Politecnico di Bari

Department of Electrical and Information Engineering

INDUSTRY 4.0

Ph.D. Program

SSD: ING-INF/05 – INFORMATION PROCESSING SYSTEMS

Final Dissertation

Vision devices and intelligent systems for monitoring the well-being of humans in healthcare and manufacturing

---

by

Laura Romeo

Supervisors:

Prof. Anna Gina Perri

Dr. Roberto Marani

Coordinator of Ph.D. Program:

Prof. Caterina Ciminelli

Course n°36, 01/11/2020-31/10/2023

Politecnico di Bari

LIBERATORIA PER L'ARCHIVIAZIONE DELLA TESI DI DOTTORATO

Al Magnifico Rettore
del Politecnico di Bari

La sottoscritta LAURA ROMEO nata a POLICORO (MT) il 03/05/1994
residente a PISTICCI (MT) in via COTUGNO, N 31 e-mail laura.romeo@poliba.it
iscritto al 3° anno di Corso di Dottorato di Ricerca in INDUSTRIA 4.0 ciclo XXXVI
ed essendo stato ammesso a sostenere l'esame finale con la prevista discussione della tesi dal titolo:
Vision devices and intelligent systems for monitoring the well-being of humans in healthcare and manufacturing

DICHIARA

1) di essere consapevole che, ai sensi del D.P.R. n. 445 del 28.12.2000, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
2) di essere iscritto al Corso di Dottorato di ricerca INDUSTRIA 4.0 ciclo XXXVI corso attivato ai sensi del "*Regolamento dei Corsi di Dottorato di ricerca del Politecnico di Bari"*, emanato con D.R. n.286 del 01.07.2013;
3) di essere pienamente a conoscenza delle disposizioni contenute nel predetto Regolamento in merito alla procedura di deposito, pubblicazione e autoarchiviazione della tesi di dottorato nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica;
4) di essere consapevole che attraverso l'autoarchiviazione delle tesi nell'Archivio Istituzionale ad accesso aperto alla letteratura scientifica del Politecnico di Bari (IRIS-POLIBA), l'Ateneo archivierà e renderà consultabile in rete (nel rispetto della Policy di Ateneo di cui al D.R. 642 del 13.11.2015) il testo completo della tesi di dottorato, fatta salva la possibilità di sottoscrizione di apposite licenze per le relative condizioni di utilizzo (di cui al sito http://www.creativecommons.it/Licenze), e fatte salve, altresì, le eventuali esigenze di "embargo", legate a strette considerazioni sulla tutelabilità e sfruttamento industriale/commerciale dei contenuti della tesi, da rappresentarsi mediante compilazione e sottoscrizione del modulo in calce (Richiesta di embargo);
5) che la tesi da depositare in IRIS-POLIBA, in formato digitale (PDF/A) sarà del tutto identica a quelle **consegnate**/inviate/da inviarsi ai componenti della commissione per l'esame finale e a qualsiasi altra copia depositata presso gli Uffici del Politecnico di Bari in forma cartacea o digitale, ovvero a quella da discutere in sede di esame finale, a quella da depositare, a cura dell'Ateneo, presso le Biblioteche Nazionali Centrali di Roma e Firenze e presso tutti gli Uffici competenti per legge al momento del deposito stesso, e che di conseguenza va esclusa qualsiasi responsabilità del Politecnico di Bari per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto il Politecnico di Bari ed i suoi funzionari sono in ogni caso esenti da responsabilità di qualsivoglia natura: civile, amministrativa e penale e saranno dal sottoscritto tenuti indenni da qualsiasi richiesta o rivendicazione da parte di terzi;
7) che il contenuto della tesi non infrange in alcun modo il diritto d'Autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta.

Luogo e data BARI, 19/12/2023 Firma Laura Romeo

La sottoscritta, con l'autoarchiviazione della propria tesi di dottorato nell'Archivio Istituzionale ad accesso aperto del Politecnico di Bari (POLIBA-IRIS), pur mantenendo su di essa tutti i diritti d'autore, morali ed economici, ai sensi della normativa vigente (Legge 633/1941 e ss.mm.ii.),

CONCEDE

- al Politecnico di Bari il permesso di trasferire l'opera su qualsiasi supporto e di convertirla in qualsiasi formato al fine di una corretta conservazione nel tempo. Il Politecnico di Bari garantisce che non verrà effettuata alcuna modifica al contenuto e alla struttura dell'opera.
- al Politecnico di Bari la possibilità di riprodurre l'opera in più di una copia per fini di sicurezza, back-up e conservazione.

Luogo e data BARI, 19/12/2023 Firma Laura Romeo

Politecnico di Bari

Department of Electrical and Information Engineering

INDUSTRY 4.0

Ph.D. Program

SSD: ING-INF/05 – INFORMATION PROCESSING SYSTEMS

Final Dissertation

Vision devices and intelligent systems for monitoring the well-being of humans in healthcare and manufacturing

---

by

Laura Romeo

Laura Romeo

Referees:

Prof. Antonio Caruso

Prof. César Domínguez Pérez

Supervisors:

Prof. Anna Gina Perri

Dr. Roberto Marani

Coordinator of Ph.D Program:

Prof. Caterina Ciminelli

Course n°36, 01/11/2020-31/10/2023

Acknowledgements

Devo ammettere che è strano essere arrivati alla fine di questo viaggio di tre anni, un viaggio che mai nella vita avrei immaginato di compiere. Questi anni hanno segnato il mio ingresso nell'età adulta, e mi hanno fatto capire che sono circondata da persone splendide, che hanno contribuito giorno dopo giorno alla mia formazione accademica, lavorativa e personale.

Tra queste, una menzione particolare va ai miei tutor di Dottorato, Anna Gina Perri e Roberto Marani, che mi hanno seguito nel corso di questi anni, dandomi la possibilità di acquisire da loro competenze non solo a livello accademico e professionale, ma anche a livello umano. La loro presenza è stata indispensabile, e non semplicemente per il ruolo che hanno avuto nel corso del mio Dottorato. Le indicazioni che mi hanno dato si sono incastrate perfettamente con gli obiettivi che io stessa mi ero prefissata, facendo in modo che questi anni diventassero, accademicamente parlando, i più stimolanti della mia vita.

Grazie a Juergen Gall e al Computer Vision Group, che mi ha accolto presso la Bonn Universitat per sei bellissimi mesi, arricchendo ulteriormente le mie competenze.

Un ringraziamento speciale va poi a tutti i miei colleghi del gruppo ISP presso l'Istituto STIIMA del CNR di Bari, ognuno dei quali ha dato un contributo nel formarmi in questi anni. Grazie ad Antonio Petitti e Annalisa Milella per avermi "iniziato" al mondo della ricerca fin dalla mia tesi di Laurea Magistrale. Grazie a Tiziana D'Orazio e Grazia Cicirelli, le quali si sono rivelate fondamentali per la mia crescita professionale, aiutandomi ad acquisire il giusto metodo per svolgere questo lavoro. Tutti loro hanno visto la mia "vocazione" per la ricerca ancor prima che la vedessi io, e non potrò mai ringraziarli abbastanza per avermi aiutato a trovare la mia strada. Ricordo che, dopo aver partecipato alla mia prima conferenza, Tiziana e Grazia mi dissero che secondo loro ero "nata per fare ricerca", e devo dire che ancora oggi questa affermazione mi motiva a dare il massimo ogni giorno di più.

Grazie ai miei genitori e a mia sorella, senza i quali sicuramente non sarei arrivata dove sono ora. Grazie a tutte quelle persone che ho la fortuna immensa di chiamare amiche e amici, sempre pronte ad ascoltare i miei dubbi e le mie paure sul futuro.

E, infine, il grazie più grande va alla persona a cui ho deciso di dedicare questa tesi, che ha sempre, sempre, sempre creduto nelle mie capacità, e che spero continui ad avere la giusta pazienza per rimanere accanto a me (e al nostro gatto, Ade) per tutta la vita.

A Nico,
la mia famiglia

Abstract

The present PhD research explores the integration of vision devices and intelligent systems to monitor and enhance human well-being in healthcare and manufacturing contexts, starting from the standards proposed in Industry 4.0 and aiming to follow the principles of the novel Industry 5.0. Depth sensors and deep learning technologies have been exploited to address the critical aspects of human mobility assessment and action segmentation in real, non-simulated scenarios. The Microsoft Azure Kinect, a state-of-the-art depth sensor, has been selected as a key instrument for data collection, and innovative camera calibration methods have been developed to ensure the accuracy and reliability of the gathered data.

Within the realm of healthcare, the research activity addresses the substantial challenges posed by neurodegenerative diseases in the well-being of older individuals. This part of the study focuses on monitoring and assessing the mobility of elderly patients, aiming to support remote diagnosis and improve their quality of life. Traditional mobility tests, administered by healthcare professionals, are essential for evaluating movement skills. Nevertheless, such techniques often suffer from human subjectivity, which could lead to errors in the assessments. To address such issues, video-based systems have been studied, aiming to remotely monitor and objectively evaluate mobility, reducing the burden on elderly patients.

In the context of manufacturing, human actions are pivotal in enhancing operational efficiency, productivity, and safety in manufacturing environments. Such challenges have led to the increasing use of industrial robotic solutions, mainly including collaborative robots, which can share a common workspace with humans, carrying out their respective tasks simultaneously. This part of the research delves into the segmentation of human tasks for intelligent manufacturing systems, exploring the integration of vision devices and deep learning technologies to improve the efficiency and accuracy of manufacturing processes. In general, the study of such systems is aimed at creating comfortable work environments, adaptable to the needs and abilities of individual people, increasing the well-being of operators in a human-centered factory concept.

The main goal of the present study is to evaluate the effectiveness of machine learning and deep learning models for mobility assessment and action segmentation, to determine their suitability for human monitoring. However, a notable gap in the literature is identified: the absence of datasets representing human actions in realistic environments. To bridge this gap, the research includes the creation and validation of datasets capturing human actions in healthcare and manufacturing scenarios, emphasizing the importance of generalization across different locations. By addressing the unique challenges in both healthcare and manufacturing, this study contributes to the development of intelligent systems that promote human well-being and enhance operational efficiency, aiming to align with the paradigms of Industry 5.0.

Contents

1 Introduction 1
- 1.1 Human Mobility Assessment in Healthcare . . . 2
- 1.2 Human Action Segmentation in Manufacturing . . . 5
- 1.3 Experimental Approach and Objectives . . . 7

2 System analysis for Human Monitoring 9
- 2.1 Introduction . . . 9
- 2.2 Azure Kinect Body Tracking analysis . . . 10
  - 2.2.1 Design of Experiments . . . 10
  - 2.2.2 Performance analysis results . . . 16
  - 2.2.3 Discussion . . . 19
- 2.3 Microsoft Azure Kinect Calibration . . . 20
  - 2.3.1 Calibration Methodology . . . 21
  - 2.3.2 Experimental setup . . . 26
  - 2.3.3 Calibration analysis . . . 28
  - 2.3.4 Discussion . . . 33

3 Video Data Acquisition for Human Monitoring 35
- 3.1 Introduction . . . 35
- 3.2 Data Acquisition in elderly facilities: SPPB Dataset . . . 36
  - 3.2.1 Tests Definition . . . 37
  - 3.2.2 Methodology . . . 38
  - 3.2.3 Dataset Evaluation . . . 39
  - 3.2.4 Discussion . . . 40
- 3.3 Data acquisition in manufacturing: HA4M Dataset . . . 40
  - 3.3.1 Study Design . . . 43
  - 3.3.2 Acquisition Setup . . . 44
  - 3.3.3 Study Participants . . . 46
  - 3.3.4 Data Annotation . . . 46

3.3.5 Technical Validation 47

4 Machine Learning and Deep Learning methodologies for Human Mobility Assessment in elderly facilities **52**
4.1 Introduction 52
4.2 Case Study Description 53
4.3 Methodology 55
4.3.1 Feature Extraction 56
4.3.2 Deep Neural Network Architectures 60
4.3.3 Data Augmentation 63
4.4 Experiments 65
4.4.1 Data Acquisition and Processing 65
4.4.2 Classification 66
4.4.3 Regression 70
4.4.4 Conv-BiLSTM Classifier: in-depth analysis 71
4.5 Discussion 73

5 Deep Learning methodologies for Human Action Segmentation in manufacturing scenarios **75**
5.1 Introduction 75
5.2 Methodological Approach 76
5.2.1 Feature Extraction 77
5.2.2 Dataset Splittings 78
5.2.3 New Data Collection 79
5.2.4 Semi-supervised Learning 79
5.2.5 Deep Learning models Selection 79
5.2.6 Evaluation Metrics 81
5.3 Experiments 82
5.3.1 Cross-Subject Evaluation 83
5.3.2 Cross-Location Evaluation 85
5.3.3 New Data Evaluation 86
5.4 Discussion 91

6 Conclusion **94**

References **113**

1 Introduction

Human action recognition and segmentation are active topics of research in computer vision [1, 2] and machine learning [3, 4], and vast research work has been carried out in the last decade, as it can be seen in the existing literature [5]. Such fields aim to understand human activities occurring in video sequences, offering valuable insights across various applications. More specifically:

- Human action recognition refers to the process of identifying and classifying specific actions or activities performed by individuals within a trimmed video sequence. The main focus is to recognize the overall human actions, giving as output a classification label for the entire action in the video.
- Human action segmentation refers to the task of dividing a continuous video sequence into segments, which correspond to a distinct action or activity. The main scope is to pinpoint the exact temporal boundaries for each action, returning temporal segments that indicate frame-wise when each action or sub-action occurs.

In this context, the recent widespread of low-cost video camera systems, including depth-cameras [6], has strengthened the development of observation systems in a variety of application domains such as video-surveillance, safety, smart home security, ambient assisted living, health-care and manufacturing. However, little work has been done in human action recognition and segmentation for manufacturing assembly tasks and elderly human mobility assessment [7, 8, 9], and the poor availability of public datasets limits the study, development,

and comparison of new methods. This is mainly due to challenging issues such as between-action similarity, complexity of actions, and availability of setups that guarantee real, non-simulated data.

As technological innovations continue to reshape the boundaries of human capabilities and potential, the integration of vision devices and intelligent systems has emerged as one of the most advanced solutions, particularly with the advent of Industry 5.0[10]. Such solutions have found various applications across different domains, but they impact particularly on the well-being of humans in healthcare and manufacturing scenarios. Both domains rely heavily on computer vision techniques, mostly regarding the extraction of video (i.e. RGB, Depth, IR, and RGB-D data) and skeleton information, and employ deep learning methodologies to unlock new dimensions in monitoring and improving human well-being.

In recent years, the need for trustworthy RGB-D sensors has increased importance in many fields [11, 12, 13]. Among RGB-D devices, the Microsoft Azure Kinect [14] (Redmond, Washington, US), released in 2019, is a Time-of-Flight (ToF) sensor [15] that offers considerably higher accuracy than other commercially available devices [16] at low cost. In addition, the possibility of exploiting the Azure Kinect Software Development Kit (SDK), even for the extraction of skeletal joints with the Azure Kinect Body Tracking SDK, represents a further step beyond the previous Kinect versions [17]. Such perks make the Azure Kinect one of the most reliable cameras used in many research fields [18, 19, 20], including healthcare and manufacturing.

The following Sections 1.1 and 1.2 deepen two vital domains where vision devices and intelligent systems have the potential to revolutionize human well-being: healthcare and manufacturing. First, the critical role of human movement analysis in healthcare is explored, with a specific focus on the elderly population and the significance of monitoring their mobility to improve their quality of life. Subsequently, the importance of temporal action segmentation and intelligent vision systems in manufacturing scenarios is unraveled, emphasizing how it can optimize operational efficiency, foster human-robot collaboration, ensure worker well-being, and boost productivity in manufacturing settings. Section 1.3 marks the concluding segment of this Chapter, introducing the experimental approaches and the goals that the present study aims to achieve.

1.1 Human Mobility Assessment in Healthcare

In the healthcare context, the analysis of human movements has allowed the realization of various functions such as remote diagnosis, support in the surveillance of fragile patients, recognition of anomalous events, etc. Many products and services have been developed for Ambient Assisted Living to aid healthy, active, and happy aging. The world is experiencing a rapid increase in the number of older people, which is expected to double over the next

three decades [21]. Furthermore, there is an increasing spread of neurodegenerative diseases that heavily affect the well-being and healthy aging of the elderly population [22]. As a consequence, elderlies need periodic monitoring to assess their movement skills. However, they are often unwilling to visit health clinics regularly, because of disabilities or logistical limitations, such as living in remote areas, thus wasting time, effort, and travel costs.

In this scenario, the analysis and control of people's motion and cognition abilities are fundamental in improving their social and clinical living conditions. Several studies demonstrate a strict link between cognitive impairment and motion dysfunction, including deficits in gait and balance [23], [24]. So, the study of human movements by video analysis can significantly help assess people's motion abilities, providing objective evaluations and supporting remote diagnosis. Well-defined mobility tests exist in clinical contexts to assess people's mobility [25]. They consist of postural stability exercises, usually administrated and observed by physicians or specialized physiotherapists to measure people's functional mobility. Automatic video-based systems could greatly help to monitor these exercises in both home and clinical environments, obtaining objective and quantitative evaluations to support both expert personnel and medical diagnosis.

In the existing literature, various instrumented systems have been proposed for real-time assessment of older people's mobility [23, 26, 27, 24].

Several works propose wearable sensors based on Inertial Measurement Units [28], or Inertial and Magnetic Measurement Systems for the evaluation of the physical functions of individuals [29, 30]. These sensors include accelerometers, gyroscopes, and magnetometers that measure the acceleration or angular velocity of the body segments to which they are attached. Although wearable sensors return valid information related to the movement of people, their output strictly depends on their position and orientation, and the activities to be monitored. Furthermore, older people, especially those suffering from neurological disorders, do not easily accept unfamiliar devices.

Contrary to wearable sensors, non-wearable ones are non-invasive for people, as they are placed in the environment. Among the most commonly used for evaluating motion abilities, there are vision-based systems characterized by cameras that acquire video information of the human body and then, by using image processing techniques, extract relevant parameters useful for the analysis of motion abilities [31]. Marker-based Motion Capture Systems (MCSs), consisting of several cameras and a set of retro-reflective markers attached to the body of the monitored subjects, are an example of vision systems beneficial for capturing human movements with reliable accuracy [32]. However, high installation costs, expertise to set up and operate the system, and marker placement and calibration, limit their use in the home, and clinical environments [33]. Furthermore, the need for markers placed on the body brings out the same drawbacks of wearable systems. Typically, MCSs are used primarily in research laboratories or controlled environments to validate other sensory systems, such as webcams

or RGB-D cameras, due to their high accuracy [34].

The limitations of marker-based systems have led to the development of markerless vision-based systems for human motion analysis [35]. In the last few years, the progress in new and low-cost optical technologies, together with the development of new and accurate pattern recognition approaches, has led to an increase in vision-based research works in this context [36, 37]. Monocular RGB cameras, stereo cameras, thermal cameras, and the recently developed RGB-D cameras, such as Microsoft Kinect or Intel RealSense [38], are the most commonly used systems to capture body movements and postural stability for assessing physical dysfunctions [39, 40].

A Kinect camera is used in [41] to observe older people while performing the Sit-to-Stand test to quantify the time taken to perform the test and to discriminate between elderly fallers and non-fallers in both laboratory and home assessments. A Kinect-based system has also been used to calculate the postural sway of older adults, estimating the variation of the center of mass of the body to provide a risk assessment of falls [42] or discriminate postural abnormalities [43].

In general, gathering data by observing people is not enough to assess the postural stability problem of human beings. Such information must be processed and elaborated through proper advanced systems to extract as much information as possible regarding the health of the elderly. In recent years, machine learning techniques for assessing movement skills are gaining more and more interest in the healthcare field [44, 45]. In particular, deep learning methodologies prove to be fundamental in health informatics. The development of automatic methods can lead to the generation, processing, and evaluation of complex data, which is difficult to deal with without the aid of technological systems.

Several deep learning architectures have been used to process different types of data. Among them, the Convolutional Neural Networks (CNNs) are usually of significant impact in pattern recognition, from image to voice processing [46]. In [47], two types of CNN architecture, designed to analyze footprint pressure images from an instrumented walkway, have been compared to classify Huntington's disease severity. Similarly, a CNN was used in [48] to classify three severity stages of Alzheimer's disease using accelerometer data records. Considering the complexity of the classification problem and the presence of complex pattern sequences of mixed length, CNN seems suitable for managing this type of data and obtaining high accuracy rates for the three classes.

Alternatively, Recurrent Neural Networks (RNNs) are widely used for the analysis of time series in applications where the outputs depend on the previous computations, such as the analysis of text, speech, and movements. In [49], an RNN processes accelerometer signals to detect falls and estimate corresponding risks in real time, reaching high efficiency and accuracy.

An evolution of the RNN is the Long Short-Term Memory (LSTM) network, which adds

cell states to the network to expand the memory of the RNN [50]. In [51], the LSTM network has been applied to sequences of spatiotemporal gait parameters to capture both temporal variations and asymmetries in gait in patients with Parkinson's disease. LSTM network, taking advantage of remembering long-term dependencies within the data, achieves high accuracy rates [52].

In general, deep learning methods have several shortcomings. They typically have very complex architectures and time-intensive training phases. Furthermore, they need a large amount of data to reveal good performance. As a result, algebraic operations involving dense matrices, matrix products, and convolutions require equally enormous resources. Therefore, they must be transferred to Graphic Processing Units (GPUs) to accelerate machine learning processes [53]. However, compared to traditional methods, deep learning methods automatically learn hierarchical feature representations that capture their spatial and temporal correlations. In addition, such methods can approximate complex non-linear functions by composing several transformations of feature representations among the network layers from one level to more abstract levels.

1.2 Human Action Segmentation in Manufacturing

The segmentation of human actions in the context of intelligent manufacturing is of great importance for various purposes, such as improving operational efficiency[8, 54], promoting human-robot cooperation [55], assisting operators [56], supporting employee training [57, 9], increasing productivity and safety [58], and promoting workers' good mental health [59].

In this context, in Human-Robot Interaction (HRI) and Human-Robot Collaboration (HRC), operator confidence plays a fundamental role in optimal interaction and collaboration [60, 61]. Monitoring devices and systems can be integrated into the shared workspace, aiming to lead the robot to fully adapt to the operator, guaranteeing the well-being of humans, thus reducing those factors that can be marked as risky or harmful to operators, both physically and cognitively. The integration of vision devices can drastically improve the behavior and efficiency of both humans and robots [62].

In recent years, various robotic solutions based on computer vision have been implemented, which significantly improve the efficiency and accuracy of manufacturing processes. Computer vision is associated with deep learning methodologies, which enable proper processing and elaboration of the gathered data [63]. In the context of industrial applications, computer vision can be implemented for various tasks, such as object recognition and tracking [64], robot navigation and localization [65], HRI and HRC [66, 67].

HRC and HRI can benefit from temporal action segmentation methodologies [68, 67], which divide a continuous stream of human activity into distinct segments, each corresponding to a semantically meaningful action. In manufacturing contexts, such algorithms allow

robots to understand and respond to the actions of human operators. These systems are implemented using techniques such as motion capture [69] or depth sensors [70], and can be used to increase the efficiency and safety of the production processes.

The current literature outlines that state-of-the-art models for temporal action segmentation are not used for analyzing which information and technique represent the best solution for the development of a system that segments the action of an operator performing a task in manufacturing scenarios.

Action segmentation methodologies using video and skeletal data with different levels of supervision have been widely addressed in the literature [71]. In temporal action segmentation approaches based on video data, RGB and Depth information are often considered for training deep learning models, since such data give additional information about environments and objects [72], which can be helpful in action segmentation tasks. On the other hand, skeletal data provide information about the pose and movement of a human body over time, which can be particularly useful for action segmentation algorithms addressing human tasks [73].

Several deep learning algorithms can be used for action segmentation, including approaches such as Convolutional Neural Networks (CNNs) [74, 75, 76] and Recurrent Neural Networks (RNNs) [77] models. These algorithms are trained on datasets of labeled video sequences to learn how to identify and segment the actions or events in the video involving various human actions. For instance, the work in [78] depicts a network model created to be built on top of existing action segmentation models, aiming to learn the relation of multiple action segments. Such model has been validated on egocentric [79, 80] and third-person [81, 82] datasets, all representing humans performing daily actions.

Another approach considered in the literature for action segmentation includes using Hidden Markov Models (HMMs) [83, 84], probabilistic models that can be used to represent temporal sequences to identify patterns and transitions between different actions. As an example, in [85], the authors present a weakly supervised action segmentation model using a hybrid RNN-HNN system. Here, the RNN is used as basic recognition model, while the HMM is used to model each action as a combination of subactions. The model has been validated on a dataset representing people making breakfast [81], and on a dataset filled with video sequences of Hollywood movies [86].

Looking closely at the literature, it is clear that there is a lack of work where action segmentation algorithms are applied to and evaluated on manufacturing tasks. The state-of-the-art models for action segmentation are used to assess datasets where humans perform various actions [87, 88]. However, such datasets [89, 82, 81, 90, 86] do not cover human actions in manufacturing environments, or while performing assembling tasks in production processes. [91] presents Assembly101, a multi-view dataset composed of people assembling and disassembling toy vehicles in a singular scenario, which has been validated on MS-TCN++

[92] and C2F-TCN [93]. However, the Assembly101 dataset is not performed in a manufacturing environment, or with a manufacturing object. Furthermore, the dataset has been captured at a single location and can thus not be used to measure the generalization capabilities across different locations, which is very important for manufacturing. [94] evaluates a custom deep learning model on an action segmentation dataset containing 24 atomic actions from video data, in a realistic robotics assembly production line. The proposed model is also compared with other models, such as MS-TCN [74]. It must be noticed, however, that the dataset is not publicly available. Furthermore, the extraction process of the features used for training follows a specific pipeline. Such features are focused only on the hand movements, thus there is not any information about the complete body of the operators. Furthermore, the structure of the features makes the methodology complex to be generalized on other tasks.

1.3 Experimental Approach and Objectives

The proposed thesis aims at monitoring the well-being of humans in the context of healthcare and manufacturing. More specifically, the goal is to prove how vision devices and intelligent systems, such as depth sensors combined with deep learning methodologies, can massively help in gathering and elaborating information about human mobility in order to guarantee the well-being of humans. Such information is crucial in both healthcare and manufacturing domains, where the ability of a man/woman to move is strictly correlated to his/her physical and cognitive conditions. The analysis and experiments defined and performed in the present work have been validated by several publications [95, 96, 97, 98, 99, 63].

The main contribution of this thesis is three-fold:

- It introduces and implements novel camera calibration methodologies based on RGB and Infrared data, with or without the associated Depth information. The Microsoft Azure Kinect has been chosen as the sensor used to carry out the experiments. The results have been discussed considering a preliminary analysis of the skeletal joints data obtained from the body tracking system.

- A video acquisition campaign has been carried out in rea-world settings, including elderly care facilities and manufacturing environments. This approach aims at capturing data from real patients and operators in action, rather than relying on simulated scenarios. Patients performing specific motion exercises and operators assembling industrial objects have been recorded using multi-camera systems involving both RGB and RGB-D sensors.

- Machine Learning and Deep Learning methodologies are developed and applied for the assessment and segmentation of human mobility tasks in the aforementioned sce-

narios. Information grabbed from real patients suffering from neurodegenerative diseases and operators performing assembling tasks with collaborative robots have been used to evaluate mobility performance while recognizing and segmenting different types of actions.

The remainder of this thesis is structured as follows. Chapter 2 presents the sensors used for Human Monitoring, focusing on the analysis of the Azure Kinect camera, and on the calibration techniques developed to properly perform 2D and 3D calibration. Chapter 3 defines the acquisition campaign carried out for the gathering of Datasets using RGB and RGB-D camera systems, in healthcare and manufacturing domains. Chapter 4 focuses on Machine Learning and Deep Learning methodologies for human mobility assessment in elderly facilities, while Chapter 5 deepens the Deep Learning methodologies for human action segmentation in manufacturing scenarios. Finally, Chapter 6 draws the conclusions.

2

System analysis for Human Monitoring

2.1 Introduction

Nowadays, the need for reliable and low-cost multi-camera systems is increasing for many potential applications, such as localization and mapping, human activity recognition, hand and gesture analysis, and object detection and localization [63]. The exact position of the humans can be easily inferred from RGB-D cameras, whose output can be processed by body tracking modules to produce exact pose estimations in real-time. However, a precise camera calibration approach is mandatory for enabling further applications that require high precision.

The present Chapter is divided into two main Sections. Section 2.2 sheds light on the alteration of measurement uncertainty in quasi-static acquisitions of human bodies. This work, which has been published in [95], uses an experimental setup made of an Azure Kinect camera to obtain data by changing intrinsic and extrinsic parameters. Section 2.3 presents different calibration methodologies using 2D and 3D approaches, all exploiting the functionalities within the Azure Kinect devices. Such work has been published in [96], and its goal is to obtain a guideline for calibrating multiple Azure Kinect RGB-D sensors to achieve the best alignment of point clouds in both color and infrared resolutions and skeletal joints returned by the Microsoft Azure Body Tracking library.

2.2 Azure Kinect Body Tracking analysis

This Section experimentally explores the performance of the affordable Microsoft Azure Kinect RGB-D camera and its body-tracking library. A parametric analysis of the uncertainty of the estimation of the skeleton joints is performed by changing the ambient light conditions, the presence of occlusions, the infrared camera resolution, and the human-camera distance. The acquired data are processed by the Azure Kinect Body Tracking SDK to highlight the worst operating conditions that may significantly affect the reliability of the output data.

The Section is structured as follows: in Section 2.2.1, the acquisition setup is presented together with the design of experiments; corresponding results are then in Section 2.2.2; discussion and remarks are finally shown in Section 2.2.3.

2.2.1 Design of Experiments

In [15], a deep study on the performance of the Azure Kinect and its body tracking SDK for gait analysis of several subjects is presented. The comparison of results with the Vicon motion capture system is performed displaying mean and standard deviations of the Euclidean distances between 3D joints computed by the Kinect sensor and the Vicon system. On the contrary, the proposed work only focuses on the standard deviation, directly linked to measurement uncertainty, computed by changing:

- Intrinsic parameter: Depth resolution.
- Extrinsic parameters: Ambient light conditions, body occlusions, subject-camera distance.

The following subsections will present the proposed setup and the processing procedures for the performance assessment of body tracking.

Setup Definition

The proposed investigation is performed using the experimental scheme of Fig. 2.1. Here, an Azure Kinect sensor is placed at a distance d from the user, ranging from 1 to 3 m by steps of 1 m. At the same time a halogen light source, having a power of 300 W, illuminates the scene, directly towards the user. Two different lighting conditions can be determined as the light source is switched on or off. Specifically, when the lamp is on, the illuminance E_v at 1 m of distance from the source is equal to 1750 lux, whereas this value is down to about 10 lux when the source is off. In both cases, camera exposure has been set to auto with framerate

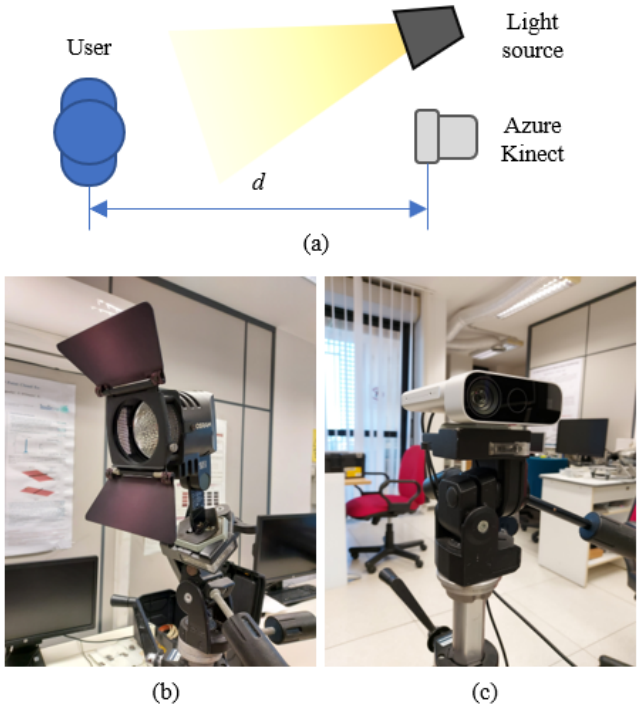

Figure 2.1: (a) Sketch of the experimental setup made of (b) a halogen light projector and (c) the Azure Kinect camera. d is the distance between the camera and the user.

priority. As a consequence, the maximum exposition time can be equal to the inverse of the camera framerate.

Fig. 2.2 shows the two different lighting conditions of the experiments. It is worth noticing that the low-light condition is not realistic in industrial environments. Anyway, it has been considered to bring the operating condition of the Azure Kinect to the limit and to better highlight its different behavior.

As stated previously, skeletons are computed either having the full user body insight ($Occl = w/o$) or with an opaque obstacle that occludes the lower part of the user body ($Occl = w/$), i.e. from the legs down. Moreover, the Azure Kinect offers two depth modes, wide ($Res = W$) and narrow ($Res = N$), which differ in the field of view ($120° \times 120°$ and $75° \times 65°$, respectively) and depth resolution (512×512 and 640×576 pixels, respectively). Both configurations will be explored in the next experiments.

By mixing all attributes, 24 videos have been acquired by the Azure Kinect sensor and processed by the Azure Body Tracking SDK (v 1.0.1) to obtain 24 skeletons framing a single user,

Figure 2.2: Comparison of lighting conditions: (a) $E_v = 1750$ lux; (b) $E_v = 10$ lux.

represented by 32 joints, whose index mapping is in [100]. All videos have a duration of 60 s and a framerate of 15 fps. In the next lines, skeletons will be named as $Sk(Res, Occl, Ev, d)$. For instance, $Sk(N, w/o, 1750, 2)$, which refers to a narrow depth resolution (640×576 pixels), without occlusions, high illuminance E_v, and a user-camera distance of 2 m, is shown in Fig. 2.3.

Processing Procedures

All the acquisitions produce skeletons Sk of 32 joints, whose 3D position is $J[j,t] = (x_1[j,t], x_2[j,t], x_3[j,t])$, where $j = 0, \ldots, 31$ is the joint index, and $t = 1, \ldots, T$ is the time-dependent sample index. Here, the reference system (x_1, x_2, x_3) is aligned to the camera coordinates (x, y, z)[100], whereas $T = 900$ resulting from 60-s-long acquisition at 15 fps. As shown in the previous sections, the proposed experiments aim at assessing the measurement uncertainty. In all the acquired videos, the user stands still, spreading his arms and keeping his feet together. He holds this pose while the camera is grabbing for 60s. As a result, joint positions are collected in 3D, leading to the scatter plot of Fig. 2.4, which shows all the joints, accumulated over time, from $Sk(N, w/o, 1750, 1)$.

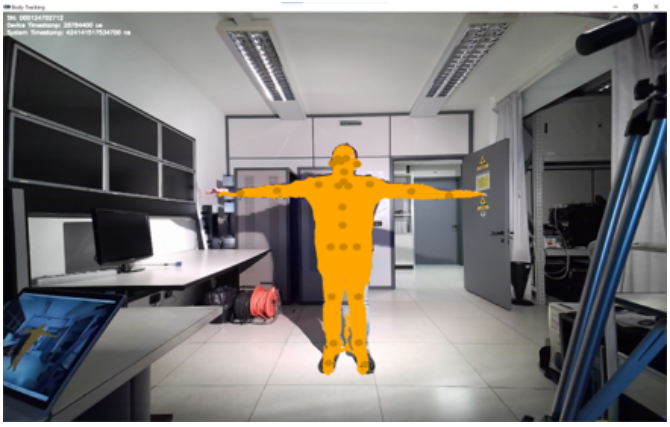

Figure 2.3: Result of body tracking: $Sk(N, w/o, 1750, 2)$. The orange points represent the estimated positions of the skeletal joints in the 2D image plane.

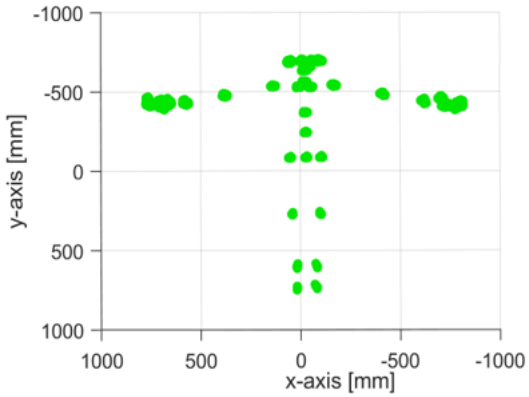

Figure 2.4: Skeletal joints in three dimensions from $Sk(N, w/o, 1750, 1)$.

Despite the user's effort to keep his pose, the body slightly fluctuates (quasi-static conditions). This is much more evident for the most peripherical parts, i.e. the hands. Fig. 2.4 proves the fluctuation of the left hand and torso joints, whose 3D coordinates suffer from (i) high-frequency oscillations, due to measurement noise and processing errors, and (ii) a low-frequency bias due to unavoidable body fluctuations. This investigation targets the evaluation of the high-frequency contribution, which produces the final uncertainty of the whole body tracking module. Further bias contributions, i.e. at low frequency, must be neglected through proper statistic evaluations.

Within these lines, quasi-static acquisitions are solved by computing the average value of the Euclidean distances of each joint coordinate from a corresponding centroid. This cen-

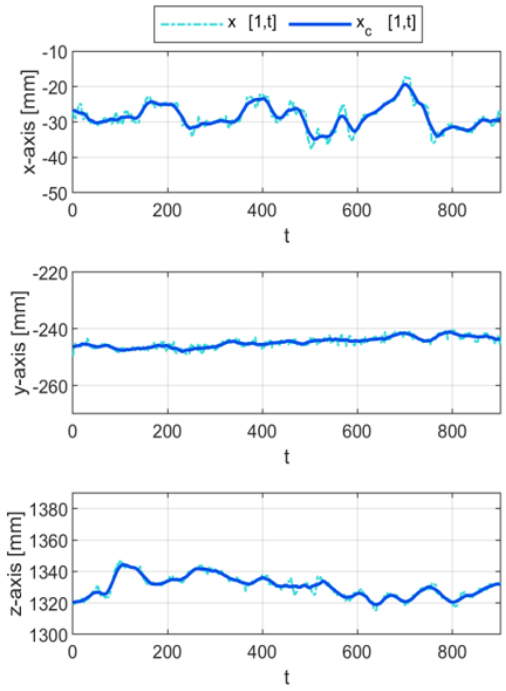

Figure 2.5: Comparison of input positions of the torso joint (dashed cyan line) and the corresponding centroid position (solid blue line) over time. Data are from $Sk(N, w/o, 1750, 1)$.

troid has a position $C[j,t] = (x_{c,1}[j,t], x_{c,2}[j,t], x_{c,3}[j,t])$, where:

$$X_{c,i}[j,t] = \frac{1}{2N+1} \sum_{p=t-N}^{t+N} x_i[j,p], \qquad i = 1,2,3 \tag{2.1}$$

At the steady-state, i.e. after N samples, this information is the result of an unweighted moving average, computed over a window of $2N+1$ samples, centered around the t-th sample of interest. It is worth noticing that the moving average is also computed at the boundaries of the input vectors $J[j,t]$, namely at $t < N+1$ and $t > TN$. In these cases, the window length is limited accordingly with the existing entries of $J[j,t]$. The results of the moving average on the coordinates of the cluster centroid of the torso ($j = 1$, SPINE_NAVAL[100]) and left-hand ($j = 8$, HAND_LEFT[100]) joints are in Figs. 5 and 6. In all the experiments, N is set to 15, which corresponds to a window length of about 2 s at 15 fps.

As expected, the inspection of Figs. 2.5 and 2.6 reveal that body fluctuations affect the hands more than the torso joint, which remains static in its position. Anyway, this different

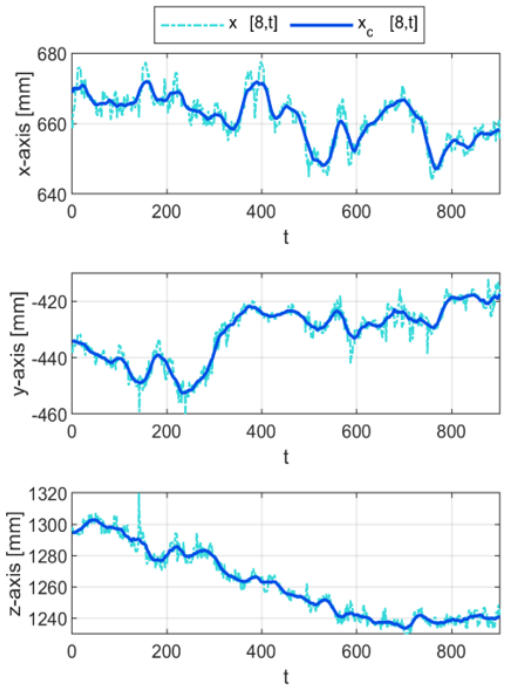

Figure 2.6: Comparison of input positions of the left-hand joint (dashed cyan line) and the corresponding centroid position (solid blue line) over time. Data are from $Sk(N, w/o, 1750, 1)$.

contribution of body fluctuations due to quasi-static acquisition will be ignored by computing the squared error ($SE[j,t]$) as follows:

$$SE[j,t] = \sum_{i=1}^{3} (x_i[j,t] - x_{c,i}[j,t])^2 \tag{2.2}$$

Accordingly, the Euclidean distance can be finally computed from SE[j,t] and then averaged over the time samples, returning the Mean Distance Error ($MDE[j]$) of the j-th joint:

$$MDE[j] = \frac{1}{T} \sum_{t=1}^{T} \sqrt{SE[j,t]} \tag{2.3}$$

An example of $MDE[j]$ from $Sk(N, w/o, 1750, 1)$ is shown in Fig. 2.7. In this case, it is possible to notice that the highest MDEs are those describing the hands, namely the hand centers ($j = 8$, HAND_LEFT[100] and $j = 15$, HAND_RIGHT[100]), the hand thumbs ($j = 10$, THUMB_LEFT[100] and $j = 17$, THUMB_RIGHT[100]), and the hand tips

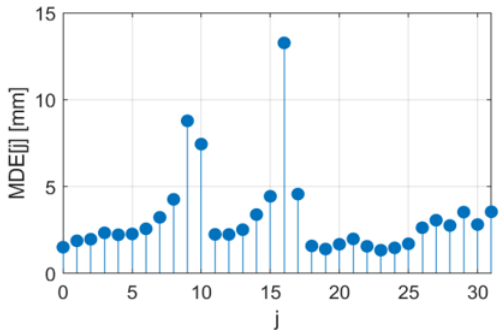

Figure 2.7: Mean Distance Error of the 32 joints[100] computed by the Azure Kinect and its body-tracking library from $Sk(N, w/o, 1750, 1)$.

($j = 9$, HANDTIP_LEFT[100] and $j = 16$, HANDTIP_RIGHT[100]). In these experiments, the hand tips and thumbs are not of interest since their extraction is typically needed for gesture recognition. The analysis of their reliability is out of the scope of this work, which focuses on people segmentation for real-time and safe control of cobots. For this reason, although the parametric analysis regards all the joints of the skeleton produced by the Azure Kinect, the next section will focus only on the MDEs of four representative joints: the head ($j = 26$, HEAD[100]), the pelvis ($j = 0$, PELVIS[100]), the left hand ($j = 8$, HAND_LEFT[100]), and the right foot ($j = 25$, FOOT_RIGHT[100]). Without any loss of generality, the results obtained for left or right joints will be replicable also for the opposite body parts.

2.2.2 Performance analysis results

As stated previously, the acquired videos release information about the 3D position of the joints of the skeleton at each frame while the participant stands for 60 s. The 24 acquisitions have been analyzed considering the four significant joints: head, pelvis, hand left, and foot right.

As a first step, 2.1 reports the values of the *MDE* of the four considered joints computed without any occlusions, varying the depth resolution of the camera (*Res*), the ambient light, defined through the E_v value, and the subject-camera distance d. As a first result of the analysis of 2.1, the *MDE* values grow as the distance increases. For better understanding, Fig. 8 shows the *MDEs* of the head, pelvis, left hand, and right foot versus d. This behavior is expected since the longer the distance, the lower the resolution of the depth estimation.

As expected, 2.1 confirms that, among the four joints, the left hand is estimated with the highest *MDE* in all cases, regardless of the operating conditions of the tests. This result is quantitatively proven by the average values of the MDEs of the head, pelvis, hand left, and

Table 2.1: *MDE* of the Head, Pelvis, Left hand, and Right foot by changing the input conditions. In all cases, experiments are run without occlusions. Entries are in millimeters.

Input Conditions	Head	Pelvis	Hand left	Foot right
$Sk(N, w/o, 10, 1)$	2.49	1.71	4.53	1.83
$Sk(N, w/o, 10, 2)$	2.35	1.70	6.66	1.54
$Sk(N, w/o, 10, 3)$	2.94	1.77	8.83	2.66
$Sk(W, w/o, 10, 1)$	2.67	2.04	10.46	1.15
$Sk(W, w/o, 10, 2)$	3.03	2.86	9.83	2.82
$Sk(W, w/o, 10, 3)$	6.92	3.90	20.97	7.73
$Sk(N, w/o, 1750, 1)$	2.63	1.51	4.27	1.71
$Sk(N, w/o, 1750, 2)$	2.66	1.51	8.82	1.82
$Sk(N, w/o, 1750, 3)$	4.74	2.42	17.63	3.57
$Sk(W, w/o, 1750, 1)$	3.21	2.73	12.99	6.67
$Sk(W, w/o, 1750, 2)$	7.70	4.84	21.51	10.65
$Sk(W, w/o, 1750, 3)$	16.63	7.08	35.84	9.38

foot right, which are equal to 4.83, 2.84, 13.53, and 4.29 mm, respectively. This result is also highlighted in Fig. 2.8, where the left-hand joint displays the highest *MDEs* at any user-camera distance.

Moreover, the analysis of 2.1 shows that setting the depth resolution of the camera to wide ($Res = W$) increases the uncertainty of the one achievable from the videos with $Res = N$. On average, the MDEs of the head, pelvis, hand left, and foot right are 2.06, 2.17, 2.28, and 2.96 times higher when *Res* is set to wide (W) instead of narrow (N). 2.1 also proves that using a light source to increase the ambient illuminance globally increases the *MDE*. Although *MDE* values obtained with different illuminations are comparable at short distances, this behavior is enhanced as the user-camera distance increases. Quantitatively, the *MDEs* of the head, pelvis, hand left, and foot right with $E_v = 1750$ lux are on average 1.66, 1.33, 1.57, and 2.37 times higher than setting $E_v = 10$ lux, respectively. This is due to the kind of illumination, which is directly pointed at the user. As a consequence of direct illumination, the corresponding depth maps have more noise contributions, thus producing an increase in the *MDE* values. In contrast, low but diffused light limits the input noise on the depth maps, thus returning lower uncertainty in joint estimation.

The parametric analysis is also meant to give information about how partial occlusions of the user can alter the *MDE* values. This aspect is very important as occlusions are typical of industrial workspaces. A worker doing some manufacturing tasks, such as part assembly, can be occluded to the camera by lots of volumes, such as a bulky instrument, a conveyor belt or a desk, the manufactured good, or the cobot itself. 2.2 shows the *MDE* of three reference

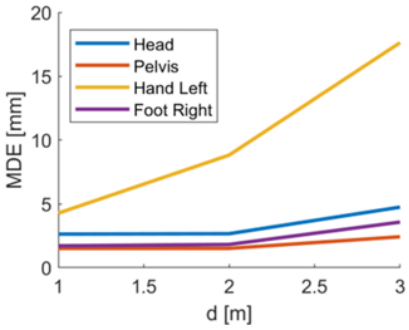

Figure 2.8: Mean Distance Error of the four joints of interest as a function of the user-camera distance d. Data are from $Sk(N, w/o, 1750, d)$.

joints (head, pelvis, and hand left) grabbed from the skeleton acquired occluding the lower body half. The joint of the right foot is not considered in this analysis, since it is occluded during the acquisitions.

All the outcomes obtained from 2.1 are still valid in the case of occlusions depicted in 2.2. In summary:

- The *MDEs* increase as the user-camera distance grows. It is valid under all the working conditions of the proposed setup;
- $Res = W$ increases the *MDE* of the three joints of interest, which is about tripled of the case with $Res = N$;
- $E_v = 1750$ lux, obtained with direct illumination, in general downs the performance of the body tracking, with an increase of the *MDE* which is on average 1.44 times higher than the corresponding obtained for $E_v = 10$ lux. However, in the case of $Sk(N, w/, E_v, 1)$ and $Sk(W, w/, E_v, 3)$, low-light ($E_v = 10$ lux) *MDEs* are comparable (or even higher) to the corresponding under direct illumination ($E_v = 1750$ lux);
- The head and the pelvis outperform the left-hand joint, which shows the worst *MDEs* under all the testing conditions.

  The comparison of 2.1 and 2.2 points out that the joints show higher values of uncertainty when the skeleton is partially occluded. From a quantitative point of view, it is possible to estimate this increase of uncertainty by computing, for each joint, the average of the *MDEs* of all the acquisitions made with or without occlusion. The result of this analysis demonstrates that occlusions increase the *MDEs* of the head, pelvis, and left-hand joints by 75.37%, 49.66%, and 47.86%, respectively. The *MDE* of the pelvis, and thus its estimation uncertainty, has the highest increase. It is due to the greater

Table 2.2: MDE of the Head, Pelvis, and Left hand by changing the input conditions. Foot estimation is not applicable as the joint is occluded. In all cases, experiments are run with occlusions. Entries are in millimeters.

Input Conditions	Head	Pelvis	Hand left
$Sk(N, w/, 10, 1)$	1.82	1.36	8.08
$Sk(N, w/, 10, 2)$	2.24	1.71	8.43
$Sk(N, w/, 10, 3)$	4.93	2.03	10.01
$Sk(W, w/, 10, 1)$	2.24	1.79	5.21
$Sk(W, w/, 10, 2)$	4.59	2.88	13.25
$Sk(W, w/, 10, 3)$	30.23	15.65	67.15
$Sk(N, w/, 1750, 1)$	2.20	1.30	4.83
$Sk(N, w/, 1750, 2)$	2.86	1.89	6.53
$Sk(N, w/, 1750, 3)$	6.77	4.20	19.96
$Sk(W, w/, 1750, 1)$	3.67	2.75	9.52
$Sk(W, w/, 1750, 2)$	11.07	6.82	33.22
$Sk(W, w/, 1750, 3)$	29.04	8.61	53.85

complexity of the body tracking module in making inferences on the data available close to the pelvis joint, which are fewer due to the occlusion. Anyway, the head and left-hand joint, which are far from the occlusion, show a significant increase in their estimation uncertainty. This means that all joints are estimated with greater uncertainty regardless of where the occlusion is. The position of the occlusion only affects the entity of the increase of uncertainty. In any case, considering all the intrinsic and extrinsic parameters considered, the MDE of the joints considered in this parametric analysis oscillates from a minimum of about 1 mm to a maximum of about 53 mm, with an average value of 8 mm, and a standard deviation of 6 mm.

2.2.3 Discussion

In Section 2.2, a parametric analysis of measurement uncertainty in body tracking has been proposed. Specifically, the performance of the Microsoft Azure Kinect in extracting skeletal joints has been investigated by changing both intrinsic and extrinsic conditions, namely the camera resolution, the ambient illumination, the user-camera distance, and adding occlusions to the user sight. The results of the analysis prove that (i) the estimation of the hand joints always suffers from the highest uncertainty, (ii) the skeletons acquired with wide depth resolution always have higher uncertainty than those with narrow depth resolution, and (iii) this uncertainty grows as the user-camera distance increases. Moreover, direct illumination degrades the depth maps and, thus,

the accuracy of the skeletal joints. Finally, the presence of occlusions increases the uncertainty of all the skeletons, also for joints far from the occlusion. The knowledge of the uncertainty of skeleton extraction of body tracking as a function of the working conditions will be of fundamental importance to improve the safety of real-time control of cobots cooperating with humans, as well as for a better understanding of the movements of patients in elderly facilities, aiming to avoid the risk of falls. Future works will focus on the analysis of further parameters, both intrinsic and extrinsic, such as the brilliance of the image, the eventual presence of multiple users, and the pose of the user's coronal plane relative to the camera.

2.3 Microsoft Azure Kinect Calibration

This work compares different calibration methodologies and suggests a guideline of the best methods to properly calibrate multiple Azure Kinect cameras, according to the data that must be processed and the measures needed. The proposed methodologies all start by analyzing a 2D target, i.e. a chessboard. This target is detected and processed in both RGB and IR images to estimate its corners. In a 3D approach, these points are projected in the 3D space, taking advantage of ToF principles. The chessboard becomes a "2.5D pattern" [101], as its planar features (corners) are directly computed from the depth map, using the intrinsic functionalities of the ToF RGB-D camera [102].

The main contributions of this work are:

- A two-camera system composed of Azure Kinects has been considered and the specific physical characteristics of these sensors have been studied to devise different calibration methodologies.
- Four different methodologies based on the data coming from color cameras and infrared cameras with or without the associated depth information have been compared in two real scenarios (dense point clouds of real objects for measures analysis, and people skeletal joints extracted from SDK Body tracking algorithm).
- A careful analysis of results provides a guideline for the best calibration techniques according to the element to be calibrated, i.e., point clouds with color or infrared resolutions and skeletal joints.

The Section is structured as follows. In Section 2.3.1, the proposed calibration methodologies are outlined. Section 2.3.2 defines the experimental setup in which the described calibration techniques are used. Section 2.3.3 analyzes the reliability of the proposed calibration methodologies applied to point clouds and skeletal joints. Finally, Sections 2.3.4 draws a final discussion.

2.3.1 Calibration Methodology

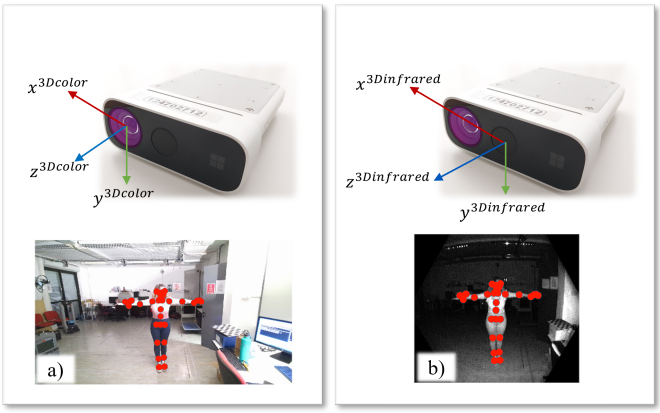

Figure 2.9: Representation of the internal Kinect sensors that produce: a) color images with a resolution of 3840x2160 and b) IR images with resolution 640x576. The origin of the coordinate systems is placed at the focal point of each sensor [14]. The skeletal joints extracted by the Body Tracking SDK are superimposed on the images.

The two-camera calibration methodologies discussed in this work consider Microsoft Azure Kinect. Such kind of device consists of an RGB camera and an infrared (IR) camera, with the latter providing depth information implementing ToF principles. Therefore, the Kinects output RGB images, IR images, and depth maps. The Azure Kinect is equipped with two software development kits for the management of all data that can be recorded by the internal cameras: the general Azure Kinect SDK and the Azure Kinect Body Tracking SDK [103]. In particular, the data provided by the Azure Kinect sensor can be represented in two different geometries: the geometry of the color camera or the geometry of the infrared camera. The term geometry, related to the RGB or IR sensors of the Azure Kinect camera, refers to a set of sensor properties, including the coordinate system, its resolution, and all intrinsic transformations. A set of routines in the general SDK allows the transformation of images or depth maps from one geometry to another. The Body Tracking SDK implements Deep Learning and Convolutional Neural Networks algorithms [104] to extract all the possible information for people segmentation, people tracking, and skeletal joint extraction. In Figure 2.9 the two cameras that produce RGB and IR images are shown. In the example images, the skeletal joints extracted by the body tracking SDK are superimposed.

Following the procedure in Figure 2.10, the depth maps acquired by the IR camera can be converted into point clouds by using the SDK functions [105]. Starting from the IR image, depth data can be converted directly, in the geometry of the infrared camera obtaining the $P^{infrared}$ point cloud. Otherwise, the point cloud can be represented in the geometry of the color camera. In this case, the SDK provides a transformation $\mathcal{T}_{intr}$, that uses also intrinsic

camera parameters, to convert the depth map into a point cloud with color geometry. The result of this step is a P^{color} point cloud.

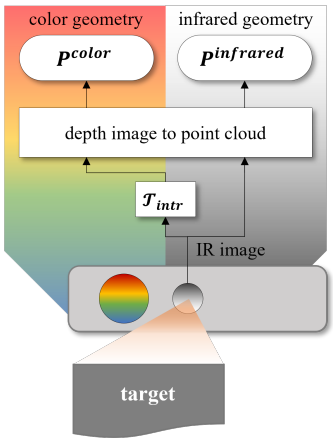

Figure 2.10: Schematic representation of the point cloud realization with color and infrared geometries, using the Azure Kinect SDK.

The proposed techniques consider a two-camera setup made of a Reference and a Template camera. Nevertheless, the system can be suited for multiple Azure Kinect calibrations. Without any loss of generality, for multiple K cameras, the calibration has to be repeated $(K-1)$ times to align the outputs of $(K-1)$ Template cameras onto the Reference one. All calibration methodologies use a 2D target that will be captured simultaneously by the RGB and IR sensors of each of the two cameras. This target is a 2D chessboard made up of m rows and n columns of black and white squares with side lengths of S. The structured geometry of the chessboard guarantees robustness and accuracy for the corner detection and processing algorithms [106]. F frames of the chessboard are acquired by moving the target to different positions and orientations in the FOVs of both cameras.

The transformation matrix that relates the two coordinate systems of the Reference and Template Cameras is defined in the following Eq. 2.4.

$$T = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \tag{2.4}$$

where $R \in \mathbb{R}^{3\times3}$ represents the rotation matrix and $t \in \mathbb{R}^{3\times1}$ the translation vector. The whole T matrix is estimated by evaluating the correspondences among the corners of the chessboard observed by the two cameras.

In Figure 2.11 the proposed calibration methodologies are graphically summarized. In this figure, $\mathcal{T}_{intr}$ and $\mathcal{T}_{intr,Ref}$ correspond to the intrinsic transformations that convert the data from the geometry of the infrared camera to the one of the color camera. On the other hand,

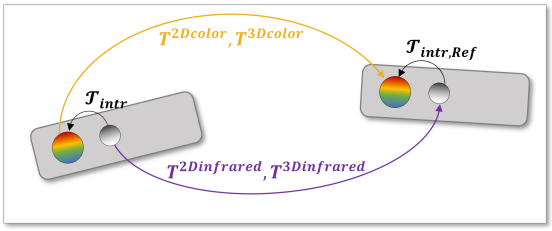

Figure 2.11: Conceptual meaning of the application of the transformation matrices obtained from the proposed calibration techniques. The transformation process using matrices with color geometry is marked in yellow, whereas the transformation process using matrices with infrared geometry is marked in purple.

four calibration matrices can be obtained comparing different camera sensors, namely RGB or IR sensors, and calibration procedures, namely 2D and 3D calibrations. Specifically, when chessboard corners are processed directly to estimate the transformation matrix, the calibration works with mere 2D image coordinates. Therefore, it ends with the following:

- $T^{2Dcolor}$ if the chessboard corners are extracted from RGB images, i.e. with the geometry of the color cameras;
- $T^{2Dinfrared}$ if the chessboard corners are detected in the IR images, i.e. with the geometry of the infrared cameras.

However, since the Azure Kinect computes depth maps of the environment, the same chessboard corners can be projected in 3D coordinates in the reference system of each camera. In this case, two further procedures working with 3D points can be defined to produce:

- $T^{3Dcolor}$ if the chessboard corners are taken from RGB images and then projected in the 3D space, using the geometry of the color camera;
- $T^{3Dinfrared}$ if the chessboard corners are extracted from the IR images and then projected in 3D, using the geometry of the infrared camera.

In the following subsections, the methodologies used to generate the 2D and 3D calibration matrices will be explained in detail.

2D calibration procedures

A schematic pipeline of the 2D calibration methodology is shown in Figure 2.12.

Let $(I^{color}_{Ref}, I^{color})$ and $(I^{infrared}_{Ref}, I^{infrared})$ generically represent the images couples from the color and infrared sensors grabbed by the Reference and Template Azure Kinect cameras, respectively. The images are input to a corner detection algorithm [107] that estimates the

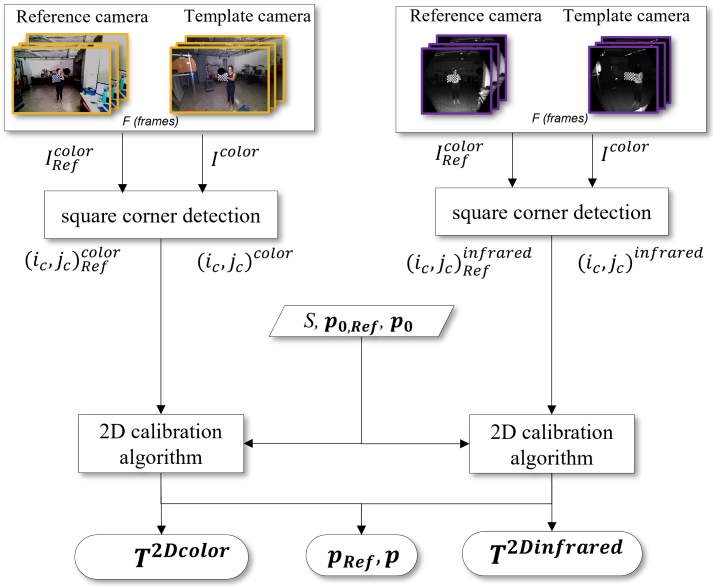

Figure 2.12: 2D calibration flow chart for the creation of the transformation matrices $T^{2Dcolor}$ and $T^{2Dinfreared}$. The frames with color resolution are marked in yellow, while the frames with infrared resolution are marked in purple.

2D coordinates of the chessboard corners, namely $((i_c,j_c)_{Ref}^{color},(i_c,j_c)^{color})$ and $((i_c,j_c)_{Ref}^{infrared}$, $(i_c,j_c)^{infrared})$, with $c = \{1,2,\ldots,(m-1)(n-1)\}$. The corners coordinates from each of the F frames acquired during calibration, together with the square size S and the trial sets of intrinsic parameters for both cameras ($p_{0,Ref}$ and p_0), feed the calibration algorithm, which finally estimates the intrinsic and extrinsic parameters of the camera [108]. The estimated intrinsic parameters include the focal length, the optical center, the skew, the Radial Distortion and the Tangential Distortion. This outcome refines the initial set of intrinsic parameters of both cameras. On the other hand, the extrinsic parameters define a rigid transformation to roto-translate the reference system of the Template camera into the reference system of the corresponding sensor of the Reference camera, as described in Eq. 2.4. As depicted in Figure 2.12, the outputs of this 2D calibration procedure are p_{Ref}, p, and the matrices $T^{2Dcolor}$ or $T^{2Dinfrared}$, depending on which sensor acquires the chessboard.

3D CALIBRATION PROCEDURES

A schematic pipeline of the 3D calibration procedures is shown in Figure 2.13. Even in this case, the first step involves detecting the corner coordinates of the chessboard in the image reference system. The same methodology explained for 2D calibration returns again, for each frame acquired during calibration by the Reference and Template cameras, $(i_c,j_c)_{Ref}^{color}$

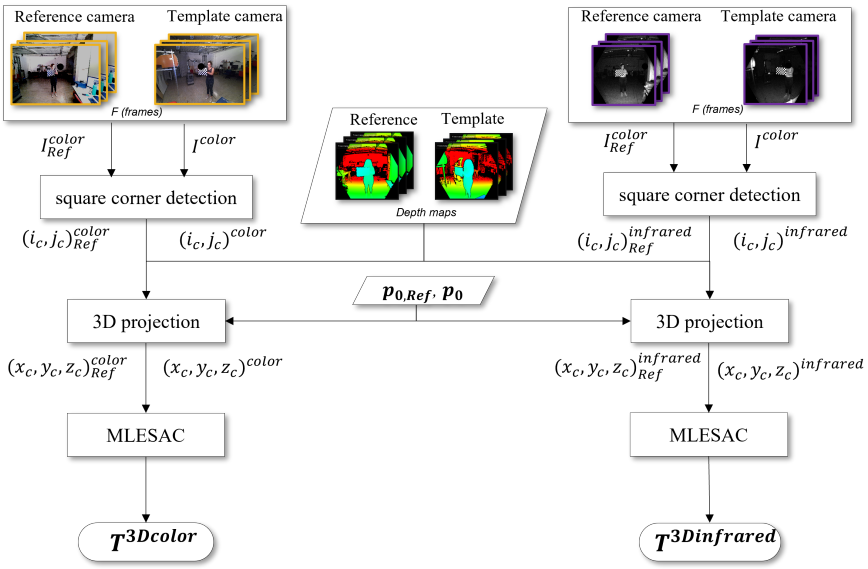

Figure 2.13: 3D calibration flow chart for the creation of the transformation matrices $T^{3Dcolor}$ and $T^{3Dinfrared}$. The frames with color resolution are marked in yellow, while the frames with infrared resolution are marked in purple.

and $(i_c, j_c)^{color}$, or $(i_c, j_c)_{Ref}^{infrared}$ and $(i_c, j_c)^{infrared}$, depending on the considered sensor of the Azure Kinect. The 3D projection procedure converts the generic pixel coordinates (i, j) in world coordinates (x, y, z), defined in the corresponding reference system of the sensor. This transformation is performed at SDK level knowing the intrinsic parameters of both cameras $p_{0,Ref}$ and p_0 (factory settings), and the corresponding depth maps. In particular, the latter is the result of the ToF measurement, performed by the IR sensor in its own geometry. The 3D projection generates points in 3D coordinates, namely $(x_c, y_c, z_c)_{Ref}^{color}$ and $(x_c, y_c, z_c)^{color}$, or $(x_c, y_c, z_c)_{Ref}^{infrared}$ and $(x_c, y_c, z_c)^{infrared}$. These points are the 3D positions of the chessboard corner, referred to in the geometries of the color and infrared cameras, respectively.

The 3D coordinates feed into the Maximum Likelihood Estimation Sample Consensus (MLESAC) estimator [109], which is a generalization of the Random Sample Consensus (RANSAC) algorithm [110]. RANSAC is an iterative method used for coordinate sets. In the first iteration, the algorithm selects random samples from the initial correspondences and finds the transformation matrix relative to the selected samples. This step is repeated iteratively, and the transformation returning the maximum number of matches, named inliers, is considered the optimal transformation matrix. All the other non-matched correspondences are considered outliers. One of the problems of the RANSAC algorithm is the setting of the threshold for correct matches. The MLESAC algorithm combines RANSAC with the Maximum Likelihood Estimation (MLE) method to find inliers. The goal of MLE is to find

the optimal way to fit a distribution to the data [111]. By applying the MLE to the initial correspondences of each iteration, the noise dips are eliminated, thus excluding from the iterations those outliers that would be included if the samples were selected randomly. Hence, the estimate of the matching points provided by the MLESAC algorithm can be more precise and closer to the true solution, even requiring a reduced number of iterations to reach the optimal solution. In the specific case of interest, the MLESAC algorithm estimates the 3D transformation between the set of 3D points of the chessboard, collected from all the acquired frames. As a result, the calibration procedure determines the final calibration matrices $T^{3Dcolor}$ and $T^{3Dinfrared}$, as in Eq. 2.4, depending on the sensor that acquires the chessboard images.

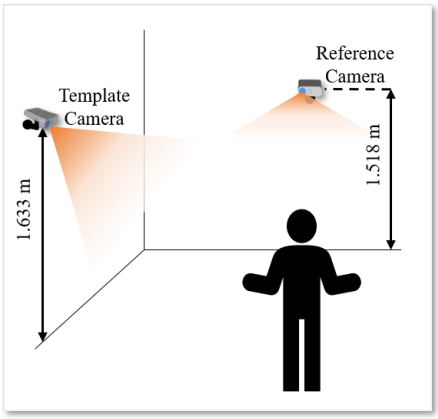

Figure 2.14: Depiction of the experimental setup considering two Azure Kinects.

2.3.2 Experimental setup

The real-case scenario in which experiments have been performed is shown in Figure 2.14. $K = 2$ Azure Kinect sensors have been placed to have an extended overlapping area and the vision of the full body of people in the scene. The calibration methodologies have been evaluated in two different cases: (i) to assess the ability to reconstruct a target object by the combination of point clouds, and (ii) to estimate the robustness of the people skeleton alignment. Figure 2.15 shows the considered workspace grabbed by both Kinect sensors. The images show that the RGB camera has a field of view wider than that of the IR camera. In addition, the RGB camera has been set with a resolution of 3840x2160, while the IR camera has been set with a resolution of 640x576, to produce depth maps with narrow FOV [14].

In particular, two experiments have been performed:

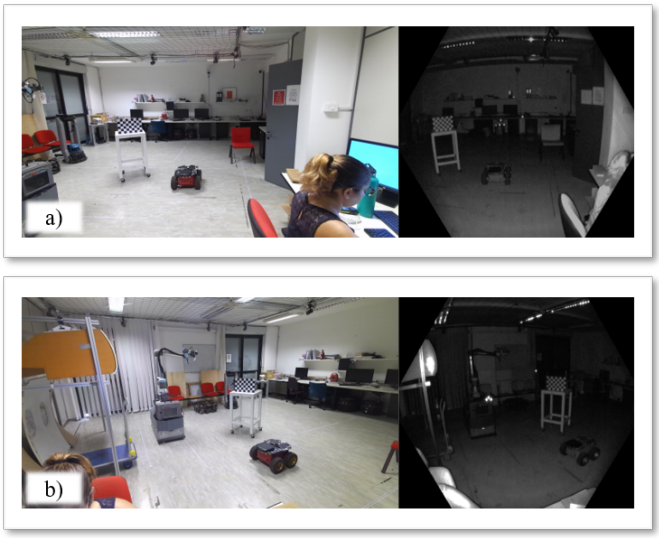

Figure 2.15: Views of the two Azure Kinects used during the experimentation, where a) represents the Reference Camera, and b) represents the Template Camera. More specifically, the images on the left in both a) and b) show the frames grabbed from the RGB sensors, while the images on the right show the frames grabbed from the IR sensors.

- To state the capability in aligning point clouds, two analyses have been proposed: in a static scenario, a still object is placed in the two camera FOVs; in a dynamic scenario, a moving target is framed simultaneously by the two cameras. After the calibration phase, the point clouds in both infrared and color geometries, grabbed by the Template camera, are transferred into the coordinate system of the Reference.

- A subject stands still with open arms in front of the two cameras, and the corresponding skeletal joints are extracted from the Azure Kinect Body Tracking library. The skeleton from the Template camera is transferred into the coordinate system of the Reference. In this case, ten consecutive frames have been collected to calculate the average position of each joint to reduce intrinsic errors [14] and average involuntary movements of the subject.

To have a clear visualization and avoid light reflections or color alterations of the chessboard due to the natural light or backlight effects, the workspace has been artificially illuminated using a light projector placed behind the Azure Kinects.

In the proposed configuration, the selected chessboard has $m = 6$ rows and $n = 9$ columns of black and white squares of side length $S = 45\,mm$. $F = 200$ frames of the chessboard have been acquired. Figure 2.16 shows some examples of the RGB and IR images acquired during the experiments.

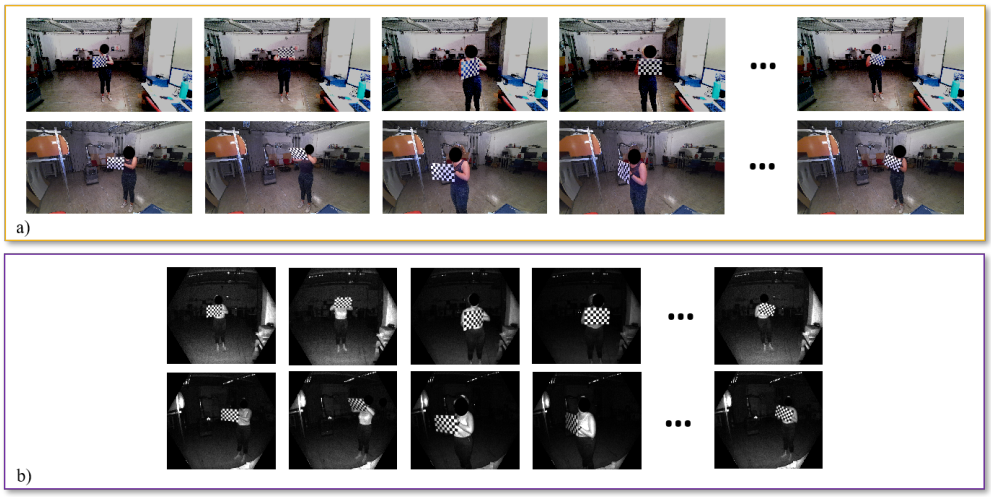

Figure 2.16: a) RGB and b) IR image samples of the chessboard. Several positions and orientations have been considered to optimize the results of the calibration.

2.3.3 Calibration analysis

The calibration methodologies have been evaluated considering the Root Mean Square Error (RMSE), defined as follows:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{J}(d_j - \hat{d}_j)^2}{J}} \tag{2.5}$$

where:

- d_j, $\hat{d}_j$ in the point cloud experiment are the 3D coordinates of points in correspondence taken from the Reference point cloud and the Template one after the application of estimated transformation. J is the total number of points in correspondence.
- d_j, $\hat{d}_j$ in the skeleton experiment are homologous 3D joint coordinates in the same reference system. Here, $J = 32$ is the total number of the joints.

Low RMSE values indicate that points (or skeletal joints) are correctly transformed in the same reference system. The value of the RMSE has been calculated for each pair of point clouds and skeletons. Subsequently, the average of all RMSEs ($\overline{RMSE}$) and their standard deviation (σ_{RMSE}) were calculated to assess the proposed calibration techniques.

Point cloud experiment

Table 2.3 shows the quantitative results of the proposed calibration methodologies in the point cloud experiment considering a static target, i.e. a robot. Overall, 38 pairs of point clouds have been considered. RMSE values are computed comparing pairs of point clouds in the geometry of the color camera (P^{color} column) or in the geometry of the infrared camera ($P^{infrared}$ column). Then, the mean of such values is computed, along with the standard deviation.

Table 2.3: Mean and standard deviation of RMSEs calculated between aligned and reference point clouds with static target, where calibration techniques have been applied [mm]. The best results of $\overline{RMSE}$ are underlined.

	P^{color}		$P^{infrared}$	
	$\overline{RMSE}$	σ_{RMSE}	$\overline{RMSE}$	σ_{RMSE}
$T^{2Dcolor}$	24.427	0.918	45.485	0.804
$T^{2Dinfrared}$	37.283	0.955	20.162	0.592
$T^{3Dcolor}$	<u>21.426</u>	0.608	36.833	0.735
$T^{3Dinfrared}$	33.194	0.758	<u>9.872</u>	0.268

The mean of the RMSE values obtained in the alignment of the point clouds P^{color} demonstrate that the best calibration matrix is $T^{3Dcolor}$, which produces an $\overline{RMSE}$ value equal to 21.426 mm. The worst result is obtained with the $T^{2Dinfrared}$ matrix which provides an $\overline{RMSE}$ value of 37.283 mm. Even σ_{RMSE} values confirm this analysis, since the variability of the RMSEs does not exceed 1 mm in any case. In addition, the calibration matrices that produce the lowest $\overline{RMSE}$s, also produce the lowest σ_{RMSE}.

Figure 2.17 provides a qualitative evaluation of the reconstructed point clouds in color geometry P^{color} obtained after the above calibrations. The images show the reconstruction of a static target, at 3.13 m from the Reference camera, resulting from the alignment of two point clouds considering the transformation matrices $T^{3Dcolor}$ and $T^{2Dinfrared}$. In the first case, the shape of the target is clearly visible, and its appearance is coherent and consistent with its expected shape. In the latter case, which underperforms the other calibrations, the target appears duplicated, and its 3D dense reconstruction fails.

In Table 2.3, the lowest value of $\overline{RMSE}$ calculated for the alignment of $P^{infrared}$ point clouds in infrared geometry is 9.872 mm, obtained by $T^{3Dinfrared}$, while the worst is 45.485 mm, obtained by $T^{2Dcolor}$. Figure 2.18 shows the results of the alignment of the same static target of Figure 2.17, but modeled in the $P^{infrared}$ point clouds in infrared geometry. Alignments are made by applying the best and worst calibration methodologies in Table 2.3. Specifically, the $T^{3Dinfrared}$ calibration matrix produces a coherent reconstruction of the target, while the

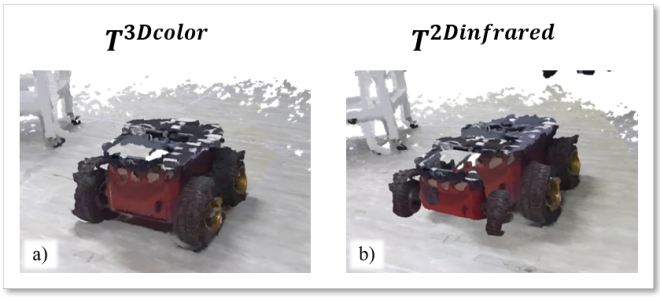

Figure 2.17: Visual representation of the a) best ($T^{3Dcolor}$) and b) worst ($T^{2Dinfrared}$) alignment of point cloud in color geometry P^{color}. The input point clouds are captured at the same timestamp from both the Azure Kinect cameras.

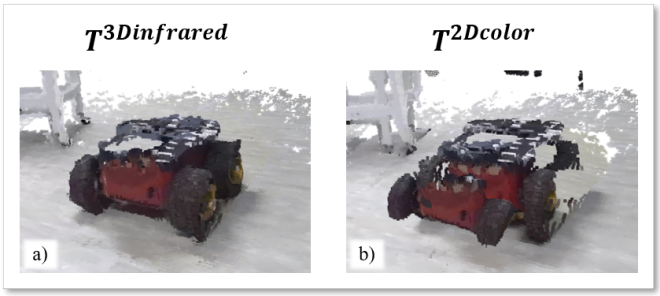

Figure 2.18: Visual representation of the a) best ($T^{3Dinfrared}$) and b) worst ($T^{2Dcolor}$) alignment of $P^{infrared}$ point clouds. he input point clouds are captured at the same timestamp from both the Azure Kinect cameras.

application of $T^{2Dcolor}$ returns an altered version of the target shape, which seems shrunk in the front while its silhouette is not complete.

A careful analysis of the quantitative results of Table 2.3 highlights that the experiments carried out considering the calibration matrices resulting from the 2D calibration method give the worst results than the 3D calibration ones. The reason lies in the fact that $T^{2Dcolor}$ and $T^{2Dinfrared}$ are generated from matches between 2D data, while 3D calibration $T^{3Dcolor}$ and $T^{3Dinfrared}$ consider matches between sets of 3D coordinates that contain more information with the introduction of depth data. This result is not straightforward, since the computation of the depth maps, which is the basis of 3D calibration procedures, can suffer from implicit errors. However, such negative contributions do not influence 3D approaches, which always outperform 2D ones.

On the other side, it is possible to notice that the P^{color} presents the lowest $\overline{RMSE}$ values when using the $T^{3Dcolor}$ calibration matrix, computed starting from the chessboard corners in RGB images. At the same time, the alignment of $P^{infrared}$ point clouds in infrared geometry has the lowest $\overline{RMSE}$ with the calibration made by matching corners from IR images. These results can be explained considering the process that the Kinect sensor uses to produce the

two point clouds in color or infrared geometries, as in Figure 2.10. The point clouds are always generated by the IR camera, but the transformation of the point cloud in the geometry of color camera requires an interpolation process that uses the intrinsic camera parameters. At the end of this process, the size of the point cloud greatly increases. In conclusion, the calibrations performed in the same space after the same transformations are those that perform better.

To better evaluate the proposed methodologies, the same calibration matrices have been applied to pairs of point clouds extracted from videos that frame a dynamic scene with a moving target. For this evaluation, 128 pairs of point clouds have been considered.

Table 2.4: Average and standard deviation of the RMSEs calculated between aligned and reference point clouds with dynamic target, where the calibration techniques have been applied [mm]. The best results of $\overline{RMSE}$ are underlined.

	P^{color}		$P^{infrared}$	
	$\overline{RMSE}$	σ_{RMSE}	$\overline{RMSE}$	σ_{RMSE}
$T^{2Dcolor}$	25.340	0.666	36.683	2.383
$T^{2Dinfrared}$	39.446	2.299	13.046	0.765
$T^{3Dcolor}$	<u>20.868</u>	1.233	33.122	2.198
$T^{3Dinfrared}$	34.039	2.024	<u>7.429</u>	0.606

Table 2.4 shows the avarage and the standard deviation of the RMSEs obtained in comparing each couple of point clouds, in both color and infrared geometries. The results are highly comparable with the one observed in Table 2.3. The standard deviations show slightly higher values, as attributed to the presence of the moving target. Nevertheless, in all cases, σ_{RMSE} values do not exceed 2.4 mm.

Skeleton experiment

The $\overline{RMSE}$ values resulting from the comparison between the skeletal joints of the Template camera, aligned in the reference system of the Reference one for all the proposed procedures are reported in Table 2.5, together with the corresponding σ_{RMSE} values. For such evaluation, 15 pairs of skeletal joints have been aligned. Each pair contains the average values of the skeletal joints grabbed from both Template camera and Reference camera, performed within 10 frames. Hence, 150 frames have been considered in total. Observing the table, it is clear that the best result is obtained using the calibration matrix $T^{3Dinfrared}$ with the lowest $\overline{RMSE}$ of 35.410 mm. The calibration performed using $T^{2Dcolor}$, instead, gives the highest $\overline{RMSE}$ value, equal to 124.602 mm. As expected, the results are in accordance with those obtained for the point cloud in infrared geometry, shown in Table 2.3, since the skeletal joints are also gener-

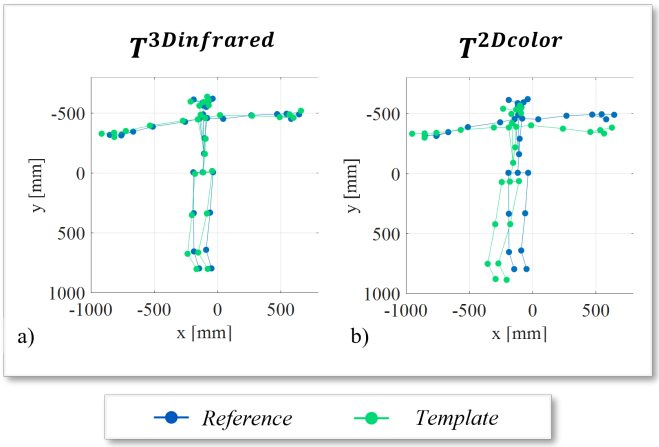

Figure 2.19: Graphic representation of the a) best ($T^{3Dinfrared}$) and b) ($T^{2Dcolor}$) worst skeleton alignments. In both graphs, the aligned template skeleton is in green, while the reference skeleton is in blue.

ated in the IR environment, using the same IR camera of $P^{infrared}$: the calibrations obtained in the same geometry produce a better overlap of the two skeletons.

Table 2.5: Average and standard deviation of the RMSEs calculated between aligned and reference skeletal joints, where calibration techniques have been applied [mm]. The best result of $\overline{RMSE}$ is underlined.

	Joints of the Skeleton	
	$\overline{RMSE}$	σ_{RMSE}
$T^{2Dcolor}$	124.602	1.349
$T^{2Dinfrared}$	44.256	4.640
$T^{3Dcolor}$	111.247	1.889
$T^{3Dinfrared}$	<u>35.410</u>	5.490

The graphs reported in Figure 2.19 allow a qualitative evaluation of the effects of the best and worst calibration procedures on skeleton alignment. In Figure 2.19 a), the results after the application of the $T^{3Dinfrared}$ matrix are shown, while in b) the skeletons are registered using the $T^{2Dcolor}$ matrix. The graphs confirm the results of the $\overline{RMSE}$ values. In Figure 2.19 a) the two skeletons are very close, while in Figure 2.19 b) some joints of the skeletons, especially those corresponding to the extreme joints of legs and arms, are very distant.

This result is very important if skeleton extraction is the target of a multi-camera setup. This goal is of increasing interest, since capturing humans from different points of view can lead to robust people tracking, even in case of camera occlusions and/or estimation errors. Furthermore, σ_{RMSE} values confirm that the skeletal joints alignment is repeatable over the

frames, as in all techniques they do not exceed 5.5 mm. Having a correct and continuous knowledge of where somebody is within a volume of interest is of critical importance to guarantee safety, for instance in human-robot collaboration, and even for action recognition tasks. In these scenarios, misalignment of the skeletons once referred to as a common coordinate system can lead to even huge and more dangerous errors. For this reason, performing a reliable camera calibration becomes mandatory.

2.3.4 Discussion

The camera calibration problem has been extensively addressed in the literature as the importance of having coherent data extracted from different sensors in a single reference system is widely recognized. However, with the availability of multi-modal sensors that provide different types of data, it is necessary to study calibration methods that take into account the specificity of the sensors and the type of data extracted. In this context, the presented work has filled the gap about the need for calibration methods specific to the Microsoft Azure Kinect cameras. Here, calibration methods have been developed starting from raw images from both the color and the infrared sensors. This choice guarantees a higher reliability in applying calibration to skeletons and point clouds, particularly with respect to [112]. Overall, the experiments have proven the efficiency of 3D-based techniques, which take advantage of the specifics of the Azure Kinect cameras. It must be noticed that the techniques here presented can be useful in calibrating a system composed of multiple Azure Kinects, as the alignment can be performed to indefinite pairs of point clouds and/or skeletons.

The main points of the proposed calibration methodologies are the following:

- In general, 3D procedures outperform 2D ones as depth information is added to the calibration. This is due to the effectiveness of depth estimation and intrinsic transformations used to project 2D image points in the 3D space.
- The alignment of point clouds in the geometry of the color camera has the lowest error value when using a calibration procedure working in 3D starting from RGB images, since both the point cloud in color geometry and the chessboard corner coordinates enable the calibration to follow the same interpolation procedures carried out by the general SDK functions.
- The alignment of the point clouds in the geometry of the infrared camera has the lowest error when the calibration works starting from IR images. Even in this case, the calibration performed in the same geometry of the point cloud produces the best result.

- The alignment of skeletons shows the best result while calibration is performed in 3D starting from IR images. It further confirms the previous statement.
- In all experiments, the standard deviations of the RMSE values state that the variability in error computations is always lower than the improvement in aligning both point clouds and skeletal joints.

The results are significant in systems with two or more cameras, mainly when low-cost sensors, such as Azure Kinects, can be efficiently used for several applications to have full 3D representations of targets and environments. For example, 3D characterization is helpful in many computer vision applications, such as 3D reconstruction, 3D localization, and 3D pose estimation. Building a proper 3D scene can allow a highly accurate assessment of a 3D map for pursuing, for instance, the reconstruction of an industrial object. Furthermore, estimating human 3D movements is required in various scenarios, which may need to detect specific activities performed by the framed subjects. To segment and recognize human movements, a properly calibrated camera system can provide a complete reconstruction of human posture, overcoming any occlusion that may limit the view of one of the cameras. Such calibration processes can be useful in surveillance, where it is crucial to know what a person is doing and where he/she is going. Also, a calibrated system that provides a complete set of 3D skeletal joints, or a dense point cloud, can easily represent a subject executing a specific task. Such data may widely facilitate segmentation and, thus, recognition of the actions needed for any assignment.

3

Video Data Acquisition for Human Monitoring

3.1 Introduction

As said in Chapter 1, the scientific community has found increasing interest in technological systems for the evaluation of the mobility performance of the elderly population. The reduced quantity of datasets for gait and balance analysis of elderly people is a serious issue in studying the link between cognitive impairment and motor dysfunction, particularly in people suffering from neurodegenerative diseases. The need for real and comprehensive datasets is also a very important topic in manufacturing domains, particularly regarding assembling tasks in production processes.

In this context, the following Chapter presents two datasets acquired in elderly facilities and manufacturing environments. More specifically, Section 3.2 defines the SPPB Dataset provided in [97], which contains skeletal information of people aged 60 years and older, while they perform well-established tests for stability assessment. Subjects have been observed and evaluated by clinical therapists while executing three motion tests, namely balance, sit-to-stand, and walking. The stability postural and gait control of each subject has been analyzed using a video-based system, made of three low-cost cameras, without the need for wearable and invasive sensors. On the other hand, Section 3.3 depicts the HA4M Dataset, introduced

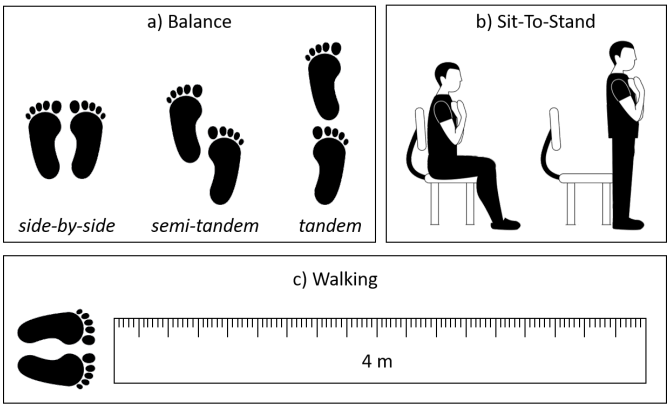

Figure 3.1: SPPB tests: a) Balance test, which is composed of three exercises where the patient has to stand with the feet in side-by-side, then in semi-tandem and finally tandem positions; b) Sit-To-Stand test, which consists in sitting and standing up 5 times while keeping the arms crossed on the chest; c) Walking test, which consists in the patient covering a path of 4 meters.

in [98], which represents a collection of multi-modal data relative to actions performed by different subjects building an Epicyclic Gear Train (EGT). Data were collected in a laboratory scenario using a Microsoft® Azure Kinect which integrates a Depth camera, an RGB camera, and InfraRed (IR) emitters. The information within both Datasets represents a good foundation to develop and test advanced action recognition and segmentation systems in several fields, including Computer Vision and Machine Learning.

3.2 Data Acquisition in elderly facilities: SPPB Dataset

It has been shown that there is a strict link between cognitive impairment and motor dysfunction such as deficits in gait and balance [113, 114]. Furthermore, functional assessment measure protocols can help to qualify the gait and posture of the patients. In this scenario, the Short Physical Performance Battery (SPPB) represents a well-established means to assess physical performance status and evaluate functional capabilities [115], to monitor and prevent the risk of falls. Such functional assessment measure is composed of three tests to assess lower body function, namely Balance Test (BT), Sit-To-Stand Test (STST) and Walking Test (WT), which instructions are represented in Fig. 3.1.

The risk of fall is qualitatively evaluated by expert clinical personnel with respect to the execution of the SPPB tests, in agreement with the medical protocols. Despite the high professional competence of the operators, the need for developing innovative technological systems is of great interest, since human-based assessment can be susceptible to drifts and biases. For this purpose, the need for datasets containing physical performance status information is be-

coming an issue of increasing interest, as the development of new technological systems that can support clinical personnel strictly depends on both the quality and quantity of available data.

In literature, various datasets related to the evaluation of the motion skills of elderly people are presented [116, 117, 118], yet none of them gives skeletal information specifically to the SPPB protocol. Most of the datasets outlined in literature provide information only regarding the static analysis of the patient, without releasing information about the dynamic aspect, which is fundamental in evaluating the risk of falls. Moreover, even when the patient's skeleton is analyzed, the dataset often concerns only a singular type of exercise, thus producing a non-heterogeneous amount of data.

This work provide a complete dataset of age, sex and skeletal information of people aged 60 years and older, while they perform all the three tests included in the SPPB protocol. A complete vision-based system, made of three low-cost cameras, has been developed for accurately measuring stability postural control, without the need for wearable and invasive sensors. The exercise videos, grabbed from two nursing institutes, have been normalized and synchronized to extract the most significant features from the skeletons, which carry information about balance, gait, and strength, to properly evaluate the risk of falls. Such features, along with sex, age, and the skeletal information of the patient itself, have been added to the dataset. The reliability of the dataset has been tested using the features extracted in the BT as input of a classifier [97]. Final results have proven a good estimation of the risk of fall of people under analysis.

3.2.1 Tests Definition

The proposed work aims to establish the risk of fall of elderly people and patients affected by neurodegenerative diseases, through the analysis of the tests included in the SPPB. Several patients, housed at the two nursing institutes of the study, have been selected for the postural and stability analysis. Each patient has been instructed to perform first the BT, then the STST, and finally the WT. For each test, a specialized therapist observes the patients and measures their time execution using a stopwatch. Such tests are then evaluated following an appropriate score system, shown in Table 3.1.

In the following, the SPPB tests are defined:

- **Balance Test:** The test of standing balance includes side-by-side, semi-tandem and tandem positions. The patient is instructed to maintain each position for 10 seconds, measured by a clinical therapist. If a patient fails to complete the test within ten seconds, the elapsed time is measured anyway.

- **Sit-To-Stand Test:** The STS test consists of sitting and rising from a chair placed against the wall for safety purposes. The patient is asked to fold her/his arms across

Table 3.1: Classification method for each test. Each exercise is assessed based on its duration.

Test	0	1	2	3	4
Balance	*side-by-side*	$0-9\,s$ *semi-tandem*	$0-2\,s$ *tandem*	$3-9\,s$ *tandem*	$10\,s$ *tandem*
Sit-To-Stand	*incapable*	$> 7.5\,s$	$7.4-5.4\,s$	$5.3-4.1\,s$	$< 4.1\,s$
Walking	*incapable*	$> 16.6\,s$	$16.6-13.7\,s$	$13.6-11.2\,s$	$< 11.2\,s$

her/his chest, and to stand up and sit down from the chair 5 times. A clinical therapist times the exercise starting from the initial sitting position to the final standing position.

- **Walking Test:** During the walking test, the patient is instructed to follow a path of 4 meters with no obstructions. A clinical therapist is in charge of timing the exercise.

3.2.2 Methodology

Camera setup and Video pre-processing

The whole setup consists of three low-cost cameras, namely the HIKVision [119], usually used for video-surveillance. The three cameras have been installed in fixed position, along the sides of a volume of interest. As stated previously, two setups have been designed and installed in two nursing homes, under different condition of lighting, acquiring 720×480 resolution videos.

As the output videos are not suitable for image processing in their raw form, a pre-processing phase is mandatory to prepare the videos to the following feature extraction procedure. In detail, the pre-processing stage is a sequence of selection and conversion algorithms, namely:

- **Frame per second (FPS) conversion:** As the videos from the three cameras have variable framerates, the lowest framerate among the three videos has been selected, projecting the time axes on a common reference, sampled with a unique framerate to achieve uniformity.

- **Video shifting:** A start signal, given the clinical therapists with a remote control, triggers the three video acquisitions, which however start with non-negligible relative de-

lays. To overcome such issue, the early-started videos are shifted of a number of frames equal to the relative delays.

- **Video trimming:** As most of the videos are long streams, the input streams are trimmed in exercise-related sub-videos.
- **Video Calibration:** As the videos suffer from image distortion, the extrinsic parameters have been extracted from the cameras of both setups to properly calibrate them.

Features Extraction

The complete knowledge of the position in space, or equivalently in the image plane, of the skeletal joints of the patients is enough to infer postural information. For this reason, the feature extraction process starts with the detection of the skeleton of the patients under analysis.

Skeleton detection is performed by means of the OpenPose library [120], which gives a real-time multi-person 2D pose estimation, aiming to represent both position and orientation of human limbs. For this work, the COCO training model has been implemented. It allows the identification of 18 skeletal joints from each person.

Different features have been chosen depending on the type of exercise, aiming to extract the most relevant information according to the test under analysis. As a matter of fact, each test provides different, yet relevant information regarding the posture and stability of the patient. Therefore, it reveals to be fundamental to properly select the highly-discriminant features with respect to each test, in order to suitably gather an amount of information about the patient as heterogeneous as possible.

3.2.3 Dataset Evaluation

The proposed work has been developed to provide sex, age, skeleton information and highly-discriminant features of patients performing SPPB tests. 20 patients suffering from a neurodegenerative disease and 27 healthy people perform the tests.

As a first step, all the acquired videos of the exercises performed by the patients have been studied, to evaluate their validity. Then, the preprocessing phase has been carried out to prepare the videos for the skeleton and features extraction, via the application of the OpenPose library. Finally, the dataset is completed with a vector of evaluation scores given by clinical therapists for each test.

Examples of patients performing BT, STST, and WT are shown in Fig. 3.2. To properly evaluate the efficiency of the dataset*, the information grabbed from patients performing BTs are considered. The highly-discriminant features extracted from the skeletons are used to feed

*The dataset will be shortly uploaded on the website: http://cms.stiima.cnr.it/isp/.

Figure 3.2: SPPB tests performed by different patients. Namely, a) Balance Test, b) Sit-To-Stand Test, and c) Walking Test.

a decision tree classifier, which has been trained to label patients into 5 classes of increasing risk of falls, shown in Table 3.1. The final score given by clinical therapists has been compared to estimated one. The good accuracy of the system (equal to 79.1%) shows the effectiveness of the provided dataset.

3.2.4 Discussion

In this Section, a complete dataset composed of sex, age, skeletal information and relevant features of elderly people performing SPPB protocol has been presented. Subjects have been grabbed by a system of three low-cost surveillance cameras. Then, proper video processing techniques have been used to highlight the skeletal joints of the subjects and to extract highly-discriminant features. It has been proved the high efficiency of the proposed dataset in the assessment of the patient's stability and posture skills, and their consequent risk of fall.

In the future, further semantic analysis of the videos will be investigated, to analyze more relevant features to be extracted from the skeletons, and to assess the progress of the neurodegenerative disease of patients observed during long periods.

3.3 Data acquisition in manufacturing: HA4M Dataset

In this Section, the Human Action Multi-Modal Monitoring in Manufacturing (HA4M) dataset is presented, which is a multi-modal dataset acquired by an RGB-D camera during the assembly of an Epicyclic Gear Train (EGT) (see Figure 3.3).

The HA4M dataset provides a good base for developing, validating and testing techniques and methodologies to recognize assembly actions. Literature is rich in RGB-D datasets for human action recognition [122, 123, 124] prevalently acquired in indoor/outdoor unconstrained settings. They are mostly related to daily actions (such as walking, jumping, waving, bending, etc.), medical conditions (such as headache, back pain, staggering, etc.), two-person interactions (such as hugging, taking a photo, finger-pointing, giving object, etc.), or gam-

ing actions (such as forward punching, tennis serving, golf swinging, etc.). Table 3.2 reports some of the most famous and commonly used RGB-D datasets on human action recognition describing their principal peculiarities.

To the best of the authors' knowledge, few vision-based datasets exist in the context of object assembly. Researchers usually build their own datasets on private video data [7, 136]. Table 3.3 compares the proposed HA4M dataset with existing datasets on assembly action recognition. As shown in Table 3.3, the proposed HA4M features various main contributions:

- **Data Variety:** The HA4M dataset provides a considerable variety of multi-modal data compared to existing datasets. Six types of simultaneous data are supplied: RGB frames, Depth maps, IR frames, RGB-to-Depth-Aligned frames, Point Clouds and Skeleton data. These data allow the scientific community to make consistent comparisons among processing approaches or machine learning approaches by using one or more data modalities.

- **Action Variety:** The HA4M dataset presents a wide variety in the action execution considering the high number of subjects (41) performing the task, the high number of actions (12), the different order followed by the subjects to perform the actions, and the interchangeably use of both hands.

- **Fine-grained Actions:** Actions present a high granularity as there is a subtle distinction between parts to be assembled and between actions that appear visually similar.

- **Challenging Issues:** The components to be assembled and the actions are very similar and symmetrical, implying a high level of context understanding and a significant

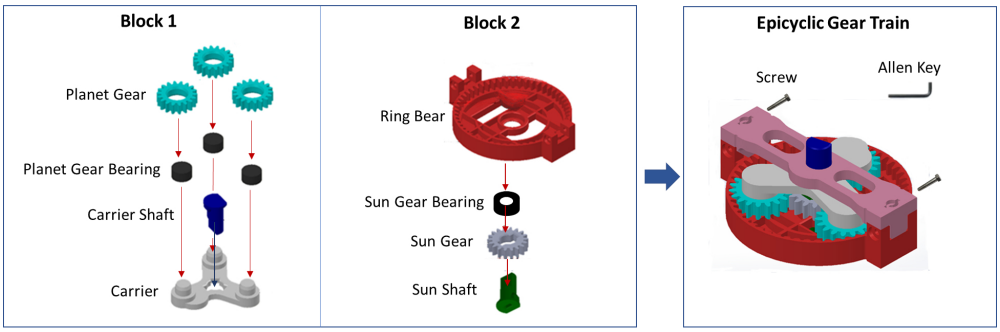

Figure 3.3: Components involved in the assembly of the Epicyclic Gear Train. The CAD model of the components is publicly available at [121].

Dataset	Sensors	Environment	Data Modalities	Actions
NTU RGB+D 120 [125, 126]	Microsoft Kinect v2	Cluttered Indoor	RGB Videos, Depth Sequences, 3D Skeleton Joints, IR Frames	Daily, Medical, Two People Interaction
SYSU 3DHOI [127]	Microsoft Kinect v1	Cluttered Indoor	RGB Videos, Depth Sequences, 3D Skeleton Joints	Daily
Drive&Act [128]	Five NIR cameras and One Microsoft Kinect	Static Driving Simulator	RGB, IR and Depth Data	Driver Behaviors
UE-HRI [129]	Two RGB cameras and one 3D sensor	Cluttered Indoor	RGB and Depth Frames	Human Robot Interaction
MoCa [130]	Three RGB cameras and Vicon Motion Capture System	Laboratory	RGB, 3D Skeleton Joints	Cooking
Grasping Dataset [131]	GoPro Hero 4 Camera, SoftKinetic Camera and IMU sensors	Living Room and Kitchen	RGB, Dept and IMU Data	Cooking, Housework
MSR-Action3D [132]	Microsoft Kinect v1	Cluttered Indoor	Depth Sequences, 3D Skeleton Joints	Daily
MSR Daily Activity 3D [133]	Microsoft Kinect v1	Cluttered Indoor	RGB Videos, Depth Sequences, 3D Skeleton Joints	Daily
UT-Kinect [134]	Microsoft Kinect v1	Cluttered Indoor	RGB Videos, Depth Sequences, 3D Skeleton Joints	Daily
RGBD-HuDaAct [135]	Microsoft Kinect v1	Laboratory	RGB Videos, Depth Sequences	Daily

Table 3.2: Some popular publicly available RGB-D Datasets for 3D Action Recognition. They prevalently collect RGB, Depth and 3D skeleton joints information relative to actions from daily activities conducted in indoor environments such as office-like, laboratory environments, or living rooms.

ability to track objects. Furthermore, unlike standard action recognition in unconstrained scenarios, the environment does not provide any information about the current action. All data have been acquired in a laboratory setting that does not change its background over time.

Dataset	Visual Sensors	Environment	Data Modalities	Task
Assembly101 [137]	Eight RGB Cameras mounted on a scaffold around a table and four monochrome cameras mounted on an headset	Laboratory	RGB frames, 3D hand poses	Assembly and Disassembly of toy vehicles
Meccano [138]	One Intel RealSense SR300 camera mounted on an headset	Laboratory	RGB videos	Assembly of a toy motorbike
IKEA-ASM [139]	Three Microsoft Kinect v2	Offices, Labs and Family Homes	RGB videos, Depth videos, 3D Skeleton Joints	Furniture Assembly
HA4M	**Microsoft Azure Kinect**	**Laboratory**	**RGB frames, Depth maps, IR frames, RGB-Depth-Aligned frames, Point Clouds, Skeleton Data**	**Assembly of an EGT**

Table 3.3: Comparison between the proposed HA4M dataset and existing vision-based datasets on assembly actions. For each dataset, information about the cameras used for data acquisition, the type of environment where acquisitions were made, the type of provided data and the assembly task are given.

3.3.1 Study Design

In the proposed dataset, a Microsoft Azure Kinect [140, 141] camera acquires videos during the execution of the assembly task. The Azure Kinect camera offers improved accuracy than other affordably RGB-D sensors implementing Time of Flight (ToF) principles [142], making the Azure Kinect one of the best solutions for indoor human body tracking in manufacturing scenarios.

With reference to Figure 3.3, the assembly of an EGT involves three phases: first, the assembly of Block 1 and Block 2 separately and then the final building of both blocks. The EGT is made up of a total of 12 components divided into two sets: the first eight components to build Block 1 and the remaining four components to build Block 2. Finally, two screws are fixed with an Allen key to assemble the two blocks, thus obtaining the EGT. Table 3.4 lists the individual components and the actions necessary for assembling Block 1, Block 2 and the whole EGT, respectively. The total number of actions is 12, divided as follows: four actions for building Block 1; four actions for building Block 2; and four actions for assembling the two blocks and completing the EGT. As can be seen in Table 3.4, some actions are performed more times as there are more components of the same type to be assembled: ac-

Components			Actions	
	Quantity	**Description**	**Action ID**	**Action Description**
Block 1	3	Planet Gear	1	Pick up/Place Carrier
	3	Planet Gear Bearing	2	Pick up/Place Gear Bearings (×3)
	1	Carrier Shaft	3	Pick up/Place Planet Gears (×3)
	1	Carrier	4	Pick up/Place Carrier Shaft
Block 2	1	Ring Bear	5	Pick up/Place Sun Shaft
	1	Sun Gear Bearing	6	Pick up/Place Sun Gear
	1	Sun Gear	7	Pick up/Place Sun Gear Bearing
	1	Sun Shaft	8	Pick up/Place Ring Bear
EGT	1	Block 1	9	Pick up Block 2 and place it on Block 1
	1	Block 2	10	Pick up/Place Cover
	2	Screws	11	Pick up/Place Screw (×2)
			12	Pick up Allen Key, Turn both screws, Return Allen Key and the EGT

Table 3.4: List of components and actions needed to build Block 1, Block 2 and EGT, respectively. First, the assembly of Block 1 (action IDs 1 to 4), then Block 2 (action IDs 5 to 8) and finally the EGT (action IDs 9 to 12).

tions 2 and 3 are executed three times, while action 11 is repeated two times. Finally, a "don't care" action (ID=0) has been added to include transitions or unexpected events such as the loss of a component during the assembly.

3.3.2 Acquisition Setup

The experiments took place in two laboratories (one in Italy and one in Spain). The acquisition setup is pictured in Figure 3.4. A Microsoft Azure Kinect® was placed on a tripod in front of the operator at an height $h = 1.54\,m$ and a distance $d = 1.78\,m$. The camera is tilted down to an angle $\alpha = 17°$(see Figure 3.4(b)). As shown in Figure 3.4(a), the individual components to be assembled are spread on a table in front of the operator and are placed according to the order of assembly. The operator can pick up one component at a time to perform the assembly task.

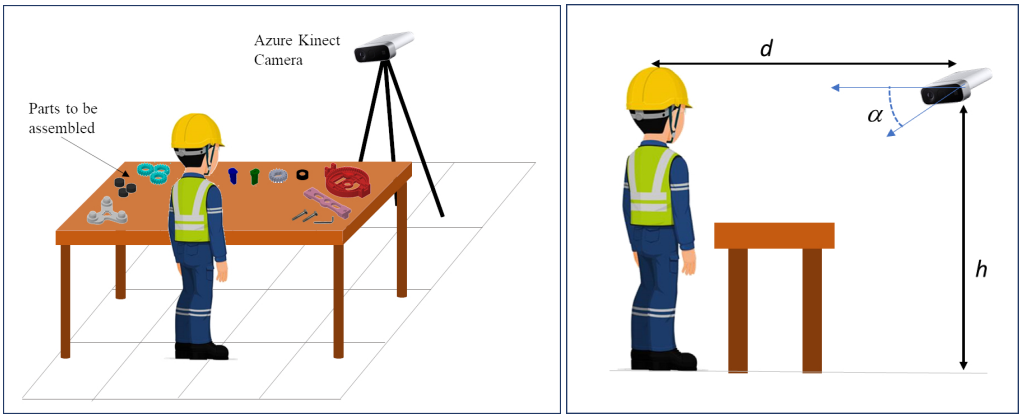

Figure 3.4: Sketch of the acquisition setup: (a) a Microsoft® Azure Kinect is placed in front of the operator and the table where the components are spread over; (b) setup specifications.

Figure 3.5: Typical video frames acquired by the RGB-D camera in the (a) "Vision and Imaging Laboratory" of STIIMA-CNR in Bari (Italy) and at the (b) "Department of Mathematics and Computer Science", Universidad de La Rioja, Logroño (Spain).

Two typical RGB frames captured by the camera in each laboratory are shown in Figure 3.5. The working table is covered by a uniform table cloth, while the components are arranged into boxes or spread over the table. In any case, two supports are fixed on the table to facilitate the assembly of Block 1 and Block 2. Block components can be in white or black color.

3.3.3 Study Participants

The HA4M dataset contains 217 videos of the assembly task performed by 41 subjects (15 females and 26 males). Their ages ranged from 23 to 60 years. All the subjects participated voluntarily and were provided with a written description of the experiment. The subjects were first instructed about the sequence of actions to perform to build the EGT. However, where possible, differences in assembly order were allowed. As an example, actions 2 and 3 can be performed three times in sequence (i.e. 2, 2, 2, 3, 3, 3) or alternatively (i.e. 2, 3, 2, 3, 2, 3). Furthermore, each subject was asked to execute the task several times and to perform the actions as preferred (e.g. with both hands), independently of their dominant hand.

3.3.4 Data Annotation

Data annotation concerns the labeling of the different actions in video sequences. The annotation of the actions has been manually done by observing the RGB videos frame by frame, and cross-checked by two researchers having different backgrounds, engineering or computer science. The start frame of each action is identified as the subject begins to move the arm to the component to be grasped. The end frame, instead, is recorded when the subject releases the component, so that the next frame becomes the start of the subsequent action. The total number of actions annotated in this study is 4123 (see Table 3.4).

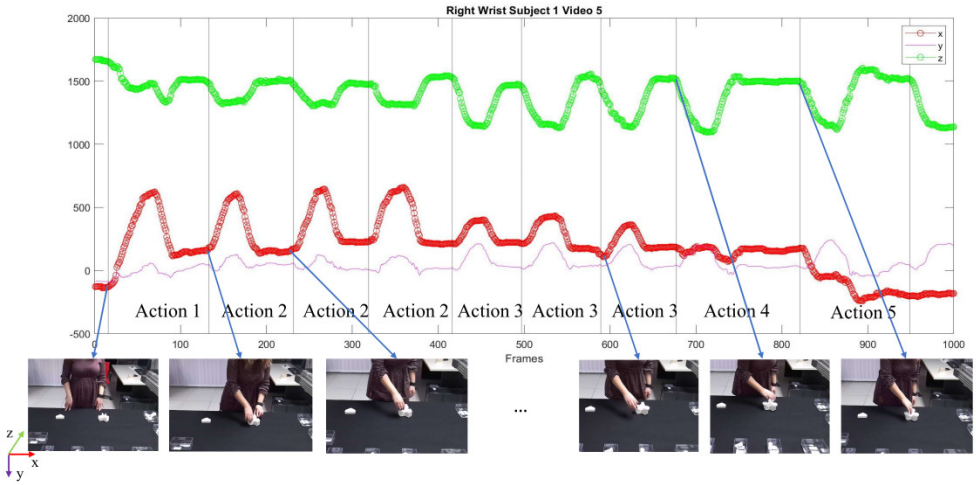

Figure 3.6: Check of annotation procedure. The plot reports the trajectories of the (x, y, z) coordinates of the right wrist of a right-handed subject in the first 1000 frames of an acquired video. The vertical lines identify the start frame of the actions annotated manually. Some relative RGB frames are also displayed. Frames have been cropped for visualization purposes.

Once the manual annotation was completed, the wrist joints of both hands were analyzed to further check the manual labeling. Referring to Figure 3.6, which shows the movement of the right wrist during the first 1000 frames of a sample video, local points of curvature variation of the x and z coordinates of the wrist joints can be considered as the points of action change. These points coincide with the start frame of each action (vertical lines in Figure 3.6) obtained by manual video annotation. It is worth noticing that the y coordinate does not give information for annotation check since it represent the joint height, typically constant and close to the table height during all actions.

3.3.5 Technical Validation

This section provides a statistical evaluation of the acquired data and an insight into some scientific issues that can be explored by using the HA4M dataset.

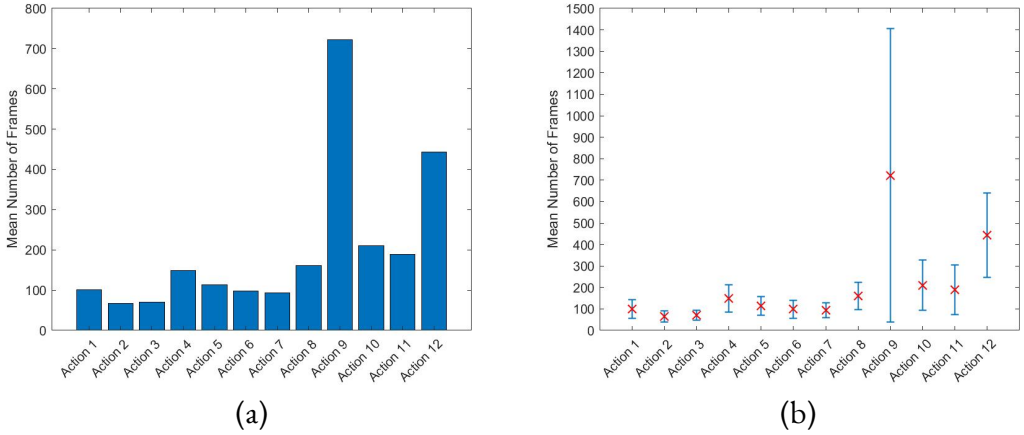

Figure 3.7: (a) Mean number of frames and (b) relative standard deviation for each action, estimated on the entire dataset.

Data Assessment

As a first characterization of the data, the variance of action durations is first assessed. Then, a spatial analysis of the 3D position of the wrist joints is also explored to further characterize the data. Notice that the "don't care" action is not considered in this evaluation study as it does not contribute to the assembly of the EGT.

Temporal Analysis

Figure 3.7(a) and 3.7(b) show the mean number of frames with the relative standard deviation for each action over all the recorded videos. For completeness, Table 3.5 numerically lists

the same time statistics with additional details. As can be seen, actions that require more time have a high variance in execution times. These actions can be more complex such as action 9 (assembly of Block 1 and 2), or can involve a longer activity such as action 12 (screw tightening). Furthermore, the subjects perform the task at their comfortable self-selected speed, so high time variance can be noticed among the different subjects. Figure 3.8 compares the mean number of frames for each action evaluated in the videos of two different subjects (number 2 and number 27) and the total dataset. As can be noticed, subject 2 executes the actions at a lower speed than subject 27, which, on the contrary, is very fast in task execution, even with respect to the total mean. This is mainly due to the different abilities of subjects in assembling the EGT or by accidental events, such as the loss and recovery of a component.

Action ID	Action Instances	Min Length	Max Length	Mean Length	Variance
1	217	8	263	100.23	42.99
2	651	22	207	66.29	26.01
3	651	25	210	70.27	23.71
4	217	63	632	148.57	62.92
5	217	48	264	113.88	42.52
6	217	37	384	98.47	42.32
7	217	38	254	93.67	35.10
8	217	54	415	161.23	63.05
9	217	114	4984	722.35	682.27
10	217	40	843	210.35	116.40
11	434	50	918	188.48	115.71
12	217	134	1488	443.70	197.60

Table 3.5: Some statistics about the actions: Action Identification Number (*Column 1*); Number of the manual annotated instances (*Column 2*); Minimum Length (*Column 3*), Maximum Length(*Column 4*), Mean Length (*Column 5*) and Variance (*Column 6*) of each action in terms of number of frames.

Spatial Analysis

The analysis of the spatial movement of both wrists of all subjects is useful for getting information about the main direction and spatial displacement of each action. Figures 3.9 (a) and (b) show the standard deviation of the coordinates (X, Y, Z) of the right wrist joint and the left wrist joint of all subjects and for each action, respectively. As can be noticed, different categories of actions can be identified according to the spatial properties: for instance, actions from 1 to 7 mainly evolve along the Z direction, whereas action 8 and 10 along the X direction. Finally, actions 9, 11 and 12 present comparable movements along the three directions as these actions require more spatial manipulations of the EGT. It is worth noticing that this spatial analysis can be biased by the way the subjects performed the tasks, since no precise rules were imposed to have the highest variability of the dataset. Accordingly, some subjects

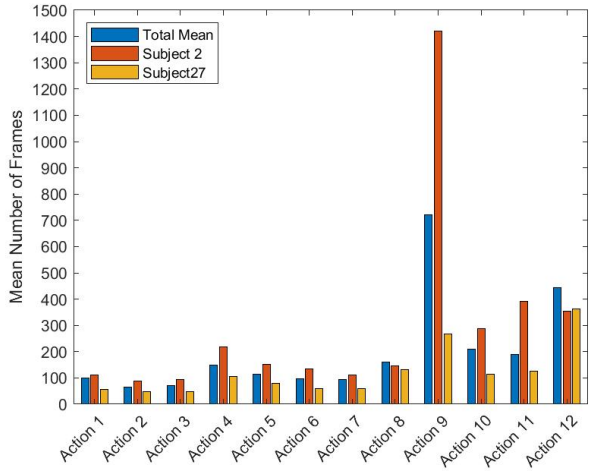

Figure 3.8: Comparative analysis of the performance of two subjects. Histograms show the mean number of frames for each action executed by subject 2 and subject 27 compared with the mean number of frames evaluated over the total dataset.

used their dominant hand while others used both hands interchangeably.

Scientific Issues

This section discusses some issues that can be explored using the proposed HA4M dataset in several application contexts.

Human centered approach in Industry 5.0
In the last years, the focus of smart manufacturing has been mainly on the transformation of manufacturing systems into new models with improved operational properties and new technologies. More recently, the focus has changed to a new perspective that puts workers at the center of the digital transformation, where technology must facilitate or improve human physical or cognitive abilities instead of replacing them [143]. As a consequence, the scientific community is very active in this domain by studying and developing intelligent systems to monitor workers to determine how they work, their pain points, and the challenges they face. So, observing the movements of human operators in the real scenario of an assembly task is very important to recognize their capabilities/competencies, especially in collaborative tasks with robots. Moreover, one of the main points of smart factory solutions is the inclusion of impaired people or people with different manual skills in production processes. The HA4M dataset represents a testbed for analysing the operative conditions of different subjects having varying skill levels. In the dataset, people with distinct ages and abilities execute complex

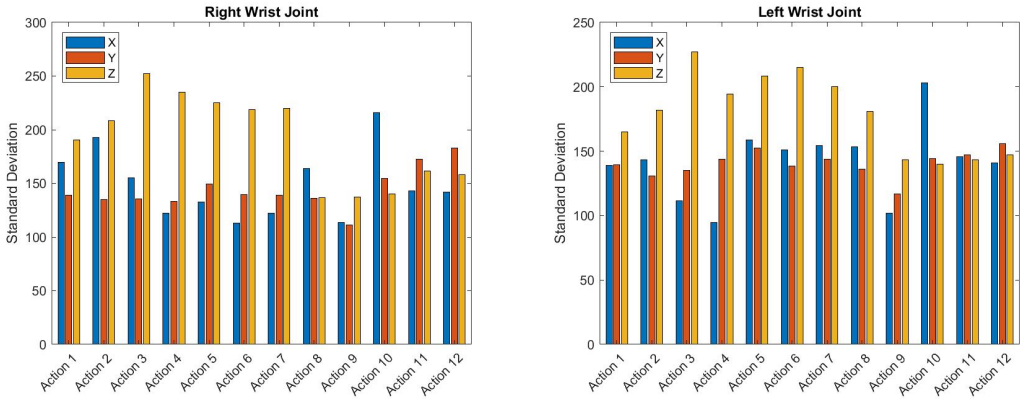

Figure 3.9: (a) Standard deviation of the coordinates (X, Y, Z) of (a) right wrist joint and (b) left wrist joint of all subjects and for each action.

actions in very different ways. One challenging task is the development of time-invariant action recognition methodologies capable of recognizing very different executions of the same actions. The spatial and temporal analysis of the actions presented in the previous section demonstrates the high variability of the execution of the actions, which is correlated not only to the speed of execution but also to the subjects' ability in handling the EGT parts.

Multi-modal data analysis
For years, human action recognition literature has been dominated by vision-based approaches using monocular RGB videos, making action representations difficult in 3D space. Moreover, challenging issues that commonly appear in the scene, such as illumination variations, clutter, occlusions, background diversity, must be tackled to have robust recognitions. The development of low-cost technologies has made available further sensory modalities to overcome some of the challenges mentioned above [144]. The HA4M dataset provides several types of data such as depth, infrared, or point cloud extracted using the Azure Kinect sensor. Therefore, the dataset allows the research in multi-modal data integration to take advantage of the peculiarity of each sensor (RGB and IR) and overcome their intrinsic limitations.

Temporal action segmentation
Literature is rich of works on action recognition methodologies successfully applied to short videos analysis. In recent years, the focus has been on temporal segmentation of actions in long untrimmed videos [145]. In Industry 4.0 domain, where collaborative tasks are performed by humans and robots in highly varying conditions, it is imperative to recognize the exact beginning and ending of an action. The HA4M dataset contains long videos with mul-

tiple instances of actions performed in different ways and in different orders. Therefore, the analysis of these videos requires the recognition of action sequences. Here, the problem of the temporal segmentation of the action aims to capture and classify each action segment into an action category.

Human-object interaction
The analysis of videos of human-object interactions involves understanding human movements, recognizing and locating objects, and observing the effects of human movements on those objects [146]. Traditional approaches to object classification and understanding of actions relied on shape features and movement analysis. In the context of assembly tasks, the relationships between movements and handled objects can help with action recognition. Sequences of actions that manipulate similar objects (such as inserting the planet gear onto the planet gear bearing) can be aggregated to create a higher level of semantic actions. The presence of RGB images and point clouds in the HA4M dataset could allow the recognition of tools and parts with pattern recognition approaches and their relative manipulation to improve the target of action classification.

4

Machine Learning and Deep Learning methodologies for Human Mobility Assessment in elderly facilities

4.1 Introduction

This Chapter presents the work published in [99], which proposes a vision-based system that observes elderly people while performing three well-defined mobility tests and automatically categorizes their mobility performance. In particular, the main contributions of this work are the following:

- The proposed system emulates the complex decision process of the expert physiotherapists in the evaluation of the mobility tests.
- The system processes real data acquired using low-cost commercial RGB cameras, typically implemented for video surveillance applications. The cameras were installed in two nursing homes that house older people who are healthy and affected by neurodegenerative diseases. The video data have been augmented and then processed to select the most informative features to provide a better-generalized model and enhance the decision process.

- Four classifiers with deep neural network architectures, based on Long-Short Term Memories (LSTMs) and Bidirectional Long-Short Term Memories (BiLSTMs), are proposed to classify the acquired data. The presented deep neural network architectures have also been rearranged to develop also regression models to further compare results with those from the classification task. Besides, comparisons with various traditional machine learning methodologies have also been conducted.

The Chapter is structured as follows. Section 4.2 gives details about the case study. Section 4.3 defines the different steps of the applied methodology; experimental results are in Section 4.4, while final remarks are in Section 4.5.

4.2 Case Study Description

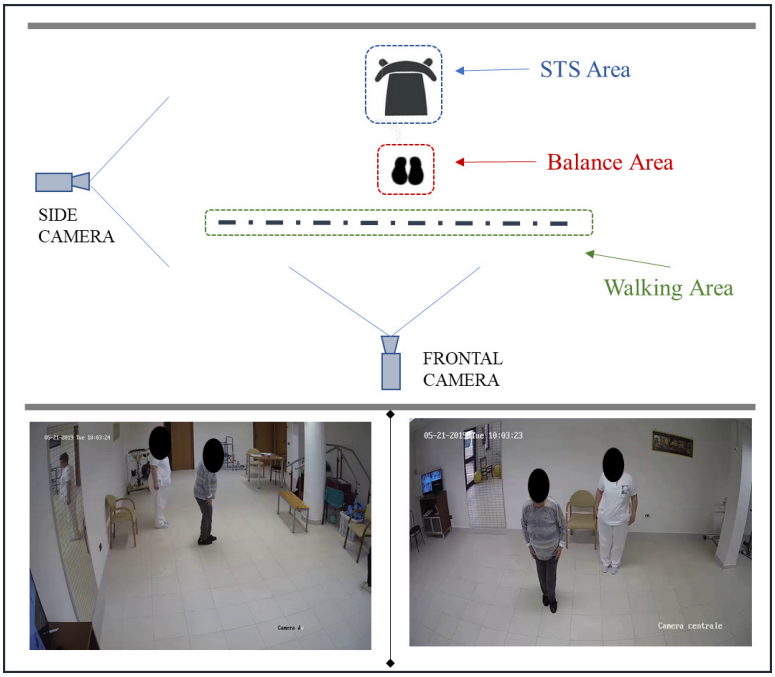

Figure 4.1: The camera setup used for video acquisition during the execution of the motion protocol.

The system setup used for data acquisition was made up of two low-cost RGB monocular cameras installed in two nursing institutes. One frontal camera and one side camera were placed in the gym of the institutes, where people usually execute mobility tests as shown in Figure 4.1.

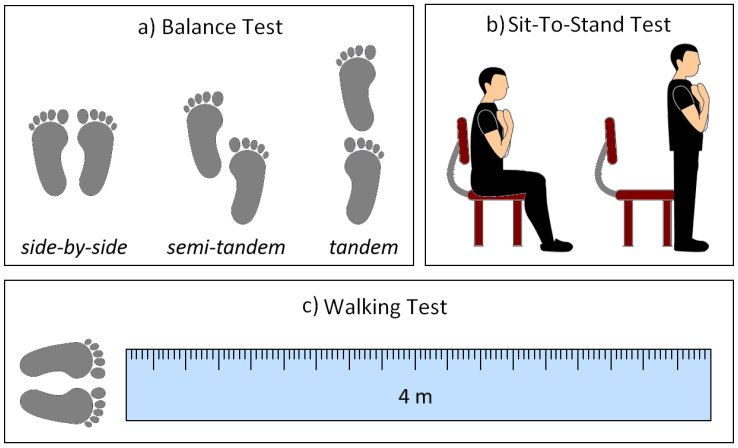

Figure 4.2: Representation of the three SPPB tests: a) Balance test: the patient stands with the feet side-by-side, then in semi-tandem and tandem positions; b) Sit-To-Stand Test: the patient sits down and stands up five times with the arms crossed on the chest; c) Walking Test: the patient walks for four meters.

The motion protocol, defined by medical staff and used in this work, consists of three mobility tests included in the so-called Short Physical Performance Battery (SPPB) [147]: the Balance Test (BT), the Walking Test (WT) and the Sit to Stand Test (STST). Figure 4.2, shows a representative scheme of these tests. Specifically, in the Balance Test, the person stands with the feet side-by-side, then in a semi-tandem position and then in a tandem position, trying to stay in each of the listed positions for ten seconds (Figure 4.2 a)). In the Sit-To-Stand test, the person sits down and stands up five times with the arms crossed on the chest (Figure 4.2 b)). In the Walking Test, the person walks a four-meter linear path, free of obstacles, and returns to the starting point (Figure 4.2 c)).

The SPPB is usually administered to people by a physiotherapist to evaluate their mobility level as it releases information regarding body posture, balance, strength, and stability. The physiotherapist evaluates the execution of each test, giving a score value between 0 and 3, representing the mobility class. The classes range from the bad one (0 value), when the person cannot execute the test, to the best one (3 value) when, instead, the person succeeded.

All the older people, and their families, where needed, gave their written informed consent to participate in this study. There were 20 people affected by neurodegenerative diseases in the early stages and 27 healthy people, all in the range of age of 60 to 95 years. The subjects were recorded while performing the tests included in the SPPB in two separate acquisition campaigns three months apart. Several difficulties emerged during the data acquisition phase as the sample of people who participated in the first acquisition campaign was reduced in the follow-up as some were no longer able to perform the SPPB tests independently.

Table 4.1: Statistical analysis of the videos of each test BT, WT, and STST, respectively.

Test	Total nr. of videos	Total nr. of Frames	Mean nr. of Frames	Standard Deviation
BT	74	19416	262.37	32.03
WT	76	10515	138.35	75.01
STST	96	39509	411.55	143.56

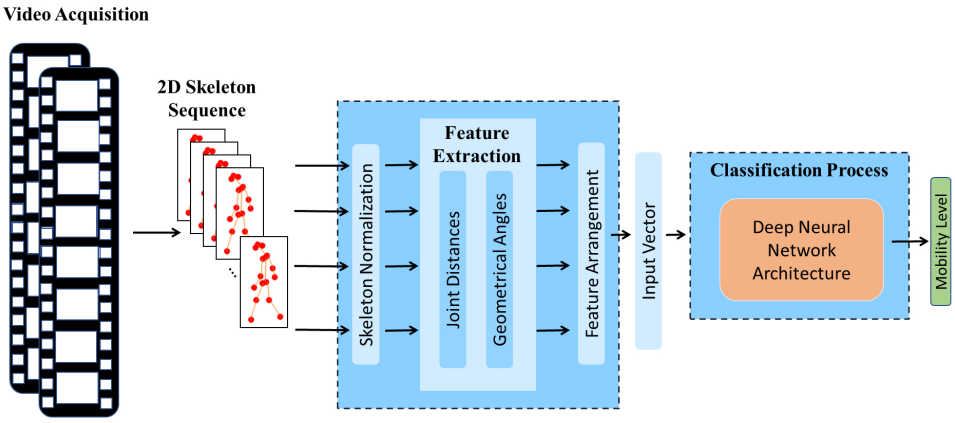

Figure 4.3: Pipeline of the proposed approach for classifying people's mobility level.

Once the video sequences of RBG images were acquired, they were appropriately processed to extract bidimensional skeletal data that have been made publicly available [97, 148]. Table 4.1 gives some information about the acquired videos. In particular, 74, 76, and 96 videos have been captured for the BT, WT, and STST, respectively. As proved by the standard deviation values, the number of frames varies considerably among the three tests. For this reason, the duration of tests is not enough discriminant to achieve mobility assessment: a qualitative evaluation of test execution is mandatory to classify people's mobility.

4.3 Methodology

The proposed system assesses people's mobility in the same classes defined by physiotherapists, but in a completely automatic and objective way, without human bias. The main steps involved in the proposed methodology are reported in Figure 4.3:

1. Commercial low-cost RGB cameras for video surveillance capture videos of test execution;

2. A preliminary processing extracts skeletal joints to evaluate complex details related to body postures, inclinations, and orientations of body parts;

3. A data augmentation technique enlarges the dataset made of the temporal evolution of joints in the image plane;

4. Significant features are extracted to construct input vectors to feed neural networks.

As primary output, this work proposes deep neural network architectures for classification based on Long-Short Term Memory (LSTM) and Bidirectional Long-Short Term Memory (BiLSTM). Following an ablation study, preliminary convolutional blocks are added for feature mixing to improve classification results. Further comparison with standard classifiers from shallow learning (Decision Tree, Naive Bayes, SVM, KNN) and deep neural network architectures for regression, i.e. labeling people's mobility with continuous scores, are also presented. The next subsections will better detail the feature extraction process (Section 4.3.1), the network architectures used for classification (Section 4.3.2), and the data augmentation technique (Section 4.3.3).

4.3.1 Feature Extraction

In this work, the well-known OpenPose library [149] is used to extract human skeletons from RGB frames. OpenPose efficiently detects the 2D pose of multiple people in an image, representing both the position and orientation of human limbs. The implemented model identifies 18 skeletal joints and 17 links between joints, as shown in Figure 4.4. Joint positions are not directly used to model people's mobility. Instead, a set of features is designed in agreement with clinicians to characterize anomalies during the SPPB tests. These features are based on 2D pairwise joint distances, normalized to body height, and geometrical angles between consecutive body segments (i.e. bones) to highlight posture variations and walking or balance problems. Features are evaluated at each frame and then put together in time series.

Figure 4.4 and Table 4.2 show the features of each SPPB test, providing detailed descriptions and indicating the camera used for their extraction. In the following, each SPPB test is analyzed together with the related features.

- *Balance/Walking Test:*
  Both balance and walking tests are administered to people to assess their static and dynamic skills. In the two cases, the following features have been considered:

  – Distance between feet (*(i)* in Figure 4.4). This distance can help evaluate the patient's confidence in following the predefined path of the WT.

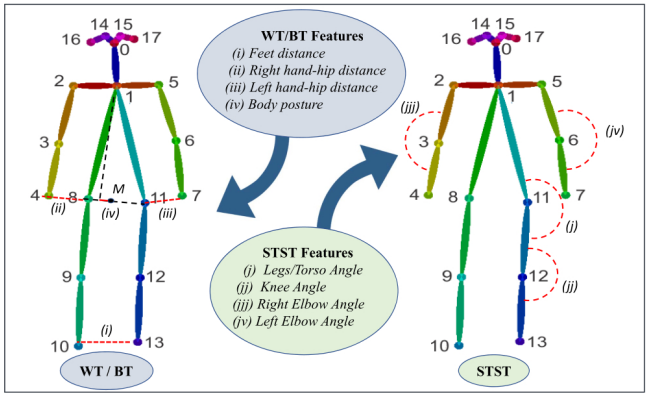

Figure 4.4: Features defined for the Walking and Balance Test (WT/BT) and Sit to Stand Test (STST), respectively.

- Distance between the right (or left) hand and the right (or left) hip from the frontal camera (*(ii)* or *(iii)* in Figure 4.4). This feature is fundamental for evaluating an eventual loss of balance and, in this case, for restoring balance with the help of the arms.
- Body posture, i.e. the column projection of the distance vector that connects the neck and the middle point M between the hips (*(iv)* in Figure 4.4). It provides information on people's torso inclination, indicating whether they keep their back straight.

It is worth noticing that, in the case of BT, the side-by-side, semi-tandem, and tandem tests are captured in three different videos. Homologous features are thus concatenated in vectors of increased lengths.

- *Sit-To-Stand Test:*
  The STST is slightly different from the previous two tests, as it provides a method to quantify the functional strength of the lower limbs and/or to identify how a person completes transitional movements between sitting and standing. In this case, the features are:

  - The angle between the legs and the torso (*(j)* in Figure 4.4) and the knee angle (*(jj)* in Figure 4.4). Both angles describe the action of sitting as captured by the side camera.
  - The angle at the right (or left) elbow from the frontal camera (*(jjj)* or *(jv)* in Figure 4.4). These features characterize people's confidence while performing the STST.

Table 4.2: Description of the features for each test (BT, WT, and STST), with specified the camera used for their extraction. The joint numbers in the table are shown in Figure 4.4.

Test	Feature	Description	Side Camera	Frontal Camera
BT	Feet distance	Distance between joints 10 and 13	✓	✓
	Right hand-hip distance	Distance between joints 11 and 7		✓
	Left hand-hip distance	Distance between joints 4 and 8		✓
	Body posture	Distance between the projection of the joint 1 and the midpoint of the segment connecting joints 8 and 11		✓
WT	Feet distance	Distance between joints 10 and 13	✓	✓
	Right hand-hip distance	Distance between joints 11 and 7	✓	
	Left hand-hip distance	Distance between joints 4 and 8	✓	
	Body posture	Distance between the projection of the joint 1 and the midpoint of the segment connecting joints 8 and 11	✓	
STST	Legs/Torso Angle	Angle between the legs and the torso	✓	
	Knee Angle	Angle at knee	✓	
	Right Elbow Angle	Angle at the right elbow		✓
	Left Elbow Angle	Angle at the left elbow		✓

To highlight how the defined features represent the different situations that occur when the SPPB tests are performed, Figures 4.5, 4.6 and 4.7 show features plots for each SPPB test and in both cases of one person who performed the test correctly and one who failed. In Figure 4.5a), for example, the graphs of the hand-hip distances show the poor postural stability of the subject. Significant fluctuations in the graphs represent the subject's attempts to maintain balance. In contrast, Figure 4.5b) shows minor fluctuations in the graph, as the subject maintains balance while performing the test.

In the case of WT, in Figure 4.6a) the feature plots clearly describe a person who needs more time to walk the path, as shown by the sequence of small steps in the walking width graph. On the other hand, the subject in Figure 4.6b) performs the WT with more confidence, without

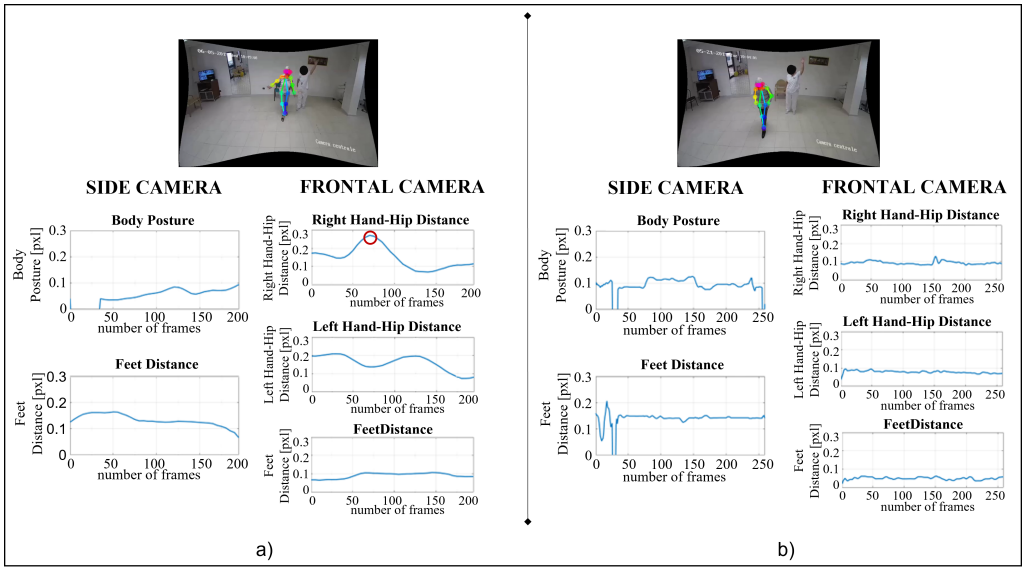

Figure 4.5: Plots of features extracted from the skeletons of two people performing the Tandem position of the BT: a) people of class 0 (unable to maintain balance); b) people of class 3 (correct body posture). The red circle on the feature plot of the Right Hand-Hip Distance indicates the subject's attempt to maintain balance by moving the right arm.

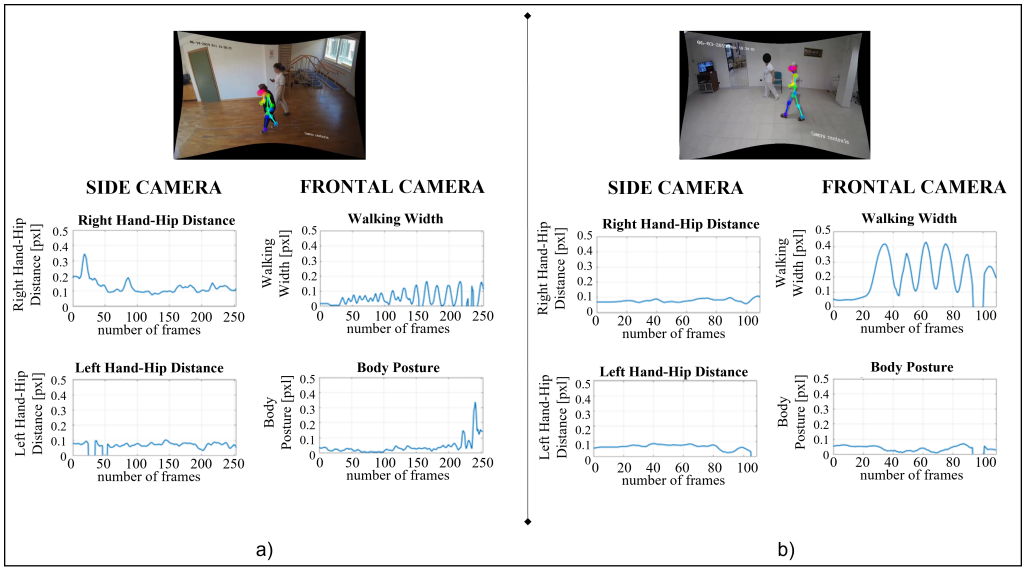

Figure 4.6: Plots of features extracted from the skeletons of two people performing the WT: a) people of class 1 (long execution time); b) people of class 3 (high walking confidence). The graphs of the hand-hip distances and walking widths show the different behavior of the two people in performing the test.

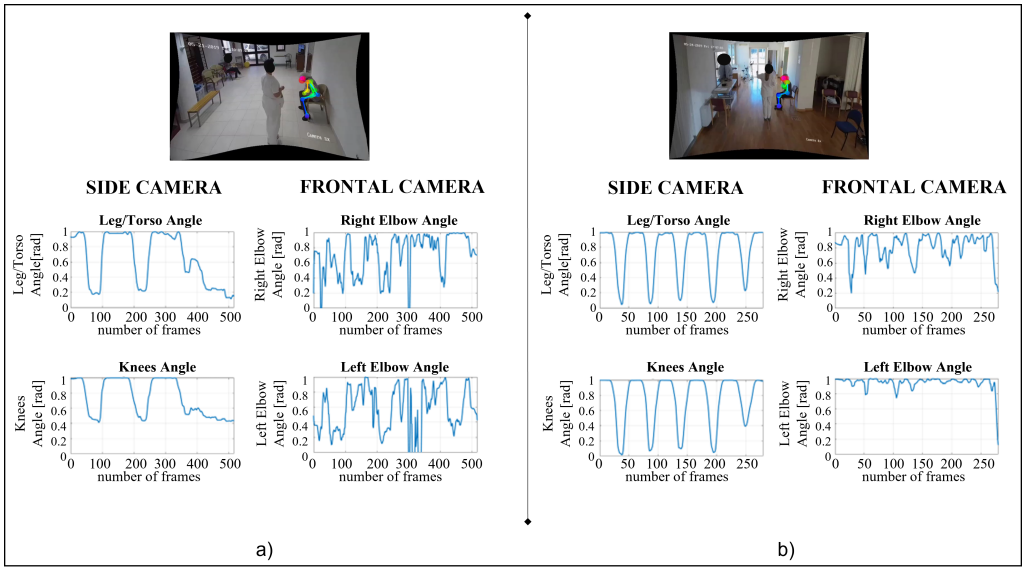

Figure 4.7: Plots of features extracted from the skeletons of two people performing the STST: a) people of class 0 (unable to perform the STST); b) people of class 3 (stands up and sits down 5 times). The left graphs of Leg/Torso and Knee Angles demonstrate that the subject succeeds only two times in standing up. The elbow angles further show the inability to keep the arms crossed on the chest.

balancing with the arms.

Finally, in the case of STST, it is evident by the feature plots shown in Figure 4.7a) how the subject succeeds only two times in standing up. Furthermore, the subject does not keep his arms crossed on the chest, thus failing the test. On the contrary, Figure 4.7b) shows the case of correct execution of the STST.

4.3.2 Deep Neural Network Architectures

In this work, the four deep neural networks in Figures 4.8 and 4.9 are compared to evaluate the best configuration. The input layer builds the feature vectors by concatenating the features $(f_1, f_2, ..., f_k)$ from all the frames contained in the video of the SPPB test, where k depends on the test under examination (see Table 4.2). Taking into account both side and frontal views, in the case of WT $k = 5$, for STST $k = 4$, while for BT $k = 15$, since BT involves three tests (side-by-side, semi-tandem, and tandem).

The deep network architectures are based on LSTM and BiLSTM models. An LSTM neural network is an extension of a recurring neural network (RNN), suitable for processing time series [150]. Its core is the LSTM block, shown in Figure 4.8a), which captures essential input features and preserves them over a long period, learning which information is worth

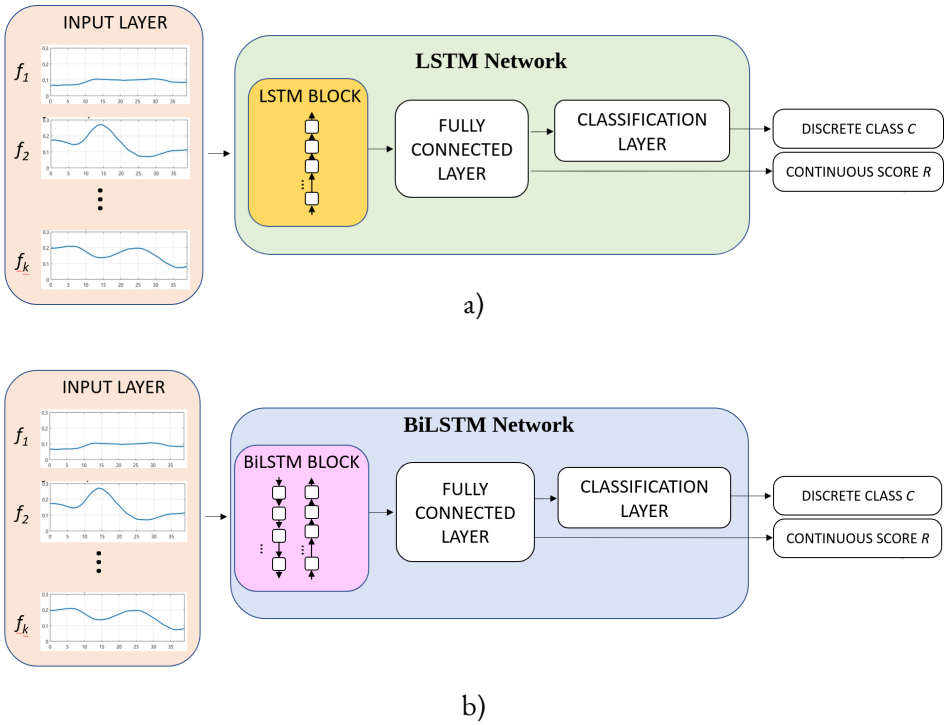

Figure 4.8: Architecture of a) the LSTM network and b) the BiLSTM network.

storing or erasing through a gating mechanism that controls the memorizing process. In the Bidirectional LSTM (BiLSTM) neural network of Figure 4.8b), a Backward LSTM and a Forward LSTM cooperate to capture past and future information by letting the data flowing forward and backward [151]. BiLSTM is well-known to achieve better performance than LSTM by modeling the sequences along both directions. In the proposed experiments, both blocks have 100 hidden units.

In the proposed work, deep neural networks are designed for two purposes: classification and regression.

- Classification: the input features are processed to select a discrete class C among four classes of interest ($C \in 0, \ldots, 3$). The result is the same as for the physiotherapists, who assign $C = 3$ to successful tests and $C = 0$ to indicate complete inability. The architecture is then completed by a Fully Connected layer, which mixes the information returning a normalized vector, and a Classification layer, which converts the output of the Fully Connected layer into probabilities through a Softmax function and compares them to minimize the cross-entropy.

- Regression: the networks process the input features to produce a continuous score R. This output is strictly dependent on the target class, but, for its nature, can estimate intermediate mobility levels. In this case, the networks are still completed by a Fully Connected layer, whose output is directly interpretable as the final regression result R. During training, the networks try to minimize the half mean-squared error loss, based on the same example of the classification task, i.e. using discrete targets to generalize then and predict continuous scores.

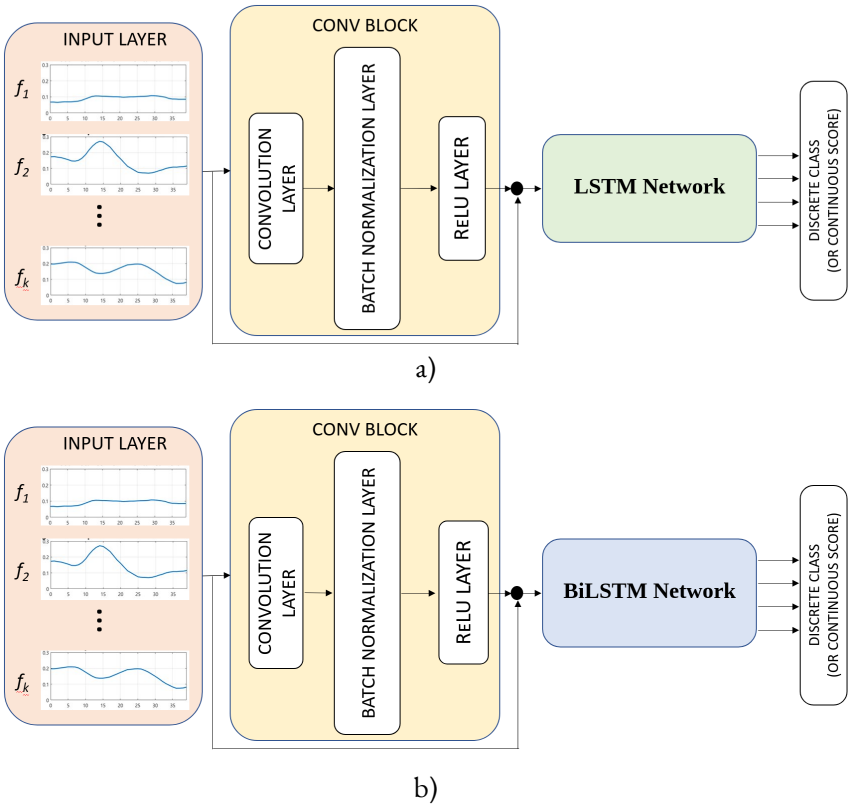

Figure 4.9: Architecture of a) the Conv-LSTM network and b) the Conv-BiLSTM network.

To enhance the correlations among features at each time step, a Convolutional Block (Conv-Block) is introduced, as shown in Figure 4.9, obtaining the so-called Conv-LSTM and Conv-BiLSTM networks. The Conv-block consists of a Convolution Layer, a Batch Normalization Layer, and a ReLU Layer, as shown in the yellow box in Figure 4.9. The Convolution Layer applies several convolutions having $k \times 1$ kernels to the sets of input features at each time step. The Batch Normalization Layer then normalizes the output vectors and is finally

rectified using the ReLU function. This block generates a new representation of the input time series to feed the recurrent networks LSTM and BiLSTM of Fig. 4.8.

4.3.3 Data Augmentation

One of the most frequent problems in machine learning, especially in deep learning, is the lack of a sufficient amount of training data or uneven class balance within the datasets. This problem is even more stringent in this work, where the amount of real data is limited for several reasons (see Section 4.2).

Data augmentation encompasses a suite of techniques that enhance the size and quality of training datasets to build better deep-learning models. In the context of image data, data augmentation includes classical image transformations such as rotation, cropping, zooming, histogram-based methods, color space augmentations, image mixing, and so on [152]. However, these image-based transformations, performed before the skeleton extraction, can induce artifacts in 2D body reconstruction. For this reason, the proposed procedure directly augments the dataset by working on the position of the joints. With reference to Figure 4.10, data augmentation is made by a set of A rigid geometric transformations of human skeletons, which create different views of the same people, maintaining the relationships between the joints. It is worth noticing that data augmentation is performed after splitting the data into the training and testing sets, to avoid having the augmented features of the same subject in different sets.

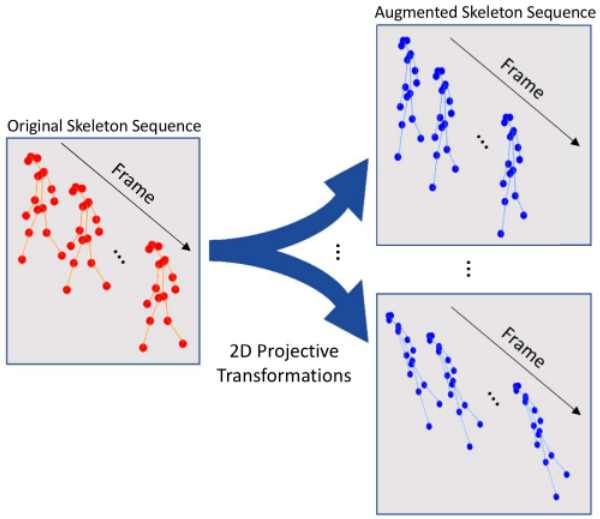

Figure 4.10: Skeleton augmentation process: 2D projective transformations are applied to the original skeletons, obtaining new sequences of augmented skeletons.

Let indicate the joint points in 2D coordinates as $J_p = [x_p, y_p]^T \in \mathbb{R}^2$ in the camera coordinate system, with $p = 0, \cdots, 17$. In general, a point $P = [x, y]^T \in \mathbb{R}^2$ in the 2D Euclidean plane can be described in homogeneous coordinates H as follows [153]:

$$H = \begin{bmatrix} wx \\ wy \\ w \end{bmatrix} \in \mathbb{P}^2 \qquad w \in \mathbb{R} - \{0\} \tag{4.1}$$

where $\mathbb{P}^2$ is the 2D projective space defined as $\mathbb{P}^2 = \mathbb{R}^3 - [0, 0, 0]^T$. For the sake of simplicity, w is typically equal to 1 to have direct transformations between 2D Euclidean and 3D homogeneous coordinates ($P = [x, y]^T \leftrightarrow H = [x, y, 1]^T$).

Let $T_i (i = 1, \cdots, A)$, the non-singular 3×3 matrix designed to produce the 2D projective transformation:

$$T_i = \begin{bmatrix} 1 & 0 & E_i \\ 0 & 1 & F_i \\ 0 & 0 & 1 \end{bmatrix} \qquad i = 1, \cdots, A \tag{4.2}$$

where E_i and F_i are discrete values representing the influence of the vanishing point to the final projection. Large values of E_i and F_i induce close-to-the-origin vanishing points, i.e. parallel lines converging faster. For this reason, these couples of values have been kept small (between 0.001 and 0.01), in the experimental phase, to guarantee reasonable augmentations. Therefore, the new homogeneous 2D coordinates $J'_{p,i}$ of the p-th joint are:

$$J'_{p,i} = \begin{bmatrix} x'_{p,i} \\ y'_{p,i} \\ 1 \end{bmatrix} = T_i \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} \qquad i = 1, \cdots, A \tag{4.3}$$

Since each transformation applies equivalently to all the skeletons, i.e. to all the frames of each video, the size of the resulting dataset after augmentation is A times higher than the initial one in terms of the number of frames. The A parameter has been fixed heuristically by evaluating the performance of the classifiers varying it.

Figure 4.11 reports the plots of one feature, the knee angle, extracted from the acquired video of a subject performing the STST (Fig. 4.11a)) and that of four skeletons (Fig 4.11b)) obtained by applying four different 2D projective transformations. From a first qualitative analysis, the features extracted from the transformed skeletons are still coherent in magnitude and time with the original ones. This aspect is of enormous importance since the features extracted from both original and transformed skeletons differ numerically, ensuring sufficient dataset variability. Still, they refer to the same target class that resembles the same high-level behavior. In the experimental section, the classification accuracy is reported, demonstrating the quantitative evidence of the proposed data augmentation procedure.

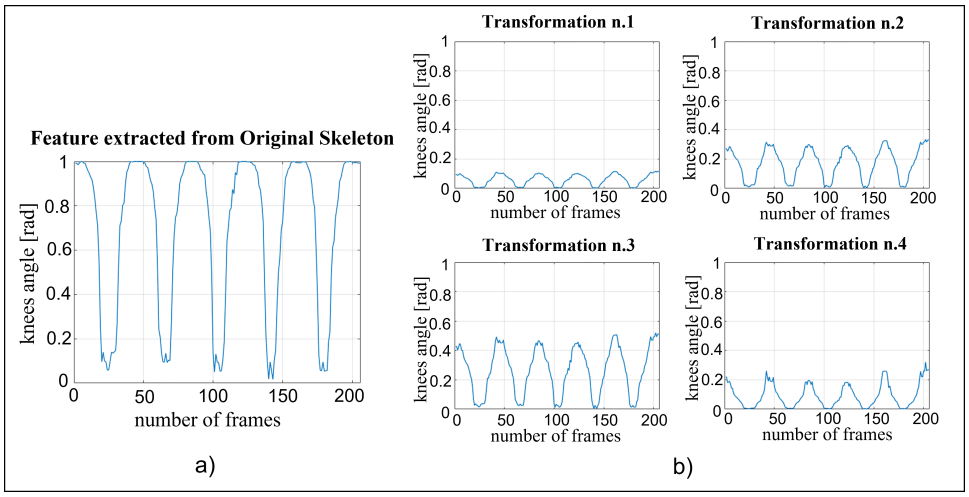

Figure 4.11: a) Plot of a sample feature (knee angle) extracted from the skeleton in an acquired video of STST. b) Different plots of the same feature extracted from the transformed skeletons by applying four different 2D projective transformations.

4.4 Experiments

This section describes the experimental results and details the different data processing steps: data acquisition and classification. All computations have been performed on a 64-bit HP Z840 Workstation, with Intel® Xeon® E5-2699v3 CPU @ 2.30 GHz processor and 256 GB of RAM. To accelerate the training process, all operations have been transferred to a NVIDIA® Quadro® K5200 GPU.

4.4.1 Data Acquisition and Processing

The cameras used for data acquisition are low-cost 4k cameras by HIKVision with 3849×2160 resolution at 20 fps, usually used in video surveillance applications. Due to the dimensions of the gym of the nursing institutes, where videos were acquired, the frontal camera had a focal length of $2.8mm$, whereas the side camera had a focal length of $4mm$. The videos are 246 in total, 74 videos relative to BT, 76 to WT and 96 to STST (see Table 4.1). Video durations can vary, depending on the test and the participant.

Due to the low-level setup of the cameras, the acquired videos presented some limitations, such as a lack of camera synchronization, slightly different camera frame rates, and misalignment of video frames. Therefore, the videos acquired by the side and frontal cameras were first projected on the same timeline, based on the lowest recorded frame rate to improve video uniformity. Then, the couples of videos were manually shifted by as many frames as the delay

between the two cameras to achieve synchronization. A signal given by the physiotherapist at the start of each SPPB test was used for this aim. Furthermore, the videos were trimmed to extract only the clips containing the execution of the tests. Finally, a camera calibration procedure was applied to remove image distortion. The OpenPose library is then applied to extract the skeletons. A skeleton tracking procedure has also been developed to detect only the skeleton of the person performing the test, discarding the skeletons of other subjects present in the scene, such as physiotherapists. The first frame of each video is manually labeled by the user to identify who is running the test. Then, for every frame, a Region of Interest (ROI) of 30×30 pixels is selected around each joint of all the subjects in the scene. With an automated process, each joint-related-ROI is compared with the corresponding ones of the subject of interest at the previous frame. Following a voting mechanism, the skeletons of other people in the scenes are discarded, while one of the subjects performing the test is retained.

Then, the obtained dataset of skeletons has been augmented by applying the data augmentation procedure described in Section 4.3.3.

4.4.2 Classification

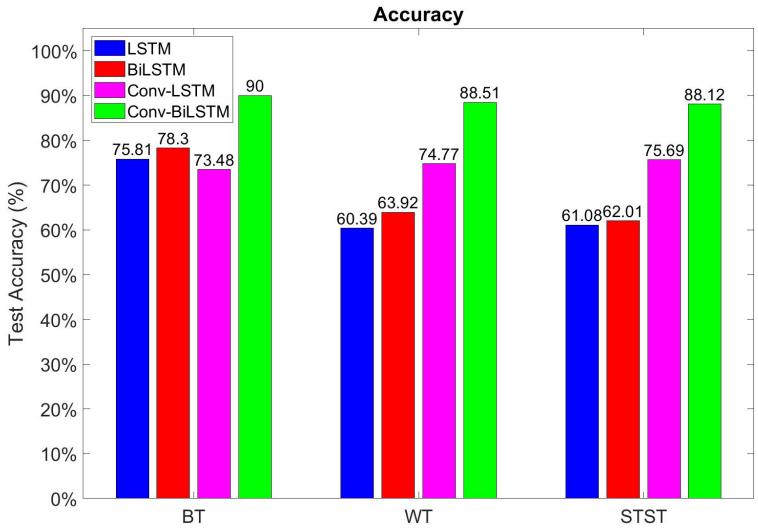

Figure 4.12: Percentages of weighted mean Accuracy of the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM and Conv-BiLSTM) for each SPPB test (BT, WT, and STST).

This section presents the classification results obtained by applying the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM, and Conv-BiLSTM) described in sec-

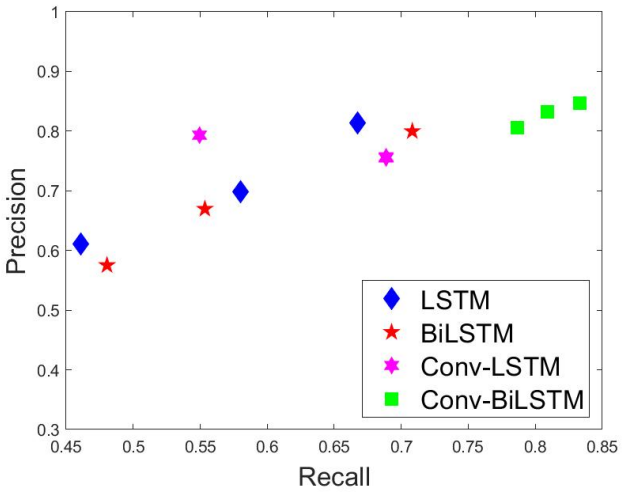

Figure 4.13: Weighted mean values of Precision vs Recall of the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM, and Conv-BiLSTM). The three markers for each classifier refer to the three SPPB tests.

tion 4.3.2. In the following, classifiers will be compared in terms of Accuracy, Precision, and Recall, whose definitions are in Table 4.3. These metrics are computed by reducing the multi-class problem to multiple binary problems in a *OneVsAll* strategy. Each metric is thus computed four times to assess the classification of each class against the others. The final evaluation metrics are then computed as the arithmetic mean of the four results, weighted by the population of the corresponding class (weighted average).

Table 4.3: Accuracy, Precision, and Recall. These quantities are evaluated starting from the computation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy	**Precision**	**Recall**
$\frac{TP+TN}{TP+FP+TN+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$

In the learning phase, the dataset of the extracted features has been divided into training, validation, and test sets. The samples included in the training and validation sets have been exclusively used for the learning phase. The validation set has been used to assess the model's convergence and stop training when accuracy does not increase for eight consecutive epochs. The test set has been used to evaluate the network's performance in labeling unknown input data. The learning phase results from optimizing a cross-entropy loss function, performed

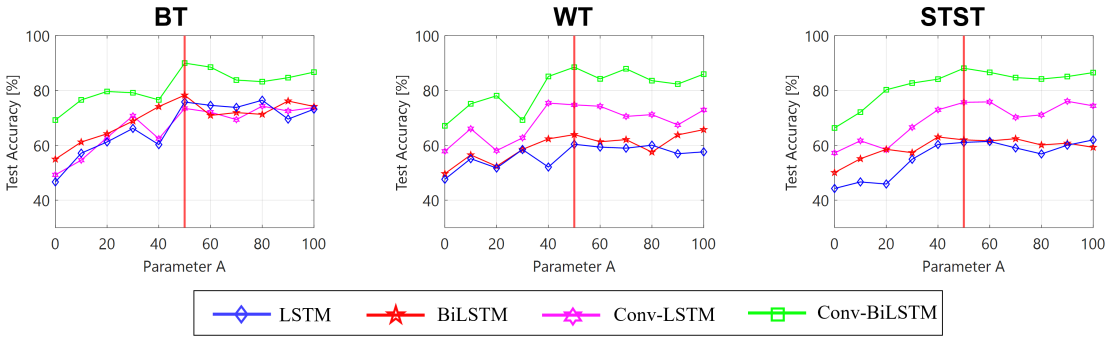

Figure 4.14: Percentages of weighted mean Accuracy of the proposed classifiers varying the A parameter, for BT, WT, and STST respectively.

using the Adam optimizer. A 5-fold cross-validation technique has been applied to verify the generalization ability of the networks. For each SPPB test, after cross-validation, only the models with the highest accuracy have been selected to classify elderly people into the four classes (0, 1, 2, 3) defined in Section 4.3.2.

Figure 4.12 shows the percentages of weighted mean accuracy of the deep neural network architectures (LSTM, BiLSTM, Conv-LSTM, and Conv-BiLSTM), for the three tests BT, WT e STST, respectively. Among the proposed deep architectures, those implementing BiLSTM produce better results than those using LSTM. For example, BiLSTM increases accuracy by an average improvement of 2.31%, considering all three tests of the SPPB. Similarly, the Conv-BiLSTM classifier outperforms the Conv-LSTM one with an average improvement of 14.23%. These results confirm that taking input in forward and backward directions increase the amount of available information, capturing the complex variability of the features. At the same time, introducing the Convolutional Block before the LSTM/BiLSTM networks produces a more significant enhancement of the classification accuracy. In particular, the Conv-LSTM network increases performance in dynamic tests (WT and STST) compared to the results of the LSTM one, as well as the Conv-BiLSTM over the BiLSTM with an average improvement of 19.91%. Indeed, applying convolutional kernels to the input feature vectors transforms the data into new vectors that better characterize the features' spatial correlation, improving the final classification ability. In Figure 4.13, the weighted averages of Precision vs Recall are reported for each deep neural network architecture and each SPPB test. Precision/Recall metrics also confirm that Conv-BiLSTM architecture outperforms the others.

Additional experiments have been conducted to evaluate how data augmentation affects the classifiers' performance. Figure 4.14 shows the resulting weighted mean accuracy of the

deep classifiers for each SPPB test when A ranges between 0 and 100. At first glance, increasing the dataset size by data augmentation results in an improvement in the average accuracy for any network and any mobility test. However, adding more data leads to longer training time. Consequently, $A = 50$ produces the best trade-off between accuracy and dataset size. It should be noted that Conv-BiLSTM always performs better regardless of the size of the dataset defined by the parameter A.

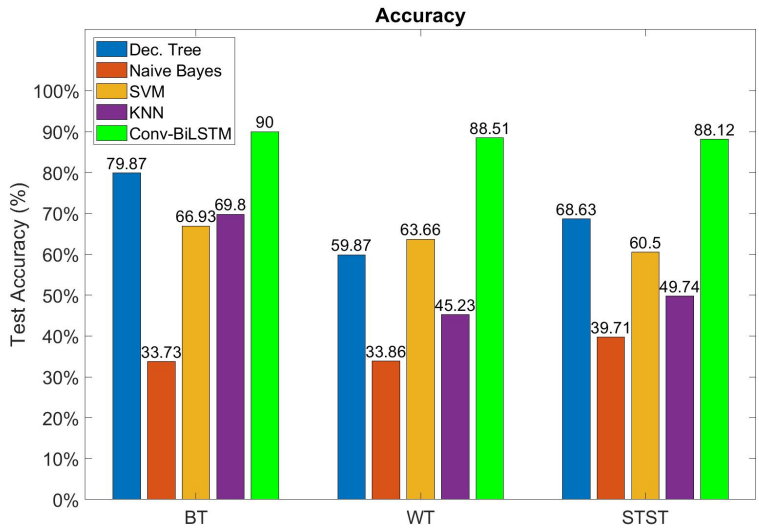

Figure 4.15: Percentages of weighted mean Accuracy of the traditional Machine Learning classifiers compared with the Classification Conv-BiLSTM network for each SPPB test (BT, WT, and STST).

For the sake of completeness, several machine learning classifiers have been also considered, namely Decision Tree, Naive Bayes, SVM, and KNN classifiers [154]. Also in this case, a 5-fold cross-validation technique has been applied during the learning phase, while the configuration with the maximum accuracy has been selected for the test phase.

Figure 4.15 shows that the considered traditional machine learning approaches perform worse than the Conv-BiLSTM approach, thus proving the need for a deep model. Only Decision Tree has good accuracy performance for what concerns the BT. In this case, the Decision Tree sets its first levels to find the end of the test, setting close-to-zero thresholds at specific samples of the input feature vectors. Accordingly, the tree classifies the input focusing only on the duration of the test, i.e. how long the subject stands in the same position. The performance of the Decision Tree emphasizes how the duration of the exercise is also an implicit feature that this specific model uses. This quantitative analysis allows for a good accuracy value compared to the other standard models. However, this is still below the best accuracy

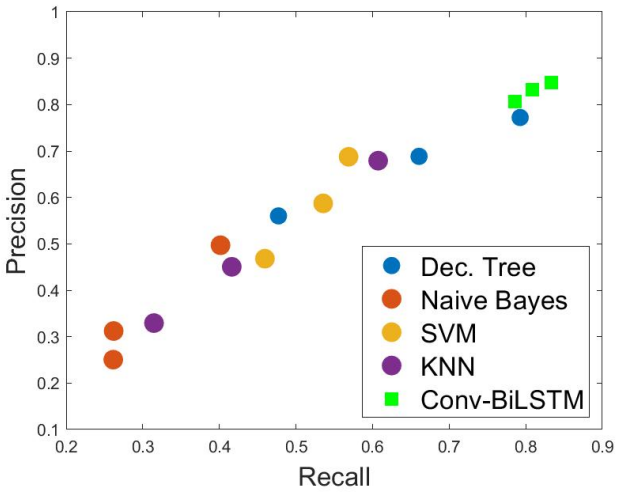

Figure 4.16: Weighted mean values of Precision vs Recall of the traditional neural network architectures (Dec. Tree, Naive Bayes, SVM, KNN) compared with those of the classification model Conv-BiLSTM. The three markers for each classifier refer to the three SPPB tests.

achievable by deep models, which even consider the quality of execution. This point is much more significant for the WT and the STST, whose classification is much more dependent on the quality of the execution. For this reason, the classification accuracies of WT and STST of the Decision Tree are 28.64% and 19.49% lower than the corresponding values, out of the Conv-BiLSTM model. In Figure 4.16, the plot of the weighted averages of Precision vs Recall leads to the same conclusion as for the accuracy plot: the Conv-BiLSTM keeps the best performances for the three tests. The Decision Tree classifier has a comparable value only for the BT.

4.4.3 Regression

As presented in Section 4.3.2, the four deep neural networks have been designed also for regression tasks. To have a proper comparison between Classification and Regression networks, the Root Mean Square Error (RMSE) has been calculated. In classification output, these metrics are computed between discrete integers (expected classes and predicted ones). In contrast, for regression models, they are computed between discrete expected classes and predicted regression values R. RMSEs are summarized in Figure 4.17.

As first remark, the best result of the regression, i.e. the lowest RMSEs, is achieved with the Conv-BiLSTM architecture. This result is in agreement with what has been found for classification, since the use of the preliminary convolutional block can help the BiLSTM net-

Figure 4.17: Graphs representing the RMSEs values from the Regression Models (orange) vs the Classification Models (blue). Each plot is relative to the exercises within the SPPB test, i.e. BT, WT, and STST.

work by aggregating features at each frame. At the same time, regression networks always perform worse than their classification counterparts. In principle, this result can be unexpected, as treating the mobility assessment to produce continuous scores rather than discrete ones should prevent heavy misclassifications, e.g. from class 3 to 0, and give results of higher quality. All these considerations would be verified if the training set was actually designed with examples from a regression scenario. However, the initial labeling of the dataset, made by physiotherapists in discrete classes, reduces the ability of regressive networks to create successful models.

4.4.4 Conv-BiLSTM Classifier: in-depth analysis

This subsection presents a detailed analysis of the performance of the Conv-BiLSTM network architecture for each class of SPPB tests. These experiments help understand the practical ability of the proposed deep architecture to recognize the classes of people that need particular attention.

Tables 4.4, 4.5 and 4.6 list the Accuracy, Precision, Recall and the resulting weighted averages for each SPPB test and for each class. These results demonstrate that the proposed Conv-BiLSTM, in most cases, can predict the correct class of mobility level for each SPPB test (BT, WT, and STST). With more detail, the weighted mean accuracy is 90% in the case of BT, while it is 88.51% and 88.12% for WT and STST, respectively.

Table 4.4: Accuracy, Precision, and Recall of the Conv-BiLSTM classifier for each output class in the case of Balance Test.

BT	Accuracy	Precision	Recall
Class 0	96.92%	87.18%	66.67%
Class 1	90.48%	85.42%	80.39%
Class 2	90.76%	62.68%	87.25%
Class 3	88.52%	90.32%	86.27%
Weighted Mean	90.00%	84.75%	83.33%

Table 4.5: Accuracy, Precision, and Recall of the Conv-BiLSTM classifier for each output class in the case of Walking Test.

WT	Accuracy	Precision	Recall
Class 0	85.78%	73.51%	97.06%
Class 1	87.38%	93.56%	71.24%
Class 2	94.49%	89.04%	63.73%
Class 3	94.12%	75.47%	78.43%
Weighted Mean	88.51%	83.22%	80.88%

Table 4.6: Accuracy, Precision, and Recall of the Conv-BiLSTM classifier for each output class in the case of Sit To Stand Test.

STST	Accuracy	Precision	Recall
Class 0	94.53%	73.74%	73.00%
Class 1	88.44%	89.84%	81.95%
Class 2	85.04%	80.38%	69.61%
Class 3	89.27%	60.61%	91.50%
Weighted Mean	88.12%	80.57%	78.64%

Precision, also called positive predictive value, measures how many predictions of a class are true. In our context, it proves the ability of the system to assign the correct mobility level to the person. The weighted averages of Precision are 84.75%, 83.22%, and 80.57%, for the BT, WT, and STST, respectively.

Recall, also known as sensitivity, measures the ability to recognize samples of a specific class. This aspect is fundamental in our experimental context, as it is necessary to be confident of which people need more attention than others. The Recall values in Tables 4.4, 4.5 and 4.6 outline the good performance of the proposed classification model. More precisely, the weighted averages of Recall reach 83.33%, 80.88%, and 78.64% for BT, WT, and STST,

respectively.

It is essential to highlight that the obtained results are satisfactory in the particular healthcare context addressed in this work. A system that makes decisions emulating the decision-making ability of human experts for assessing people's mobility has been developed. It is crucial to notice that only specialized physiotherapists with specific competencies can make these evaluations. So developing such an automatic system is of great help for supporting clinicians to identify people with mobility limitations objectively.

Finally, concerning the computational costs, it is straightforward to acknowledge that a longer mobility test leads to longer videos, which require more time for training the corresponding network. In this case, the training time of the architectures for modeling the WT and the STST is higher than that for modeling the BT, although the numbers of training epochs are comparable (26, 28, and 30 for BT, WT, and STST, respectively). The same consideration is still valid for the test phase. The average times for a single video classification are 28 *ms* for the BT, 37 *ms* for the WT, and 49 *ms* for the STST. These last durations are computed on the setup described previously, exploiting the huge capabilities of a GPU implementing Nvidia CUDA drivers. However, the same classifications have been repeated on the single CPU of the same processing unit, leading to average times of 290 *ms* for the BT, 344 *ms* for the WT, and 416 *ms* for the STST. Although CPU processing takes more time than GPU processing, classification times are always much shorter than required for performing every mobility test. This paves the way for future implementations of the trained model on low-resource platforms, such as apps for mobile phones or tablets, towards a fully-integrated telehealthcare system.

4.5 Discussion

In recent years, the increase in the elderly population and the need to support diagnostic issues in retirement residences have brought considerable interest in developing telehealthcare systems. This work deals with the complex problem of the motion ability evaluation of older people. In literature, several automatic systems, both invasive (based on wearable sensors) and noninvasive, have been proposed to measure specific parameters related to gait or posture. On the contrary, few works explore only partially the analysis of people's movement. Currently, the evaluation of motion abilities is carried out by experienced medical personnel who observe people performing some mobility tests through a defined protocol and evaluate their mobility level according to defined rank. Despite the high professionalism of physiotherapists, this evaluation can be affected by their subjectivity, confidence, and experience. Therefore, the development of automatic systems can significantly help medical personnel improve diagnostic accuracy and the elderly themselves by limiting the number of visits to health clinics.

The main contributions of this work have been:

- The feasibility of developing an automated system to assess the motion skills of older people while performing a specific mobility test protocol has been demonstrated. The proposed system is noninvasive for people. It consists of low-cost visual cameras that record videos of people performing the tests and a complete processing framework that extracts significant features and builds models that classify the test executions emulating the complex decision process of physiotherapists.

- The proposed system has been validated using real video data acquired in two nursing institutes hosting elderly people, both healthy and affected by neurodegenerative diseases. Significant features have been extracted from the skeletal representations of the subjects observed. To increase the dataset dimensionality, a data augmentation technique has been applied to the extracted skeletons. Finally, the proposed deep neural network, based on BiLSTM, has been used to classify the observed people's mobility levels. Numerical experiments have been analyzed quantitatively in terms of Accuracy, Precision, and Recall metrics, demonstrating the improvement of results due to preliminary processing made by a convolutional block.

- Several machine learning methods for automatically classifying the motion functionalities of older adults have been compared. Once again, the deep neural network classifier with convolutional filters and a BiLSTM model provides the best performance among all the implemented techniques.

- The proposed deep neural network architectures have also been tuned to perform regression. The results show an improvement in RMSE due to convolutional blocks. However, the input labels (discrete classes) do not constitute a significant dataset for training regression models, which perform worse than the Conv-BiLSTM designed for the classification of patients performing the SPPB test.

The proposed system reveals the mobility levels of people, supporting clinicians to timely detect mobility anomalies, and preventing dangerous conditions such as falls or worsening health conditions. Furthermore, the development of mobile apps that collect video of people performing mobility tests, extract data, and transmit them to medical staff, could provide good support to increase telehealthcare functionalities. Telehealthcare systems will be a valid instrument for remote monitoring of older adults often unwilling to visit health clinics periodically, reducing time, costs, and efforts.

5

Deep Learning methodologies for Human Action Segmentation in manufacturing scenarios

5.1 Introduction

In this Chapter, the state-of-the-art models for temporal action segmentation on the novel Human Action Multi-Modal Monitoring in Manufacturing (HA4M) dataset [98] are evaluated, analyzing which features are more suitable for realizing a system that recognizes the actions of an operator.

More specifically, five state-of-the-art architectures, namely MS-TCN [74], MS-TCN++ [92], BCN [75], C2F-TCN [93], and ASFormer [155], are trained on different input modalities including skeletal data and video features extracted by the Inflated 3D ConvNet model (I3D) [156] from RGB and/or Depth data. As HA4M includes videos of operators performing an industrial task in different manufacturing scenarios, the dataset has been split first considering a Cross-Subject approach, then a Cross-Location approach. Furthermore, the model trained on the Cross-Subject splitting has also been tested on a new set of data, which considers different locations where new subjects perform the same task as in HA4M. Furthermore, a semi-supervised learning setting is proposed.

The main contribution of this work is threefold:

- Is is shown that the I3D model is very effective in extracting features not only from RGB data, but also from Depth data, and from the frames where the RGB data have been aligned to the Depth data;
- The effectiveness of the features extracted from the novel HA4M dataset for addressing temporal action segmentation in manufacturing scenarios is evaluated, considering multiple deep-learning models at the state-of-the-art. Different splittings for training and testing sets have been selected, to assess the generalization of both models and data;
- A new set of videos to test the trained models is considered, which is recorded with the same standards of the HA4M dataset, in different simulated industrial locations and with different subjects. The new set has also been assessed by performing the training of the models with a semi-supervised learning approach, first considering 60%, and then 30% of the initial training set as labeled.

The results demonstrate the validity of HA4M, as the models selected manage to properly segment the actions of the assembly task. Such outcome can lead to a new perspective in developing systems for HRI and HRC, as observing the movements of human operators while performing an assembly task is fundamental to detect their capabilities, particularly in collaborative tasks with robots.

The Chapter is structured as follows. Section 5.2 describes the evaluation protocol, including a description of the newly collected data, the feature extraction, and the Cross-Subject and Cross-Location splittings. In Section 5.3, the outcomes of all the experiments are presented and discussed. Finally, Section 5.4 draws the discussion.

5.2 Methodological Approach

In the HA4M dataset [98], 41 subjects are recorded using an Azure Kinect [95] while they assemble an EGT in two realistic manufacturing scenarios. For sake of clarity, the term Loc1 refers to the first location where 22 subjects perform the assembly task of a white EGT over a black tablecloth. The term Loc2 refers to the second location where the remainder 19 subjects perform the assembly task of black and white EGTs over white and black tablecloths, respectively. To assemble the industrial object, the subjects conduct 12 actions, some of them more than once. Frames that do not show any of the 12 assembly-related actions are labeled by the class "0".

The present work follows the pipeline depicted in Figure 5.1. RGB, RGBA, Depth and Skeletal information have been gathered from the HA4M dataset, as it will be described in

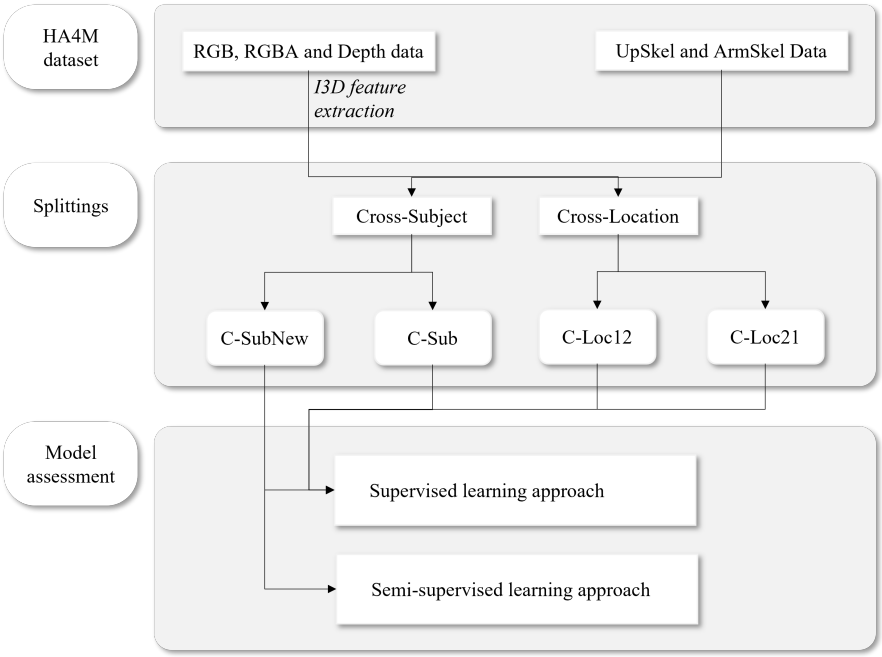

Figure 5.1: Pipeline of the proposed analysis. Cross-Subject refers to the train-test splitting obtained considering different subjects from the train and test sets. In C-Sub, the test set is within HA4M, while in C-SubNew, the test set is from a new set of data collected. Cross-Location refers to the train-test splitting obtained considering Loc1 for the train set and Loc2 for the test set (C-Loc12), and vice versa (C-Loc21).

Section 5.2.1. Such features have then been split following a Cross-Subject (C-Sub) and Cross-Location sorting (C-Loc12 and C-Loc21), which will be described in Section 5.2.2. State-of-the-art deep learning models for temporal action segmentation have been trained using a supervised learning approach. The best ones have been tested considering a Cross-Subject splitting with new data gathered from new videos (C-SubNew). This splitting is also used to evaluate a semi-supervised learning procedure, which will be described in Section 5.2.3.

5.2.1 Feature Extraction

The I3D model [156] has been used to extract features from RGB, RGBA and Depth frames. Intending to allow the data to be as much heterogeneous as possible, RGB, RGBA, and Depth sets of features have been mixed up, creating 4 additional sets of features, i.e. [RGB + RGBA], [RGB + Depth], [RGBA + Depth], [RGB + RGBA + Depth]. As for the skeletal data, two sets of features have been taken into account, as depicted in Figure 5.2. The first set of features has been labeled as ArmSkel, which contains the 3D coordinates of the joints

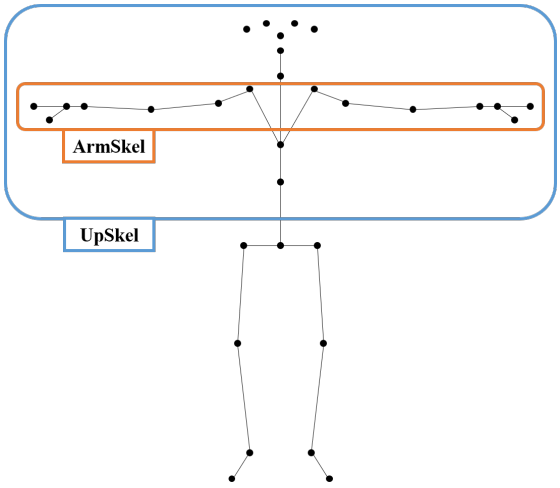

Figure 5.2: Representation of the skeletal joints extracted from the Azure Kinect. The blue box and the orange box represent the joints considered for the features UpSkel and ArmSkel, respectively.

composing the arms of each skeleton. The second set of features has been labeled UpSkel, and considers all the 3D coordinates of the upper-body joints of each subject.

The extracted sets of features can be represented as follows, where F_i^{I3D} corresponds to the sets of features extracted using the I3D model, while F_i^{Sk} corresponds to the sets of features extracted from the skeletal data:

$$F_i^{I3D} = \begin{bmatrix} f_{1,1} & f_{2,1} & \cdots & f_{N_i,1} \\ f_{1,2} & f_{2,2} & \cdots & f_{N_i,2} \\ \cdots & \cdots & \cdots & \cdots \\ f_{1,D} & f_{2,D} & \cdots & f_{N_i,D} \end{bmatrix}, F_i^{Sk} = \begin{bmatrix} f_{1,1}^{x_1} & f_{2,1}^{x_1} & \cdots & f_{N_i,1}^{x_1} \\ f_{1,1}^{y_1} & f_{2,2}^{y_1} & \cdots & f_{N_i,1}^{y_1} \\ \cdots & \cdots & \cdots & \cdots \\ f_{1,D}^{z_D} & f_{2,D}^{z_D} & \cdots & f_{N_i,D}^{z_D} \end{bmatrix} \tag{5.1}$$

with $i = \{1, 2, \ldots, I\}$. I represents the number of videos, while N_i represents the number of frames of the i-th video. D represents the number of features extracted from each video. Such value differs depending on which type of data is considered.

5.2.2 Dataset Splittings

The I videos have been split considering a Cross-Subject and a Cross-Location approach. In the C-Sub splitting, the training set $I_{train}^{\text{C-Sub}}$ with videos and the testing set $I_{test}^{\text{C-Sub}}$ with videos have been separated by subjects. This protocol allows to evaluate the generalization performance of a model across subjects but not across locations.

Cross-Location uses the splitting based on the two different locations Loc1 and Loc2. As a result, two Cross-Location splittings have been obtained, namely C-Loc12 and C-Loc21.

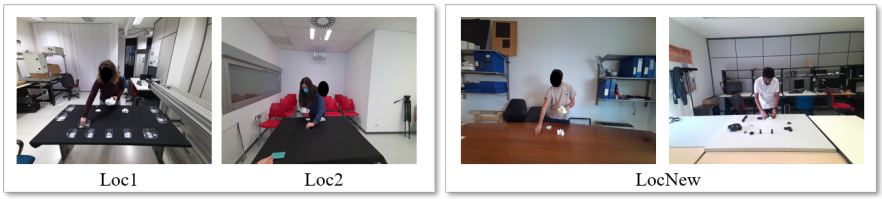

Figure 5.3: Images representing the locations considered, namely Loc1 and Loc2, which are included in the HA4M dataset, and LocNew, which represents the new locations for the extended data gathered.

The training set of C-Loc12 is represented as $I^{\text{C-Loc1}}$, while the training set of C-Loc21 is represented as $I^{\text{C-Loc2}}$, and vice versa. Splitting the dataset according to different locations, in addition to different subjects, allows further analysis in assessing the generalization performance of a model across locations. Since each location differs in lighting conditions, background clutter, and other factors, the Cross-Location protocols are very challenging.

5.2.3 New Data Collection

While the HA4M dataset consists of videos of 41 different subjects, the dataset has been extended by collecting additional test videos of subjects, which form the new testing set I_{test}^{LocNew}. The videos were collected using an Azure Kinect following the same standard as HA4M. Figure 5.3 shows the different locations considered for the analyzed data. For evaluation on the new dataset, the models are trained with the $I_{train}^{\text{C-Sub}}$ training set. The additional Cross-Subject protocol has been denoted by C-SubNew.

5.2.4 Semi-supervised Learning

Besides evaluating the approaches using fully-supervised learning, a semi-supervised protocol was also proposed. In this setting, the training data has been further divided into two subsets. The first subset contains the labeled training videos and the second subset contains training videos without any annotations. Two cases have been considered. In the first case, 60% of the training videos are labeled, while in the second case, only 30% are labeled. For semi-supervised learning, the models are first trained on the labeled videos and then predict the labels on the unlabeled training videos. Subsequently, the newly-labeled data are added to the already-labeled data to train the model.

5.2.5 Deep Learning models Selection

The models considered for the evaluation of the HA4M dataset represent the state-of-the-art in the assessment of temporal action segmentation tasks [74, 92, 75, 93, 155]. Temporal Con-

Table 5.1: Definition of the models from the literature considered for the analysis.

Approach	Definition of the Architecture
MS-TCN[74]	Multi-Stage hierarchical temporal convolutional network. Each stage is composed of multiple temporal convolutional layers with 1D dilated convolutions.
MS-TCN++[92]	Improvement of the MS-TCN model. A Dual Dilated layer is introduced, which combines convolutions with small and large dilation factors.
BCN[75]	Improves the performance of Multi-Stage segmentation models by using a cascading paradigm to enable the model to have adaptive receptive fields and more confident predictions for ambiguous frames.
C2F-TCN[93]	Temporal encoder-decoder model with a coarse-to-fine ensemble of decoding layers, which aims at resolving the problem of sequence fragmentation.
ASFormer[155]	Transformer-based model with an encoder and several decoders, which combine temporal convolutions and local attention blocks.

volutional Networks (TCNs) capture long range dependencies using temporal convolutional filters. In particular, Encoder-Decoder TCNs, such as C2F-TCN [93], are intended to shrink and expand the temporal resolution with layer-wise pooling and upsampling, while Multi-Stage architectures (MS-TCNs), such as [74, 92], expand the temporal receptive field with constant temporal resolution using progressively larger dilated convolutions. Multi-Stage segmentation algorithms have also been used with a Barrier Generation module [75], which enables the later stages to focus on ambiguous frames by introducing a pooling operator to smooth noisy boundary predictions. Finally, temporal modeling has been recently addressed using Transformer architectures [157]. The ASFormer model [155] combines dilated temporal convolutions with local transformer blocks. The used approaches for temporal action segmentation are summarized in Table 5.1. While these models have been previously applied to pre-computed RGB features like I3D, also other features have been investigated, including depth and skeletal data, as described in Section 5.2.1.

5.2.6 Evaluation Metrics

The most common evaluation metrics for temporal action segmentation are frame-wise Accuracy, Segmentation Edit Score, and F1-Score. The Accuracy is defined as follows:

$$\text{Accuracy} = \frac{\hat{N}_{correct}}{N} \tag{5.2}$$

where N represents the number of frames of all considered videos, while $\hat{N}_{correct}$ represents the number of all correctly predicted frames. Accuracy is widely used for the evaluation of temporal action segmentation approaches. However, it is not reliable when the action frame distribution is not well-balanced, which happens in most datasets, and it does not penalize over-fragmentation. It is thus not suitable for manufacturing tasks, but it is still reported due to consistency with previous works.

The Segmentation Edit Score, also known as Edit Score, is computed starting from the Levenshtein distance [158], and it quantifies how similar two sequences are to each other. Such metric is obtained starting from the accumulative distance value, defined as follows:

$$lev(\hat{s}, s) = \begin{cases} 0 & \text{if } \hat{s} = s = 0 \\ \hat{s} & \text{if } s = 0 \text{ and } \hat{s} > 0 \\ s & \text{if } \hat{s} = 0 \text{ and } s > 0 \\ min \begin{cases} lev(\hat{s}-1, s) + 1 \\ lev(\hat{s}, s-1) + 1 \\ lev(\hat{s}-1, s-1) + 1(\hat{S}[\hat{s}] \neq S[s]) \end{cases} & \text{otherwise} \end{cases} \tag{5.3}$$

$\hat{S}$ and S denote the ordered list of predicted and ground truth action segments, respectively, while $\hat{s}$ and s represent their indices. The indicator function 1 denotes the cost for a substitution. To obtain the Edit Score, the maximum length of the ground truth and the corresponding predicted sequences are normalized and computed as:

$$\text{Edit} = \frac{1 - lev(|\hat{S}|, |S|)}{\max(|\hat{S}|, |S|)} \cdot 100 \tag{5.4}$$

The Edit Score is frequently used for assessing temporal action segmentation approaches, as it measures how well the model predicts the ordering of an action sequence, but it does not measure the duration and timing of the predicted actions.

The F1-score is obtained by comparing the Intersection over Union (IoU) of each predicted segment with respect to the corresponding ground truth segment. A segment is correctly predicted, denoted as True Positive (TP), if IoU is above a specified threshold. Predic-

Table 5.2: Number of videos considered for each splitting.

Type of Splitting	Name	Train Set	Test Set
Cross-Subject	C-Sub	152	53
	C-SubNew	152	10
Cross-Location	C-Loc12	109	96
	C-Loc21	96	109

tions that do not match any ground truth segment are False Positives (FP) and missed action segments are False Negatives (FN). The F1-score is then computed by

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{5.5}$$

5.3 Experiments

For evaluation, all 205 videos within the HA4M dataset [98] are used. In the Cross-Subject (C-Sub) splitting, the $I_{train}^{\text{C-Sub}}$ set includes 152 videos, while the $I_{test}^{\text{C-Sub}}$ set includes 53 videos. As for the Cross-Location splitting, the $I^{\text{C-Loc1}}$ set is composed of 96 videos, while the $I^{\text{C-Loc2}}$ set includes 109 videos. The models trained using the C-Sub training set have also been evaluated on a newly collected testing set I_{test}^{LocNew} composed of 10 videos, considering fully-supervised and semi-supervised learning approaches. All the splittings are summarized in Table 5.2. 9 sets of features have been evaluated, which have been extracted from each video and are summarized in Table 5.3.

The models were trained using a machine equipped with an NVIDIA GeForce GTX 1080 Ti. The analysis of training execution times reveals that, on average, the training durations were 3 hours for MS-TCN, 4 hours for MS-TCN++, 6 hours for BCN, 8 hours for C2F-TCN, and 13 hours for ASFormer.

The inference analysis has been focused mainly on the F1-score value at the highest threshold, which is considered the most complete metric for assessing models for temporal action segmentation. As discussed in Section 5.2.6, the F1-score takes into account Precision and Recall, which are both highly important in evaluating action segmentation models. Furthermore, the use of overlapping thresholds allows for a more nuanced evaluation of the performance of the models. Such assessment can be of critical importance in models for action segmentation, as the boundaries of action segments may not be precisely defined, and different levels of overlapping actions may be acceptable. The thresholds have been set to $\tau = \{60, 70, 80\}$ for Cross-Subject evaluation and to $\tau = \{10, 25, 50\}$ for Cross-Subject evaluation since this task is much more challenging.

Table 5.3: Definition of the sets of features used for the evaluation.

Set of Features	Definition
RGB RGBA Depth	Sets of features extracted using the I3D model, resulting in N_i features of dimension 1024.
RGB + RGBA RGB + Depth RGBA + Depth	Sets of features extracted using the I3D model, resulting in N_i features of dimension 2048.
RGB + RGBA + Depth	Set of features extracted using the I3D model, resulting in N_i features of dimension 3072.
UpSkel	3D coordinates of 23 skeletal joints representing arms, hands, chest and head, resulting in N_i features of dimension 69.
ArmSkel	3D coordinates of 14 skeletal joints representing arms and hands, resulting in N_i features of dimension 42.

In the following subsections, the outcomes of the performed analysis are presented and discussed.

5.3.1 Cross-Subject Evaluation

In Figure 5.4, the F1-score results for each model trained with I3D and Skeletal features are presented in bar plots. Overall, the model that performs best is ASFormer trained on the I3D features, specifically [RGB + RGBA + Depth]. For this particular model, the Accuracy reached 95.77%, the Edit Score reached 97.69%, and the F1@{60, 70, 80} reached 94.72%, 93.14%, and 89.24%, respectively. Surprisingly, ASFormer performs worst for the skeletal data and the best performance has been obtained by MS-TCN++ trained on UpSkel. This model reached an Accuracy value of 94.92%, an Edit Score of 94.28%, and a F1@{60, 70, 80} at 92.57%, 88.57%, and 81.85%, respectively. This indicates that the transformer-based ASFormer tends to overfit on the low-dimensional Skeletal features, while the methods based on temporal convolutional networks are less prone to overfitting. However, the results for different thresholds 60, 70, and 80 are compared, it has been observed that ASFormer mainly struggles to infer accurate action segment boundaries in case of Skeletal features since the gap to the other methods increases as the threshold increases. This shows that there is not a single approach that performs best for all input modalities.

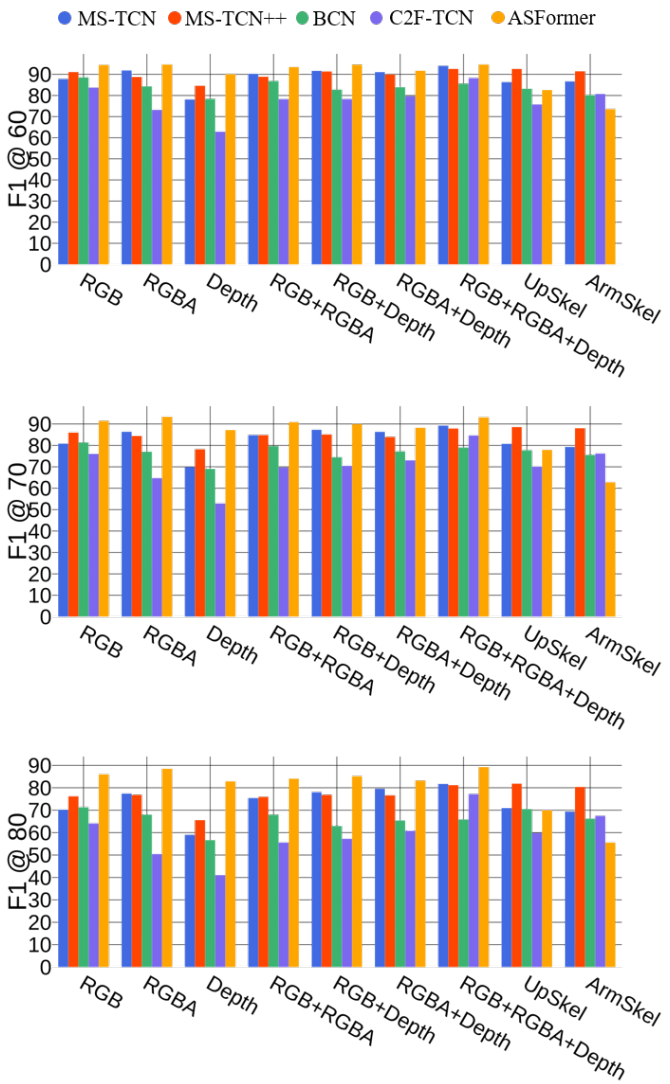

Figure 5.4: F1@$\tau = \{60, 70, 80\}$ results obtained by the analyzed models trained with skeletal and I3D features, considering C-Sub splitting.

The results clearly show that, overall, the I3D features give the highest performances when all the sets are considered, particularly the ones including RGB information. As for the skeleton data, the models trained with UpSkel seem to be more effective in segmenting the actions. Furthermore, aside from the ASFormer architecture, there is not a considerable gap between the outcomes of I3D and Skeletal features, even though the firsts slightly outperform the seconds. Among the features, the Depth features alone perform worst for all methods ex-

cept ASFormer, which performs worst for the UpSkel features. While the results show that RGB information is very useful, Skeletal features preserve privacy and the results show that MS-TCN++ in combination with UpSkel features provides a privacy-friendly approach that achieves only a slightly lower accuracy than ASFormer with RGB features.

5.3.2 Cross-Location Evaluation

The Cross-Location splitting is much more challenging than the Cross-Subject splitting. This setting is also very interesting since other datasets, such as Assembly101 [91], have been recorded at a single location and do not allow to evaluate how the features and methods generalize to other locations. The F1-score values for all models trained with I3D and Skeletal features are presented in Figures 5.5 and 5.5. More precisely, Figure 5.5 shows three bar plots representing F1@{10, 25, 50} results for each model trained with each set of features considering the C-Loc12 splitting. Figure 5.5 presents the same outcomes for the models considering the C-Loc21 splitting.

Observing all the outcomes, it is clear that Depth and Skeletal features generalize best across locations. This outcome is expected since, due to the Cross-Location splitting, the RGB features result in an overfitting on the environment where the assembly task occurred. In case of Depth and Skeletal features, the environment information is weak or even missing and the learning process thus focuses on the task itself.

Having a closer look at Figure 5.5, it is possible to notice that ASFormer performs best, considering both I3D and Skeletal features. The ASFormer model trained using Upkel as input reached an Accuracy of 61.29%, an Edit Score of 74.13%, and an F1@{10, 25, 50} at 69.32%, 60.46%, and 41.33%, respectively. While ASFormer also performs much worse when I3D features instead of Skeletal features are used, it still achieves the best results among the methods, particularly when considering Depth features. In this case, the Accuracy score reached 48.40%, the Edit score reached 60.38%, and the F1@{10, 25, 50} reached 52.55%, 45.49%, and 26.09%, respectively.

Figure 5.6 reports the results for the C-Loc21 splitting. It shows that the ASFormer architecture fed with UpSkel features as inputs outperforms all others methods. With this configuration, the model reached an Accuracy of 84.38%, an Edit Score of 77.39%, and an F1@{10, 25, 50} of 83.71%, 82.45%, and 72.56%, respectively. The results show that the C-Loc12 splitting contains more difficult videos in the test set than the C-Loc21 splitting. Also among the I3D features, the ASFormer model trained using [RGBA + Depth] features performed best. It gave an Accuracy of 50.82%, an Edit Score of 57.86%, and an F1@{10, 25, 50} of 43.91%, 38.10%, and 26.55%, respectively.

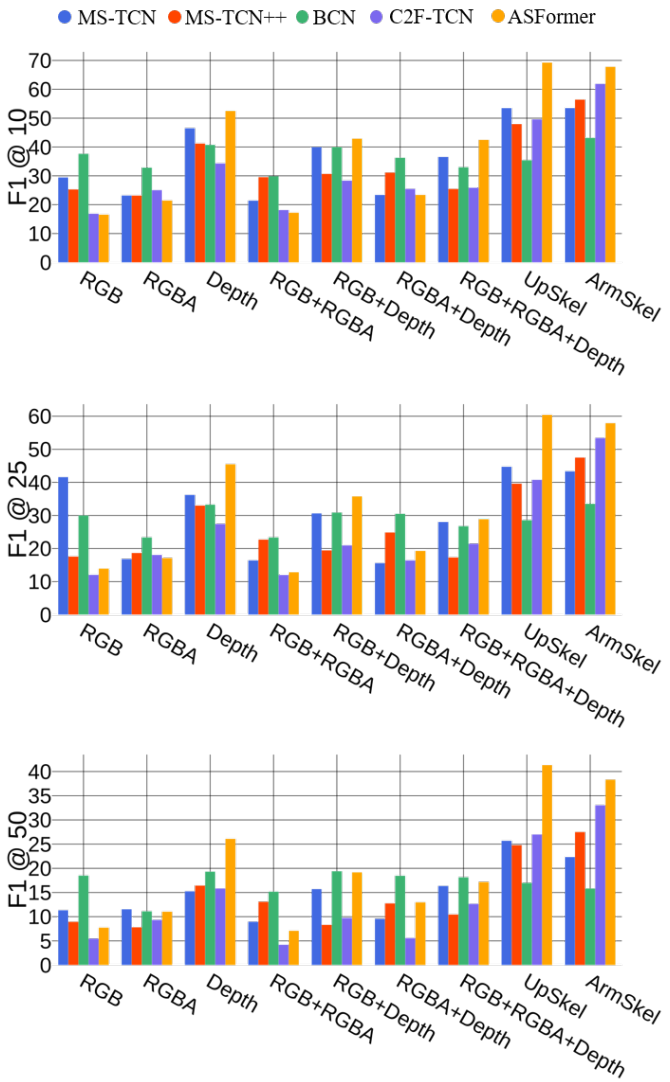

Figure 5.5: F1@$\tau = \{10, 25, 50\}$ results obtained by the analyzed models trained with skeletal and I3D features, considering C-Loc12 splitting.

5.3.3 New Data Evaluation

The Cross-Subject analysis showed the efficiency of MS-TCN++ and ASFormer architectures in properly evaluating the assembly task performed in HA4M when trained using Skeletal or I3D features, respectively. To further assess the mentioned models, in the new acquisition campaign the assembling task has been recorded in 2 new simulated manufacturing locations, collecting a total of $I_{test}^{new} = 8$ videos performed by 3 new subjects.

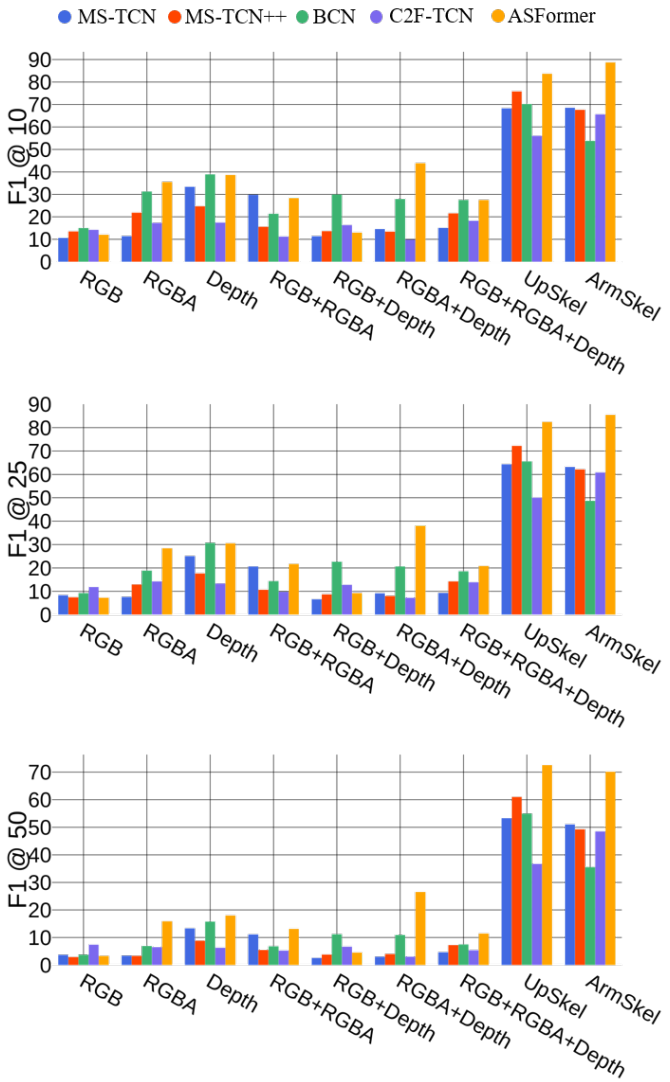

Figure 5.6: F1@$\tau = \{10, 25, 50\}$ results obtained by the analyzed models trained with skeletal and I3D features, considering C-Loc21 splitting.

The new set of data has been evaluated first considering a fully-supervised learning approach, thus testing the set on the models trained in the Cross-Subject splitting analysis. Then, the data has been assessed considering a semi-supervised learning approach, as depicted in Figure 5.7. To this end, the $I_{train}^{\text{C-Sub}}$ set has been divided first into 60/40%, then into 30/70%, where 60% (30%) of the videos are annotated and the remaining 40% (70%) are unlabeled. For semi-supervised learning, the models are first trained on the labeled videos and then used to

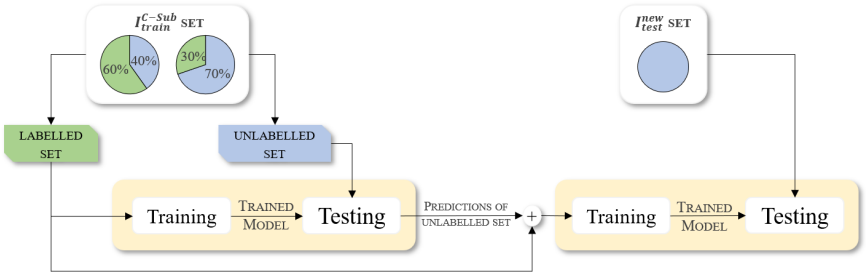

Figure 5.7: Representative scheme of the semi-supervised learning approach. The C-Sub training set has been split first into 60/40%, then into 30/70%. For both types of splitting, the green part in the scheme has been considered labeled and used as training set. Then, the trained model predicts the labels on the blue part of the original training set, which has been considered unlabeled. The estimated labels are then used as ground-truth labels. With this new labeling, both green and blue sets have been used as unique training sets. The newly trained model has then been evaluated on the testing set, i.e., I_{test}^{LocNew}.

infer the temporal segmentation on the unlabeled training videos. The models have then been retrained, adding the newly-labeled videos. Finally, the trained models are evaluated on the new data I_{test}^{LocNew}.

Fully-supervised learning approach

Figure 5.8 shows the results of the MS-TCN++ and ASFormer models trained with $I_{train}^{\text{C-Sub}}$ and tested on I_{test}^{LocNew}. As expected, the MS-TCN++ architecture performed best when trained with Skeletal features, i.e., UpSkel. More precisely, this model reached an Accuracy of 87.48%, an Edit Score of 94.57%, and an F1@{10, 25, 50} of 92.09%, 91.16%, and 83.72%, respectively. The model trained with I3D features gave poor results, and the best performing has been trained considering the [RGBA + Depth] set of features. Here, the Accuracy reached 62.27%, the Edit Score reached 62.96%, and the F1@{10, 25, 50} reached 63.70%, 58.87%, and 44.35%, respectively.

Similarly, the ASFormer architecture gave the best outcomes when trained with Skeletal features, although the RGBA set of features returned satisfying results. In this case, the Accuracy reached 75.44%, the Edit Score reached 82.99%, and the F1 @ {10, 25, 50} reached 82.72%, 79.09, and 60.90%, respectively. On the other hand, the model trained using the UpSkel features achieved an Accuracy of 88.85%, an Edit Score of 85.49%, and an F1@{10, 25, 50} of 88.78%, 87.85%, and 80.37%, respectively. In contrast to $I_{test}^{\text{C-Sub}}$, the best performances are obtained by Skeletal features. While $I_{test}^{\text{C-Sub}}$ is very similar to $I_{train}^{\text{C-Sub}}$ since the data has been recorded at the same location and time, the new dataset I_{test}^{LocNew} presents a more realistic scenario since the recording differs more from the training set.

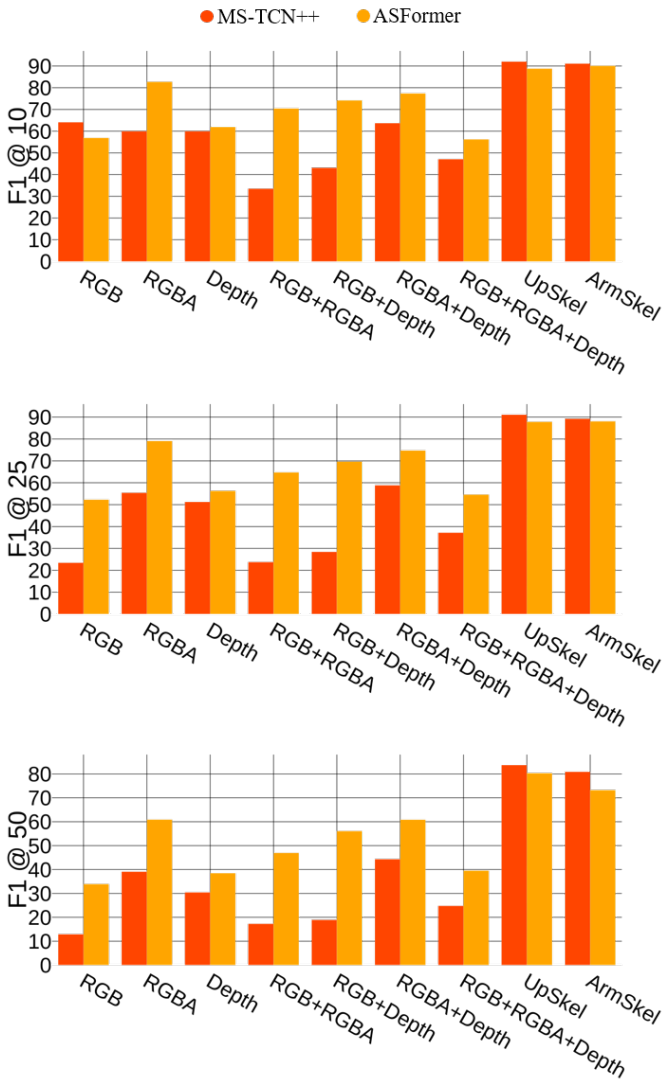

Figure 5.8: F1@$\tau = \{10, 25, 50\}$) results obtained by the MS-TCN++ and the ASFormer models trained on $I_{train}^{\text{C-Sub}}$ with Skeletal and I3D features, and tested using the I_{test}^{LocNew} set.

Semi-supervised learning approach

Table 5.4 presents the results of the best models trained with Skeletal or I3D features, comparing semi-supervised learning with fully-supervised learning. More precisely, the first part of the table shows the results of the models trained using 60% of the $I_{train}^{\text{C-Sub}}$ set as labeled. Contrariwise to the I3D features, UpSkel gave particularly good results, reaching for the MS-TCN++ architecture an Accuracy of 87.00%, an Edit Score of 89.74%, and an F1 @{10, 25,

Table 5.4: Comparison of the fully-supervised approach (100%) and different levels of semi-supervised learning, i.e., 30% and 60% of labeled videos. The table shows the best results (according to F1@$\tau = 50$) obtained by the MS-TCN++ and the ASFormer models trained with Skeleton or I3D features from $I^{\text{C-Sub}}_{train}$, and tested using the I^{LocNew}_{test} set.

Models	Features	Acc	Edit	F1 @ {10, 25, 50}		
100%						
MS-TCN++	RGBA	53.90	58.25	60.00	55.45	39.09
	RGBA+Depth	62.27	62.96	63.70	58.87	44.35
	UpSkel	87.48	94.57	92.09	91.16	83.72
	ArmSkel	86.02	94.13	91.16	89.30	80.93
ASFormer	RGBA	75.44	82.99	82.72	79.09	60.90
	RGBA+Depth	73.58	83.27	77.39	74.78	60.86
	UpSkel	88.85	85.49	88.78	87.85	80.37
	ArmSkel	80.97	91.96	90.09	88.11	73.26
60%						
MS-TCN++	RGBA	62.55	69.91	62.65	55.42	39.35
	RGBA+Depth	59.01	66.01	66.66	57.14	30.31
	UpSkel	87.00	89.74	89.77	89.70	83.55
	ArmSkel	83.85	92.31	88.78	87.85	76.63
ASFormer	RGBA	74.83	79.76	78.81	76.27	64.40
	RGBA+Depth	68.32	79.51	72.24	66.96	50.22
	UpSkel	84.00	84.97	87.47	84.65	74.41
	ArmSkel	77.99	91.96	93.06	86.13	70.29
30%						
MS-TCN++	RGBA	51.58	56.14	53.81	47.53	33.18
	RGBA+Depth	46.74	45.18	47.96	33.48	20.81
	UpSkel	80.55	82.71	85.71	83.92	75.89
	ArmSkel	78.32	90.40	86.79	83.01	69.81
ASFormer	RGBA	69.57	69.10	70.07	65.35	48.81
	RGBA+Depth	71.05	80.82	78.11	72.10	53.21
	UpSkel	76.18	91.96	89.10	88.11	66.33
	ArmSkel	69.57	91.96	87.12	81.18	57.42

50} of 89.77%, 89.70%, and 83.55%, respectively. The same Skeleton features used in the ASFormer architecture performed similarly, giving an Accuracy 84.00%, an Edit Score of 84.97%, and an F1 @{10, 25, 50} of 87.47%, 84.65%, and 74.41%, respectively.

The results of the second semi-supervised learning analysis, where the $I^{\text{C-Sub}}_{train}$ set has been divided considering 30% as labeled, are also shown in Table 5.4. The outcomes follow a trend

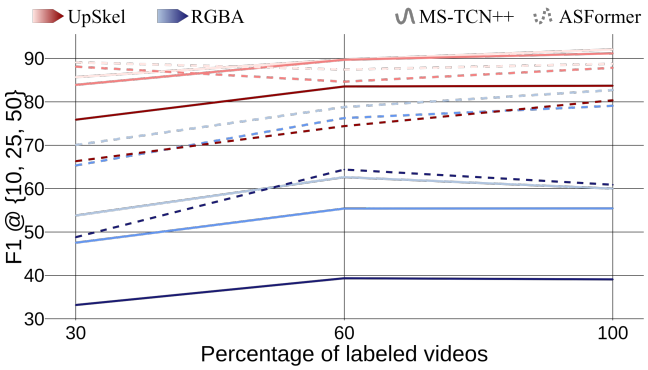

Figure 5.9: F1@$\tau = \{10, 25, 50\}$ results obtained by MS-TCN++ (solid line) and ASFormer (dashed line) trained with UpSkel (in red) and RGBA (in blue), considering the train set $I_{train}^{\text{C-Sub}}$ with 30%, 60% and 100% of labeled videos.

similar to the ones from the first semi-supervised learning analysis. In fact, the best models are the ones trained using Skeletal features, namely UpSkel. In particular, the MS-TCN++ architecture reached an Accuracy of 80.55%, and Edit Score of 82.71% and an F1@{10, 25, 50} of 85.71%, 83.92%, and 75.89%, respectively. The model built on the ASFormer architecture performed similarly, giving 76.18% as Accuracy, 91.96% as Edit Score, and 89.10%, 88.11%, and 66.33% as F1@{10, 25, 50}, respectively.

For better clarity, Figure 5.9 shows a line graph that represents the best Skeletal and I3D features used for training MS-TCN++ and ASFormer models, for each percentage of labeled videos. The graph clearly proves the superiority of the Skeletal features in training both models. Overall, the MS-TCN++ architecture proved to return the best model in successfully segmenting human actions when addressing both semi-supervised learning approaches. It is clear, though, that ASFormer performed best when the models are trained with I3D features. Such outcome is expected since the ASFormer architecture gave far the best performance in Cross-Subject analysis when trained with I3D features.

To further support the obtained outcomes, Figure 5.10 reports the qualitative results of a sample video within the I_{test}^{LocNew} set. It is clear that the MS-TCN++ model trained with I3D features, more specifically with RGBA, gave poor results in all the levels of supervision. On the contrary, the same feature used to train ASFormer returned proper segmentation. Nevertheless, the Skeletal features proved to be the best data for feeding the models, as both MS-TCN++ and ASFormer achieve high segmentation results when trained using UpSkel.

5.4 Discussion

In this work, five state-of-the-art models for temporal action segmentation have been analyzed in combination with nine different feature sets based on Skeletal features, Depth fea-

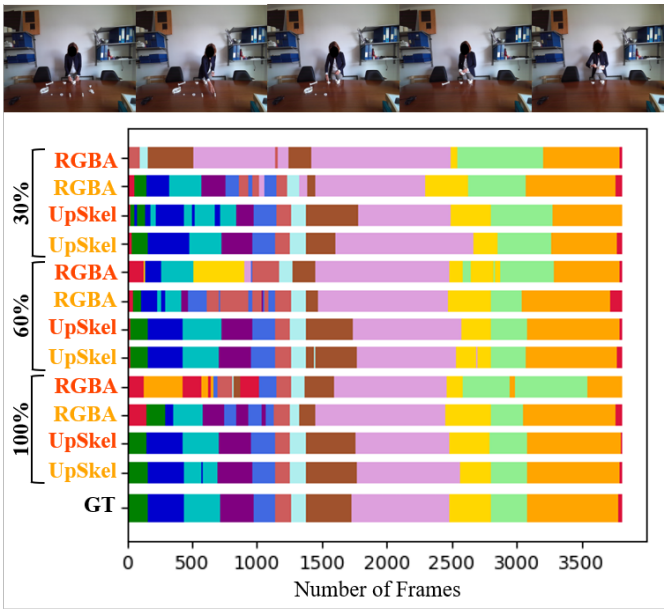

Figure 5.10: Qualitative results for the temporal action segmentation task. The graph depicts the segmentation results of a sample video within the I_{test}^{LocNew} set, compared with the ground truth (GT). Such results are obtained from the ASFormer models (in yellow) and the MS-TCN++ models (in orange), both trained with RGBA and UpSkel features at different levels of supervision, i.e. 30% 60% and 100%.

tures and RGB features, starting from the HA4M dataset. The aim of the study was to analyze which features are more suitable for realizing a system that segments the actions of an operator performing a task in manufacturing scenarios.

The data has been extended from the HA4M dataset by newly captured sequences, and defined new evaluation protocols, considering Cross-Subject and Cross-Location splittings. In the Cross-Subject analysis, all the models performed very well, managing to correctly segment the actions of the operators performing the assembly task. The MS-TCN++ architecture distinguished among the others for performing best when privacy-preserving Skeletal features are used. Such data are also faster to gather, as the Azure Kinect Body Tracking DK allows a real-time and automatic extraction of the skeletal joints. Nevertheless, the ASFormer architecture outperformed the other models when trained with I3D features. It also performed best in the Cross-Location analysis, where two splittings for training and testing sets have been considered. In this part of the study, the best models proved to be the ones trained with Skeletal features. Among the models trained with I3D features, the Depth proved to be the best set of features. While it is expected that Skeletal and Depth features show better generalization to different environments where the task is performed, it also shows that Cross-Location evaluation is very important, which is currently missing in assembly datasets

like Assembly101.

Further analyses were addressed considering a new set of videos gathered following the same standards as HA4M, extracting the same sets of features. The evaluation has been focused on the two best-performing approaches, namely MS-TCN++ and ASFormer. The models trained in the Cross-Subject analysis have been tested on the new dataset, considering also semi-supervised learning approaches. The results show the superiority of MS-TCN++ in segmenting the actions of the operators by using Skeletal features whereas ASFormer gave the best performance when trained with I3D features.

6 Conclusion

With the spreading of Industry 5.0, human action recognition and segmentation in healthcare and manufacturing have emerged as pivotal applications of computer vision and deep learning technologies. The goal of this thesis was to explore these domains, addressing the challenges of human monitoring assessment and human action segmentation. State-of-the-art vision devices such as the Microsoft Azure Kinect depth camera were deepened, using such sensors to gather data. Advanced machine learning and deep learning methodologies were developed and applied to the acquired dataset, involving real case studies.

The key contributions of the presented thesis are the following:

- A meticulous examination of camera calibration methodologies for RGB and Infrared data was provided, with or without associated depth information. Results proved the effectiveness of the proposed calibration techniques, mainly regarding the 3D procedures which outperform 2D ones.

- A comprehensive video acquisition campaign was conducted in healthcare and manufacturing environments, aiming to gather real, non-simulated data. Elderly patients undergoing specific motion exercises and operators engaged in manufacturing assembly tasks were recorded using multi-camera systems, considering both RGB and RGB-D sensors. Such data was gathered into two datasets, namely SPPB and HA4M, which represented the bedrock for the subsequent experiments and methodologies.

- The main contribution of the presented thesis lies in the development and application of machine learning and deep learning methodologies for human mobility assessment and human action segmentation, which perfectly adhere to the principles of the novel Industry 5.0.
  - In the healthcare domain, advanced algorithms were developed and applied to assess human mobility exercises, particularly among patients grappling with neurodegenerative diseases. This work illuminated the potential of technology to objectively evaluate the mobility of individuals, supporting medical diagnosis and ultimately advancing their well-being.
  - In the context of manufacturing, deep learning methodologies were applied to segment human actions in real-world manufacturing scenarios. These cutting-edge approaches were fed with various discriminant features extracted from video data, aiming to understand and respond to the actions of human operators, while emphasizing the importance of human-robot interaction and collaboration. These contributions enlightened the key role of vision devices and intelligent systems in boosting the well-being of both operators and the broader manufacturing ecosystem.

With an aging global population, it is fundamental to increase the demand to support diagnostic issues in retirement residences and beyond. The experiments conducted in this thesis involved the complicated topic of motion ability assessment for elderly individuals, enlightening a path toward telehealthcare systems. In healthcare, several systems, both invasive and non-invasive, have previously been considered to measure specific parameters related to gait and posture. However, such systems often only address partially the mobility assessment issue. Currently, the evaluation of motion abilities primarily relies on experienced medical personnel, who meticulously observe individuals while performing mobility exercises.

The first system proposed in this research performs mobility assessment avoiding any human subjectivity, lack of experience, or confidence, by offering a comprehensive, data-driven approach. By exploiting the power of deep learning methodologies, the proposed system enhances the diagnostic accuracy of healthcare professionals while protecting the elderly from the risk of physical injuries due to falls. The poor quantity of data available was overcome by creating a real dataset acquired in elderly facilities, where the patients were recorded while performing specific exercises, using a calibrated system composed of three cameras. The dataset included the skeletal data of the patient's movements, and it was further enlarged by using a data augmentation technique. Such augmentation involved applying rototranslation matrices to create new skeletal data, for a complete and comprehensive training process. The study also included an in-depth analysis of the most relevant features to extract from the skeleton

data, which led to the obtained results. The machine learning and deep learning models performed successfully, particularly the ones including the LSTM network architecture. Hence, the method has been validated to be effective in correctly predicting the motion assessment of elderly patients affected by neurodegenerative diseases. Such system is intended to reduce frequent clinic visits, providing a convenient and effective alternative for health assessment, which can be performed also remotely. As we progress into an era where telehealthcare systems become more and more crucial, the presented research lays the foundation for a future where healthcare is proactive, individualized, and embraces the well-being of elderly patients as its primary focus.

On the other hand, the advent of Industry 5.0 and the increasing complexity of manufacturing processes have underlined the importance of building a safe, efficient, and well-functioning manufacturing environment. The experiments conducted in this thesis have deepened into the heart of the manufacturing domain, where human operators play a pivotal role. The integration of vision depth devices and deep learning technologies has unlocked new dimensions in monitoring and enhancing the well-being of individuals working in manufacturing settings. It is clear that current manufacturing demands precision and efficiency, but it is also driven by a commitment to defending the mental and physical health of operators. In this context, the presented experiments set the baseline for a new way of perceiving manufacturing tasks, addressing the complex challenge of action segmentation in industrial assembling lines. Furthermore, performing such segmentation on a real dataset, including color, depth, and skeletal information, increased the relevance of the analysis. All the state-of-the-art deep learning models successfully segmented the actions for completing the assembly task. In particular, the Transformer architecture was the most suited to processing video data, while the Multi-Stage architecture gave the best results when processed with skeletal information. These outcomes imply that data generalization is mandatory in obtaining efficient results with heterogeneous models. Such statement is further validated by the cross-subject and cross-location analysis performed, particularly regarding the new set of data collected in new manufacturing scenarios. The models segmented correctly and efficiently the actions required for completing the task in both fully and semi-supervised learning circumstances, meaning that the dataset is successful in correctly generalizing the assembling in any scenario.

By using deep learning methodologies to predict and segment human actions in real-world manufacturing environments, the second system proposed in this research aims to enhance the objectivity and accuracy of production performance, avoiding cognitive and physical impairment of the operators. The development of automatic systems presents a new disruptive potential for setting the operator at the core of manufacturing environments, enabling robotic systems to fully adapt to humans, and not the other way around.

The presented research paves the path to new challenges and topics, which can be investi-

gated in the future. Some of such issues are presented as follows.

- The acquisition of more consistent datasets regarding the number of observed subjects, setups, and environments will allow a deeper validation of the proposed automatic systems. For both healthcare and manufacturing fields, the need for new realistic sets of data is neverending, and the more real data is gathered, the more deep learning technologies can be trained and improved at their best, guaranteeing proper generalization.

- As for the system developed for the assessment of elderly motion functions, future improvements may involve the use of user-friendly sensors such as phone cameras or smart devices. With these tools, the evaluation of fall risk and balance can be seamlessly extended to home environments, allowing users to proactively monitor their well-being and health with complete autonomy. With this approach, patients can independently record themselves while performing the planned exercises, while the system can autonomously and easily evaluate their motion capabilities. Such user-centric setup promotes a sense of control and convenience in health monitoring, potentially reducing the need for frequent clinical visits.

- In the context of human action segmentation for manufacturing tasks, the ability to observe the movements of human operators during an assembly assignment is of critical importance for understanding their capabilities in collaborative tasks with robots. The outcomes provided by this work can lead to new studies integrating action segmentation models into robotic scenarios, aiming to develop real-time systems for Human - Robot interaction and collaboration. Such systems can detect the actions of human operators performing a specific task, and allow the robot to respond accordingly in real-time. The camera setup may be improved by adding a second depth sensor and acquiring new data in a calibrated environment. The effectiveness of the analyzed models can be further tested in industrial environments where collaborative robots share the workspace with operators, aiming to assess their robustness and generalization. Enhancing collaborative robots with such deep learning algorithms guarantees high efficiency and precision while placing an invisible protective shield around the operators, promoting their safety.

This research echoes a vision where technology, guided by empathy, serves as a guardian of human well-being, following the tenets of Industry 5.0. It offers a glimpse into a future where human-machine collaboration is enhanced by intelligent systems, fostering a nurturing environment in which the well-being of humans is paramount.

References

[1] T. Özyer, D. S. Ak, and R. Alhajj. Human action recognition approaches with video datasets — a survey. *Knowledge-Based Systems*, 222:1–36, 2021.

[2] U. Mahbub and M. A. R. Ahad. Advances in human action, activity and gesture recognition. *Pattern Recognition Letters*, 2021.

[3] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang, J. X. Du, and D. S. Chen. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors*, 19(5):1–20, 2019.

[4] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[5] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:1–17, 2020.

[6] A. Sarkar, A. Banerjee, P.K. Singh, and R. Sarkar. 3D Human Action Recognition: Through the eyes of researchers. *Expert Systems With Applications*, 193:116424, 2022.

[7] M. Al-Amin, R. Qin, M. Moniruzzaman, Z. Yin, W. Tao, and M. C. Leu. An individualized system of skeletal data-based CNN classifiers for action recognition in manufacturing assembly. *Robotics of Intelligent Manufacturing*, July 2021, 2021.

[8] C. Chen, T. Wang, D. Li, and J. Hong. Repetitive assembly action recognition based on object detection and pose estimation. *Journal of Manufacturing Systems*, 55:325–333, 2020.

[9] M. A. Zamora-Hernandez, J. A. Castro-Vergas, J. Azorin-Lopez, and J. Garcia-Rodriguez. Deep learning-based visual control assistant for assembly in industry 4.0. *Computers in Industry*, 131:1–15, 2021.

[10] Praveen Kumar Reddy Maddikunta, Quoc-Viet Pham, B Prabadevi, Natarajan Deepa, Kapal Dev, Thippa Reddy Gadekallu, Rukhsana Ruby, and Madhusanka Liyanage. Industry 5.0: A survey on enabling technologies and potential applications. *Journal of Industrial Information Integration*, 26:100257, 2022.

[11] Matteo Lavit Nicora, Elisabeth André, Daniel Berkmans, Claudia Carissoli, Tiziana D'Orazio, Antonella Delle Fave, Patrick Gebhard, Roberto Marani, Robert Mihai Mira, Luca Negri, et al. A human-driven control architecture for promoting good mental health in collaborative robot scenarios. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 285–291. IEEE, 2021.

[12] Grazia Cicirelli, Roberto Marani, Antonio Petitti, Annalisa Milella, and Tiziana D'Orazio. Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population. *Sensors*, 21(10):3549, 2021.

[13] Bilal SA Alhayani et al. Visual sensor intelligent module based image transmission in industrial manufacturing for monitoring and manipulation problems. *Journal of Intelligent Manufacturing*, 32(2):597–610, 2021.

[14] Microsoft Azure Kinect SDK. Azure Kinect SDK v1.4.1.

[15] Justin Amadeus Albert, Victor Owolabi, Arnd Gebel, Clemens Markus Brahms, Urs Granacher, and Bert Arnrich. Evaluation of the pose tracking performance of the azure kinect and kinect v2 for gait analysis in comparison with a gold standard: A pilot study. *Sensors*, 20(18):5104, 2020.

[16] Josh Brown Kramer, Lucas Sabalka, Ben Rush, Katherine Jones, and Tegan Nolte. Automated depth video monitoring for fall reduction: A case study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 294–295, 2020.

[17] Laura Romeo, Roberto Marani, Matteo Malosio, Anna G Perri, and Tiziana D'Orazio. Performance analysis of body tracking with the microsoft azure kinect. In *2021 29th Mediterranean Conference on Control and Automation (MED)*, pages 572–577. IEEE, 2021.

[18] Chanhwi Lee, Jaehan Kim, Seoungbae Cho, Jinwoong Kim, Jisang Yoo, and Soonchul Kwon. Development of real-time hand gesture recognition for tabletop holographic display interaction using azure kinect. *Sensors*, 20(16):4566, 2020.

[19] James McGlade, Luke Wallace, Bryan Hally, Andrew White, Karin Reinke, and Simon Jones. An early exploration of the use of the microsoft azure kinect for estimation of urban tree diameter at breast height. *Remote Sensing Letters*, 11(11):963–972, 2020.

[20] Ádám Uhlár, Mira Ambrus, Márton Kékesi, Eszter Fodor, László Grand, Gergely Szathmáry, Kristóf Rácz, and Zsombor Lacza. Kinect azure–based accurate measurement of dynamic valgus position of the knee—a corrigible predisposing factor of osteoarthritis. *Applied Sciences*, 11(12):5536, 2021.

[21] United Nations. World Population Ageing 2020 Highlights: Living arrangements of older persons. Technical Report ST/ESA/SER.A/451, Department of Economic and Social Affairs, United Nations, New York, 2020.

[22] World Health Organization). *Global action plan on the public health response to dementia 2017-2025*. 2017. ISBN: 978-92-4-151348-7.

[23] C. Buckley, L. Alcock, R. McArdle, R. Z. U. Rehman, S. Del Din, C. Mazzà, A. J. Yarnall, and L. Rochester. The Role of Movement Analysis in Diagnosing and Monitoring Neurodegenerative Conditions: Insights from Gait and Postural Control. *Brain Sciences*, 9(2):1–21, 2019.

[24] G. Cicirelli, D. Impedovo, V. Dentamaro, R. Marani, G. Pirlo, and T. D'Orazio. Human Gait Analysis in Neurodegenerative Diseases: a Review. *IEEE Journal of Biomedical and Health Informatics*, 26(1):229–242, 2022.

[25] R. Soubra, A. Chkeir, and J.L. Novella. A Systematic Review of Thirty-One Assessment Tests to Evaluate Mobility in Older Adults. *BioMed Research International*, pages 1–17, June 2019.

[26] G. Grossi, R. Lanzarotti, P. Napoletano, N. Noceti, and F. Odone. Positive technology for elderly well-being: A review. *Pattern Recognition Letters*, 137:61–70, 2020.

[27] G. Cicirelli, R. Marani, A. Petitti, A. Milella, and T. D'Orazio. Ambient Assisted Living: A Review of Technologies, Methodologies and Future Perspectives for Healthy Aging of Population. *Sensors*, 21(10):1–22, 2021.

[28] L. Brognara, P. Palumbo, B. Grimm, and L. Palmerini. Assessing Gait in Parkinson's Disease Using Wearable Motion Sensors: A Systematic Review. *Diseases*, 7(1):1–14, 2019.

[29] J. Howcroft, J. Kofman, and E. D. Lemaire. Prospective Fall-Risk Prediction Models for Older Adults Based on Wearable Sensors. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1812–1820, 2017.

[30] A. Cereatti, U. Della Croce, and A. M. Sabatini. *Three-Dimensional Human Kinematic Estimation Using Magneto-Inertial Measurement Units*, pages 1–24. Handbook of Human Motion. Springer, Cham, 2017.

[31] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. *Sports Medicine-Open*, 4(24):1–15, 2018.

[32] R. Rucco, V. Agosti, F. Jacini, P. Sorrentino, P. Varriale, M. De Stefano, G. Milan anad P. Montella, and G. Sorrentino. Spatio-temporal and Kinematic Gait Analysis in Patients with Frontotemporal Dementia and Alzheimer's Disease through 3D Motion Capture. *Gait & Posture*, 52:312–317, 2017.

[33] N. K. Mangal and A. K. Tiwari. A review of the evolution of scientific literature on technology-assisted approaches using rgb-d sensors for musculoskeletal health monitoring. *Computers in Biology and Medicine*, 132:1–15, 2021.

[34] F. Wang, E. Stone, M. Skubic, J. M. Keller, C. Abbott, and M. Rantz. Towards a Passive Low-Cost In-Home Gait Assessment System for Older Adults. *Journal of Biomedical and Health Informatics*, 17(2):346–355, 2013.

[35] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos. A review of 3D human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding*, 212:103275, 2021.

[36] C. P. Hensley, D. Millican, N. Hamilton, A. Yang, J. Lee, and A. H. Chang. Video-Based Motion Analysis Use: A National Survey of Orthopedic Physical Therapists. *Physical Therapy*, 100(10):1759–1770, 2020.

[37] L. Romeo, R. Marani, N. Lorusso, M. T. Angelillo, and G. Cicirelli. Vision-based assessment of balance control in elderly people. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2020.

[38] N. Kour, Sunanda, and S. Arora. Computer-Vision Based Diagnosis of Parkinson's Disease via Gait: A Survey. *IEEE Access*, 7:156620–156645, 2019.

[39] R. A. Clark, B. F. Mentiplay, E. Hough, and Y. H. Pua. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. *Gait & Posture*, 68:193–200, February 2019.

[40] Peter Fermin Dajime, Heather Smith, and Yanxin Zhang. Automated classification of movement quality using the microsoft kinect v2 sensor. *Computers in Biology and Medicine*, 125:104021, 2020.

[41] A. Ejupi, M. Brodie, Y. J. Gschwind, S. R. Lord, W. L. Zagler, and K. Delbaere. Kinect-Based Five-Times-Sit-to-Stand Test for Clinical and In-Home Assessment of Fall Risk in Older People. *Gerontology*, 62:118–124, 2016.

[42] O. Mazumder, S. Tripathy, S. Roy, S. Chakravarty, D. Chatterjee, and A. Sinha. Postural Sway based Geriatric Fall Risk Assessment using Kinect. In *IEEE Sensors*, pages 1–3, Glasgow, Scotland, UK, Nov. 2017.

[43] F. Romano, P. Colagiorgio, A. Buizza, F. Sardi, and S. Ramat. Extraction of traditional cop-based features from com sway in postural stability evaluation. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3715–3718, 2015.

[44] D. Raví, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Yang. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[45] Y. Chen, Y. Tian, and M. He. Monocular human pose estimation: A survey of deep learning-based method. *Computer Vision and Image Understanding*, 192:102897, 2020.

[46] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a Convolutional Neural Network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE, 2017.

[47] S. Zhang, S. K. Poon, K. Vuong, A. Sneddon, and C. T. Loy. *A Deep Learning-Based Approach for Gait Analysis in Huntington Disease*, volume 264 of *Studies in Health Technology and Informatics*, pages 477–481. IOS Press, 2019.

[48] S. Bringas, S. Salomón, R. Duque, J. L. Monta a, and C. Lage. A Convolutional Neural Network-Based Method for Human Movement Patterns Classification in Alzheimer's Disease. *Proceedings*, 31(1):1–9, 2019.

[49] Francisco Luna-Perejón, Manuel Jesús Domínguez-Morales, and Antón Civit-Balcells. Wearable Fall Detector Using Recurrent Neural Networks. *Sensors*, 19(22):1–18, 2019.

[50] A. Graves. *Supervised sequence labelling with Recurrent Neural Networks*. Springer, 2012.

[51] C. Tunca, G. Salur, and C. Ersoy. Deep learning for fall risk assessment with inertial sensors: Utilizing domain knowledge in spatio-temporal gait parameters. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1994–2005, July 2020.

[52] X. Shu, J. Tang, G. J. Qi, W. Liu, and J. Yang. Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1110–1118, 2021.

[53] Kyoung-Su Oh and Keechul Jung. GPU Implementation of Neural Networks. *Pattern Recognition*, 37(6):1311–1314, 2004.

[54] Laura Romeo, Antonio Petitti, Roberto Marani, and Annalisa Milella. Internet of robotic things in smart domains: applications and challenges. *Sensors*, 20(12):3355, 2020.

[55] L. Wang, R. Gao, J. Vancza, J. Krüger, X.V. Wang, and S. Makris. Symbiotic human-robot collaborative assembly. *CIRP Annals - Manufacturing Technology*, 68:701–726, 2019.

[56] W. Tao, M. Al-Amin, H. Chen, M. C. Leu, Z. Yin, and R. Qin. Real-Time Assembly Operation Recognition with Fog Computing and Transfer Learning for Human-Centered Intelligent Manufacturing. *Procedia Manufacturing*, 48:926–931, 2020.

[57] J. Patalas-Maliszewska, D. Halikowski, and R. Damaševičius. An Automated Recognition of Work Activity in Industrial Manufacturing Using Convolutional Neural Networks. *Electronics*, 10:1–17, 2021.

[58] T. Kobayashi, Y. Aoki, S. Shimizu, K. Kusano, and S. Okumura. Fine-grained Action Recognition in Assembly Work Scenes by Drawing Attention to the Hands. In *15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, page 440–446, 2019.

[59] Matteo Lavit Nicora, Elisabeth André, Daniel Berkmans, Claudia Carissoli, Tiziana D'Orazio, Antonella Delle Fave, Patrick Gebhard, Roberto Marani, Robert Mihai

Mira, Luca Negri, Fabrizio Nunnari, Alberto Peña Fernandez, Alessandro Scano, Gianluigi Reni, and Matteo Malosio. A human-driven control architecture for promoting good mental health in collaborative robot scenarios. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 285–291, 2021.

[60] Anita Pollak, Mateusz Paliga, Matias M Pulopulos, Barbara Kozusznik, and Malgorzata W Kozusznik. Stress in manual and autonomous modes of collaboration with a cobot. *Computers in Human Behavior*, 112:106469, 2020.

[61] Quan Liu, Zhihao Liu, Wenjun Xu, Quan Tang, Zude Zhou, and Duc Truong Pham. Human-robot collaboration in disassembly for sustainable manufacturing. *International Journal of Production Research*, 57(12):4027–4044, 2019.

[62] Yingzhong Tian, Guopeng Wang, Long Li, Tao Jin, Fengfeng Xi, and Guangjie Yuan. A universal self-adaption workspace mapping method for human–robot interaction using kinect sensor data. *IEEE Sensors Journal*, 20(14):7918–7928, 2020.

[63] Cristina Brambilla, Roberto Marani, Laura Romeo, Nicola Matteo Lavit, Fabio A. Storm, Gianluigi Reni, Matteo Malosio, Tiziana D'Orazio, and Alessandro Scano. Azure kinect performance evaluation for human motion and upper limb biomechanical analysis. *Helyion*, 9(11):e21606, 2023.

[64] Jing Yang, Shaobo Li, Zheng Wang, Hao Dong, Jun Wang, and Shihao Tang. Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. *Materials*, 13(24):5755, 2020.

[65] Anca Morar, Alin Moldoveanu, Irina Mocanu, Florica Moldoveanu, Ion Emilian Radoi, Victor Asavei, Alexandru Gradinaru, and Alex Butean. A comprehensive survey of indoor localization methods based on computer vision. *Sensors*, 20(9):2641, 2020.

[66] Eloise Matheson, Riccardo Minto, Emanuele GG Zampieri, Maurizio Faccio, and Giulio Rosati. Human–robot collaboration in manufacturing applications: A review. *Robotics*, 8(4):100, 2019.

[67] Debasmita Mukherjee, Kashish Gupta, Li Hsin Chang, and Homayoun Najjaran. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing*, 73:102231, 2022.

[68] Riccardo Caccavale, Matteo Saveriano, Alberto Finzi, and Dongheui Lee. Kinesthetic teaching and attentional supervision of structured tasks in human–robot interaction. *Autonomous Robots*, 43(6):1291–1307, 2019.

[69] Nida Khalid, Munkhjargal Gochoo, Ahmad Jalal, and Kibum Kim. Modeling two-person segmentation and locomotion for stereoscopic action identification: a sustainable video surveillance system. *Sustainability*, 13(2):970, 2021.

[70] Zoë Moore, Carter Sifferman, Shaniah Tullis, Mengxuan Ma, Rachel Proffitt, and Marjorie Skubic. Depth sensor-based in-home daily activity recognition and assessment system for stroke rehabilitation. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1051–1056. IEEE, 2019.

[71] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern technique. *arXiv preprint arXiv:2210.10352*, 2022.

[72] Bangli Liu, Haibin Cai, Zhaojie Ju, and Honghai Liu. Rgb-d sensing based human action and interaction analysis: A survey. *Pattern Recognition*, 94:1–12, 2019.

[73] Pallabi Ghosh, Yi Yao, Larry Davis, and Ajay Divakaran. Stacked spatio-temporal graph convolutional networks for action segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 576–585, 2020.

[74] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.

[75] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020.

[76] Shijie Li, Jinhui Yi, Yazan Abu Farha, and Juergen Gall. Pose refinement graph convolutional network for skeleton-based action recognition. *IEEE Robotics and Automation Letters*, 6(2):1028–1035, 2021.

[77] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 754–763, 2017.

[78] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14024–14034, 2020.

[79] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[80] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018.

[81] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[82] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[83] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European conference on computer vision*, pages 36–52. Springer, 2016.

[84] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.

[85] Hilde Kuehne, Alexander Richard, and Juergen Gall. A hybrid rnn-hmm approach for weakly supervised temporal action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):765–779, 2020.

[86] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

[87] Ahmad Jalal, Shaharyar Kamal, and Cesar A Azurdia-Meza. Depth maps-based human segmentation and action recognition using full-body plus body color cues via

recognizer engine. *Journal of Electrical Engineering & Technology*, 14(1):455–461, 2019.

[88] Qing Lei, Ji-Xiang Du, Hong-Bo Zhang, Shuang Ye, and Duan-Sheng Chen. A survey of vision-based human action evaluation methods. *Sensors*, 19(19):4129, 2019.

[89] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.

[90] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019.

[91] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[92] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. arXiv, 2020.

[93] Dipika Singhania, Rahul Rahaman, and Angela Yao. Coarse to fine multi-resolution temporal convolutional network. *arXiv preprint arXiv:2105.10859*, 2021.

[94] Matthew Kent Myers, Nick Wright, A Stephen McGough, and Nicholas Martin. Hand guided high resolution feature enhancement for fine-grained atomic action segmentation within complex human assemblies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 471–480, 2023.

[95] Laura Romeo, Roberto Marani, Matteo Malosio, Anna G. Perri, and Tiziana D'Orazio. Performance analysis of body tracking with the microsoft azure kinect. In *2021 29th Mediterranean Conference on Control and Automation (MED)*, pages 572–577, 2021.

[96] Laura Romeo, Roberto Marani, Anna Gina Perri, and Tiziana D'Orazio. Microsoft azure kinect calibration for three-dimensional dense point clouds and reliable skeletons. *Sensors*, 22(13):4986, 2022.

[97] Laura Romeo, Roberto Marani, Antonio Petitti, Annalisa Milella, Tiziana D'Orazio, and Grazia Cicirelli. Image-based mobility assessment in elderly people from low-cost systems of cameras: A skeletal dataset for experimental evaluations. In *Ad-Hoc,*

Mobile, and Wireless Networks: 19th International Conference on Ad-Hoc Networks and Wireless, ADHOC-NOW 2020, Bari, Italy, October 19–21, 2020, Proceedings 19, pages 125–130. Springer, 2020.

[98] Grazia Cicirelli, Roberto Marani, Laura Romeo, Manuel García Domínguez, Jónathan Heras, Anna G Perri, and Tiziana D'Orazio. The ha4m dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. *Scientific Data*, 9(1):745, 2022.

[99] Laura Romeo, Roberto Marani, Tiziana D'Orazio, and Grazia Cicirelli. Video based mobility monitoring of elderly people using deep learning models. *IEEE Access*, 11:2804–2819, 2023.

[100] Azure kinect dk documentation.

[101] Chenyang Zhang, Teng Huang, and Qiang Zhao. A new model of rgb-d camera calibration based on 3d control field. *Sensors*, 19(23):5082, 2019.

[102] David Ferstl, Christian Reinbacher, Gernot Riegler, Matthias Rüther, and Horst Bischof. Learning depth calibration of time-of-flight cameras. In *BMVC*, pages 102–1, 2015.

[103] Microsoft Azure Kinect SDK. Azure Kinect Body Tracking SDK v1.1.0.

[104] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):1066–1078, 2020.

[105] Microsoft Azure Kinect SDK. Azure Kinect SDK functions Documentation.

[106] Valsamis Douskos, Ilias Kalisperakis, and George Karras. Automatic calibration of digital cameras using planar chess-board patterns. In *Proceedings of the 8th Conference on Optical*, pages 9–12. Citeseer, 2007.

[107] Andreas Geiger, Frank Moosmann, Ömer Car, and Bernhard Schuster. Automatic camera and range sensor calibration using a single shot. In *2012 IEEE international conference on robotics and automation*, pages 3936–3943. IEEE, 2012.

[108] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[109] Liang Zhang, Houman Rastgar, Demin Wang, and André Vincent. Maximum likelihood estimation sample consensus with validation of individual correspondences. In *International Symposium on Visual Computing*, pages 447–456. Springer, 2009.

[110] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Joint Pattern Recognition Symposium*, pages 236–243. Springer, 2003.

[111] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pages 6448–6457. PMLR, 2019.

[112] Sang-ha Lee, Jisang Yoo, Minsik Park, Jinwoong Kim, and Soonchul Kwon. Robust extrinsic calibration of multiple rgb-d cameras with body tracking and feature matching. *Sensors*, 21(3):1013, 2021.

[113] Jeffrey Schlicht, David N Camaione, and Steven V Owen. Effect of intense strength training on standing balance, walking speed, and sit-to-stand performance in older adults. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56(5):M281–M286, 2001.

[114] Rita Pavasini, Jack Guralnik, Justin C Brown, Mauro Di Bari, Matteo Cesari, Francesco Landi, Bert Vaes, Delphine Legrand, Joe Verghese, Cuiling Wang, et al. Short physical performance battery and all-cause mortality: systematic review and meta-analysis. *BMC medicine*, 14(1):215, 2016.

[115] Jack M Guralnik, Eleanor M Simonsick, Luigi Ferrucci, Robert J Glynn, Lisa F Berkman, Dan G Blazer, Paul A Scherr, and Robert B Wallace. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *Journal of gerontology*, 49(2):M85–M94, 1994.

[116] Oliver Perkin, Polly McGuigan, and Keith Stokes. Dataset for'exercise snacking to improve muscle function in healthy older adults: A pilot study'. 2019.

[117] Tung Wai Auyeung, H Arai, LK Chen, and Jean Woo. Normative data of handgrip strength in 26344 older adults-a pooled dataset from eight cohorts in asia. *The journal of nutrition, health & aging*, 24(1):125–126, 2020.

[118] Trong-Nguyen Nguyen and Jean Meunier. Walking gait dataset: point clouds, skeletons and silhouettes. *DIRO, University of Montreal, Tech. Rep.*, page 1379, 2018.

[119] HIKVision. Products and solution.

[120] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[121] D. F. Redaelli, F. A. Storm, and G. Fioretta. MindBot Planetary Gearbox, 2021.

[122] J. Zhang and W. Li and P. O. Ogunbona and P. Wang and C. Tang. RGB-D-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016.

[123] P. Wang and W. Li and P. Ogunbona and J. Wan and S. Escalera. RGB-D-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 171:118–139, 2018.

[124] A. Lopes, R. Souza, and H. Pedrini. A Survey on RGB-D Datasets. Preprint at https://arxiv.org/abs/2201.05761, 2022.

[125] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *IEEE Computer Society Conference Computer Vision Pattern Recognition (CVPR)*, page 1010–1019, Los Alamitos, CA, USA, 27-30 June 2016.

[126] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2684–2701, 2020.

[127] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang. Jointly learning het-erogeneous features for RGB-D activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(11):2186–2200, 2017.

[128] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Rei, M. Voit, and R. Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2801–2810, 2019.

[129] A. B. Youssef, C. Clavel, S. Essid, M. Bilac, and M. Chamoux. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *ACM International Conference on Multimodal Interaction*, pages 464–472, 2017.

[130] E. Nicora and G. Goyal and N. Noceti and A. Vignolo and A. Sciutti and F. Odone. The MoCA dataset, kinematic and multi-view visual streams of fine-grained cooking actions. *Scientific Data*, 7(432), 2020.

[131] A. Saudabayev and Z. Rysbek and R. Khassenova1 and H. A. Varol. Human grasping database for activities of daily living with depth, color and kinematic data streams. *Scientific Data*, 5(180101), 2018.

[132] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14, San Francisco, CA, USA, 13-18 June 2010.

[133] J.Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 16-21 June 2012.

[134] L. Xia, C. C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 16-21 June 2012.

[135] B. Ni, G. Wang, and P. Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 1147–1153, Barcelona, Spain, 6-13 Nov. 2011.

[136] J. Zhang, P. Wang, and R. X. Gao. Hybrid machine learning for human action recognition and prediction in assembly. *Robotics and Computer-Integrated Manufacturing*, 72:102184, 2021.

[137] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. Preprint at https://arxiv.org/pdf/2203.14712.pdf, 2022.

[138] F. Ragusa, A. Furnari, S. Livatino, and G. M. Farinella. The MECCANO dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 1569–1578, 2021.

[139] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould. The IKEA ASM dataset: Understanding people assem- bling furniture through actions, objects and pose. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 847–859, 2021.

[140] Microsoft. Azure Kinect DK documentation, 2021. https://docs.microsoft.com/en-us/azure/kinect-dk/, Accessed March 2022.

[141] L. Romeo, R. Marani, M. Malosio, A. G. Perri, and T. D'Orazio. Performance analysis of body tracking with the microsoft azure kinect. In *2021 29th Mediterranean Conference on Control and Automation (MED)*, pages 572–577, 2021.

[142] J. A. Albert and V. Owolabi and A. Gebel and U. Granacher and B. Arnrich. Evaluation of the Pose Tracking Performance of the Azure Kinect and Kinect v2 for Gait Analysis in Comparison with a Gold Standard: A Pilot Study. *Sensors*, 20(18), 2020.

[143] F. Longo and L. Nicoletti and A, Paodovano. New perspectives and results for Smart Operators in industry 4.0: A human-centered approach. *Computers & Industrial Engineering*, 163:107824, 2022.

[144] S. K. Yadav and K. Tiwari and H. M. Pandey and S. AliAkbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

[145] Y. Li and Z. Dong and K. Liu and L. Feng. L. Hu and J. Zhu and L. Xu and Y. wang and S. Liu. Efficient Two-Step Networks for Temporal Action Segmentation. *Neurocomputing*, 454:373–381, 2021.

[146] O. Moutik and S. Tigani and R. Saadane and A. Chehri. Hybrid Deep Learning Vision-based Models for Human Object Interaction Detection by Knowledge Distillation. *Procedia Computer Science*, 192:5093–5103, 2021.

[147] Jessica Fish. *Short Physical Performance Battery*, pages 2289–2291. Encyclopedia of Clinical Neuropsychology. Springer, NY, 2011.

[148] L. Romeo, R. Marani, T. D'Orazio, and G. Cicirelli. SPPBdataset, 2020.

[149] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[150] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[151] Yi Bin, Yang Yang, Fumin Shen, Xing Xu, and Heng Tao Shen. Bidirectional Long-Short Term Memory for video description. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 436–440, 2016.

[152] C. Shorten and T. M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(60):1–48, 2019.

[153] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[154] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, Second Edition, 2018. ISBN: 978-0387-31073-2.

[155] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

[156] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[157] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[158] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.