.

Model	**Movement**	**Data type**	***Inference [mm]***	***Fully [mm]***	***Outputs [mm]***
			$\mathbf{RMSE_m \pm \rho_m}$ **(CI)**	$\mathbf{RMSE_m \pm \rho_m}$**(CI)**	$\mathbf{RMSE_m \pm \rho_m}$**(CI)**
Body	Lifting	MoCap	$39 \pm 2(38, 41)$	$15 \pm 2(13, 17)$	$16 \pm 1(14, 17)$
Arm	Lifting	MoCap	$31 \pm 4(29, 34)$	$9 \pm 1(8, 11)$	$11 \pm 1(10, 11)$
Body	Picking	RGB	$104 \pm 14(96, 112)$	$26 \pm 1(24, 28)$	$46 \pm 4(42, 50)$
Arm	Picking	RGB	$160 \pm 41(137, 182)$	$95 \pm 13(84, 106)$	$97 \pm 6(91, 103)$

Table 4.2 – Prediction error of **Body Model** and **Arm Model** marker augmenter models for asymmetric handling movements (Lifting task) and industrial handling and picking movements (Picking task). Average RMSE ($\mathrm{RMSE_m}$) and corresponding standard deviations (ρ_m) and 95% confidence interval **(CI)** are given in millimeters. Prediction error is given when using ***Inference***, and ***Fully*** and ***Outputs*** fine tuning strategies.

For the Picking task, we obtained similar important decrease of the prediction error when using fine tuning compared to directly applying the pretrained model. For the **Body Model**, the pretrained models leaded to 104 mm and 160 mm errors for the **Body Model** and **Arm Model** respectively. Let us recall here that the Picking task involved real video and HPE as input of the system, and that the ground truth was obtained with Xsens sensors. These data may differ from those obtained to evaluate the Opencap marker augmenter models for Lifting task. For the **Body Model**, this error decreased down to 26 mm and 46 mm for the ***Fully*** and ***Outputs*** fine tuning strategies respectively. For the **Arm Model**, the error decreased from 160 mm to 95 mm and 97 mm for the ***Fully*** and ***Outputs*** fine-tuning strategies, respectively. However, even with this reduction, the fine-tuned **Arm Model** still exhibits relatively high prediction errors under these experimental conditions.

Let us consider now the computing performance of the two fine tuning strategies in the various experimental conditions. Table 4.3 reports the training time (in minutes), the number of Epochs used to converge, and the amount of parameters that were trained by both the ***Fully*** and the ***Outputs*** fine tuning strategies. As the ***Fully*** strategy retrains all the layers of the model, it leads to adapt a large amount of parameters compared to the ***Outputs*** strategy.

Model	Movement	Data type	Time	Epoch	Params	Data
Fully						
Body	Lifting	MoCap	78	166	504394	120000
Arm	Lifting	MoCap	193	241	607832	120000
Body	Picking	RGB	35	68	504394	160000
Arm	Picking	RGB	16	26	607832	160000
Outputs						
Body	Lifting	MoCap	34	162	19587	120000
Arm	Lifting	MoCap	73	300	3696	120000
Body	Picking	RGB	47	141	19587	160000
Arm	Picking	RGB	24	64	3696	160000

Table 4.3 – Performance indicators of the training process in all the test conditions: training time in minutes, number of epochs, number of trained parameters, and training data size for different tested fine-tuning strategies.

Figure 4.5 illustrates the comparison of the estimated **RSHO** (right acromion) position during inference stage and both fine tuning strategies, and compares it to the ground truth, during Lifting Task.

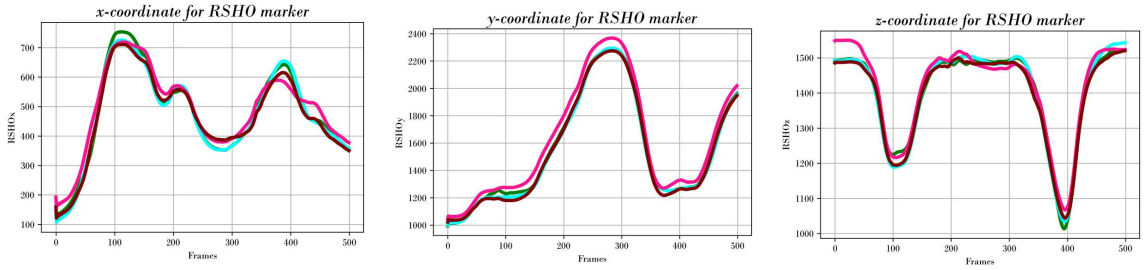


Figure 4.5 – Estimated 3D trajectory (in mm) of the **RSHO** (right acromion) anatomical landmarker using ***Inference*** (fuchsia) and the two fine-tuning strategies for Lifting Task: ***Fully*** (green) and ***Outputs*** (maroon). Ground truth value is depicted in Cyan.

To evaluate the impact of the quantity of training data that were used for training, on the fine tuning results, we also trained the **Body Model** and the **Arm Model** with less data. For the Lifting Task, the data from 5 subjects among the 11 wad used for training, and the data from one of the remaining 11-5=6 subjects was used for testing. These tests were carried-out on with the ***Fully*** and the ***Outputs*** fine tuning strategies. The results (see table 4.4) show an increase of error when using this lower quantity of training data, for all the strategies, and all the models.

Model	***Fully [mm]***		***Outputs [mm]***	
	50% dataset	100% dataset	50% dataset	100% dataset
Body	$19 \pm 3(16, 22)$	$15 \pm 2(13, 17)$	$31 \pm 2(29, 33)$	$16 \pm 1(14, 17)$
Arm	$13 \pm 2(11, 16)$	$9 \pm 1(8, 11)$	$25 \pm 1(24, 27)$	$11 \pm 1(10, 11)$

Table 4.4 – Prediction error of **Body Model** and **Arm Model** marker augmenter models for asymmetric handling movements (Lifting Task) when training with all the data or half of the dataset.

4.3.2 Joint angles estimation

The estimated anatomical markers of the lifting tasks data were used to compute the joint angles of a biomechanical model, using the Custom Software. Table 4.5 reports the average RMSE (RMSE_{jc}) and corresponding standard deviation ρ_{jc}, between the predicted joint angles and the one obtained with ground truth anatomical markers. (RMSE_{jc}) is given for the ***Inference***, ***Fully*** and ***Outputs*** fine tuning strategies. Overall, both fine tuning strategies showed improvements over the inference method in most joint estimations. For instance, in the **Right Hip F/E** joint, ***Fully*** reduced (RMSE_{jc}) from 8.2° down to 6.9°. ***Outputs*** strategy further decreased (RMSE_{jc}) down to 7.3°. Similar trends were observed across other joints. Figure 4.6 provides a visual representation of joint angles over time in all the conditions. In this figure, the right and left hip, ankle, and elbow joints are depicted.

Joint Angles	***Inference*** *[°]*	***Fully*** *[°]*	***Outputs*** *[°]*
Metric	$\mathbf{RMSE_{jc} \pm \rho_{jc}(CI)}$	$\mathbf{RMSE_{jc} \pm \rho_{jc}(CI)}$	$\mathbf{RMSE_{jc} \pm \rho_{jc}(CI)}$
Right Hip F/E	8.2 ± 1.6	6.9 ± 2.0	7.3 ± 0.9
Right Hip A/A	5.4 ± 0.5	3.4 ± 0.4	5.3 ± 1.0
Right Hip I/E	15.6 ± 1.0	7.3 ± 0.4	11.0 ± 2.5
Right Knee F/E	8.5 ± 1.0	4.0 ± 1.7	7.9 ± 1.4
Right Ankle F/E	5.2 ± 0.4	4.0 ± 1.1	5.1 ± 1.2
Right Ankle I/E	16.5 ± 2.8	8.5 ± 1.0	11.1 ± 2.2
Right Clavicle P/R	22.5 ± 2.1	20.4 ± 2.7	22.9 ± 2.6
Right Clavicle D/E	13.4 ± 2.6	$11.1 \pm 1,8$	13.0 ± 1.6
Right Clavicle I/E	36.5 ± 3.4	34.7 ± 3.4	37.6 ± 2.2
Right Glenohumeral PoE	132.9 ± 26.3	$118.3 \pm 44,8$	132.7 ± 63.2
Right Glenohumeral D/E	57.5 ± 12.8	55.0 ± 23.1	57.8 ± 22.6
Right Glenohumeral nPoE	132.9 ± 26.3	118.3 ± 44.8	132.7 ± 63.2
Right Glenohumeral I/E	20.2 ± 2.3	15.4 ± 0.4	17.6 ± 2.4
Left Hip F/E	8.6 ± 1.6	7.0 ± 2.5	7.4 ± 1.8
Left Hip A/A	4.8 ± 0.3	3.2 ± 0.4	5.0 ± 1.2
Left Hip I/E	11.9 ± 2.9	6.5 ± 1.1	10.8 ± 2.8
Left Knee F/E	9.0 ± 1.3	3.9 ± 1.8	7.7 ± 1.0
Left Ankle F/E	6.3 ± 0.6	4.1 ± 1.4	5.2 ± 1.5
Left Ankle I/E	14.6 ± 2.6	7.1 ± 1.3	10.0 ± 2.0
Left Clavicle P/R	22.6 ± 4.0	21.8 ± 3.5	22.5 ± 2.5
Left Clavicle D/E	12.9 ± 1.3	10.9 ± 1.1	11.6 ± 1.0
Left Clavicle I/E	38.2 ± 5.4	34.5 ± 5.7	35.9 ± 4.5
Left Glenohumeral PoE	132.7 ± 27.3	136.3 ± 47.8	150.4 ± 32.4
Left Glenohumeral D/E	50.0 ± 15.4	47.1 ± 11.4	49.3 ± 13.0
Left Glenohumeral nPoE	132.7 ± 27.3	136.3 ± 47.8	150.4 ± 32.4
Left Glenohumeral I/E	20.6 ± 2.6	14.7 ± 0.7	17.1 ± 1.4
Right Elbow F/E	14.6 ± 1.9	6.8 ± 0.8	8.1 ± 0.5
Right Forearm P/S	23.6 ± 6.5	14.6 ± 2.9	15.8 ± 2.9
Left Elbow F/E	13.9 ± 2.0	7.5 ± 1.6	8.0 ± 0.6
Left Forearm P/S	26.2 ± 2.4	14.9 ± 5.1	14.8 ± 4.8
Lumbar Spine F/E	16.4 ± 4.1	15.2 ± 1.7	15.6 ± 1.3
Lumbar Spine LF/LE	9.2 ± 1.0	8.4 ± 0.4	9.4 ± 0.7
Lumbar Spine I/E	51.6 ± 4.3	53.2 ± 6.4	52.3 ± 5.0
Thoracic Spine F/E	13.2 ± 0.9	13.2 ± 0.6	13.7 ± 0.7
Thoracic Spine LF/LE	13.4 ± 0.2	12.7 ± 0.6	13.9 ± 1.4
Thoracic Spine I/E	49.2 ± 4.0	49.5 ± 4.7	47.8 ± 3.9
Mean Error (All joint angles)	$32.5 \pm 5.6(29, 35)$	$28.8 \pm 7.7(22, 35)$	$31.8 \pm 7.9(24, 38)$
Mean Error (OpenCap joint angles)	$12,8 \pm 1,7(11, 13)$	$9.5 \pm 1.6(8, 10)$	$11.4 \pm 1.8(9, 13)$

are expressed in degrees.

Table 4.5 – The average error in joint angles estimation using ***Inference***, ***Fully*** and ***Outputs*** conditions for Lifting tasks. Average RMSE ($RMSE_{jc}$) and corresponding standard deviation (ρ_{jc}) and 95% confidence interval **(CI)**

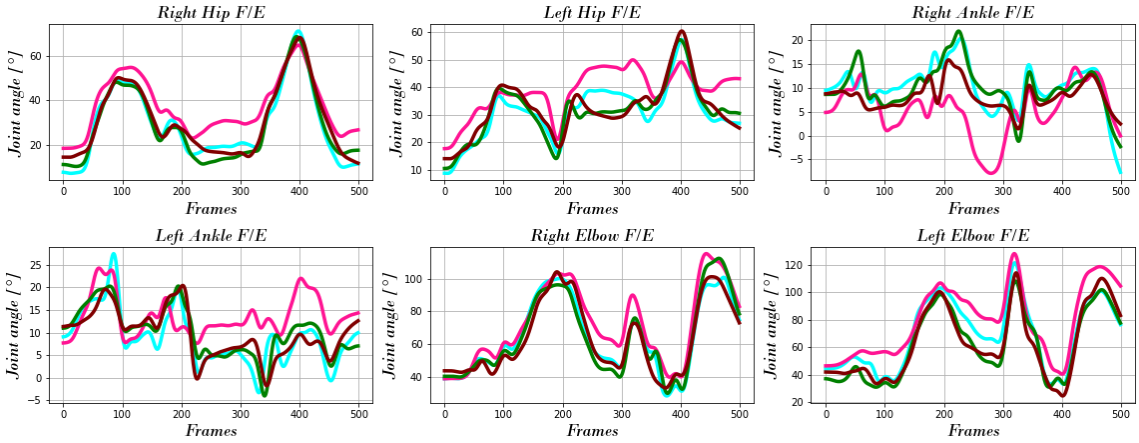


Figure 4.6 – The estimated joint angles (in degrees) for the **Right Hip F/E**, **Left Hip F/E**, **Right Ankle F/E**, **Left Ankle F/E**, **Right Elbow F/E**, and **Left Elbow F/E** joints are presented using three different methodologies: ***Inference*** (represented in fuchsia), and two fine-tuning strategies for the Lifting Task: ***Fully*** (green) and ***Outputs*** (maroon). The ground truth values are depicted in Cyan.

4.4 Discussion

The aim of this chapter was to evaluate two strategies to generalize the Opencap pre-trained marker augmenter models for the body and the arms. The two major challenges are 1) the new type of motions that was not present in the initial Opencap training set, and 2) the differences in output markersets. Subsection 4.4.1 proposes a discussion about the results obtained when predicting the position of markers which do not completely fit the original Opencap markerset. The impact of the joint angles computation using inverse kinematics is discussed in subsection 4.4.2.

4.4.1 3D anatomical markers

Considering the lifting tasks, based on Mocap data, the **Body Model** consistently demonstrated higher error values across all tested strategies compared to the **Arm Model**. Additionally, our results indicated that the ***Fully*** strategy yielded the lowest root mean square error values for both models, compared to the ***Outputs*** strategy. It is consistent with the fact that the ***Outputs*** strategy has less parameters to tune and, thus, less possibilities to find an accurate solution. However, the ***Fully*** strategy requires a huge amount of parameters (>500K parameters for both the Arm and the Body models), which may lead to overfitting as the training dataset is not big enough.

For both strategies, the error (between 9 mm and 16 mm) yielded in similar range compared to those observed after an inverse kinematics step, in classical motion capture analyses (Begon et al., 2018; Lund et al., 2015; Muller et al., 2015; Puchaud et al., 2020) (errors ranging from 4 mm to 40 mm). These results suggest that such models can be

used to generate inputs for classical inverse kinematics methods, leading to a similar level of uncertainty on the joint angles. One can think that weighting markers showing the highest ($RMSE_{jc}$) may be a good way to minimize their impact on the inverse kinematics outputs (Livet et al., 2023).

For the Picking tasks, based on RGB input videos, the **Body Model** better estimated the anatomical markers, compared to the **Arm Model**, for all learning strategies. The two fine tuning strategies enhanced performance, but the ***Fully*** strategy obtained better improvements: $RMSE_m$ decreased from almost 75% for the **Body Model**, and 40% for the **Arm Model**. However the residual error was still high (26 to 95 mm) compared to the results obtained with the Mocap data (Lifting Task). A first explanation leads to the use of a different HPE in this work compared to the original Opencap chapter. Hence, the first input layers have been trained with a slightly different definition and distribution of 3D keypoints. Only the ***Fully*** strategy can retrain the input layer, which is supported by clearly more accurate marker prediction than the ***Outputs*** strategy. Hence, it would be interesting to evaluate a new strategy, denoted ***Inputs*** strategy, that would introduce a new input layer, and retrain the two resulting input layers while freezing the remaining of the architecture. Moreover, in the Picking dataset, the reference data was obtained from XSens motion capture clips, which estimated surface anatomical markers based on inertial sensor data and a calibrated skeleton. Consequently, the way these markers were estimated was different from original motion capture data used by Opencap, to train the models. In addition, similarly to the Mocap data, the nature of the motion itself may lead to a different distribution of input-output samples, and may need additional data to properly handle this new distribution.

To conclude, although fine tuning enables us to significantly decrease the estimation error compared to using inference directly, the results obtained from RGB data for the Picking task do not seem usable as inputs for classical inverse kinematics methods, with average errors going up to 97 mm.

Although we found a larger error in the ***Outputs*** strategy compared to the ***Fully*** strategy in all cases, the number of trainable parameters was much bigger for the ***Fully*** strategy (see table 4.3). As there is an unbalance between the high number of parameters and the small size of training data, the ***Fully*** strategy may fall to overfitting (Goodfellow et al., 2016). In Alwosheel et al., 2018, authors suggest that a dataset of about 10 times the number of parameters could be enough for classical deep learning training to decrease the risk of overfitting. In our case, this rule was not respected for the ***Fully*** approach whereas it was the case for the ***Outputs*** one. This should be balanced by the fact that fine-tuning does not impact the loss in a similar manner as a full training.

We also have demonstrated that decreasing the size of the training dataset, using only 5 subjects among 11 for training, leaded to more important errors. This drop of performance is especially true for the ***Outputs*** strategy, which is tuning less parameters than the ***Fully***

strategy. This suggests that the number of trial data actually significantly affect the fine tuning performance, and should be considered for future use of this approach.

4.4.2 Joint angles

For the Lifting tasks based on optoelectronic motion capture data, the results show that the angle prediction error for **Right Hip I/E** reduced more compared to the same angle on the other body side **Left Hip I/E**. Ankle joints estimation error showed small improvements after fine tuning. The estimation error for the forearm P/S exhibited the highest $RMSE_{jc}$ values, indicating greater difficulty in accurately estimating these angles. Since the **Arm Model** was responsible for predicting the elbow and wrist markers, there was a need for targeted improvements in this specific model. Indeed, the input of the marker augmenter lacks of information about the hand and explains the poor accuracy in the estimation of the forearm joint angles. Similarly, the lack of information about the head position is an issue that may impact the lumbar and thoracic joint angles prediction. This issue suggest the development of more advanced HPE methods, able to track additional anatomical markers on the head and the hands of the subjects, which are very relevant information in ergonomics.

Figure 4.6 illustrate the resulting joint angles obtained from augmented data. Both ***Fully*** and ***Outputs*** show better results than the ***Inference*** method. Compared to previous works, we observed varying levels of accuracy among different joint angle estimation approaches, particularly in the context of walking and bipedal locomotion tasks. Previous work (Kanko, 2020) reported that Theia system has a mean angular error of 6.4°, with a range spanning from 3.3° to 11°. In contrast, Pose2Sim exhibited a mean error of 4.9°, with confidence intervals between 3.1° and 6.6° (Pagnon et al., 2021), indicating a more consistent accuracy compared to Theia, with less estimated degrees of freedom. Similarly, (Needham et al., 2022) reported a mean error of 4.9°, with slightly tighter confidence intervals from 2.9° to 6.0°, underscoring the reliability of their system. Opencap, matched Needham's system in mean error (4.9°) and confidence intervals (2.9°–6.0°). In the current study, the joint mean error on the same set of joints was 9.6±1.6° showing a slightly less accurate result but still acceptable in ergonomics for posture assessment (Plantard, Shum, Le Pierres, et al., 2017; Rodrigues et al., 2022).

Inverse kinematics, as expressed in the current chapter, is affected by several factors: soft tissue artifacts (STA), kinematic mismatch due to limited degrees of freedom (DoFs) in the model, experimental marker misplacement, geometrical calibration of the model, and measurement noise. In addition, marker augmentation through Opencap generates additional uncertainty: the learned augmented anatomical positions are inaccurate, and may be affected by postures far from the ones used to train the model. In our case, these issues may explain that the highest joint angle differences are reached for internal/external

rotations, that are the most affected by small uncertainties on the marker positions.

The biomechanical model should be questionned as well. First, the shoulder joint angle errors are very high, but this result should be considered with caution: the glenohumeral joint is modeled with a redundancy of the plane of elevation to avoid Gimbal lock issues that generates an infinity of solutions to get the proper orientation of the humerus with regard to the thoraco-scapular complex. Thus, the reconstruction error remains low, but the algorithm proposes an alternative angle sequence to place the humerus. As well, the outputs of the marker augmentation gives a limited set of information out of the sagittal plane for the trunk, leading in particularly high errors in joint angles quantifying internal/external rotations. All of those restrictions are confirmed by the fact that Opencap was evaluated using mainly lower limbs joint angles.

The calibration of the model should be taken with caution as well. Indeed, the calibrated model is based on the marker augmentation that suffers from the inaccuracy of the segment lengths, issued from the joint centers estimation. Therefore, the calibrated model may be far from the one obtained directly from the motion capture data.

4.4.3 Applicability in ergonomics and perspectives

Calibration-free approaches for ergonomic assessments (Plantard, Shum, Le Pierres, et al., 2017) rely on skeletal data that lack the precision necessary for accurate joint angle computation according to ISB standards. These methods also exhibit significant errors during occlusions. However, they offer real-time implementation. In contrast, methods such as Opencap and Pose2Sim, which require calibration, although potentially longer to use, due to the need for precise calibration processes, can better incorporate biomechanical constraints, leading to more accurate assessments. The trade-off between speed and accuracy must be carefully considered when selecting an approach for real-time ergonomic assessment. Furthermore, deploying these systems in industrial contexts requires careful consideration of factors such as the number of camera views and robustness to occlusions. At a minimum, two camera views are recommended to ensure comprehensive coverage and reliability (Uhlrich et al., 2023).

When dealing with real conditions, such as cluttered environments, occlusions, clothes, lighting conditions..., capturing the operator's motion generally leads to sparse and noisy data. In the same way, according to the complexity of the task, the operator biomechanical model may or not have some simplifications. This variability of experimental and modelling conditions may complicate the task of the pre-trained DL Opencap marker augmenter. It may also lead to important errors that may not be compatible with traditional inverse kinematics and dynamics frameworks, such as OpenSim (Delp et al., 2007), Anybody (Damsgaard et al., 2006), or Custom (Muller, Pontonnier, Puchaud, et al., 2019). In this chapter, the tested markersets were different, and the studied tasks mostly involved

upper-limb movements, contrary to the original data used to train the Opencap marker augmenter. To exploit this approach to a new output markerset, the idea supported in this chapter is to add a new output layer which contains as neurons as the 3D coordinates of the studied markerset (i.e. 3 times the number of markers). In this chapter, in the ***Outputs*** strategy, we proposed to re-train the two output layers (the original and the additional ones), which leads to 3696 parameters for the **Arm Model**, and 19587 parameters for the **Body Model**. (Alwosheel et al., 2018) consider that the ratio between the number of observations and the number of weights of an artificial neural network should be higher than 10 to limit the risk of overfitting. It means that 36960 and 195870 poses should be required to retrain the **Arm Model** and the **Body Model** respectively. Hence, for a new type of motions, it suggests to collect similar ground truth and accurate data in laboratory conditions, using for example IMU-bases or optoelectronic systems. Once these data are collected, the Opencap marker augmenter can be re-trained offline before being used with new on-site Mocap data. We also quantified the decrease of accuracy when using a much smaller set of data for training, demonstrating an important limitation of this DL based approach. For companies which develop such RGB-based ergonomic tools, it involves regularly collecting new data, with ground truth motion capture, to improve their models, or adapt to specific needs of their customers.

The results reported in this chapter tend to show that input 3D keypoints obtained with computer vision systems lead to less accurate results compared to using reference Mocap systems. Future works would be needed to evaluate the relevance of applying the same strategy for input data: adding a new input layer which is re-trained according to the new types of inputs. However, this would also require to jointly capture these 3D keypoints with a reference and the on-site systems concurrently. To take the on-site conditions into account (such as occlusions or lighting problems), it would require to move the reference system on-site, which might be difficult. Future works will explore how to optimize the re-training strategies in this condition. By retraining the input keypoints, we could expect an increase of the accuracy, as improvements observed for the output layers.

Computation time needed for training with such a dataset leads to 16 to 193 minutes according to the conditions and the fine tuning strategy. However, this computation is performed offline, which does not affect the inference computation time used to exploit the re-trained Opencap marker augmenter.

4.5 Conclusion

This study highlighted the potential of using DL-based methods, such as Opencap, for estimating joint angles from sparse 3D keypoints. While these methods showed promising results in enhancing sparse 3D video keypoints for inverse kinematics analysis, their generalization capabilities across different types of tasks and markersets remains diffi-

cult. The main contribution of this chapter is to propose and evaluate methods to retrain Opencap to new experimental conditions, including new poses and new markerset. It provides companies and researchers with guidelines to efficiently adapt Opencap to their motion capture protocols and methods. Our findings indicated that while pretrained models, such as Opencap, could provide valuable insights, they might require fine tuning on task-specific datasets to achieve optimal performance. However, it is important to notice that fine tuning comes with its own set of limitations, such as the risk of catastrophic forgetting (Arora et al., 2019), where the model might lose previously learned information when adapting to new tasks. We showed that retraining the very last output layers only, provides very promising results, with a limited set of examples for training. We also showed that the accuracy of such marker augmenter decreases when using real RGB data and HPE as inputs, compared to reference Mocap data. It opens new questions about the interest of applying the same fine tuning strategy to retrain the first input layers, in order to adapt to new HPE specifications. However, this is more difficult to handle, especially for collecting relevant training data with video. The ability to accurately estimate reliable joint angles from on-site RGB videos opens up new opportunities for research and practical applications to exploit on-site RGB videos to estimate joint torques and forces using standard inverse dynamics framework. Further exploration of fine-tuning techniques and expansion of training datasets could enhance the reliability and applicability of these methods in diverse real-world scenarios.

Conclusion

In this thesis, we tackle the challenge of quantifying joint torques in an industrial setting marked by low-frequency motion data, imprecisions, and occlusions. The study is guided by two main objectives.

The first objective examines the accuracy and robustness of various motion capture systems used in physical ergonomic assessments. Previous studies, such as those conducted by (Menolotto et al., 2020) and (Humadi et al., 2021), have validated these systems' applications in ergonomics; however, they predominantly evaluated earlier Kinect systems, including the Kinect V2 depth camera, which is no longer commercially available. Moreover, recent research has investigated the potential of monocular cameras (L. Li et al., 2020; McKinnon et al., 2022; Nayak & Kim, 2021; Yuan & Zhou, 2023) and multiple RGB cameras (W. Kim et al., 2021) for postural assessments. These studies have primarily assessed these systems in isolation, without performing comparative analyses under simulated work task conditions. Furthermore, no study has yet compared various motion capture systems—such as monocular RGBD/RGB systems, hybrid systems that combine inertial measurement units (IMUs) with RGBD/RGB systems (e.g., KIMEA (Moovency, 2024) and VIMU (Adjel et al., 2023))—within simulated real-world working environments. This gap raises significant questions about the practical applicability of these systems. Our first contribution aims to address this gap by evaluating these systems under simulated working conditions, specifically examining the performance of the THEIA system, the Kinect Azure DK system, the KIMEA system, the KIMEA Cloud system. The results revealed that hybrid systems scored RULA in over 80% of cases, with systems using multiple RGB cameras achieving a success rate of 86%. While depth-based systems demonstrated potential—particularly in measuring most joint angles—certain joint angles, such as those of the wrist, posed challenges in achieving precision. Hybrid systems that integrated wearable sensors for wrist measurements effectively addressed this limitation, rendering them more suitable for comprehensive ergonomic assessments. Nevertheless, we identified several limitations. The XSens motion capture system was selected for its ability to avoid infrared interference issues prevalent in optoelectronic systems; however, this choice introduced specific measurement errors, including drift and inaccuracies. Additionally, a key limitation in biomechanical analysis is the variability in biomechanical models across motion capture systems, particularly in anatomical landmarks, degrees of freedom, joint center definitions, kinematic assumptions, and model complexity, which often differ from those recommended by the International Society of Biomechanics (ISB) (Wu et al., 2005). This variability can significantly affect joint angle calculations. To ad-

dress this issue, we propose implementing a standardized calibration phase to minimize errors arising from these model differences. Furthermore, while automatic ergonomic scoring systems, such as RULA, have demonstrated considerable potential, improvements are necessary—particularly in wrist angle estimations—to enhance the accuracy of ergonomic risk assessments. Future systems could benefit from incorporating confidence intervals for risk scores, enabling more nuanced and data-driven evaluations of musculoskeletal disorder risks.

The second objective focuses on evaluating learning-based approaches for estimating joint torques and determining whether they can provide an accurate alternative to traditional inverse dynamics in human motion analysis. Our approach involved developing and evaluating a learning-based torque estimation model, validated against data from optoelectronic motion capture systems for upper limb torque estimation during one-handed load carrying tasks. Our findings suggest that these learning models, when applied to tasks involving specific handling heights and specific loads during dynamic phases, provide torque estimates that align with reference data. However, a limitation remains in the model's generalizability across different movement types and subject-specific variations. To address these limitations and improve model accuracy and robustness, we propose two key strategies grounded in recent advancements. First, we recommend the generation of synthetic data using techniques such as Generative Adversarial Networks (GANs) (J. Zhang et al., 2022) or inverse methods-based musculoskeletal analysis pipelines. These pipelines ensure that generated poses adhere to biomechanical plausibility, respecting physical constraints like body segment lengths and joint angle limits. This would significantly enhance the diversity and quality of training datasets, enabling better generalization across varied movements. Second, we advocate for the integration of prior physical knowledge into the loss functions of the learning models. This includes maintaining invariant body segment lengths, preserving appropriate ratios between body parts, and constraining joint angles within physiological limits. By embedding these constraints, we can reduce learning bias and ensure that the predicted torques remain physically valid. Such enhancements build on the physics-informed approaches outlined by (Banerjee et al., 2024), which have shown promise in improving both accuracy and physical realism in learning models.

To address the challenges of generalizing/fine-tuning models to new tasks, movements, and populations, we explored the capacity of Opencap's learning models—originally pretrained on a bipedal locomotion dataset—to extend their applicability to bi-manual manipulation and object-picking tasks, as well as to novel marker sets. Previous works have made progress in this area (Falisse, Uhlrich, Chaudhari, et al., 2024) by creating a much larger and more diverse training dataset; however, a key unresolved question remains: *can pre-trained models reliably predict 3D positions of anatomical markers for tasks outside of their training scope?* To address this question, we explored several fine-tuning

techniques, including retraining the entire model, modifying only the final layers, and incorporating task-specific output layers. Our findings indicated a reduction in error following fine-tuning, even when using a limited training dataset. This result is particularly promising for applications involving previously unseen measurement conditions, such as industrial environments. Despite this progress, our study focused solely on estimating joint coordinates, leaving the question of accurately predicting joint torques unanswered. Estimating joint torques requires accurate estimation of external forces, which can be achieved either through learning models, as demonstrated in recent studies (Faisal et al., 2024; Louis et al., 2022), or by jointly minimizing external forces and joint torques using a motion-based prediction method (Morin et al., 2021). A key technical challenge that remains is identifying contact frames—specific moments when the body interacts with the environment—within video data. However, recent advancements in computer vision algorithms have made this increasingly feasible (Y. Chen et al., 2023). In addition, a hybrid approach in the literature combines closed-chain inverse dynamics with a deep learning-assisted wearable sensor network to enable accurate real-time biomechanical analysis of manual material handling tasks. This approach implements an algorithm to analyze the biomechanics of the human musculoskeletal system using inverse dynamics, alongside a method for estimating the load and its distribution through an egocentric camera and deep learning-based object recognition. Kinematic data, as well as foot contact information, are provided by a fully wearable sensor network consisting of inertial measurement units. Finally, the performance of the fine-tuned Opencap models can be further enhanced by analyzing the noise levels in Opencap's input data relative to our training datasets. This analysis is crucial for minimizing observation bias and selecting the optimal input layers for fine-tuning.

In conclusion, this thesis has provided valuable insights into the evaluation of computer vision-based systems for physical ergonomic assessment and the use of learning-based approaches for joint torque estimation in industrial environments. Future research should prioritize the development of methodologies that enable the creation of diverse and representative training datasets for motion analysis. Such advancements are essential to improve the accuracy and robustness of learning algorithms, expanding their applicability to real-world industrial tasks. Additionally, further work is needed to refine the quantification of physical effort using biomechanical variables, integrating joint stress, fatigue, and repetitive strain for a comprehensive assessment of operator workload over time. This could lead to more reliable ergonomic evaluations and enhance safety and efficiency in industrial settings.

LIST OF PUBLICATIONS

International Journals

Ouadoudi Belabzioui, H., Plantard, P., Pontonnier, C., Dumont, G., & Multon, F (2024). Generalization of Inverse Kinematics Frameworks Based on Deep Learning to New Motor Tasks and Markersets. *International Journal of Industrial Ergonomics*

Ouadoudi Belabzioui, H., Plantard, P., Pontonnier, C., Dumont, G., & Multon, F (2024). Comparison of Computer Vision-Based Motion Capture Systems for Ergonomic Assessment in Work Conditions, *The paper has been submitted to the International Journal of Industrial Ergonomics (Under review).*

International Conference with Proceedings

Ouadoudi Belabzioui, H., Pontonnier, C., Dumont, G., Plantard, P., & Multon, F. (2023, July). Estimation of Upper-Limb Joint Torques in Static and Dynamic Phases for Lifting Tasks. In *International Conference on Digital Human Modeling* (pp. 71–80). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37848-5_8

National Conference with Proceedings

Ouadoudi Belabzioui, H., Pontonnier, C., Dumont, G., Plantard, P., & Multon, F. (2024, October). Generalization of Marker Augmenter Learning Models on a Motion Base for Inverse Kinematics. Multidisciplinary Biomechanics Journal, Vol 1. https://doi.org/10.46298/mbj.14520

APPENDICES

.1 Generalization of Inverse Kinematics Frameworks Based on Deep Learning to New Motor Tasks and Markersets

.1.1 Inputs and Outputs of Marker Augmenter Models

This section explores the inputs and outputs associated with marker augmenter models, as depicted in figures 7 and 8. Initially, in section .1.1, we examine the original inputs and outputs defined by Opencap (Uhlrich et al., 2023). Subsequently, section .1.1 presents our adapted inputs and outputs for lifting tasks. Lastly, section .1.1 discusses the adapted inputs and outputs for picking tasks. For both lifting and picking tasks, our outputs focus on a subset of the OpenCap marker set. Specifically, markers such as `r_calc`, `r_thigh1`, `r_thigh2`, `r_thigh3`, `L_thigh1`, `L_thigh2`, `L_thigh3`, `r_sh1`, `r_sh2`, `r_sh3`, `L_sh1`, `L_sh2`, `L_sh3`, `RHJC`, and `LHJC` are excluded from both inference error estimation and fine-tuning training and error estimation phases.

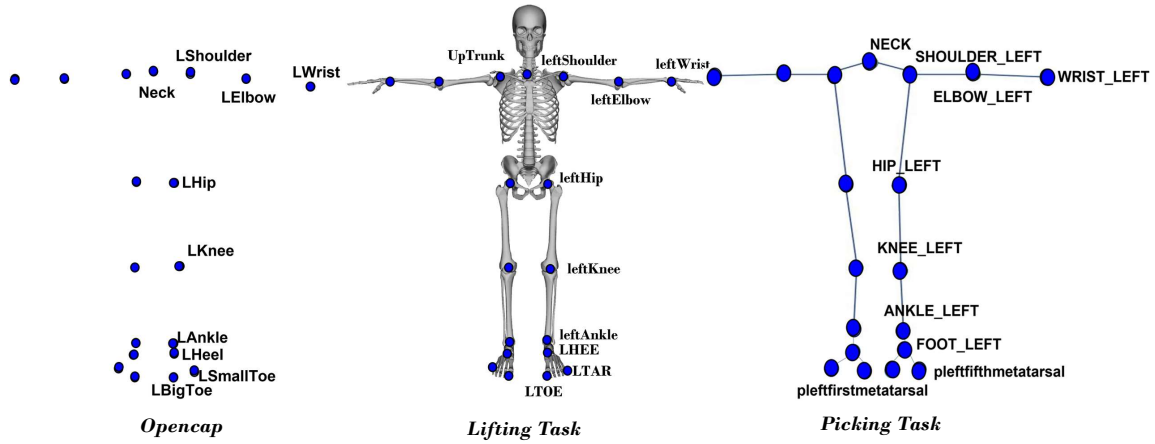


Figure 7 – **OpenCap, Lifting Task (MoCap data), and Picking Task (RGB data)** are compared in terms of their 3D keypoints.

Opencap original inputs/outputs

The method detailed by (Uhlrich et al., 2023) employs a synthetic technique to create datasets by matching 3D video keypoints with corresponding 3D anatomical markers, as

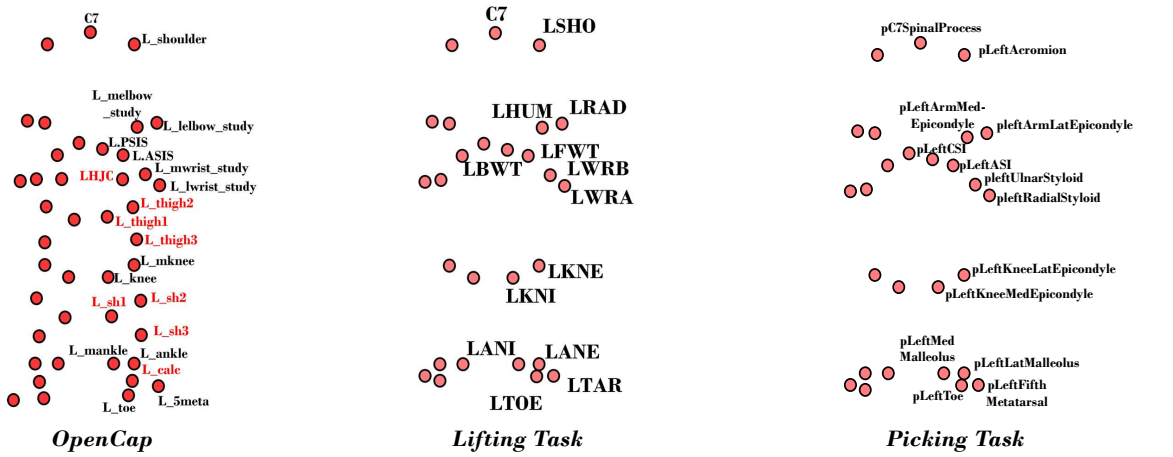


Figure 8 – Anatomical markers in **OpenCap, Lifting Task (MoCap data), and Picking Task (RGB data)** are compared. Markers in the OpenCap marker set highlighted in red are excluded from inference error estimation and are not considered during fine-tuning training or error estimation.

depicted in figures 7 and 8. These datasets are derived from 108 hours of motion capture data, which had previously been processed using OpenSim software (Seth et al., 2018) and compiled from various published biomechanics studies. Opencap emphasizes generating 3D anatomical markers using this synthetic approach, based on the same motion capture data and biomechanics studies processed with OpenSim software. For the **Body Model**, the markers included r.ASIS, L.ASIS, r.PSIS, L.PSIS, r_knee, r_mknee, r_ankle, r_mankle, r_toe, r_5meta, r_calc, L_knee, L_mknee, L_ankle, L_mankle, L_toe, L_5meta, L_calc, r_shoulder, L_shoulder, C7, r_thigh1, r_thigh2, r_thigh3, L_thigh1, L_thigh2, L_thigh3, r_sh1, r_sh2, r_sh3, L_sh1, L_sh2, L_sh3, RHJC, and LHJC. For the **Arm Model**, the markers include r_lelbow_study, r_melbow_study, r_lwrist_study, r_mwrist_study, L_lelbow_study, L_melbow_study, L_lwrist_study, and L_mwrist_study.

Lifting tasks adapted inputs/outputs

To emulate the 3D video keypoints from MoCap data, we implemented the following two steps:

1. Transforming the MoCap data from the world reference frame to the pelvis reference frame.
2. Estimating the 3D keypoints in the Opencap global reference frame (See the figure 7).

1. **Transforming the MoCap data from the world reference frame to the pelvis reference frame:** Our local reference frame was represented by a transformation matrix that converts coordinates from the world reference frame to the

pelvis reference frame. The following sub-steps outline the process for defining a local reference frame:

(a) **Identify the anatomical landmarks:** We used the pelvis as an anatomical landmark to define a local reference frame. This pelvic reference is established by an anatomically accurate local reference frame centered on the pelvis. The anatomical markers **RFWT** (right anterior superior iliac spine), **LFWT** (left anterior superior iliac spine), **RBWT** (right posterior superior iliac spine), and **LBWT** (left posterior superior iliac spine) are utilized as anatomical landmarks. To ensure a consistent local reference frame using anatomical markers that may move during motion analysis, the local reference frame is determined in each frame.

(b) **Define the local axes:** The X-axis vector is defined as:

$$\mathbf{x} = 0.5\left((\mathbf{LFWT} + \mathbf{RFWT}) - (\mathbf{LBWT} + \mathbf{RBWT})\right) \tag{2}$$

For the Y-axis vector, we first compute:

$$\mathbf{z}' = \mathbf{RBWT} - \mathbf{LBWT} \tag{3}$$

Then, we calculate:

$$\mathbf{y} = \mathbf{z}' \wedge \mathbf{x} \tag{4}$$

The Z-axis vector is determined by:

$$\mathbf{z} = \mathbf{x} \wedge \mathbf{y} \tag{5}$$

The origin is given by:

$$O = 0.25 \times (\mathbf{LFWT} + \mathbf{RFWT} + \mathbf{RBWT} + \mathbf{LBWT}) \tag{6}$$

The figure 9 below illustrates the local reference frame details.

(c) **Calculate the orthonormal vectors:** The transformation matrix can be constructed using the orthonormal vectors that define the pelvis reference frame. Orthonormal vectors indicate the direction of the X, Y, and Z axes within the pelvis reference frame relative to the global reference frame. Our transformation matrix can be constructed by placing these vectors as columns of a 3×3 matrix. The vectors $\mathbf{f_x} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$, $\mathbf{f_y} = \frac{\mathbf{y}}{\|\mathbf{y}\|}$, $\mathbf{f_z} = \frac{\mathbf{z}}{\|\mathbf{z}\|}$ represent the orthonormal

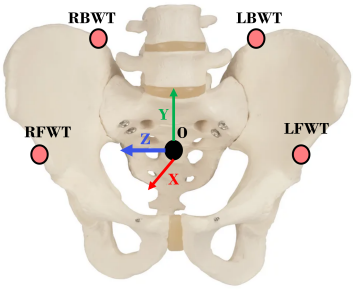


Figure 9 – Local reference frame details

basis vectors for the pelvis reference frame.

(d) **Create the transformation matrix:** To convert marker positions from the global reference frame to the pelvis reference frame, we need to define the transformation matrix that relates the two coordinate systems. The transformation matrix is:

$$ {}^{0}\mathrm{T}_{\mathrm{P}} = \begin{bmatrix} f_{z0} & f_{x0} & f_{y0} & O_z \\ f_{z1} & f_{x1} & f_{y1} & O_x \\ f_{z2} & f_{x2} & f_{y2} & O_y \\ 0 & 0 & 0 & 1 \end{bmatrix} $$

(e) **Apply the transformation matrix:** We applied the transformation matrix ${}^{0}T_{P}$ to convert coordinates from the world reference frame to the pelvis reference frame, represented by ${}^{P}K = {}^{P}T_{0}\,{}^{0}K$. Here, ${}^{P}K$ denotes the marker position in the pelvis reference frame, ${}^{P}T_{0}$ is the inverse of ${}^{0}T_{P}$, and ${}^{0}K$ represents the marker position in the world reference frame.

2. **Estimating the 3D keypoints in the Opencap global reference frame:** After expressing the 3D positions of anatomical markers in the pelvis reference frame, we used the following regression equations to estimate 3D keypoints:
Up trunk *according to* Reed et al., 1999:

$$\begin{aligned} \mathrm{UpTrunk}_z &= \mathrm{C7}_z \\ \mathrm{UpTrunk}_x &= \mathrm{C7}_x + \cos(8 \times \pi/180) \times 0.55 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \\ \mathrm{UpTrunk}_y &= \mathrm{C7}_y + \sin(8 \times \pi/180) \times 0.55 \times \mathrm{norm}(\mathrm{CLAV} - \mathrm{C7}) \end{aligned}$$

Shoulders *according to* Reed et al., 1999:

$$
\begin{aligned}
\text{rightShoulder}_z &= \text{RSHO}_z \\
\text{rightShoulder}_x &= \text{RSHO}_x + \cos(11 \times \pi/180) \times 0.43 \times \text{norm}(\text{CLAV} - \text{C7}) \\
\text{rightShoulder}_y &= \text{RSHO}_y - \sin(11 \times \pi/180) \times 0.43 \times \text{norm}(\text{CLAV} - \text{C7}) \\
\text{leftShoulder}_z &= \text{LSHO}_z \\
\text{leftShoulder}_x &= \text{LSHO}_x + \cos(11 \times \pi/180) \times 0.43 \times \text{norm}(\text{CLAV} - \text{C7}) \\
\text{leftShoulder}_y &= \text{LSHO}_y - \sin(11 \times \pi/180) \times 0.43 \times \text{norm}(\text{CLAV} - \text{C7})
\end{aligned}
$$

Ankles:

$$
\begin{aligned}
\text{rightAnkle} &= (\text{RANE} + \text{RANI}) \times 0.5 \\
\text{leftAnkle} &= (\text{LANE} + \text{LANI}) \times 0.5
\end{aligned}
$$

Knees:

$$
\begin{aligned}
\text{rightKnee} &= (\text{RKNE} + \text{RKNI}) \times 0.5 \\
\text{leftKnee} &= (\text{LKNE} + \text{LKNI}) \times 0.5
\end{aligned}
$$

Hips *according to* Leardini et al., 1999:

$$
\begin{aligned}
\text{rightHip}_z &= ((\text{LFWT}_z + \text{RFWT}_z) \times 0.5) + 0.38 \times \text{norm}(\text{RFWT} - \text{LFWT}) \\
\text{rightHip}_x &= ((\text{LFWT}_x + \text{RFWT}_x) \times 0.5) - 0.31 \times \text{norm}[((\text{LFWT} + \text{RFWT}) \times 0.5) - ((\text{LBWT} + \text{RBWT}) \times 0.5)] \\
\text{rightHip}_y &= ((\text{LFWT}_y + \text{RFWT}_y) \times 0.5) - 0.096 \times [\text{norm}(\text{RANI} - \text{RKNE}) + \text{norm}(\text{RKNE} - \text{RFWT})] \\
\text{leftHip}_z &= ((\text{LFWT}_z + \text{RFWT}_z) \times 0.5) - 0.38 \times \text{norm}(\text{RFWT} - \text{LFWT}) \\
\text{leftHip}_x &= ((\text{LFWT}_x + \text{RFWT}_x) \times 0.5) - 0.31 \times \text{norm}[((\text{LFWT} + \text{RFWT}) \times 0.5) - ((\text{LBWT} + \text{RBWT}) \times 0.5)] \\
\text{leftHip}_y &= ((\text{LFWT}_y + \text{RFWT}_y) \times 0.5) - 0.096 \times [\text{norm}(\text{LANI} - \text{LKNE}) + \text{norm}(\text{LKNE} - \text{LFWT})]
\end{aligned}
$$

Elbows:

$$
\begin{aligned}
\text{rightElbow} &= (\text{RHUM} + \text{RRAD}) \times 0.5 \\
\text{leftElbow} &= (\text{LHUM} + \text{LRAD}) \times 0.5
\end{aligned}
$$

Wrists:

$$\text{rightWrist} = (\text{RWRA} + \text{RWRB}) \times 0.5$$
$$\text{leftWrist} = (\text{LWRA} + \text{LWRB}) \times 0.5$$

After estimating the 3D keypoints, which were initially expressed in a local reference frame, we converted them into the world reference frame and subsequently into the Opencap global reference frame.

For the outputs, the **Body Model** included the following markers: **RFWT, LFWT, RBWT, LBWT, RKNI, RKNE, RANE, RANI, RTOE, RTAR, LKNI, LKNE, LANE, LANI, LTOE, LTAR, RSHO, LSHO, and C7**. For the **Arm Model**, the markers included are: **RRAD, RHUM, RWRA, RWRB, LRAD, LHUM, LWRA, and LWRB**. These are detailed in table 6 and illustrated in figure 8.

Marker	Definition	Marker	Definition
RFWT	Right anterior superior iliac spine	**LKNI**	Medial condyle of the left femur
LFWT	Left anterior superior iliac spine	**LKNE**	Lateral condyle of the left femur
RBWT	Right posterior superior iliac spine	**LANI**	Left internal malleolus
LBWT	Left posterior superior iliac spine	**LANE**	Left external malleolus
RKNI	Medial condyle of the right femur	**LTOE**	Left acropodion
RKNE	Lateral condyle of the right femur	**LTAR**	Left Ankle I/E folding
RANI	Right internal malleolus	**RSHO**	Right acromion
RANE	Right external malleolus	**LSHO**	Left acromion
RTOE	Right acropodion	**C7**	Spinous process of the 7th cervical
RTAR	Right Ankle I/E folding	**RRAD**	Head of the right radius
RHUM	Medial epicondyle of the right humerus	**LRAD**	Head of the left radius
RWRA	Styloid process of the right radius	**LHUM**	Medial epicondyle of the left humerus
RWRB	Styloid process of the right ulna	**LWRA**	Styloid process of the left radius
LWRB	Styloid process of the left ulna		

Table 6 – Definitions of anatomical markers used in MoCap data for lifting tasks.

Picking tasks adapted inputs/outputs

To process the unique RGB camera, we utilized the KIMEA Cloud solution developed by Moovency. This enabled us to obtain the 3D keypoints, as illustrated in figure 7.

For the outputs, the markers for the **Body Model** included pRightASI, pLeftASI, pRightCSI, pLeftCSI, pRightKneeMedEpicondyle, pRightKneeLatEpicondyle, pRightLatMalleolus, pRightMedMalleolus, pRightToe, pRightFifthMetatarsal, pLeftKneeMedEpicondyle, pLeftKneeLatEpicondyle, pLeftLatMalleolus, pLeftMedMalleolus, pLeftToe, pLeftFifthMetatarsal, pRightAcromion, pLeftAcromion, and pC7SpinalProcess. For the **Arm Model**, the markers included pRightArmLatEpicondyle, pRightArmMedEpicondyle, pRightRadialStyloid, pRightUlnarStyloid, pLeftArmLatEpicondyle, pLeftArmMedEpicondyle, pLeftRadialStyloid, and pLeftUlnarStyloid, as shown in figure 8.

BIBLIOGRAPHY

Abobakr, A., Nahavandi, D., Hossny, M., Iskander, J., Attia, M., Nahavandi, S., & Smets, M., (2019), RGB-D ergonomic assessment system of adopted working postures, *Applied ergonomics*, *80*, 75–88.

Adjel, M., Sabbah, M., Dumas, R., Mansard, N., Mohammed, S., Watier, B., & Bonnet, V., (2023), Multi-modal upper limbs human motion estimation from a reduced set of affordable sensors, *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10926–10932.

Aghazadeh, F., Arjmand, N., & Nasrabadi, A., (2020), Coupled artificial neural networks to estimate 3d whole-body posture, lumbosacral moments, and spinal loads during load-handling activities, *Journal of biomechanics*, *102*, 109332.

Ahn, J., Choi, H., Lee, H., Lee, J., & Kim, H.-D., (2023), Novel multi-view rgb sensor for continuous motion analysis in kinetic chain exercises: a pilot study for simultaneous validity and intra-test reliability, *Sensors*, *23*, 9635.

Al Borno, M., O'Day, J., Ibarra, V., Dunne, J., Seth, A., Habib, A., Ong, C., Hicks, J., Uhlrich, S., & Delp, S., (2022), Opensense: an open-source toolbox for inertial-measurement-unit-based measurement of lower extremity kinematics over long durations, *Journal of neuroengineering and rehabilitation*, *19*, 22.

Alpaydin, E., (2020), *Introduction to machine learning*, MIT press.

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G., (2018), Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis, *Journal of choice modelling*, *28*, 167–182.

Ameli, (2019), *Les tms : définition et impact*. https://www.ameli.fr/entreprise/sante-travail/risques/troubles-musculosquelettiques-tms/tms-definition-impact

Antico, M., Balletti, N., Ciccotelli, A., Ciccotelli, M., Laudato, G., Lazich, A., Notarantonio, M., Oliveto, R., Ricciardi, S., Scalabrino, S., et al., (2021), 2vita-b physical: an intelligent home rehabilitation system based on microsoft azure kinect, *Frontiers in Human Dynamics*, *3*, 678529.

APDM, (2024), *Apdm - research-grade wearable sensors*. https://apdm.com/wearable-sensors/

Arora, G., Rahimi, A., & Baldwin, T., (2019), Does an lstm forget more than a cnn? an empirical study of catastrophic forgetting in nlp, *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, 77–86.

Banerjee, C., Nguyen, K., Fookes, C., & George, K., (2024), Physics-informed computer vision: a review and perspectives, *ACM Computing Surveys*.

Barone, A. V. M., Haddow, B., Germann, U., & Sennrich, R., (2017), Regularization techniques for fine-tuning in neural machine translation, *arXiv preprint arXiv:1707.09920*.

Begon, M., Andersen, M. S., & Dumas, R., (2018), Multibody kinematics optimization for the estimation of upper and lower limb human joint kinematics: a systematized methodological review, *Journal of biomechanical engineering*, *140*, 030801.

Belabzioui, H. O., Pontonnier, C., Dumont, G., Plantard, P., & Multon, F., (2023), Estimation of upper-limb joint torques in static and dynamic phases for lifting tasks, *8th International Digital Human Modeling Symposium*.

Benjaminse, A., Bolt, R., Gokeler, A., & Otten, B., (2020), A validity study comparing xsens with vicon, *ISBS Proceedings Archive*, *38*, 752.

Benoit, D. L., Damsgaard, M., & Andersen, M. S., (2015), Surface marker cluster translation, rotation, scaling and deformation: their contribution to soft tissue artefact and impact on knee joint kinematics, *Journal of biomechanics*, *48*, 2124–2129.

Bernard, B. P., (1997), Musculoskeletal disorders and workplace factors : a critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back, *DHHS (NIOSH)*, *97*.

Bernard, B. P., & Putz-Anderson, V., (1997), Musculoskeletal disorders and workplace factors: a critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back.

Bertram, J., Krüger, T., Röhling, H. M., Jelusic, A., Mansow-Model, S., Schniepp, R., Wuehr, M., & Otte, K., (2023), Accuracy and repeatability of the microsoft azure kinect for clinical measurement of motor function, *Plos one*, *18*, e0279697.

Bhoi, A., (2019), Monocular depth estimation: a survey.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T., (2016), Man is to computer programmer as woman is to homemaker? debiasing word embeddings, *Advances in neural information processing systems*, *29*.

Bordes, F., Balestriero, R., Garrido, Q., Bardes, A., & Vincent, P., (2022), Guillotine regularization: why removing layers is needed to improve generalization in self-supervised learning, *arXiv preprint arXiv:2206.13378*.

Borg, G., (1990), Psychophysical scaling with applications in physical work and the perception of exertion, *Scandinavian journal of work, environment & health*, 55–58.

Borg, G., (1998), *Borg's perceived exertion and pain scales.*, Human kinetics.

Breiman, L., (2001), Random forests, *Machine learning*, *45*, 5–32.

Breiman, L., (2017), *Classification and regression trees*, Routledge.

Brodie, M., Walmsley, A., & Page, W., (2008), The static accuracy and calibration of inertial measurement units for 3d orientation.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., (2020), Language models are few-shot learners, *Advances in neural information processing systems*, *33*, 1877–1901.

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H., (2022), The effects of data quality on machine learning performance, *arXiv preprint arXiv:2207.14529*.

Burdorf, A., Derksen, J., Naaktgeboren, B., & Van Riel, M., (1992), Measurement of trunk bending during work by direct observation and continuous measurement, *Applied ergonomics*, *23*, 263–267.

Burdorf, A., & Laan, J., (1991), Comparison of methods for the assessment of postural load on the back, *Scandinavian journal of work, environment & health*, 425–429.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y., (2017), Realtime multi-person 2d pose estimation using part affinity fields, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Caputo, F., Greco, A., D'Amato, E., Notaro, I., & Spada, S., (2019), Imu-based motion capture wearable system for ergonomic assessment in industrial environment, *Advances in Human Factors in Wearable Technologies and Game Design*, 215–225.

Caruana, R., (1997), Multitask learning, *Machine learning*, *28*, 41–75.

Chen, L., & Qin, G., (2010), Optimization of the collision detection technology in 3d skeleton animation, *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, *10*, V10–539.

Chen, X., & Koskela, M., (2013), Sequence alignment for rgb-d and motion capture skeletons, *International Conference Image Analysis and Recognition*, 630–639.

Chen, Y., Dwivedi, S. K., Black, M. J., & Tzionas, D., (2023), Detecting human-object contact in images, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17100–17110.

Cieza, A., Causey, K., Kamenov, K., Hanson, S. W., Chatterji, S., & Vos, T., (2020), Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019, *The Lancet*, *396*, 2006–2017.

Colombini, D., Occhipinti, E., Delleman, N., Fallentin, N., Kilbom, A., Grieco, A., et al., (2001), Exposure assessment of upper limb repetitive movements: a consensus document.

Corporation, I., (2019), Tuning depth cameras for best performance [Accessed: 2025-01-07].

Dagès, T., Cohen, L. D., & Bruckstein, A. M., (2023), A model is worth tens of thousands of examples.

Damsgaard, M., Rasmussen, J., Christensen, S. T., Surma, E., & De Zee, M., (2006), Analysis of musculoskeletal systems in the anybody modeling system, *Simulation Modelling Practice and Theory*, *14*, 1100–1111.

David, G. C., (2005), Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders, *Occupational medicine*, *55*, 190–199.

Delépine, A., Levert, C., Meyer, J., & Zana, J., (2011), Travail et lombalgie, *Du facteur de risque au facteur de soin. Édition INRS ED*, *6087*.

Delp, S. L., Anderson, F. C., Arnold, A. S., Loan, P., Habib, A., John, C. T., Guendelman, E., & Thelen, D. G., (2007), Opensim: open-source software to create and analyze dynamic simulations of movement, *IEEE transactions on biomedical engineering*, *54*, 1940–1950.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L., (2009), Imagenet: a large-scale hierarchical image database, *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., (2018), Bert: pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*.

Diego-Mas, J.-A., Alcaide-Marzal, J., & Poveda-Bautista, R., (2017), Errors using observational methods for ergonomics assessment in real practice, *Human Factors*, *59*, 1173–1187.

Dixon, M. F., Polson, N. G., & Sokolov, V. O., (2019), Deep learning for spatio-temporal modeling: dynamic traffic flows and high frequency trading, *Applied Stochastic Models in Business and Industry*, *35*, 788–807.

Dockrell, S., O'Grady, E., Bennett, K., Mullarkey, C., Mc Connell, R., Ruddy, R., Twomey, S., & Flannery, C., (2012), An investigation of the reliability of rapid upper limb assessment (rula) as a method of assessment of children's computing posture, *Applied ergonomics*, *43*, 632–636.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T., (2014), Decaf: a deep convolutional activation feature for generic visual recognition, *International conference on machine learning*, 647–655.

Egeonu, D., & Jia, B., (2024), A systematic literature review of computer vision-based biomechanical models for physical workload estimation, *Ergonomics*, 1–24.

Fagarasanu, M., & Kumar, S., (2002), Measurement instruments and data collection: a consideration of constructs and biases in ergonomics research, *International journal of industrial ergonomics*, *30*, 355–369.

Faisal, M. A. A., Mahmud, S., Chowdhury, M. E., Khandakar, A., Ahmed, M. U., Alqahtani, A., & Alhatou, M., (2024), Robust and novel attention guided multiresunet model for 3d ground reaction force and moment prediction from foot kinematics, *Neural Computing and Applications*, *36 3*, 1105–1121.

Falisse, A., Uhlrich, S., Chaudhari, A., Hicks, J., & Delp, S., (2024), Marker data enhancement for markerless motion capture, *bioRxiv*, 2024–07.

Falisse, A., Uhlrich, S. D., Chaudhari, A. S., Hicks, J. L., & Delp, S. L., (2024), Marker data enhancement for markerless motion capture, *bioRxiv*.

Falisse, A., Uhlrich, S. D., Hicks, J. L., Chaudhari, A. S., & Delp, S. L., (2023), Xix international symposium on computer simulation in biomechanics july 26th–28th 2023, kyoto marker data augmentation for robust markerless motion capture.

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., & Lu, C., (2022), Alphapose: whole-body regional multi-person pose estimation and tracking in real-time, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*, 7157–7173.

Featherstone, R., (2014), *Rigid body dynamics algorithms*, Springer.

Fu, Z., Yang, H., So, A. M.-C., Lam, W., Bing, L., & Collier, N., (2023), On the effectiveness of parameter-efficient fine-tuning, *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 12799–12807.

Géron, A., (2022), *Hands-on machine learning with scikit-learn, keras, and tensorflow*, O'Reilly Media, Inc.

Goodfellow, I., Bengio, Y., & Courville, A., (2016), *Deep learning*, MIT press.

Gorton III, G. E., Hebert, D. A., & Gannotti, M. E., (2009), Assessment of the kinematic variability among 12 motion analysis laboratories, *Gait & posture*, *29 3*, 398–402.

Goswami, S., (2020), Impact of data quality on deep neural network training, *arXiv preprint arXiv:2002.03732*.

Graves, A., & Schmidhuber, J., (2005), Framewise phoneme classification with bidirectional lstm networks, *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, *4*, 2047–2052.

Gulli, A., Kapoor, A., & Pal, S., (2019), *Deep learning with tensorflow 2 and keras: regression, convnets, gans, rnns, nlp, and more with tensorflow 2 and the keras api*, Packt Publishing Ltd.

Habehh, H., & Gohel, S., (2021), Machine learning in healthcare, *Current genomics*, *22*, 291.

Hagberg, M., (1995), Work related musculoskeletal disorders (wmsds), *A reference book for prevention*.

Haj Mahmoud, O., Pontonnier, C., Dumont, G., Poli, S., & Multon, F., (2021), A neural networks approach to determine factors associated with self-reported discomfort in picking tasks, *Human Factors*, 00187208211047640.

Han, Z., Gao, C., Liu, J., Zhang, S. Q., et al., (2024), Parameter-efficient fine-tuning for large models: a comprehensive survey, *arXiv preprint arXiv:2403.14608*.

Hardt, M., & Recht, B., (2021), Patterns, predictions, and actions: a story about machine learning.

Hastie, T., (2009), The elements of statistical learning: data mining, inference, and prediction.

Hatamizadeh, A., (2020), *Deep learning of unified region, edge, and contour models for automated image segmentation*, University of California, Los Angeles.

He, K., Zhang, X., Ren, S., & Sun, J., (2016), Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hignett, S., & McAtamney, L., (2000), Rapid entire body assessment (reba), *Applied ergonomics*, *31*, 201–205.

Hinton, G., (2015), Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531*.

Hochreiter, S., (1997), Long short-term memory, *Neural Computation MIT-Press*.

Hoerl, A. E., & Kennard, R. W., (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, *12*, 55–67.

Hsiao, C.-Y., Chang, C.-C., Chen, T.-Y., & Lin, Y.-T., (2022), Developing a computer-vision model to estimate anatomical joint coordinates during manual lifting tasks, *Physical Ergonomics and Human Factors*, *63*.

Humadi, A., Nazarahari, M., Ahmad, R., & Rouhani, H., (2021), In-field instrumented ergonomic risk assessment: inertial measurement units versus kinect v2, *International Journal of Industrial Ergonomics*, *84*, 103147.

Ioffe, S., (2015), Batch normalization: accelerating deep network training by reducing internal covariate shift, *arXiv preprint arXiv:1502.03167*.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A., (2019), Deep learning for time series classification: a review, *Data mining and knowledge discovery*, *33*, 917–963.

Jo, S., Song, S., Kim, J., & Song, C., (2022), Agreement between azure kinect and marker-based motion analysis during functional movements: a feasibility study, *Sensors*, *22*, 9819.

Kanko, R. M., (2020), *Validation of a markerless motion capture system for human movement analysis* (Master's thesis), Queen's University (Canada).

Kee, D., & Karwowski, W., (2001), The boundaries for joint angles of isocomfort for sitting and standing males based on perceived comfort of static joint postures, *Ergonomics*, *44*, 614–648.

Khan, F., Salahuddin, S., & Javidnia, H., (2020), Deep learning-based monocular depth estimation methods—a state-of-the-art review, *Sensors*, *20 8*, 2272.

Khirodkar, R., Chari, V., Agrawal, A., & Tyagi, A., (2021), Multi-instance pose networks: rethinking top-down pose estimation, *Proceedings of the IEEE/CVF International conference on computer vision*, 3122–3131.

Kim, S., & Nussbaum, M. A., (2013), Performance evaluation of a wearable inertial motion capture system for capturing physical exposures during manual material handling tasks, *Ergonomics*, *56*, 314–326.

Kim, W., Sung, J., Saakes, D., Huang, C., & Xiong, S., (2021), Ergonomic postural assessment using a new open-source human pose estimation technology (openpose), *International Journal of Industrial Ergonomics*, *84*, 103164.

Krizhevsky, A., Sutskever, I., & Hinton, G. E., (2012), Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, *25*.

Lahkar, B. K., Muller, A., Dumas, R., Reveret, L., & Robert, T., (2022), Accuracy of a markerless motion capture system in estimating upper extremity kinematics during boxing, *Frontiers in Sports and Active Living*, *4*, 939980.

Lawson, M., Naemi, R., Needham, R. A., & Chockalingam, N., (2024), Can machine learning predict running kinematics based on upper trunk gps-based imu acceleration? a novel method of conducting biomechanical analysis in the field using artificial neural networks, *Applied Sciences*, *14*, 1730.

Leardini, A., Benedetti, M., Catani, F., Simoncini, L., & Giannini, S., (1999), An anatomically based protocol for the description of foot segment kinematics during gait, *Clinical biomechanics*, *14*, 528–536.

Lebel, K., Boissy, P., Hamel, M., & Duval, C., (2013), Inertial measures of motion for clinical biomechanics: comparative assessment of accuracy under controlled conditions - effect of velocity, *PLOS ONE*, *8*, 1–9.

Leboeuf, F., Reay, J., Jones, R., & Sangeux, M., (2019), The effect on conventional gait model kinematics and kinetics of hip joint centre equations in adult healthy gait, *Journal of Biomechanics*, *87*, 167–171.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., (1998), Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, *86*, 2278–2324.

Li, D., & Zhang, H., (2021), Improved regularization and robustness for fine-tuning in neural networks, *Advances in Neural Information Processing Systems*, *34*, 27249–27262.

Li, G., & Buckle, P., (1999), Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods, *Ergonomics*, *42*, 674–695.

Li, L., Martin, T., & Xu, X., (2020), A novel vision-based real-time method for evaluating postural risk factors associated with musculoskeletal disorders, *Applied Ergonomics*, *87*, 103138.

Li, Z., & Hoiem, D., (2017), Learning without forgetting, *IEEE transactions on pattern analysis and machine intelligence*, *40*, 2935–2947.

Liang, Y., Qi, S., Xu, T., & Hu, Y., (2023), 3d gait analysis for the elderly mobility based on multiple rgb cameras, *2023 29th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, 1–5.

Liaw, A., (2002), Classification and regression by randomforest, *R news*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I., (2017), A survey on deep learning in medical image analysis, *Medical image analysis*, *42*, 60–88.

Livet, C., (2022), *Contributions algorithmiques à l'analyse musculo-squelettique: modèles et méthodes* (Doctoral dissertation), École normale supérieure de Rennes.

Livet, C., Rouvier, T., Sauret, C., Pillet, H., Dumont, G., & Pontonnier, C., (2023), A penalty method for constrained multibody kinematics optimisation using a levenberg–marquardt algorithm, *Computer Methods in Biomechanics and Biomedical Engineering*, *26*, 864–875.

Long, T., Outerleys, J., Yeung, T., Fernandez, J., Bouxsein, M. L., Davis, I. S., Bredella, M. A., & Besier, T. F., (2024), Predicting ankle and knee sagittal kinematics and kinetics using an ankle-mounted inertial sensor, *Computer Methods in Biomechanics and Biomedical Engineering*, *27*, 1057–1070.

Louis, N., Corso, J. J., Templin, T. N., Eliason, T. D., & Nicolella, D. P., (2022), Learning to estimate external forces of human motion in video, *Proceedings of the 30th ACM International Conference on Multimedia*, 3540–3548.

Lowe, B. D., (2004a), Accuracy and validity of observational estimates of shoulder and elbow posture, *Applied ergonomics*, *35*, 159–171.

Lowe, B. D., (2004b), Accuracy and validity of observational estimates of wrist and forearm posture, *Ergonomics*, *47*, 527–554.

Lu, T.-W., & O'connor, J., (1999), Bone position estimation from skin marker co-ordinates using global optimisation with joint constraints, *Journal of biomechanics*, *32*, 129–134.

Lund, M. E., Andersen, M. S., de Zee, M., & Rasmussen, J., (2015), Scaling of musculoskeletal models from static and dynamic trials, *International Biomechanics*, *2*, 1–11.

Manghisi, V. M., Uva, A. E., Fiorentino, M., Bevilacqua, V., Trotta, G. F., & Monno, G., (2017), Real time rula assessment using kinect v2 sensor, *Applied ergonomics*, *65*, 481–491.

Mao, Y., Deng, Z., Yao, H., Ye, T., Kawaguchi, K., & Zou, J., (2023), Last-layer fairness fine-tuning is simple and effective for neural networks, *arXiv preprint arXiv:2304.03935*.

Masoumian, A., Rashwan, H. A., Cristiano, J., Asif, M. S., & Puig, D., (2022), Monocular depth estimation using deep learning: a review, *Sensors*, *22*, 5353.

McAtamney, L., & Corlett, E. N., (1993), Rula: a survey method for the investigation of work-related upper limb disorders, *Applied ergonomics*, *24*, 91–99.

McKinnon, C. D., Sonne, M. W., & Keir, P. J., (2022), Assessment of joint angle and reach envelope demands using a video-based physical demands description tool, *Human Factors*, *64*, 568–578.

Menolotto, M., Komaris, D.-S., Tedesco, S., O'Flynn, B., & Walsh, M., (2020), Motion capture technology in industrial applications: a systematic review, *Sensors*, *20*, 5687.

Mitchell, T. M., & Mitchell, T. M., (1997), *Machine learning*, McGraw-hill New York.

Moghadam, S. M., Yeung, T., & Choisne, J., (2023), A comparison of machine learning models' accuracy in predicting lower-limb joints' kinematics, kinetics, and muscle forces from wearable sensors, *Scientific reports*, *13*, 5046.

Mohseni, M., Aghazadeh, F., & Arjmand, N., (2022), Improved artificial neural networks for 3d body posture and lumbosacral moment predictions during manual material handling activities, *Journal of Biomechanics*, *131*, 110921.

Moovency, (2024), *Moovency*. https://moovency.com/

Morin, P., Muller, A., Pontonnier, C., & Dumont, G., (2021), Studying the impact of internal and external forces minimization in a motion-based external forces and moments prediction method: application to fencing lunges, *ISB 2021-XXVIII Congress of the International Society of Biomechanics*, 1.

Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., & Hellier, P., (2022a), A survey on deep learning for skeleton-based human animation, *Computer Graphics Forum*, *41*, 122–157.

Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., & Hellier, P., (2022b), A survey on deep learning for skeleton-based human animation, *Computer Graphics Forum*, *41*, 122–157.

Muller, A., Germain, C., Pontonnier, C., & Dumont, G., (2015), A simple method to calibrate kinematical invariants: application to overhead throwing, *ISBS-conference proceedings archive*.

Muller, A., Pontonnier, C., & Dumont, G., (2017), Uncertainty propagation in multibody human model dynamics, *Multibody System Dynamics*, *40*, 177–192.

Muller, A., Pontonnier, C., & Dumont, G., (2019), Motion-based prediction of hands and feet contact efforts during asymmetric handling tasks, *IEEE Transactions on Biomedical Engineering*, *67*, 344–352.

Muller, A., Pontonnier, C., Puchaud, P., & Dumont, G., (2019), Custom: a matlab toolbox for musculoskeletal simulation, *Journal of Open Source Software*, *4*, 1–3.

Naeemabadi, M., Dinesen, B., Andersen, O. K., & Hansen, J., (2018), Investigating the impact of a motion capture system on microsoft kinect v2 recordings: a caution for using the technologies together, *PloS one*, *13*, e0204052.

Nayak, G. K., & Kim, E., (2021), Development of a fully automated rula assessment system based on computer vision, *International Journal of Industrial Ergonomics*, *86*, 103218.

Needham, L., Evans, M., Cosker, D. P., Wade, L., McGuigan, P. M., Bilzon, J. L., & Colyer, S. L., (2021), The accuracy of several pose estimation methods for 3d joint centre localisation, *Scientific reports*, *11*, 20673.

Needham, L., Evans, M., Wade, L., Cosker, D. P., McGuigan, M. P., Bilzon, J. L., & Colyer, S. L., (2022), The development and evaluation of a fully automated markerless motion capture workflow, *Journal of Biomechanics*, *144*, 111338.

Niswander, W., Wang, W., & Kontson, K., (2020), Optimization of imu sensor placement for the measurement of lower limb joint kinematics, *Sensors*, *20*.

Occhipinti, E., (1998), Ocra: a concise index for the assessment of exposure to repetitive movements of the upper limbs, *Ergonomics*, *41*, 1290–1311.

Oliveira, N., Park, J., & Barrance, P., (2022), Using inertial measurement unit sensor single axis rotation angles for knee and hip flexion angle calculations during gait, *IEEE Journal of Translational Engineering in Health and Medicine*, *11*, 80–86.

Orbbec, (2024), *Orbbec depth cameras*. https : / / www . orbbec . com / microsoft - collaboration/

Özsoy, U., Yıldırım, Y., Karaşin, S., Şekerci, R., & Süzen, L. B., (2022), Reliability and agreement of azure kinect and kinect v2 depth sensors in the shoulder joint range of motion estimation, *Journal of Shoulder and Elbow Surgery*, *31*, 2049–2056.

Pagliari, D., & Pinto, L., (2015), Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors, *Sensors*, *15*, 27569–27589.

Pagnon, D., Domalain, M., & Reveret, L., (2021), Pose2sim: an end-to-end workflow for 3d markerless sports kinematics—part 1: robustness, *Sensors*, *21*, 6530.

Pagnon, D., Domalain, M., & Reveret, L., (2022a), Pose2sim: an end-to-end workflow for 3d markerless sports kinematics—part 2: accuracy, *Sensors*, *22*, 2712.

Pagnon, D., Domalain, M., & Reveret, L., (2022b), Pose2sim: an open-source python package for multiview markerless kinematics, *Journal of Open Source Software*, *7 77*, 4362.

Pan, S. J., & Yang, Q., (2009), A survey on transfer learning, *IEEE Transactions on knowledge and data engineering*, *22*, 1345–1359.

Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M., (2019), 3d human pose estimation in video with temporal convolutions and semi-supervised training, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7745–7754.

Plamondon, A., Delisle, A., Larue, C., Brouillette, D., McFadden, D., Desjardins, P., & Larivière, C., (2007), Evaluation of a hybrid system for three-dimensional measurement of trunk posture in motion, *Applied Ergonomics*, *38*, 697–712.

Plantard, P., Auvinet, E., Le Pierres, A.-S., & Multon, F., (2015), Pose estimation with a kinect for ergonomic studies: evaluation of the accuracy using a virtual mannequin, *Sensors*, *15*, 1785–1803.

Plantard, P., H. Shum, H. P., & Multon, F., (2017), Filtered pose graph for efficient kinect pose reconstruction, *Multimedia Tools and Applications*, *76*, 4291–4312.

Plantard, P., Muller, A., Pontonnier, C., Dumont, G., Shum, H. P., & Multon, F., (2017), Inverse dynamics based on occlusion-resistant kinect data: is it usable for ergonomics?, *International Journal of Industrial Ergonomics*, *61*, 71–80.

Plantard, P., Shum, H. P., Le Pierres, A.-S., & Multon, F., (2017), Validation of an ergonomic assessment method using kinect data in real workplace conditions, *Applied ergonomics*, *65*, 562–569.

Plantard, P., Shum, H. P., & Multon, F., (2017), Usability of corrected kinect measurement for ergonomic evaluation in constrained environment, *International Journal of Human Factors Modelling and Simulation*, *5*, 338–353.

Poddar, S., Kumar, V., & Kumar, A., (2016), A comprehensive overview of inertial sensor calibration techniques, *Journal of Dynamic Systems, Measurement, and Control*, *139*, 011006.

Puchaud, P., Sauret, C., Muller, A., Bideau, N., Dumont, G., Pillet, H., & Pontonnier, C., (2020), Accuracy and kinematics consistency of marker-based scaling approaches on a lower limb model: a comparative study with imagery data, *Computer Methods in Biomechanics and Biomedical Engineering*, *23*, 114–125.

Qamar, T., & Bawany, N. Z., (2023), Understanding the black-box: towards interpretable and reliable deep learning models, *PeerJ Computer Science*, *9*, e1629.

Qualisys, (2024), *User-friendly mocap software*. https://www.qualisys.com/software/qualisys-track-manager/

Quinlan, J. R., (1986), Induction of decision trees, *Machine learning*, 81–106.

Reed, M. P., Manary, M. A., & Schneider, L. W., (1999), *Methods for measuring and representing automobile occupant posture* (tech. rep.), SAE Technical Paper.

Rekant, J., Rothenberger, S., & Chambers, A., (2022), Inertial measurement unit-based motion capture to replace camera-based systems for assessing gait in healthy young adults: proceed with caution, *Measurement: Sensors*, *23*, 100396.

Robert-Lachaine, X., Mecheri, H., Larue, C., & Plamondon, A., (2017), Validation of inertial measurement units with an optoelectronic system for whole-body motion analysis, *Medical & biological engineering & computing*, *55*, 609–619.

Robertson, M., Amick III, B. C., DeRango, K., Rooney, T., Bazzani, L., Harrist, R., & Moore, A., (2009), The effects of an office ergonomics training and chair intervention on worker knowledge, behavior and musculoskeletal risk, *Applied ergonomics*, *40*, 124–135.

Rodrigues, P. B., Xiao, Y., Fukumura, Y. E., Awada, M., Aryal, A., Becerik-Gerber, B., Lucas, G., & Roll, S. C., (2022), Ergonomic assessment of office worker postures using 3d automated joint angle assessment, *Advanced Engineering Informatics*, *52*, 101596.

Roetenberg, D., Luinge, H., Baten, C., & Veltink, P., (2005), Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *13*, 395–405.

Roetenberg, D., Luinge, H., & Slycke, P. J., (2009), Xsens mvn: full 6dof human motion tracking using miniature inertial sensors, https://api.semanticscholar.org/CorpusID:16142980

Russell, S. J., & Norvig, P., (2016), *Artificial intelligence: a modern approach*, Pearson.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., et al., (2022), Current progress and open challenges for applying deep learning across the biosciences, *Nature Communications*, *13*, 1728.

Sarker, I. H., (2021), Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions, *SN computer science*, *2*, 420.

Schepers, M., Giuberti, M., Bellusci, G., et al., (2018), Xsens mvn: consistent tracking of human motion using inertial sensing, *Xsens Technol*, *1*, 1–8.

SDK, K., (2024), *Kinect for windows sdk*. https://msdn.microsoft.com/en-us/library/dn799271.aspx

Seber, G. A., & Lee, A. J., (2012), *Linear regression analysis*, John Wiley & Sons.

Seth, A., Hicks, J. L., Uchida, T. K., Habib, A., Dembia, C. L., Dunne, J. J., Ong, C. F., DeMers, M. S., Rajagopal, A., Millard, M., et al., (2018), Opensim: simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement, *PLoS computational biology*, *14 7*, e1006223.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S., (2014), Cnn features off-the-shelf: an astounding baseline for recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 806–813.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., & Wang, H., (2024), Continual learning of large language models: a comprehensive survey, *arXiv preprint arXiv:2404.16789*.

Shiao, Y., Chen, G.-Y., & Hoang, T., (2024), Three-dimensional human posture recognition by extremity angle estimation with minimal imu sensor, *Sensors*, *24*, 4306.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A., (2011), Real-time human pose recognition in parts from single depth images, *CVPR 2011*, 1297–1304.

Sibson, B. E., Banks, J. J., Yawar, A., Yegian, A. K., Anderson, D. E., & Lieberman, D. E., (2024), Using inertial measurement units to estimate spine joint kinematics and kinetics during walking and running, *Scientific Reports*, *14*, 234.

Silverstein, B. A., Fine, L. J., & Armstrong, T. J., (1987), Occupational factors and carpal tunnel syndrome, *American journal of industrial medicine*, *11*, 343–358.

Simonyan, K., & Zisserman, A., (2014), Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

Simundic, A.-M., et al., (2008), Confidence interval, *Biochemia Medica*, *18*, 154–161.

Sindhu Meena, K., & Suriya, S., (2020), A survey on supervised and unsupervised learning techniques, *Proceedings of international conference on artificial intelligence, smart grid and smart city applications: AISGSC 2019*, 627–644.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., (2014), Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research*, *15*, 1929–1958.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., & Wang, J., (2019), High-resolution representations for labeling pixels and regions, *arXiv preprint arXiv:1904.04514*.

Sutton, R. S., (2018), Reinforcement learning: an introduction, *A Bradford Book*.

Tan, M., & Le, Q., (2019), Efficientnet: rethinking model scaling for convolutional neural networks, *International conference on machine learning*, 6105–6114.

Tang, W., van Ooijen, P. M., Sival, D. A., & Maurits, N. M., (2022), 2d gait skeleton data normalization for quantitative assessment of movement disorders from freehand single camera video recordings, *Sensors*, *22*, 4245.

THEIA, (2024), *Theia website*. https://www.theiamarkerless.ca/

Uhlrich, S. D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A. S., Hicks, J. L., & Delp, S. L., (2023), Opencap: human movement dynamics from smartphone videos, *PLoS computational biology*, *19*, e1011462.

Valero, E., Sivanathan, A., Bosché, F., & Abdel-Wahab, M., (2016), Musculoskeletal disorders in construction: a review and a novel system for activity tracking with body area network, *Applied ergonomics*, *54*, 120–130.

Vicon, (2024), *Vicon nexus - drive performance*. https://www.vicon.com/

Wade, L., Needham, L., McGuigan, P., & Bilzon, J., (2022), Applications and limitations of current markerless motion capture methods for clinical gait biomechanics, *PeerJ*, *10*, e12995.

Widyanti, A., (2020), Validity and inter-rater reliability of postural analysis among new raters, *Malaysian Journal of Public Health Medicine*, *1*, 161–166.

Wiktorin, C., Karlqvist, L., Winkel, J., & Group, S. M. I. S., (1993), Validity of self-reported exposures to work postures and manual materials handling, *Scandinavian journal of work, environment & health*, 208–214.

Wu, G., Cavanagh, P. R., et al., (1995), Isb recommendations for standardization in the reporting of kinematic data, *Journal of biomechanics*, *28*, 1257–1262.

Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., Whittle, M., D D'Lima, D., Cristofolini, L., Witte, H., et al., (2002), 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*, 543–548.

Wu, G., Van der Helm, F. C., Veeger, H. D., Makhsous, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A. R., McQuade, K., Wang, X., et al., (2005), 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*, 981–992.

Xsens, (2024), *Mtw awinda - wearables for research*. https://www.xsens.com/products/mtw-awinda/

Xu, G., Wang, X., Wu, X., Leng, X., & Xu, Y., (2024), Development of skip connection in deep neural networks for computer vision and medical image analysis: a survey, *arXiv preprint arXiv:2405.01725*.

Xu, T., An, D., Jia, Y., & Yue, Y., (2021), A review: point cloud-based 3d human joints estimation, *Sensors*, *21*, 1684.

Xu, X., Robertson, M., Chen, K. B., Lin, J.-h., & McGorry, R. W., (2017), Using the microsoft kinect™ to assess 3-d shoulder kinematics during computer use, *Applied Ergonomics*, *65*, 418–423.

Yamalik, N., (2007), Musculoskeletal disorders (msds) and dental practice part 2. risk factors for dentistry, magnitude of the problem, prevention, and dental ergonomics, *International dental journal*, *57*, 45–54.

Yasin, H., Ghani, S., & Krüger, B., (2023), An effective and efficient approach for 3d recovery of human motion capture data, *Sensors*, *23*, 3664.

Yasin, H., Hussain, M., & Weber, A., (2020), Keys for action: an efficient keyframe-based approach for 3d action recognition using a deep neural network, *Sensors*, *20*, 2226.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H., (2014), How transferable are features in deep neural networks?, *Advances in neural information processing systems*, *27*.

Yuan, H., & Zhou, Y., (2023), Ergonomic assessment based on monocular rgb camera in elderly care by a new multi-person 3d pose estimation technique (romp), *International Journal of Industrial Ergonomics*, *95*, 103440.

Yunus, M. N. H., Jaafar, M. H., Mohamed, A. S. A., Azraai, N. Z., & Hossain, M. S., (2021), Implementation of kinetic and kinematic variables in ergonomic risk assessment using motion capture simulation: a review, *International Journal of Environmental Research and Public Health*, *18*, 8342.

Zell, P., & Rosenhahn, B., (2017), Learning-based inverse dynamics of human motion, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 842–850.

Zell, P., & Rosenhahn, B., (2020), Learning inverse dynamics for human locomotion analysis, *Neural Computing and Applications*, *32*, 11729–11743.

Zell, P., Rosenhahn, B., & Wandt, B., (2020), Weakly-supervised learning of human dynamics, *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 68–84.

Zhang, J., Zhao, Y., Shone, F., Li, Z., Frangi, A. F., Xie, S. Q., & Zhang, Z.-Q., (2022), Physics-informed deep learning for musculoskeletal modeling: predicting muscle forces and joint kinematics from surface emg, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Zhang, X., Zhao, B., Yao, J., & Wu, G., (2023), Unsupervised monocular depth and camera pose estimation with multiple masks and geometric consistency constraints, *Sensors*, *23*, 5329.

Zhao, J., Obonyo, E., & G. Bilén, S., (2021), Wearable inertial measurement unit sensing system for musculoskeletal disorders prevention in construction, *Sensors*, *21*, 1324.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., & Shah, M., (2023), Deep learning-based human pose estimation: a survey, *ACM Computing Surveys*.

Zheng, H., Shen, L., Tang, A., Luo, Y., Hu, H., Du, B., & Tao, D., (2023), Learn from model beyond fine-tuning: a survey, *arXiv preprint arXiv:2310.08184*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q., (2020), A comprehensive survey on transfer learning, *Proceedings of the IEEE*, *109*, 43–76.

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

école normale supérieure rennes

Titre : Contributions à l'Analyse Biomécanique et Ergonomique in-situ des Postes de Travail à l'aide de l'Apprentissage Automatique et de l'Apprentissage Profond

Mot clés : Évaluation des risques biomécaniques, santé au travail, analyse du mouvement humain, ergonomie, estimation de la pose humaine, apprentissage automatique, apprentissage profond, estimation des coordonnées articulaires, estimation des couples articulaires, cinématique inverse, dynamique inverse.

Résumé :

L'évaluation du risque de troubles musculosquelettiques en milieu industriel représente un défi en raison de la complexité des processus de fabrication modernes. Ces environnements comprennent divers facteurs influençant l'activité des opérateurs, tels que les éléments organisationnels, managériaux et environnementaux, ainsi que le rythme de travail. Il est crucial d'évaluer les contraintes physiques auxquelles sont soumis les opérateurs pour prévenir ces troubles. Bien que de nombreux systèmes monitorent actuellement les mouvements des opérateurs et évaluent les contraintes posturales pour fournir un aperçu de l'activité physique, ils échouent souvent à analyser les forces physiques subies ou générées par l'opérateur. Par conséquent, il est essentiel de quantifier ces forces afin d'identifier les facteurs de risque physique liés à l'effort. Cependant, les méthodes classiques de mesure impliquent souvent des processus complexes, invasifs et peu pratiques en milieu industriel.

Cette thèse relève ces défis en évaluant des approches d'apprentissage pour estimer les contraintes physiques sans recourir à des mesures invasives, ce qui est fondamental pour améliorer les outils et les pratiques ergonomiques. Nous avons commencé par comparer la précision et la robustesse des systèmes de mesure basés sur la vision par ordinateur pour l'évaluation du RULA, en nous focalisant particulièrement sur les évaluations ergonomiques sur site. Notre analyse s'est principalement concentrée sur l'évaluation des systèmes basés sur la vision par ordinateur, y compris ceux dotés d'une ou plusieurs caméras, utilisant des images RVB ou des images de profondeur, et les systèmes qui s'appuient uniquement sur des données visuelles ou qui intègrent des capteurs portables (systèmes hybrides). Ensuite, nous avons développé et évalué plusieurs architectures d'apprentissage conçues pour émuler l'étape de la dynamique inverse dans l'analyse du mouvement. Ces dernières prédisent les couples articulaires à partir des données squelettiques de l'opérateur et son poids et la masse de la charge transportée, offrant ainsi une nouvelle alternative aux méthodes classiques de dynamique inverse. Enfin, nous avons examiné la généralisabilité des outils basés sur l'apprentissage profond, tels qu'OpenCap, dans les tâches industrielles. En utilisant le fine-tuning - une technique courante dans l'apprentissage profond pour adapter les modèles à de nouveaux ensembles de données avec des échantillons minimaux -nous avons cherché à adapter les modèles d'apprentissage d'OpenCap à un nouveau type de mouvement et à un nouvel ensemble de marqueurs.

Title: Contributions to the in-situ Biomechanical and Ergonomic Analysis of Workstations using Machine Learning and Deep Learning

Keywords: Biomechanical risk assessment, occupational health, human motion analysis, ergonomics, human pose estimation, machine learning, deep learning, joint coordinates estimation, joint torques estimation, inverse kinematics, inverse dynamics.

Abstract:

Assessing the risk of musculoskeletal disorders in industrial environments is a challenging task, given the complexity of modern manufacturing processes. These environments include various factors influencing operator activity, such as organizational, managerial and environmental elements, as well as the pace of work. Assessing the physical constraints to which operators are subjected is crucial to preventing these disorders. Although many systems currently monitor operator movements and assess postural constraints to provide an overview of physical activity, they often fail to analyze the physical forces experienced or generated by the operator. Consequently, it is essential to quantify these forces in order to identify effort-related physical risk factors. However, conventional measurement methods are often complex, invasive and impractical in industrial environments.

This thesis addresses these challenges by evaluating learning approaches for estimating physical stresses without resorting to invasive measurements, which is fundamental to improving ergonomic tools and practices. We began by comparing the accuracy and robustness of computer vision-based measurement systems for RULA assessment, focusing particularly on on-site ergonomic evaluations. Our analysis focused primarily on the evaluation of computer vision-based systems, including those with one or more cameras, using RGB or depth images, and systems that rely solely on visual data or incorporate wearable sensors (hybrid systems). Next, we developed and evaluated several learning architectures designed to emulate the inverse dynamics step in motion analysis. These predict joint torques from the operator's skeletal data and the weight and mass of the load carried, thus offering a new alternative to classical inverse dynamics methods. Finally, we examined the generalizability of deep learning-based tools, such as OpenCap, in industrial tasks. Using fine-tuning - a common technique in deep learning for adapting models to new data sets with minimal samples - we sought to adapt OpenCap's learning models to a new type of motion and a new set of markers.