Online human motion analysis in industrial context: A review

Toufik Benmessabih, Rim Slama, Vincent Havard, David Baudry

Online human motion analysis in industrial context: A review

Toufik Benmessabih[a], Rim Slama[b], Vincent Havard[c], David Baudry[d]

[a]CESI LINEACT, tbenmessabih@cesi.fr 69100, Lyon, France
[b]CESI LINEACT, rsalmi@cesi.fr 69100, Lyon, France
[c]CESI LINEACT, vhavard@cesi.fr 76800, Rouen, France
[d]CESI LINEACT, dbaudry@cesi.fr 76800, Rouen, France

Abstract

Human motion analysis plays a crucial role in industry 4.0 and, more recently, in industry 5.0 where human-centered applications are becoming increasingly important, demonstrating its potential for enhancing safety, ergonomics and productivity. Considering this opportunity, an increasing number of studies are proposing works on the analysis of human motion in an industrial context, taking advantage of the rise of artificial intelligence technologies and sensor technologies. The objective of this work is to provide a review of recent studies exploring these technologies in the analysis of human movement while specifically considering industrial context. First, a taxonomy of key human motion analysis applications is proposed, presenting statistical insights to reveal trends and highlighting lacks in current research. Furthermore, this work identifies benchmark datasets acquired in various industrial case studies and associated sensors. Many recommendations for selecting optimal sensors and valuable benchmarks are proposed. Then, the paper outlines the current trend of utilizing hybrid deep learning methodologies in human movement analysis while underscoring the performance and limitations of these proposed methods, considering industrial constraints such as real-time recognition and frugality. Finally, challenges and future works are highlighted, focusing on the opportunities to address problems related to the complex industrial environment in order to achieve reliable performances. Human motion analysis plays a crucial role in industry 4.0 and recently in industry 5.0, where human-centered applications are becoming increasingly important demonstrating its potential for enhancing safety, ergonomics and productivity. While numerous works have examined diverse approaches for human motion analysis, few studies have specifically addressed the industrial environment, which poses distinct challenges and problems. In light of this gap, the present study aims to investigate the importance of online human motion analysis across various industrial domains. It underscores the need for benchmark datasets and compares them based on different criteria such as the type of interactions, recorded views, body part analyzed and available modalities. Furthermore, it provides a thorough review of the technological advancements applied in the field of human motion analysis mainly based on MoCaps sensors. Additionally, it presents a comparative evaluation of various deep learning approaches mostly focusing on human action recognition methods on trimmed and untrimmed sequences. Finally, we discuss the limitations. These include but are not limited to imbalanced datasets and poor performance of online recognition methods. We also shed light on unresolved challenges that arise in industrial environments. These encompass the choice of sensors for data acquisition, data annotation for large benchmark datasets, and online action

~~segmentation in continuous sequences. In conclusion, the exploration of synthetic data generation and multimodal approaches is encouraged. Further investigation is called for, considering aspects such as frugality and system interoperability which help to inspire future research directions in this area.~~

Keywords:
Industry 4.0/5.0; Online human motion analysis; Data acquisition technologies; Benchmarks; Deep learning; Human-robot collaboration.

---

1. Introduction

The importance of analyzing human motion in industry 4.0 and recently in industry 5.0 is paramount, as it highlights the crucial role of humans and their impact on various industrial applications, including ergonomics, safety and productivity (Alves et al., 2023). While human motion refers to the physical movement of the human body, human action goes beyond the physical movement and involves a purpose or intention behind the motion. Human motion analysis (HMA) is a multidimensional field comprising various facets of human movement comprehension. It involves human gesture recognition (HGR), which specifically deals with identifying and categorizing human actions, especially when only the hands are captured (Sturm et al., 2023). Additionally, HMA encompasses human action recognition (HAR), focusing on identifying and classifying actions when the whole body is captured (Nazmus Sakib et al., 2022). Moreover, human action detection involves detecting the start and end points of action instances (Xu et al., 2021). Lastly, there is action anticipation or forecasting, which involves predicting future actions by anticipating the movement of the operator (Moutinho et al., 2023).

Accurate analysis and recognition of human actions hold significant importance for several purposes, including facilitating effective human-robot collaboration (HRC) (Zhou et al., 2023), optimizing industrial processes (Hernandez et al., 2021), assisting operators in their work (Moutinho et al., 2023) and safeguarding their health (Tassi et al., 2022) and security (Kwon et al., 2021). The acquisition of displacement and movement data in human motion analysis can be achieved using a range of methods and technologies. The choice of approach and technology depends on the particular context of the case study and the adopted methodology (Menolotto et al., 2020). To assess and compare various methods for human motion analysis, researchers frequently utilize benchmark datasets tailored for industrial scenarios, which are increasingly being made available. These datasets consist of annotated motion sequences that capture diverse human actions and interactions within manufacturing environments. They serve as standardized platforms, enabling the testing of approaches against real-world challenges introduced by industrial requirements. These benchmark datasets promote fair comparisons and facilitate advancements in the field (Sener et al., 2022).

As depicted in Table 1, literature has been examined concerning various components, such as technologies used for acquiring human movement data, as well as the methodologies and evaluation metrics applied in the analysis of human actions and movements. A noticeable observation highlights the absence of recent comprehensive surveys that encompass all facets of HMA in the industrial context.

Table 1: Recent surveys on motion analysis in industrial context from 2018 to 2023.
HMA: Human Motion Analysis, HAR: Human Action Recognition, DL: Deep Learning, P: Partial study of the item, G: General study of the item without a specific focus, MoCap: Motion Capture.

Year	Title ~~Method~~	Reference	Time Span	Goal	Industry	Sensors	Datasets	HMA approaches		
								Untrimmed	DL	Metrics
2023	The Expanding Role of Artificial Intelligence in Collaborative Robots for Industrial Applications: A Systematic Review of Recent Works	(Borboni et al., 2023)	2018-2022	Comparison between automated robots and cobots for tasks such as vision-based action recognition	✓			✓	✓	
2022	Human–robot collaboration in industrial environments: A literature review on non-destructive disassembly	(Hjorth and Chrysostomou, 2022)	2009-2020	Human robot collaboration disassembly using action, posture and gesture recognition	✓				✓	
2022	Temporal Action Segmentation: An Analysis of Modern Techniques	(Ding et al., 2022)	2011-2023	Comparison between datasets, approaches and metrics used for online temporal action segmentation	G		✓	✓	✓	✓
2022	Wearable Sensors and Artificial Intelligence for Physical Ergonomics: A Systematic Review of Literature	(Donisi et al., 2022)	2009-2021	Overview of sensors and approaches employed for ergonomic applications in industry	✓	✓		✓	✓	
2022	Vision-based holistic scene understanding towards proactive human-robot collaboration	(Fan et al., 2022)	2010-2021	Scene understanding considering cognition object, human and environment + visual reasoning	✓			✓	✓	
2021	Human-Robot Perception in Industrial Environments: A Survey	(Bonci et al., 2021)	2010-2020	Sensor equipment useful for human detection and action recognition with fixed and mobile robots.	✓	✓		✓	P	
2020	Motion Capture Technology in Industrial Applications: A Systematic Review	(Menolotto et al., 2020)	2011-2021	List MoCap in industry and identify their most targeted applications	✓	✓		G	G	
2019	Human Activity Recognition for Production and Logistics: A Systematic Literature Review	(Reining et al., 2019)	2009-2018	Roadmap of HAR in production and logistics	P	✓	✓	✓	P	✓
2018	Active and assisted living: a comprehensive review of enabling technologies and scenarios	(Manoj and Thyagaraju, 2018)	2005-2018	Active and assisted living via HAR	G	✓			G	
2023	Ours	-	2018-2023	Study the applications of online human motion analysis in industrial context, the technologies used to acquire motion data, benchmark datasets and approaches.	✓	✓	✓	✓	✓	✓

In fact, some surveys have examined the analysis of human motion within industrial settings, either in a broad scope encompassing various contexts such as sports and health (Ding et al., 2022; Manoj and Thyagaraju, 2018) (these surveys are denoted by 'G' in Table 1 column 'Industry') or in a narrower scope such as logistics or ergonomics limited to specific sub-fields within the industrial context (Reining et al., 2019; Donisi et al., 2022) (these surveys are denoted by 'P' in Table 1 column 'Industry').

Among the mentioned surveys, approximately half of them have not addressed the sensors utilized for data acquisition (Borboni et al., 2023; Hjorth and Chrysostomou, 2022; Ding et al., 2022; Fan et al., 2022). Only two studies (Ding et al., 2022; Reining et al., 2019) specifically addressed benchmark datasets and evaluation metrics presenting some industrial datasets such as Assembly101 (Sener et al., 2022) and HA4M (Cicirelli et al., 2022). However, in these studies, it was noted that most reviewed approaches were evaluated on general datasets like Breakfast activities (Singhania et al., 2022), which may not adequately capture the complexities encountered in manufacturing environments. Furthermore, some surveys (Hjorth and Chrysostomou, 2022; Manoj and Thyagaraju, 2018; Donisi et al., 2022) do not review benchmark datasets which makes it challenging to compare the approaches discussed in their studies.

The existing literature remains limited, particularly in terms of surveys that specifically focus on analyzing human movements within untrimmed sequences. Some of the available surveys are still primarily centered around trimmed sequences which fail to encompass the complete complexity and nuances of human motion in industrial environments (Hjorth and Chrysostomou, 2022; Manoj and Thyagaraju, 2018). Some approaches focus only on a specific architecture of deep learning (DL) approaches, as in (Reining et al., 2019; Bonci et al., 2021), only convolution neural network (CNN) based approaches are reviewed (these surveys are denoted by 'P' in Table 1 column 'DL'). On the contrary,

other works (Manoj and Thyagaraju, 2018; Menolotto et al., 2020) cover not only DL but broader methodologies such as machine learning methods (these approaches are denoted by 'G' in Table 1 column 'DL'). It is worth mentioning that deep learning techniques employed in some surveys (Borboni et al., 2023; Fan et al., 2022) exhibited limitations in terms of time efficiency. Additionally, case studies utilized in the presented approaches predominantly involved controlled environments with operators in static poses. However, in real-world scenarios, operators are often moving within the working area in a dynamic environment with changing backgrounds. More surveys (Ding et al., 2022; Vahdani and Tian, 2022; Matheson et al., 2019; Yonga Chuengwa et al., 2023; Castro et al., 2021; Sun et al., 2022) were also identified, covering general applications not specific to the industry or centered more on human-robot collaboration in general than specifically delving into human motion analysis.

To address the limitations and gaps identified in previous surveys, this study specifically focuses on the industrial context. A more up-to-date viewpoint is provided by concentrating on recently published works from 2018 onwards. In this review, human movement analysis encompasses human gesture recognition, human action recognition, human action detection and anticipation. To the best of our knowledge, this is the first review providing a comprehensive overview of online human motion analysis in an industrial context, addressing various items including applications, data acquisition technologies, data modalities, benchmark datasets and evaluation metrics. In addition to categorizing existing works, their strengths and limitations, this paper proposes various taxonomies and discussions emphasizing the opportunities to consider in future works. To resume, this study presents several contributions by:~~This paper aims to investigate applications, data acquisition technologies, data modalities, benchmark datasets and evaluation metrics. The objective of this paper is to identify and provide a comprehensive analysis of existing resources while also paving the way for future directions. The key contributions of this study can be summarized as follows:~~

- Identifying and categorizing key applications using human motion analysis, establishing a taxonomy that distinguishes main and sub-applications. Also, this study provides statistical insights, revealing trends and underscoring gaps in existing works across these application domains. ~~We provide a taxonomy of primary applications of human motion analysis within an industrial context, offering insights into their usage, significance, and impact, whether or not robots are present. Additionally, we identify emerging trends in these applications, highlighting the evolving landscape within industrial settings.~~

- Proposing a taxonomy by investigating various sensors for capturing operator movements and acquiring motion data, including both motion capture sensors and contemporary signal-based ones. Additionally, this paper provides a comparative analysis of these technologies, along with recommendations for selecting optimal sensors based on specific application requirements and the type of movement being analyzed. ~~We identify and conduct a comparative analysis of the technologies used for capturing operator movements and acquiring data. Additionally, we present a taxonomy and analyze the modalities provided by each technology.~~

- Pioneering the creation of a comprehensive in-depth view consolidating recent benchmark datasets explicitly tailored for human motion analysis within indus-

trial applications. This synthesis serves as a valuable resource for the research community, enabling results reproduction and comparison while significantly reducing time spent on data acquisition. Additionally, this paper presents a comprehensive cross-analysis to identify diverse perspectives on movement acquisition, encompassing viewpoints, modalities and analyzed body segments. ~~We pinpoint diverse benchmark datasets suitable for training and evaluating motion analysis methodologies, emphasizing their diverse features, case studies, and distinctive attributes they offer.~~

- Outlining the current trend of utilizing hybrid deep learning methodologies in human movement analysis while underscoring the critical needs for less laborious approaches, particularly those learned from abundant data suggesting the consideration of recent approaches such as self-supervised and zero-shot learning. This work emphasizes the relationship between the evolution of deep learning techniques and the necessity to integrate frugality and embedded systems. Additionally, the paper recognizes the pressing need for more mature recognition approaches capable of handling untrimmed data in industrial applications. ~~We conduct an analysis and comparison of various deep learning methods proposed in the literature for human action segmentation, recognition, and anticipation. Additionally, we identify the main metrics used to evaluate these methods across a spectrum of benchmark datasets.~~

- Proposing a roadmap for future directions and identifying critical points, starting with the role of motion analysis in Industry 4.0 and its evolution to Industry 5.0. Several opportunities are identified to enhance action recognition in the industry, considering both online recognition constraints and the need for frugal approaches. Additionally, this study suggests new opportunities with emerging sensors and data acquisition methods to address the problems related to complex industrial environments. ~~We thoroughly examine the limitations and challenges that arise in the context of human motion analysis within the industrial environment. Simultaneously, we put forward potential areas of future research and investigation that can help to address these issues and advance the field.~~

The paper contains several sections illustrated in Fig. 1, and it is structured as follows: Section 2 introduces the methodology used in selecting articles to write our paper. Section 3 reviews human motion applications in industry. Section 4 highlights the technologies utilized to acquire data and create the benchmark datasets. Human motion analysis approaches mostly human action recognition methods and metrics are listed in Section 5. Finally, the potential future development of HMA is discussed in Section 6.

2. Review methodology

In this section, we outline the methodology employed in selecting articles used in this study, inspired by (Prunet et al., 2022). First, we consider the selection of specific keywords that align with our research objectives and study scope. To identify these keywords, we relied on a set of predefined general items which are as follows:

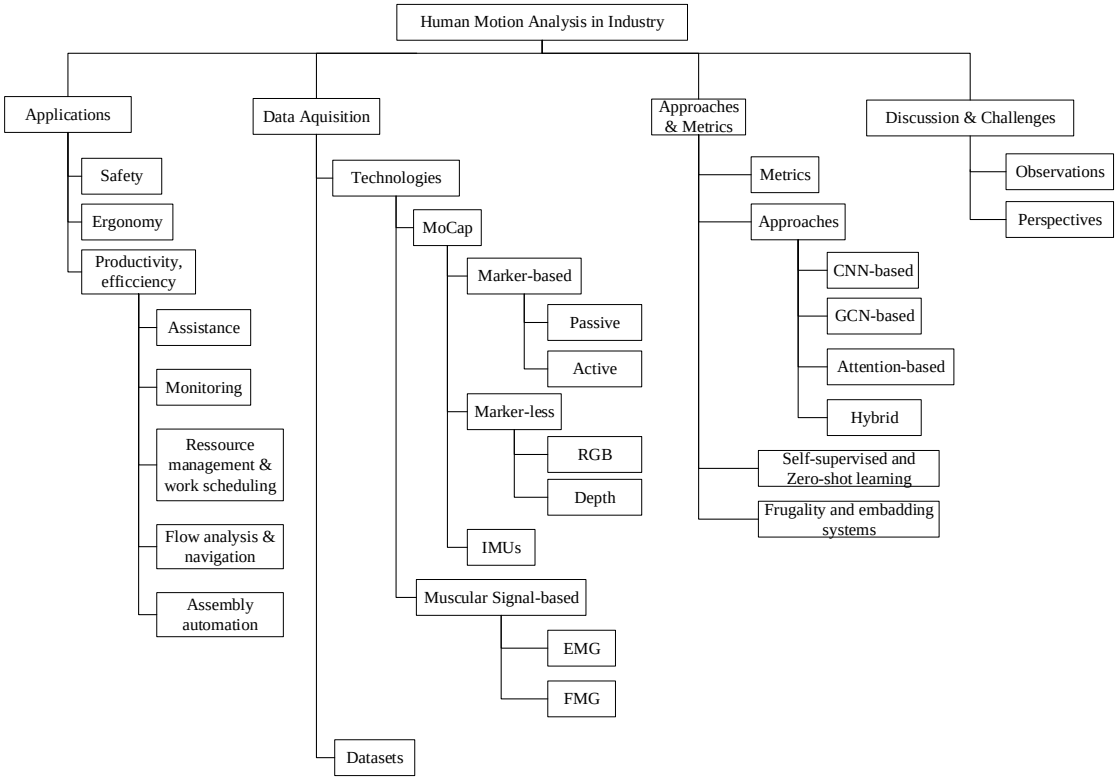


Figure 1: The review architecture.

- Human and Motion Analysis: We exclusively consider studies that analyze human movements. This category includes the analysis and recognition of human gestures and actions but also the interactions between the operator and his environment through gestures.
- Industrial Environment: The papers we study must treat the application of HMA in an industrial setting. While HMA solutions have been proposed in various contexts, such as sports and surveillance, our primary interest lies in its application in industrial environments, where it can have a significant impact.
- Data Modalities: The papers we review must explicitly specify the type of data employed for HMA. This may encompass RGB, depth and skeletal data. These are acquired using MoCaps dedicated cameras or IMUs. Understanding these modalities is important for making informed decisions regarding the choice of approaches to use.

The derived keywords from these general items are detailed in Table 2. Our review methodology entailed a specific process for paper and document selection which can be summarised as follows:

- Phase 1: We began by identifying pertinent keywords related to our study as previously mentioned. These keywords were incorporated into the following query: "(industr* OR manufactur* OR robot) AND (human OR operator) AND ("action recognition" OR "gesture recognition" OR "motion analysis" OR "movement recognition") AND (skeleton OR video OR RGB OR depth OR camera OR MoCap OR IMU)". This query was applied to both 'Scopus' and 'Scinapse' databases. We established specific criteria, including a focus on publications between 2018 and May 2023 (included), exclusivity to the English language, restriction to journal and conference papers, and relevance to the domains of engineering or computer science. Our initial search resulted in 706 records.
- Phase 2: We refined the results by eliminating duplicated entries, leaving us with 658 unique articles.
- Phase 3: We conducted a more detailed screening phase by examining titles, keywords and abstracts based on specific criteria detailed hereafter. Only works using deep-learning approaches were selected. Articles exclusively examining robot displacement analysis without considering human motion analysis were excluded. Only works specifically focused on an industrial context were included, excluding those related to other contexts such as sports and healthcare. This resulted in 270 articles.
- Phase 4: An in-depth review process was undertaken on the content of the remaining articles by applying the same criteria as in Phase 3. This process involved a more advanced examination of the paper's content to ensure alignment with the scope of our research, leading to a total of 114 selected articles.
- Phase 5: Relevant papers were identified by consulting the references and citations of the selected papers as supplementary resources. Ultimately, a total of 140 ~~133~~ papers were included as the foundation for this review.

Table 2: Used keywords in the papers selection process.

General item	Keywords
Industry	industr*, manufactur*, robot
Human	human, operator
Motion analysis	"action recognition", "gesture recognition", "motion analysis", "movement recognition"
Modalities	skeleton, video, RGB, depth, camera, IMU, MoCap

3. Applications

Industry 4.0 and more recently industry 5.0 emphasize the need for industries to exhibit resilience, durability and a human-centered approach, prioritizing the safety and well-being of individuals (Alves et al., 2023). To reach these objectives, the analysis of human motion is of utmost importance and holds significant relevance in the industry (Menolotto et al., 2020; Rana et al., 2023). These endeavors find practical application

in various industrial domains, most notably in robotics and automation, training, production, logistics and aeronautics (Al-Amin et al., 2019; Niemann et al., 2021; Moutinho et al., 2023). This study reviewed articles and grouped them into three primary applications (Ergonomy, Safety and Productivity-efficiency) with or without robot participation which are depicted in Fig. 2.

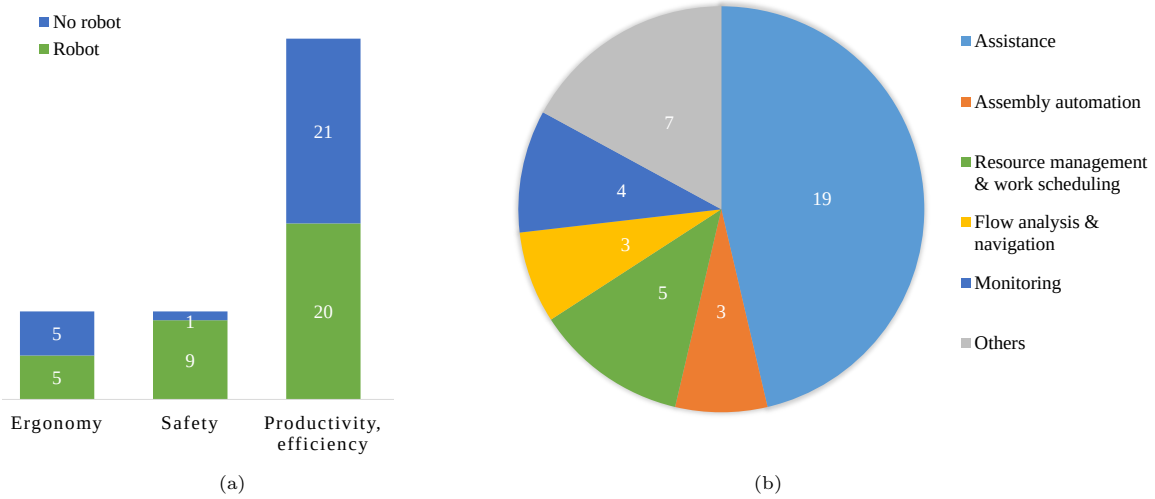


Figure 2: Human motion analysis applications in industry. (a) Main applications considering human-robot collaboration. (b) Productivity, efficiency sub-fields.

3.1. Safety

Ensuring the safety of operators in industrial settings is crucial, as accidents can occur when machines are near sensitive body parts such as the head or arms. To mitigate this risk, several approaches have been applied to robots, including collision detection and reactive motion planning techniques, commonly referred to as speed and separation monitoring (SSM) (Mohammadi et al., 2020) and power and force limitation (PFL) (Vicentini, 2021) methods, respectively. Most studies (9 out of 10 studies) ~~8 out of 9 studies~~ focus on study cases that involve the use of robots while only one study (Delamare et al., 2020) tracks the movements of operators in a workstation without robot presence. Most of the research in this field references its applications within the context of industry 4.0. These methods can be broadly categorized into two types. The first category involves the placement of sensors on robots known as presence sensors (non-visual sensors) along with cameras to monitor human activity. For instance, Abu Al-Haija and Al-Saraireh (2022) proposed an autonomous human-robot contact detection model to prevent damaging collisions during collaboration in industrial tasks. Similarly, Kwon et al. (2021) proposed a method that determines the location of the collision and its occurrence to avoid abrupt and unpredictable collisions. Birjandi et al. (2020) proposed a collision detection system using a sensor fusion setup with high accuracy and decreasing detection uncertainty. (Zhang et al., 2020) introduced a collision prevention model that predicts the future operator motion trajectory while working on engine assembly. The anticipated trajectory

guides real-time planning and execution of robot actions. The second category involves placing sensors on the operator along with additional cameras with the primary objective of capturing data utilized in collision detection algorithms. For instance, Grushko et al. (2021) attempted to predict the robot's future trajectory, rather than anticipating the operator's movement. They used a notification system that vibrates as the operator's hand approaches the robot's planned trajectory to avoid collisions. Pastor et al. (2022) presented a haptic database to estimate the grasping location precisely when grabbing a human limb, which can be applied in rescue or assistive robotics. Delamare et al. (2020) proposed a new dataset for tracking operators and gathering accurate information on worker movement to improve safety conditions in the industry.

3.2. Ergonomy

In industries where humans play a central role, ensuring the well-being of workers is of utmost importance. To achieve this, human motion analysis can be employed to estimate ergonomic indexes, which can then be used to suggest optimal working postures and correct any harmful postures that could result in musculoskeletal disorders (MSDs) (Maurice et al., 2019). HMA also plays a vital role in intelligent automation by automating time-consuming, repetitive and physically demanding activities to preserve the mental and physical health of workers (Maddikunta et al., 2022).

Numerous studies have been carried out to assess ergonomic parameters based on well-known ergonomic indexes such as Rapid Upper Limb Assessment (RULA), which is the most commonly used index (da Silva et al., 2022). RULA is employed to evaluate posture, force and movement related to sedentary tasks (Chiabert and Aliev, 2020). Rapid Entire Body Assessment (REBA), which is similar to RULA in principle, is used for full body posture studies, including the neck, hand, shoulder, leg and limb twisting (da Silva et al., 2022). Ergonomic Assessment WorkSheet (EAWS) is a comprehensive screening method that considers traditional risk factors such as work strength, frequency and repetition (da Silva et al., 2022). Ovako Working posture Analyzing System (OWAS) is a method that identifies and evaluates poor working postures and determines the urgency of corrective measures (Inkulu et al., 2022). National Institute for Occupational Safety and Health (NIOSH) was developed to determine the maximum load that can be manually handled and moved during a work activity (da Silva et al., 2022). It is worth noting that there are additional ergonomic scores beyond the ones listed here (Joshi and Deshpande, 2019). These indexes have been utilized in several studies. Some work (Havard et al., 2019; Kim et al., 2021) involved the use of robots in their study cases while the rest (Manghisi et al., 2022; Bortolini et al., 2020; Sedlar et al., 2023) focused on human ergonomics in the absence of robots. The majority of studies focus their research within the context of industry 4.0, with only a limited number expanding their scope to industry 5.0. Sedlar et al. (2023) introduced a dataset to estimate the pose of handheld tools during industrial tasks. Havard et al. (2019) studied the ergonomics of a manual workstation according to the position of a robotic arm used as a third arm to maintain the assembled product. It allows adapting the robotic arm program to the operator's profile. Kim et al. (2021) used a robot arm to hold an object that had to be polished by the human subject to estimate ergonomic parameters such as joint torque, muscle fatigue and manipulability to identify the risk of injury associated with poor posture and suggest an improved posture arrangement. Manghisi et al. (2022) proposed a tool called ErgoVR, which offers a real-time evaluation of ergonomic postural risk by providing 3D

visualization of postures. Bortolini et al. (2020) developed a motion analysis system (MAS) that measures the human skeleton movements during activity execution and uses these estimations to evaluate previously mentioned ergonomic indexes, such as OWAS, REBA, NIOSH and EAWS.

3.3. Productivity and efficiency

The optimization of productivity and efficiency constitutes a primary objective across diverse industrial sectors. To attain these objectives, human motion analysis serves as a pivotal tool. In industrial settings, identifying movements can pose a challenge, particularly when tasks are allocated among multiple actors, as is often the case in human-robot collaboration in manufacturing. In such cases, the robot assumes the role of a facilitator, assisting in tasks such as tool transfer and heavy object manipulation (Eisenbach et al., 2022). To enable effective collaboration, the robot must possess the ability to anticipate the operator's intentions and take proactive measures (Zhang et al., 2022c). While this aspect is acknowledged in industry 4.0, it is further emphasized in industry 5.0, where the enhancement of these capabilities is a primary focus (Moutinho et al., 2023). The application of HMA is viable for various sub-fields of process optimization, as indicated in the following enumeration:

- Assistance can significantly improve productivity by facilitating the transfer of parts, objects and tools during assembly processes. Furthermore, It can also aid operators in achieving better, faster and higher-quality product assembly outcomes. Numerous studies focused on the enhancement of assembly processes through human assistance and effective object handover techniques. Castro et al. (2021) and Matheson et al. (2019) conducted extensive analyses on human assistance and object handover. Collaborative robots were used for handling tasks, such as using a screwdriver, delivering assembly tools, or even grasping objects from operators and adjusting their orientations for assistance (Zhang et al., 2022a; Zhou et al., 2023; Tassi et al., 2022). In parallel, other researchers focused on the facilitation of operator activity through object holding and movement. Activities regrouped mainly assembly tasks, such as assembling an internal combustion engine, assembling a small wooden box, soldering assistance and performing shop-floor tasks (?Moutinho et al., 2023; Toichoa Eyam et al., 2021; Darvish et al., 2018; Lagamtzis et al., 2022). Finally, works such as (Liu et al., 2023b; Xu et al., 2023) considered anticipating human movements to improve operator assistance by robots. They addressed intention and behavior recognition, incorporating methods such as rule reasoning, early action detection and object recognition (Zhang et al., 2022b).

- Monitoring can enhance productivity in the industry by identifying mistakes, during assembly and guiding operators for better, faster and higher-quality product assembly. Rana et al. (2023) proposed a new dataset for monitoring suspicious activities, especially in large industries to optimize workflow during tricycle assembly. Zamora et al. (2021) evaluated the activity of the operator during the manual assembly of skateboard parts to identify errors and guide the operator, enhancing productivity in the industry. Chen et al. (2020a) proposed an approach to monitor operators during assembly tasks. It identifies repetitive assembly actions and predicts their respective operating durations. This approach can be used to minimize

potential quality issues caused by the lack of key operational steps and the irregular operation of workers during assembly activities.

- Resource management and work scheduling refers to the systematic approach of efficiently allocating and organizing available resources, such as time, space and human labor, to optimize productivity and performance in various activities. Bortolini et al. (2020) proposed a MAS that can distinguish between time and space spent for added-value or non-added-value activities. Hernandez et al. (2021) proposed a hierarchical HAR method with a dataset to measure the performance of manual labor by estimating productivity indicators, such as worker availability, worker performance and overall labor effectiveness automatically for packing and shipping orders. Wang et al. (2019) proposed a vision sensor-based HAR to improve quality in the industry. They enhanced productivity and resource management by minimizing task execution time and improving operator efficiency.

- Flow analysis and navigation refers to the process of studying and simulating regular activities that occur in industrial environments, with the aim of understanding and optimizing the movement of robots within these settings. Schreiter et al. (2022) tried to emulate regular activities performed in an industrial environment. Their work was based on Rudenko et al. (2020) who used an automated guided vehicles (AGV) type of robot as a moving obstacle in the working area. Li et al. (2021) used a robotic arm to grab, deliver and hold objects such as a toolbox while moving next to the operator.

- Assembly plans automation refers to the process of generating assembly plans automatically through the utilization of action recognition techniques applied to manufacturing assembly tasks. Upadhyay et al. published the IKEA ASM dataset to generate assembly plans automatically for assembling furniture.

  The remaining studies prioritized the optimization of solutions for complex action segmentation, recognition and anticipation. These optimized approaches have the potential to be implemented across diverse domains and applications, ultimately improving productivity and efficiency in the industrial sector (Singhania et al., 2022; Liu et al., 2023a; Zhang et al., 2021).

3.4. Summary of findings

Upon reviewing articles concerning human motion analysis applications in an industrial context, three key applications have been identified. These applications encompass safety, where the detection of physical contact and collision avoidance is pivotal for operator safety in potentially hazardous industrial environments. Ergonomics is another significant application, as it involves estimating ergonomic indices to correct the working postures of operators, ensuring their comfort and well-being. Additionally, there is a focus on productivity and efficiency, achieved through continuous monitoring, assessment of operators and the anticipation of their movements. This facilitates human-robot collaboration, ultimately enhancing manufacturing processes.
Most studies addressed their research within the context of industry 4.0, with a few extending to industry 5.0. The results highlight a clear preference for applications related to industrial productivity and efficiency. Specifically, 41 ~~36~~ studies have delved into

the utilization of HMA to enhance these aspects. However, prioritizing ergonomic and safety considerations represents a new era in industrial practices. Although industry 4.0 acknowledges this aspect, industry 5.0 places even greater focus on it as it aligns with its core pillar, emphasizing both efficiency and human well-being.

4. Data acquisition

In the field of human motion analysis, understanding and accurately capturing data is of paramount importance. In the previous section, we explored the main applications of HMA in industry, highlighting its significance across various domains such as safety, ergonomics and process optimization. Now, we delve deeper into the fundamental aspect that underlies the entire process of HMA which is data acquisition. In this section, we will introduce the technologies employed for the acquisition of human movement data, highlighting their advantages and limitations. Additionally, we enumerate the industrial datasets obtained through the use of these technologies. Finally, we provide insights into the current state-of-the-art datasets.

4.1. Technologies

In recent years, a multitude of technologies have been developed to acquire and analyze operator movements and displacements in industrial environments (Menolotto et al., 2020). Among these, motion capture (MoCap) technologies have been widely used to extract the human skeleton. Additionally, some studies utilized muscle map activity to improve the analysis of human hand displacement using signal-based sensors such as electromyography (EMG) (Al-Amin et al., 2019; Kim et al., 2021). In the following, we provide an overview of all the technologies employed for acquiring and analyzing operator movements in industrial settings. The proposed taxonomy and comparative study are depicted in Fig. 3, Fig. 4, Table 3 and Table 4. This provides a practical illustration of how these technologies can be utilized. It helps researchers and practitioners gain a better understanding of the capabilities and limitations of each technology and facilitates informed decision-making when selecting the most appropriate motion capture method for their specific case studies. Furthermore, we present a comparative study of these technologies, highlighting their advantages and limitations when used alone or in combination for specific industrial applications.

4.1.1. MoCap technologies

Motion capture technology has been increasingly used in recent years. We present a taxonomy that categorizes MoCaps into optical systems and non-optical systems, as illustrated in Fig. 3.

Optical systems:. Cameras are devices capable of capturing a diverse range of reflections, which can indirectly measure human body movements by recording images in the form of pixel arrays. These images can be used to determine the positions of various body joints, enabling the tracking of human motion (Chen et al., 2020b). Optical sensors are particularly useful in providing a comprehensive understanding of the environment, making them ideal for monitoring the workspace, ensuring personal safety and detecting the presence of objects (Maurice et al., 2019). However, their performance is greatly

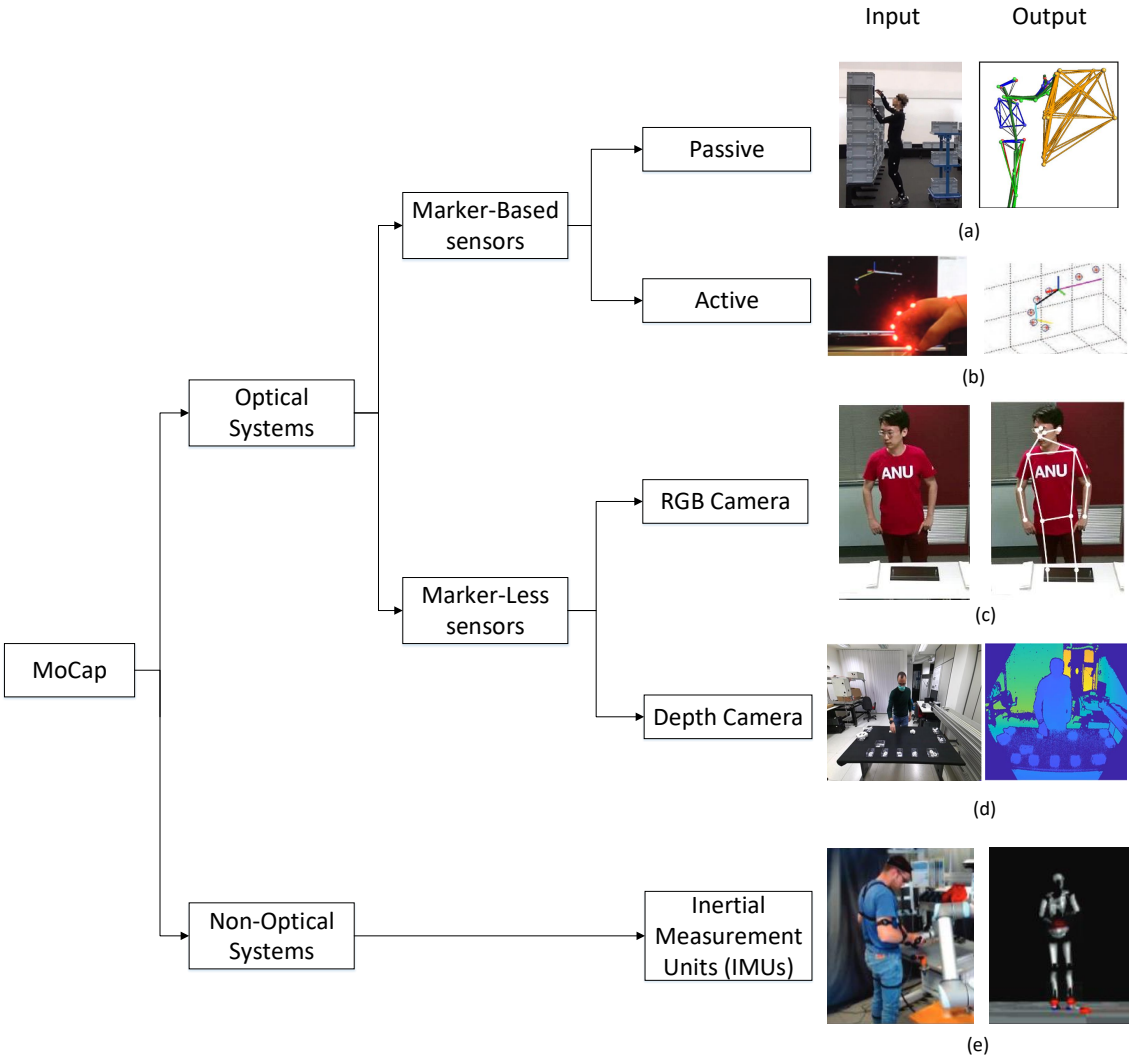


Figure 3: MoCap technologies. Examples of data acquired by: (a) Optical MoCap suit (Niemann et al., 2020) (b) Active LED markers (Yun et al., 2013) (c) Kinect V2 (Ben-Shabat et al., 2021) (d) Kinect V1 (Cicirelli et al., 2022) (e) Perception Neuron 32 (?).

influenced by environmental conditions. For example, depth sensors are sensitive to reflective surfaces, while RGBD (Red Green Blue and Depth) sensors are susceptible to strong variations in brightness (Bonci et al., 2021). Optical-based sensors can be classified into two main types:

- Marker-Based sensors: Marker-based sensors require the attachment of markers to target regions such as limbs (Khan et al., 2020), which must be visually distinguishable from the surrounding environment so that cameras can easily recognize and process their patterns. Niemann et al. (2020) used 39 reflective markers placed

on an optical MoCap suit to track human movement during logistic activities. In a separate study, they also tracked 12 different objects using optical MoCap markers to capture rigid objects (Niemann et al., 2021). Tamantini et al. (2021) employed optoelectronic MoCap to acquire data, which can be used to estimate factors that may contribute to MSDs and therefore prevent their occurrence. While camera marker-based sensors are well-suited for tracking human movements, they are not appropriate for industrial applications due to the long preparation times required for marker setup (Colyer et al., 2018).

- Marker-Less sensors: Marker-less sensors rely primarily on camera-based sensors for motion tracking. Iodice et al. (2022) utilized an RGBD Realsense D435i camera to capture a complete visual coverage of the workspace, including the operator. Ben-Shabat et al. (2021) employed three Kinect V2 cameras capturing simultaneously three RGB views (front, side and top) and one depth view (front) of the workspace during assembly tasks in real-time. Cicirelli et al. (2022) utilized a Kinect V1 sensor equipped with RGBD cameras. This type of sensor provides users with more freedom and has no setup time, making it ideal for tracking full-body movements. However, they are sensitive to environmental factors such as lighting conditions, temperature and the presence of other objects in the background, which results in lower accuracy compared to marker-based sensors (Colyer et al., 2018).

Inertial Measurement Units (IMUs):. IMUs are devices commonly used for acquiring human motion data, consisting of a set of three accelerometers and three gyroscopes (Ribeiro et al., 2020). They can be affixed to various parts of the human body, including the wrist, ankle, or waist (Attal et al., 2015). Multiple IMUs are often employed in the literature to accurately analyze human pose and movement. For instance, researchers have used IMUs to track the motion of the arms during assembly tasks (Al-Amin et al., 2019), record whole-body kinematics during various activities (Maurice et al., 2019) and generate detailed 3D models of operators while performing assembly tasks (?). IMUs are lightweight, simple and easy to deploy, making them suitable for sports training, clinical biomechanics and industrial applications (McGinnis, 2013). However, they are prone to accuracy degradation over extended periods of use due to factors such as imperfections and drift (Ribeiro et al., 2020). For instance, Al-Amin et al. (2019) used two IMUs placed on each arm to track their motion during the assembly of a 3D printer. Other researchers used wearable IMUs to generate a more detailed 3D model of the operator such as ?. They used a Perception Neuron 32 v2 system to capture skeletal data by tracking 17 major body joints of the human body while performing assembly tasks. Maurice et al. (2019) record whole-body kinematics using the Xsens MVN Link system. They equipped participants with 17 IMUs placed all over the body to measure the orientation of the body segments while performing tasks, including assembly and movement in the working area.

In Table 3, a comparison of various MoCap systems is presented. The accuracy of marker-based sensors is directly proportional to the number of markers and infrared (IR) cameras used, indicating the higher the better. However, the setup process for markers is both time-consuming and expensive, making it impractical for industrial workers. Therefore, marker-less sensors are preferred due to their ease of deployment, affordability and better representation of the environment. Nevertheless, marker-less sensors are susceptible to external environmental factors such as lighting, temperature and occlusions. To

obtain high-quality recordings, proper camera placement at suitable angles and locations is critical.

Table 3: A comparative representation of MoCap systems for human movement data acquisition in industry (inspired by (Menolotto et al., 2020)).

			Accuracy	Setup	Cost	Portability	Limitations	Modality	Case study
MoCap Systems	Optical	Marker-Based	Very high (0.1 mm and 0.5°); subject to number/ location of cameras	Time-consuming, frequent calibrations	[USD 5000 - USD 150,000]	Limited	Camera obstructions	RGB, IR, Depth, Point-Cloud, Skeleton	(Tamantini et al., 2021) (Niemann et al., 2020) (Niemann et al., 2021) (Colyer et al., 2018)
		Marker-Less	Low (static, 0.0348 m) subject to distance from camera	Checkerboard calibrations	USD 200(unit)	Yes	Camera obstructions, difficulties tracking bright/dark objects		(Iodice et al., 2022) (Ben-Shabat et al., 2021) (Cicirelli et al., 2022)
	Non-Optical	IMUs	High (0.75° to 1.5°)	Straightforward; subject to number of IMUs	[USD 50 (unit) - USD +12,000 (full-body)]	Yes	Drift, bias, random noise, factor imperfections, misalignment (Ribeiro et al., 2020)	Skeleton	(Maurice et al., 2019) (?) (Al-Amin et al., 2019)

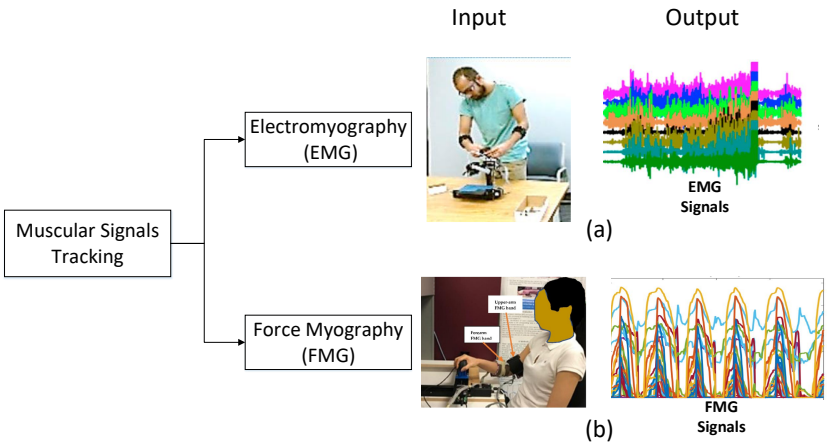


Figure 4: Muscular signal tracking sensors. Examples of data acquired using: (a) Myo armbands (Al-Amin et al., 2019) (b) Force Myography bands (Zakia and Menon, 2022).

4.1.2. Muscular Signal-Based sensors

Muscle signal-based sensors have been utilized in various domains, including healthcare, sports and more recently, the industrial sector. We present two commonly used types of signal-based sensors named surface electrocardiography and force myography (FMG), as depicted in Fig. 4.

Surface Electrocardiography (sEMG) is a method of capturing multi-point electromyography recordings of muscles in real-time during dynamic movements. These sensors map muscle activity to provide more precise data that can be used for biomechanical analysis, ergonomic risk assessment and gesture recognition (Bassani et al., 2021). Previous studies have utilized sEMG sensors, such as Al-Amin et al. (2019) who used two sensors attached to an operator's left and right arm to monitor muscular activity during assembly tasks, and Kim et al. (2021) who used sEMG sensors to estimate muscular fatigue when manipulating assembly tools. EMG provides rich information as it is a direct measure of muscle activation preceding muscle movement (Jiang et al., 2020). This is since

EMG signals are detected when the brain sends instructions to the muscles to control movement. However, there are some limitations to EMG techniques. For example, the signal is non-stationary and can be affected by noise and artifacts that may occur during physical activity or sweating (Zheng et al., 2022).

FMG is a modern and non-invasive wearable technology that can measure muscle activity during muscle contractions and expansions (Zakia and Menon, 2022; Xiao and Menon, 2019; Barioul et al., 2020; Wu et al., 2020). FMG can be used to recognize the applied hand forces of humans during physical human-robot interactions (pHRI) by detecting the force of each target muscle through pressure measurements against the contacting muscle (Jiang et al., 2020). FMG sensors have several advantages over sEMG sensors, including lower cost, smaller signal processing units, Bluetooth technology and ease of wear. Additionally, FMG provides more accurate fatigue parameters than EMG for high-speed motion (Prakash et al., 2021). Consequently, FMG sensors are an excellent choice for human-robot interaction (HRI) projects (Zheng et al., 2022). Zakia and Menon (2022) conducted an experiment to study human intentions of manipulating a linear robot/biaxial stage using FMG sensors to estimate the interactive force between the operator and the robot.

Combining EMG and FMG could potentially compensate for the limitations of each modality and improve overall performance (Jiang et al., 2020). Ke et al. (2020) presented a novel modular EMG-FMG sensor to improve the accuracy of recognition of hand gestures. This underscores the benefits of sensor fusion in human motion analysis. Each type of sensor has its own advantages and limitations. Combining multiple technologies can ensure better accuracy in data acquisition (Sun et al., 2022).

Table 4: A comparative representation of muscle signal tracking systems for human movement data acquisition in industry.

	Accuracy	Setup	Cost	Portability	Limitations	Modality	usage
EMG	77.8%	Along the longitudinal midline of desired muscle parallel to muscle fibers.	[USD 1900 - USD 2400]	Yes	Limited muscle representation	Signal	(Al-Amin et al., 2019) (Kim et al., 2021)
FMG	68.9 - 99%	Force sensing resistors arranged as a grid in a portable wearable band (Delva et al., 2020) calibration	[USD 8 - USD 144]	Yes	Sensor shifting (Xiao and Menon, 2019)		(Zakia and Menon, 2022)

4.1.3. Hybrid technologies

Hybrid datasets have been proposed to combine the benefits of IMUs and optical-based sensors such as RGB cameras for improved tracking of human movements. ? utilized an IMU-based system to capture skeletal data, while simultaneously using 3 cameras to capture RGB data from different viewpoints. Similarly, Maurice et al. (2019) combined Qualisys MoCap sensors with IMUs, an e-glove was used to monitor hand and finger flexion, and RGB cameras were utilized for video recording. Yoshimura et al. (2022) utilized IMU units to acquire acceleration data on three axes, as well as gyroscope and quaternion data, by attaching them to the subject's left and right wrists and upper arms. They also installed Kinect and LiDAR sensors as front-view cameras and an RGB camera as a top-view camera. The LiDAR sensor was considered effective in accurately tracking the subject's position when they were away from the workbench. The extracted skeleton data from these approaches has demonstrated improved recognition of coarse actions involved with arm motion.

As previously discussed, while IMUs and optical-based sensors can offer precise data, they have limitations in certain applications. For example, MoCap sensors allow us to get the ergonomic RULA score for the posture but lack the information about the load raised by the operator during activity execution. Optical sensors or IMUs cannot differentiate between actions performed with similar postures but varying load intensities. To overcome this limitation, other types of sensors are required, such as EMG and FMG sensors, which can provide valuable insights into muscle activity when used in conjunction with MoCap technologies in an ergonomic setting. Kim et al. (2021) employed a combination of RGB cameras, EMG sensors and IMUs to improve the accuracy of human motion recognition. Al-Amin et al. (2019) utilized Myo armbands, which integrate both IMUs and EMG sensors to recognize fine actions associated with finger motion with high precision. Therefore, depending on the use case specifications and available resources, different technologies can be used independently or in combination for various projects.

4.2. Datasets

Prior research has indicated a deficiency of datasets on human motion analysis in the industrial sector, which was brought to attention. However, recent developments reveal a shift in this trend as an increasing number of datasets are being made available in the industry for training and evaluating HMA techniques. A comprehensive list of these datasets is presented in Table 5.

For safety applications, Iodice et al. (2022) have proposed a 3000 video dataset for industrial HRI denoted as HRI30. The dataset consists of three distinct sets of actions, which include human-object interaction actions, actions without interaction and collaboration and finally, collaborative and end-collaborative actions. Additionally, Mohammadi et al. (2020) have presented the Physical Human-Robot Contact Detection dataset to facilitate collision detection and human movement anticipation. Munasinghe et al. (2022) have introduced the COVERED dataset, which employs multi-LiDAR systems comprising four-point cloud cameras to overcome occlusion and acquire high-resolution data for identifying elements such as robots, humans and AGVs in an industrial environment.

Several datasets have been introduced for estimating ergonomic indices. Sedlar et al. (2023) have proposed the Imitrob dataset for training and evaluating 6D object pose estimators. Tamantini et al. (2021) have proposed the WGD—Working Gesture Dataset, which aims to prevent MSDs. Moreover, Maurice et al. (2019) have presented the AnyData-lab-onePerson dataset, which focuses on industry-oriented activities where participants assume various postures to minimize operator load and enhance the working experience.

Numerous datasets have been proposed to assist operators in industry by identifying mistakes or providing tools during manufacturing tasks. ? have introduced the InHARD dataset and its virtual representation InHARD-DT (Dallel et al., 2023), based on a real use-case in an industrial environment, to generate self-labeled data. Lagamtzis et al. (2022) have presented CoAx, a collaborative action dataset for human motion forecasting in an industrial workspace. Yoshimura et al. (2022) have introduced OpenPack, a large-scale dataset for recognizing packaging works in Internet of Things (IoT) enabled logistic environments. For toy assembly tasks, Ragusa et al. (2021) have provided a challenging egocentric MECCANO dataset for human-object interactions in industry, while Sener et al. (2022) have proposed a more comprehensive multi-modal Assembly101

dataset, which investigates recognition, anticipation, temporal segmentation and mistake detection based on 3D hand poses.

In the context of flow analysis and navigation, datasets such as THOR dataset (Rudenko et al., 2020) and its extended version Magni dataset (Schreiter et al., 2022) have been introduced for motion forecasting in industrial environments. Furthermore, Delamare et al. (2020) have proposed a novel dataset utilizing ultra-wideband (UWB) and MoCap systems to monitor worker movements during tricycle assembly in an industrial setting.

To automate assembly planning, a multi-modal and multi-view dataset named the IKEA ASM dataset (Ben-Shabat et al., 2021) has been proposed. It comprises various scenarios of furniture assembly with different backgrounds, which enables robust recognition of human activities through actions, objects and poses.

Several datasets have been published to improve productivity and efficiency in industry. The listed datasets address various aspects of this goal, including action or gesture detection, recognition and anticipation. Cicirelli et al. (2022) have presented the multi-modal HA4M dataset designed for recognizing human actions in complex assembly tasks in manufacturing. The HA4M dataset includes RGB, depth, IR, skeleton, point cloud and RGB-A (RGB-depth Aligned) modalities, providing a robust foundation for developing, validating and testing methodologies for recognizing assembly actions. In the context of logistics, Niemann et al. (2020) have proposed the LARa dataset, which uses attribute representation for HAR, followed by the CAARL dataset (Niemann et al., 2021), a context-aware activity recognition in logistics dataset that focuses more on object representation to explore the potential of context information for HAR. Additionally, Sturm et al. (2023) have introduced the HAD-V1 dataset for vision-based human hand action recognition in industrial assembly.

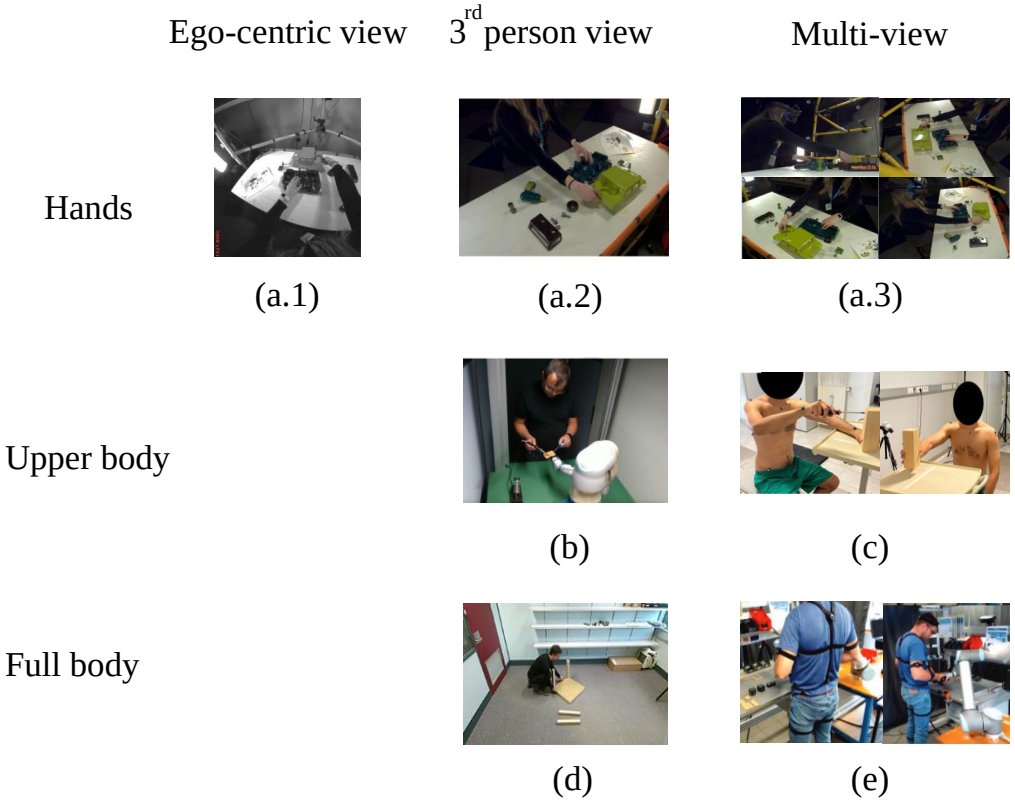


Figure 5: Illustrations from reviewed datasets grouped by tracked body parts, type of interactions and views. (a): (Sener et al., 2022) (a.1) Human-object interaction (a.2) Human-object interaction,(a.3) Human-object interaction, (b) (Lagamtzis et al., 2022) Human-robot interaction, (c) (Tamantini et al., 2021) Human-object interaction,(d) (Ben-Shabat et al., 2021) Human-object interaction, (e) (?) Human-robot interaction + human-object interaction.

Table 5: Selected datasets for industrial human motion analysis.
U: Untrimmed sequences, T: Trimmed sequences, HO: Human-Object interaction, HR: Human-Robot interaction, NI: No Interaction, P: Partial body part representation (hand, arm), Up: Upper body representation, F: Full body representation, E: Egocentric camera view, A: Allocentric (3rd-person) camera view, NA: Not Applicable, MNP: Model Not Precised.

Dataset	Year	U	T	Nbr of subjects	Nbr of classes	Nbr of samples	Nbr of views	Point of views	Human body	Technologies	Types of interaction			Modalities	Public
											HO	HR	NI		
IHAD$_V$1(*Sturmet al.*, 2023)	2023	✓		-	12	-	1	A	P	UI-326xCP-C camera	✓			RGB + Skeleton 3D	
HRI30 (Iodice et al., 2022)	2022		✓	11	30	2 940	1	A	F	Intel Realsense D435i camera	✓	✓	✓	RGB	✓
CoAx (Lagamtzis et al., 2022)	2022	✓		6	9	180	1	A	Up	Intel Realsense D435 camera	✓	✓		RGB, Depth, Pointcloud, 2D and 3D hand pose	
OpenPack (Yoshimura et al., 2022)	2022	✓		16	10	20129	1	A	Up	Kinect V1, LiDAR + Intel RealSense D534i camera	✓			Acceleration, Gyroscope, Quaternion, , Keypoints, Blood volume pulse (BVP), Electrodermal activity (EDA) data, Point cloud, Depth	
InHARD-DT (Dallel et al., 2023)	2022	✓		12	18	4 799	3	A	F	C920 camera, HTC Vive Pro Eye Digital twin (headset and controllers) + Perception Neuron 32	✓	✓		RGB + Skeleton 3D	✓
HA4M (Cicirelli et al., 2022)	2022	✓		41	12	4 124	1	A	F	Kinect V1	✓			RGB, Depth, IR, Skeleton, Point cloud, RGB-A	✓
Assembly101 (Sener et al., 2022)	2022	✓		53	1 380	1 013 523	12	E, A	P	8 RGB cam (MNP) + 4 monochrome cam (MNP) on the headset	✓			RGB, Depth, 3D hand Skeleton	✓
COVERED (Munasinghe et al., 2022)	2022	✓		-	6	218	4	A	F	Ouster OS0-128 LiDAR			✓	Point cloud	✓
CAARL (Niemann et al., 2021)	2021	✓		2	7	70	-	A	F	Optical Marker-based MoCap	✓		✓	IR + Skeleton 3D	✓
WGD (Tamantini et al., 2021)	2021		✓	8	3	-	8	A	Up	optoelectronic marker-based system	✓			RGB + Skeleton 3D	
Physical Human-Robot Contact Detection (Mohammadi et al., 2020)	2021	✓		-	5	2 125	2	A	F	Kinect V2 + robot sensors		✓	✓	RGB, Depth, Skeleton 3D	✓
LARa (Niemann et al., 2020)	2020	✓		14	8	420	NA	A	F	Optical marker-based MoCap, IMU	✓		✓	IR + Skeleton 3D	✓
InHARD (?)	2020	✓		16	14	4 804	3	A	F	C920 cams + Perception Neuron 32	✓	✓		RGB + Skeleton 3D	✓
MECCANO (Ragusa et al., 2021)	2020		✓	20	61	8 858	1	E	p	Intel RealSence SR300 camera	✓			RGB, Depth + Gaze	✓
IKEA ASM (Ben-Shabat et al., 2021)	2020	✓		48	33	17 577	3	A	F	Kinect V2	✓			RGB, Debth + Skeleton 3D	✓
AndyData-lab-onePerson (Maurice et al., 2019)	2019	✓		13	6	195	2	A	F	Xsens MVN Link, Qualisys motion capture system, e-glove + Video camera (MNP)	✓			RGB + Skeleton 3D	✓

By conducting a thorough analysis of Table 5 and evaluating the criteria outlined in existing datasets, valuable observations can be made, and a range of diverse challenges can be identified as follows:

- Trimmed vs Untrimmed: As shown in Fig. 6. (c), more datasets are published for untrimmed human action sequences than for segmented and clipped sequences. This growing trend highlights the demand for online solutions that can be applied to real-world industries.

- Variation in human action execution, time, order and expertise (skills): Human activities in the industrial context can greatly vary based on the expertise and skills of individual operators, leading to differences in the way actions are executed, the order in which tasks are performed and the time taken to complete them. Recent datasets have begun to address these complexities, such as the Assembly101 dataset (Sener et al., 2022), which takes into account the temporal ordering of tasks with a high degree of variation. Additionally, it annotates skill levels based on task execution speed and mistakes, providing a more realistic representation of industrial activities. For example, experts tend to complete tasks faster than novice workers.

- Variation in data distribution: The scarcity of industrial HMA datasets is still an issue, as most published datasets exhibit highly imbalanced data distributions that result in a significant variation in action representation. To gain insight into the distribution of datasets, we calculated the mean, variance and standard deviation (SD) of several datasets while excluding the "no action" class shown in Table 6. The selection of datasets was based on the availability of meta-data, which facilitated the computation of these measures. A higher SD indicates greater data dispersion from the mean, reflecting a highly imbalanced dataset. For instance, Assembly101 (Sener et al., 2022) has a high SD, indicating significant dispersion, with only the "pickup" and "put down" classes having the same number of samples as the sum of 18 other classes, such as "shake" and "pull". In CAARL dataset (Niemann et al., 2021), approximately 70% of the labeled actions belong to the "Handling" class. For the MECCANO dataset (Ragusa et al., 2021), only two actions, "take" and "check," account for 50% of the labeled actions. Although the InHARD dataset (?) has a better SD, actions such as "assemble system" and "picking left" account for more than 50% of the dataset samples. The HA4M (Cicirelli et al., 2022) dataset has a low SD, but 47% of the data are represented by only three actions, such as "Pick up/Place Gear Bearings," "Pick up/Place Planet Gears," and "Pick up/Place Screw," while other actions, such as "pulling" and "turning," are infrequent. We notice that the digital twin dataset (InHARD-DT) has the lowest SD with the same mean as its physical twin InHARD. In addition to these measures, the varying length of action sequences, which can range from a second to tens of seconds, such as "walking with a polisher," poses a challenge for accurately identifying actions in real-time (Koch et al., 2022).

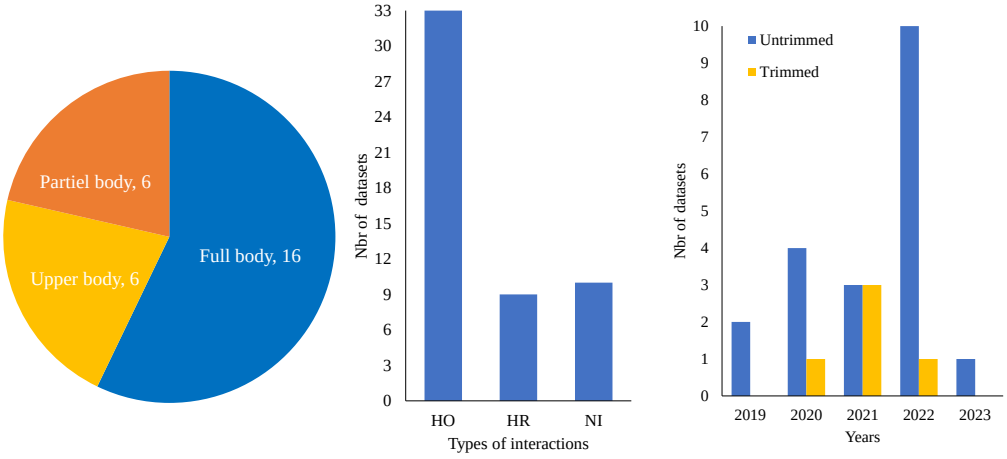
- Human body parts, type of interaction and viewpoints variation: Different configurations of HMA can be applied based on the specific case study, as illustrated in Fig. 5. For example, in static assembly tasks, only the upper body or arms are monitored using cameras or signal-based sensors, which is the case in datasets like

Table 6: Mean and standard deviation of sample numbers per class in some dataset examples.

Datasets	Mean $\pm SD$
HA4M (Cicirelli et al., 2022)	307.42 ± 165
InHARD-DT (Dallel et al., 2023)	369.15 ± 112
InHARD (?)	369.46 ± 347
MECCANO (Ragusa et al., 2021)	737.75 ± 793
CAARL (Niemann et al., 2021)	955.50 ± 1245
Assembly101 (Sener et al., 2022)	42 479.17 ± 48 666

MECCANO (Ragusa et al., 2021), WGD (Tamantini et al., 2021), Assembly101 (Sener et al., 2022) and HAR-V1 (Sturm et al., 2023). On the other hand, when the operator is moving in the working area, the whole body is monitored using IMUs or cameras, as seen in datasets like HRI30 (Iodice et al., 2022), Magni (Schreiter et al., 2022), THOR (Rudenko et al., 2020), COVERED (Munasinghe et al., 2022), LARA (Niemann et al., 2020) and CAARL (Niemann et al., 2021). The listed datasets exhibit two main types of interactions, namely human-object (HO) interactions and human-robot (HR) interactions, which are identified as verbs in some datasets. The InHARD dataset (?) solely employs verbs to recognize actions, a method that can be refined by concatenating verbs and objects as presented in the Assembly101 (Sener et al., 2022) and MECCANO (Ragusa et al., 2021) dataset. Although the existing datasets such as the IKEA ASM dataset (Ben-Shabat et al., 2021) have been experimented within various environments, recognizing human movements in an industrial environment is challenging due to occlusions caused by limited viewpoints, poor lighting conditions, moving entities in the background and machinery. To minimize occlusions and ease both recognition of human actions and gestures, some works have suggested multiple viewpoints. For example the WGD dataset (Tamantini et al., 2021) used 8 RGB cameras in front and behind the operator to capture upper body images. Assembly101 dataset (Sener et al., 2022) used 8 static cameras with 4 egocentric cameras for gesture recognition. In contrast, others monitored the entire operator body while moving in the workspace, carrying objects and doing packaging activities as in CAARL (Niemann et al., 2021) and LARa datasets (Niemann et al., 2020). Fig. 6 (a) and (b) illustrate the frequency of body parts representation and the distributions of the number of datasets regarding types of interactions during experiments of previously presented datasets.

- Human intention awareness: The datasets presented in this study have played a crucial role in facilitating the detection, recognition and anticipation of human actions across a wide range of contexts. To enhance workplace safety, hybrid approaches were proposed. These approaches integrate human action detection and recognition to enable robots to proactively anticipate human movements and prevent contact or collisions, thereby promoting faster and safer HRI. However, despite these advancements, there remains a need for further progress in enhancing the accuracy of action detection (Maurice et al., 2019; Mohammadi et al., 2020). Alternative approaches were also introduced, which involved identifying verbs, active objects



(a) (b) (c)

Figure 6: Statistics for selected datasets shown in Table 5. (a) Frequency of body parts representation. (b) Distributions of the number of datasets regarding types of interactions. (c) Distribution of the number of trimmed/untrimmed datasets per year.

and fine-grained actions to anticipate the next action. These approaches offer a more detailed understanding of human intention. For example, in the scenario of car assembly, the enumeration of wheels allowed for the inference of four successive actions of "put wheel," considering that cars typically possess four wheels. Nonetheless, it is important to acknowledge that this approach has limitations when it comes to situations where the sequential order of actions is not strictly followed (Sener et al., 2022). Moreover, several other datasets, such as LARa (Niemann et al., 2020), CAARL (Niemann et al., 2021) and IKEA ASM (Ben-Shabat et al., 2021), have focused on enhancing context awareness by monitoring both the operator and various objects like tables and tools, such as screwdrivers. This comprehensive monitoring approach has demonstrated notable improvements in the accuracy of action recognition and the generalizability of the model.

- Technologies: The datasets presented in this study utilized a diverse range of sensors for data acquisition. A majority of the datasets leveraged RGBD cameras to capture operator movements within the workspace, as demonstrated in HRI30 (Iodice et al., 2022), HA4M (Cicirelli et al., 2022) and IHADv1 (Sturm et al., 2023) datasets. Other datasets used motion capture IMUs in conjunction with cameras to acquire precise information regarding the movement and navigation of humans and robots within workstations. Examples of such datasets include InHARD (?) andyDataLab (Maurice et al., 2019), CAARL (Niemann et al., 2021) and LARa (Niemann et al., 2020) datasets. Lastly, Mohammadi et al. (2020) used torque joints embedded in robot joints to avoid collisions and contact between humans and robots. Although combining multiple sensor modalities can improve accuracy, it is important to acknowledge that this approach also introduces complexity and requires extra resources for both data acquisition and processing.

- Modalities: In the listed datasets, different modalities were utilized to capture human actions. The CAARL (Niemann et al., 2021), CoAx (Lagamtzis et al., 2022) and HRI30 (Iodice et al., 2022) datasets used RGB modality. The WGD (Tamantini et al., 2021) andy-lab datast (Maurice et al., 2019), LARa (Niemann et al., 2020), InHARD (?) and IKEA ASM (Ben-Shabat et al., 2021) datasets used RGB and skeleton modality. Skeletons are preferred over RGB for recognizing human actions, as they are more adaptable to different subjects and environments. Considering hybrid modality, OpenPack (Yoshimura et al., 2022) utilized accelerometers, skeleton and RGB data, while HA4M (Cicirelli et al., 2022) used RGB video, depth, IR and skeleton. Although using multiple modalities can improve performance, it requires additional resources for acquisition and processing, leading to increased complexity and a larger volume of data needed for prediction. We noticed a diversity of data modalities used in the listed datasets. However, fusing data from multiple modalities requires extra resources for its acquisition and processing. Also, it adds complexity and may impact performance due to the large volume of data required for the prediction.

- Digital twin and auto labeling: The process of data annotation can be a significant challenge, requiring substantial time and resources for manual labeling and evaluation. To address this issue, Dallel et al. (2023) utilized a digital twin approach to generate data, which involves the recreation of a real-world environment and the assignment of triggers to each action to facilitate precise auto-labeling of data. This approach can help to balance datasets by reproducing less frequently occurring actions and addressing the problem of imbalanced datasets. Similarly, Vysocky et al. (2022) created synthetic data for training depth images for industrial hand tracking and proposed a dataset for industrial applications of hand localization.

4.3. Summary of findings

Human motion analysis involves the examination of operator movements, interactions and physiological signals. To accomplish this, various technologies are employed, including Mocaps and signal-based sensors. These technologies yield various data modalities, such as skeleton, RGB, depth and signal data. Notably, Mocaps, particularly Kinect cameras and IMUs are the prevailing choices, with signal-based sensors being less common. A total of 16 datasets have been identified, primarily oriented toward manufacturing applications, particularly assembly operations. The majority of these datasets offers RGB and skeleton data, primarily captured using Mocaps. It's important to note that 14 of these datasets utilize multiple sensors to generate various data modalities.

5. Online human motion analysis approaches and metrics

Human motion analysis involves studying and understanding the patterns, characteristics and dynamics of human movements. It is of utmost importance in the industrial sector as it contributes significantly to enhancing productivity, efficiency, safety and ergonomy. Within this context, the analysis of human motion encompasses multiple aspects, including action detection, action recognition, hand gesture recognition and action forecasting. Action recognition specifically focuses on the identification and recognition of continuous events or actions within a video sequence (Deng et al., 2023). Numerous

models have been proposed to address the task of human action recognition (Ding et al., 2022; Donisi et al., 2022; Fan et al., 2022). We review HMA methods with a specific focus on HAR methods applied to industrial case studies.

5.1. Online metrics

The evaluation of HMA methods involves various metrics for validating and providing feedback on their performances. Table 7 summarizes the used metrics to evaluate HMA approaches in industry. Accuracy is the most commonly used metric in the studied literature. However, it should be noted that many datasets suffer from class imbalance, with certain classes having more samples than others. While reported performances often demonstrate high accuracy levels approaching 100%, a more comprehensive assessment is needed. The use of the F1-score, which computes the mean and weighted average of precision and recall, provides a fairer evaluation of HMA performance.

When evaluating action recognition algorithms on untrimmed sequences, it is important to consider metrics that account for temporal action detection and spatiotemporal action detection. For temporal action detection, mean average precision (mAP) is used which involves the calculation of average precision (AP) for individual action classes using multiple intersections over union (IoU) thresholds. Also, a variety of regression metrics are utilized for the estimation of human or robot trajectories to reduce estimation errors (Sedlar et al., 2023). These metrics serve in evaluating approaches for ergonomics and safety applications, aimed at preventing accidents and physical interaction between operators and robots in a collaborative working environment. Considering the computational cost, the metric of floating point operations (FLOPs) can be employed, but also execution time that regroups inference time, detection delay and post-processing time (Kwon et al., 2021). Finally, based on the information presented in Table 8, the predominant metrics employed for evaluating benchmark datasets are accuracy, F1 score and mAP over various IoU thresholds.

5.2. Online approaches

Arshad et al. (2022) claim that CNN has been the most widely utilized technique in 25% of the literature on human activity recognition, closely followed by long-short term memory (LSTM) at 13% and support vector machine (SVM) at 12% while other machine learning techniques are less commonly utilized. Besides, recent deep learning techniques demonstrate encouraging performances for spatio-temporal feature learning (Le et al., 2022). In addition, transformers with convolution-free networks have shown efficacy in human action recognition and other fields of computer vision (Ding et al., 2022; Le et al., 2022; Menolotto et al., 2020; Reining et al., 2019; Fan et al., 2022). Table 8 lists approaches that use deep learning techniques for HMA in industry including human action detection, gesture recognition and action anticipation with a major focus on HAR.

In this study, given the scarcity of research utilizing recurrent neural network (RNN)-based methods within the scope of this paper, we made a deliberate decision to categorize deep learning approaches into four distinct categories: convolution neural network-based, graph convolution network (GCN)-based, attention-based and hybrid methods. As shown in Fig. 7, the chart illustrates the approaches listed in Table 8 according to the year of publication. This trend reveals a gradual decline in the utilization of CNN-based methodologies with the rise of GCNs and attention-based ones. Furthermore, hybrid approaches are increasingly adopted, emerging as the predominant methodology.

Table 7: Used metrics to evaluate human motion analysis approaches.
TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, P: predicted labels, GT: Ground Truth, IoU: Intersection over Union, L: Action length, l: temporal interval between the frame of the starting GT and the frame of the predicted label, p: predicted keypoint, q: keypoint true location, d: distance, $Recall_i$: Recall at threshold i, ADD: Average distance between the corresponding predicted and reference vertices and centroid of the object bounding box, UWB: Ultra-Wide-Band, Ti: Inference time, Dd: Detection delay, Tpp: Time post-processing, P^i_{ref}: reference vertices, P^i_{pre} predicted vertices, x :real value of the ith observation, y: predicted value of the ith observation, n: total number of observations,k: total number of keypoints, v: total number of vertices.

	Metric	Formula	Goal	Reference
General Common Metrics	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Determines how often the model's predictions match the actual labels.	(?), (Munasinghe et al., 2022), (Ragusa et al., 2021), (Sener et al., 2022), (Wang et al., 2019), (Tassi et al., 2022), (Liu et al., 2023a), (Zhang et al., 2021), (Upadhyay et al.), (Hernandez et al., 2021), (Al-Amin et al., 2019), (Belay Tuli et al., 2022), (Moutinho et al., 2023), (Mohammadi et al., 2020), (Koch et al., 2022), (Dallel et al., 2022)
	Precision	$TP/(TP+FP)$	Focuses on the accuracy of positive predictions. It helps evaluate how precise and reliable the model is when it predicts a positive outcome. calculates the ratio of correct predictions for a specific category.	(Niemann et al., 2020), (Sener et al., 2022), (Kwon et al., 2021), (Zamora et al., 2021)
	Recall	$TP/(TP+FN)$	Quantifies the model's ability to find all relevant samples without missing any, minimizing the number of false negatives. measures how many relevant elements were detected.	(Niemann et al., 2020), (Upadhyay et al.), (Kwon et al., 2021), (Zamora et al., 2021)
	F1-score	$\frac{2*Precision*Recall}{Precision+Recall}$	Evaluate classification models with tow or more classes particularly when the classes are imbalanced.	(Dallel et al., 2022) (Niemann et al., 2021), (Niemann et al., 2020), (Kwon et al., 2021), (Singhania et al., 2022), (Rana et al., 2023), (Dallel et al., 2023), (Zhang et al., 2022a)
Untrimmed Metrics	IoU	$(P \cap GT)/(P \cup GT)$	Assess the quality of object localization or segmentation algorithms. measures how well the algorithm identifies and accurately delineates objects of interest in an image (spatial) or an event in a sequence (temporal).	(Dallel et al., 2022), (?), (Munasinghe et al., 2022), (Ragusa et al., 2021), (Moutinho et al., 2023)
	Mean Average Precision (mAP)	$\frac{1}{n}\sum^{n}_{Recall_i} Precision(Recall_i)$	Represents the overall performance of an object detection or instance segmentation model, taking into account both precision and recall across different confidence thresholds.	(Upadhyay et al.) (Li et al., 2021)
	Edit-score	$1-(S_{edit}(G,P)/max(M,N))$	Measures the difference $S_{edit}(G,P)$ between predicted segments $P=\{P_1,...,P_N\}$ and ground truth segments $G=\{G_1,...,G_M\}$ to penalize over-segmentation errors.	(Singhania et al., 2022)
	Percentage of Correct Predictions (PCP)	$\frac{1}{k}\sum_{i=1}^{k} d(p_i,q_i) < t$	Indicates the probability that a detected keypoint p is within a distance threshold t, given in pixel, of its true location q. used to assess accuracy.	(Lagamtzis et al., 2022)
	ADD	$\frac{1}{v}\sum_{i=1}^{v} \|P^i_{pre} - P^i_{ref}\|_2$	Estimates the accuracy of object detection for 6D pose estimation.	(Sedlar et al., 2023)
Regression Metrics	Rotation error (Erot)	$\angle(\overrightarrow{P},\overrightarrow{GT})$	Estimate the object orientation error.	(Sedlar et al., 2023)
	Translation error (Etra)	$\|P-GT\|$	Estimate the object location error.	
	Mean Absolute Error (MAE)	$(\frac{1}{n})\sum_{i=1}^{n}\|y_i-x_i\|$	Express average model prediction error by treating all errors equally.	(Pastor et al., 2022)
	Root Mean Squared Error (RMSE)	$\sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i-x_i)^2}$	Express average model prediction error by penalizing larger errors.	(Pastor et al., 2022), (Delamare et al., 2020), (Zhou et al., 2023)
	Geometric Dilution Of Precision (GDOP)	$RMSE_{loc}/RMSE_{range}$	Estimates the error caused by the relative position of other sensors such as GPS or UWB.	(Delamare et al., 2020)
Time Cost	Latency	$Dd = l/L$	Evaluate the delay to detect the action.	(Dallel et al., 2022)
	Processing time	$T_i + D_d + T_{pp}$	Assess and optimize the efficiency and performance of the system.	(Kwon et al., 2021), (Wang et al., 2019)

5.2.1. Convolution-based methods

A CNN model comprises three main types of layers: convolutional, pooling and fully connected, each with distinct functions (Morshed et al., 2023). A general CNN architecture is illustrated in Fig. 8. These models excel in 2D image analysis, leveraging their spatial feature learning capabilities (Sun et al., 2022). By utilizing convolutional layers to

Table 8: Proposed approaches of the state-of-the-art human motion analysis in industrial context. Metrics are defined in Table 7.
P: Private, AR: Action Recognition, VMM: Variable-length Markov Modeling.

	Year	Approach	Goal	Untrimmed	Trimmed	Dataset	Metrics					
							Accuracy	F1-score	Latency	IoU	mAP	Edit-score
CNN-based	2023	CNN (Rana et al., 2023)	recognize human activity from noisy indoor trajectory data using a semi-supervised learning at a tricycle assembly workshop	✓		P	-	0.81-0.95	-	-	-	-
		TSM pretrained on Kinetics-400 (Deng et al., 2023)	evaluate different datasets by supervised and self-supervised learning for action and gesture recognition		✓	InHARD (?)	0.881	-	-	-	-	-
		TSN pretrained on Kinetics-400 (Deng et al., 2023)			✓	MECCANO (Ragusa et al., 2021)	0.411	-	-	-	-	-
	2022	ResNet-50 (Tassi et al., 2022)	recognize human actions and intentions to optimize the robot's reactive behaviour and human ergonomics		✓	HRI30 (Iodice et al., 2022)	0.866	-	-	-	-	-
	2021	CollisionNet (Kwon et al., 2021)	collision detection and localisation	✓		P	-	0.965	-	-	-	-
		CNN (res-net) (Zamora et al., 2021)	evaluate activity of operator (error identification) while manual assembly	✓		P	-	-	-	-	-	-
	2020	3D-CNN (Mohammadi et al., 2020)	recognize human action and detect contact between robot and human to enhance safety	✓		P	0.997	-	-	-	-	-
		1D-CNN (Mohammadi et al., 2020)				PHR Contact Detection Dataset (Mohammadi et al., 2020)	0.96	-	-	-	-	-
	2019	Hierarchical clustering based CNN (Wang et al., 2019)	recognize human actions for work scheduling and productivity	✓		P	0.56	-	-	-	-	-
		Weighted fusion CNNs (Al-Amin et al., 2019)	develop a sensor fusion based system to recognize human actions in performing assembly tasks		✓	P	0.846	-	-	-	-	-
	2018	DCNN (Wang et al., 2018)	recognize human motions and identify the context of associated action in car engine assembly	✓		P (3 actions)	0.96	-	-	-	-	-
RNN-based	2023	LSTM (Orsag et al., 2023)	recognize human spatial-temporal activity using human skeleton	✓		InHARD (4/14 actions including background) (?)	0.38 - 0.67	0.68	-	-	-	-
GCN-based	2023	Mask-GCN (Liu et al., 2023a)	recognize complex human actions with novel motion patterns		✓	P	0.73	-	-	-	-	-
	2022	STGCN-SWMV (Dallel et al., 2022)	recognize human actions	✓		OAD (Li et al., 2016)	0.954	0.953	-	0.977	-	-
						UOW (Tang et al., 2018)	0.934	0.936	0.047	0.958	-	-
		ST-GCN (Dallel et al., 2023)	recognize human actions using both physical and self-generated data		✓	75% InHARD (?) 25% InHARD-DT (Dallel et al., 2023)	0.956	0.955	-	-	-	-
		ST-GCN (Zhang et al., 2022c)	recognize operator's assembly actions during human-robot collaboration		✓	P	0.40	-	-	-	-	-
		MTM-STGCN-SW (Koch et al., 2022)	estimate the assembly progress by extending AR methods to multi-variant processes	✓		InHARD (3/14 actions with no background) (?)	0.593	-	-	-	-	-
	2021	TL-STGCN (Li et al., 2021)	recognize human actions for robot reactive control and decision-making using transfer learning		✓	P	-	-	-	-	0.956	-
Attention-based	2022	P3DAttenNet (Upadhyay et al.)	recognize human actions and generate assembly plans	✓		IKEA ASM (Ben-Shabat et al., 2021)	0.692	0.81-0.95	-	-	0.431	-
	2020	HAMLET (Islam and Iqbal, 2020)	recognize human activities for collaborative robotic systems	✓		UTD-MHAD (Chen et al., 2015)	0.951	-	-	-	-	-
						UT-Kinect (Xia et al., 2012)	0.974	-	-	-	-	-
						UCSD-MIT (Kubota et al., 2019)	0.815	-	-	-	-	-
		SAM (Mahmud et al., 2020)	recognize human actions using features representations from body-worn sensors data	✓		PAMAP2 (Reiss and Stricker, 2012)	-	0.95-0.96	-	-	-	-
						Opportunity (Roggen et al., 2010)	-	0.61-0.67	-	-	-	-
						USC-HAD (Zhang and Sawchuk, 2012)	-	0.50-0.55	-	-	-	-
						Skoda (Stiefmeier et al., 2008)	-	0.93-0.97	-	-	-	-
Hybrid	2023	ResNest-34 + LSTM (Moutinho et al., 2023)	recognize human actions and extract high-level context of industrial assembly operations to be integrated into collaborative assembly plans	✓		P	0.966	-	-	0.941	-	-
		Attention + CNN + Bi-LSTM (Zhou et al., 2023)	estimate human motion trajectory for assembly collaboration	✓		P	-	-	-	-	-	-
		CNN-LSTM (Belay Tuli et al., 2022)	recognize human actions		✓	InHARD (9/14 actions including background) (?)	0.85-0.88	-	-	-	-	-
		C2F-TCN (Singhania et al., 2022)	segment temporal human actions and gestures using fully-supervised and semi-supervised learning	✓		Assembly101 (Sener et al., 2022)	-	0.212	-	0.5	-	0.324
	2022	ConvTransformer (Zhang et al., 2022d)	recognize human actions using knowledge of commonly used human action recognition sensors	✓		Opportunity (Roggen et al., 2010)	-	0.443-0.861	-	-	-	-
						PAMAP2 (Reiss and Stricker, 2012)	-	0.858-0.915	-	-	-	-
						DSADS (Yao et al., 2018)	-	0.846	-	-	-	-
	2021	CNN-VMM (Zhang et al., 2021)	recognize and anticipate human actions based on VMM in engine assembly tasks	✓		P	0.947	-	-	-	-	-
		GRU + CT-HMM (Hernandez et al., 2021)	recognize human tasks at tow levels, primitives and activities for measuring performance and manual labor		✓	P	0.784-0.925	-	-	-	-	-
	2020	TSN + GCN + Attention (Jiao et al., 2020)	recognize human actions in industrial workflows		✓	P	-	0.816	-	-	-	-

capture local features and pooling layers to aggregate information, CNNs can accurately learn discriminative representations for different actions. When combined with efficient

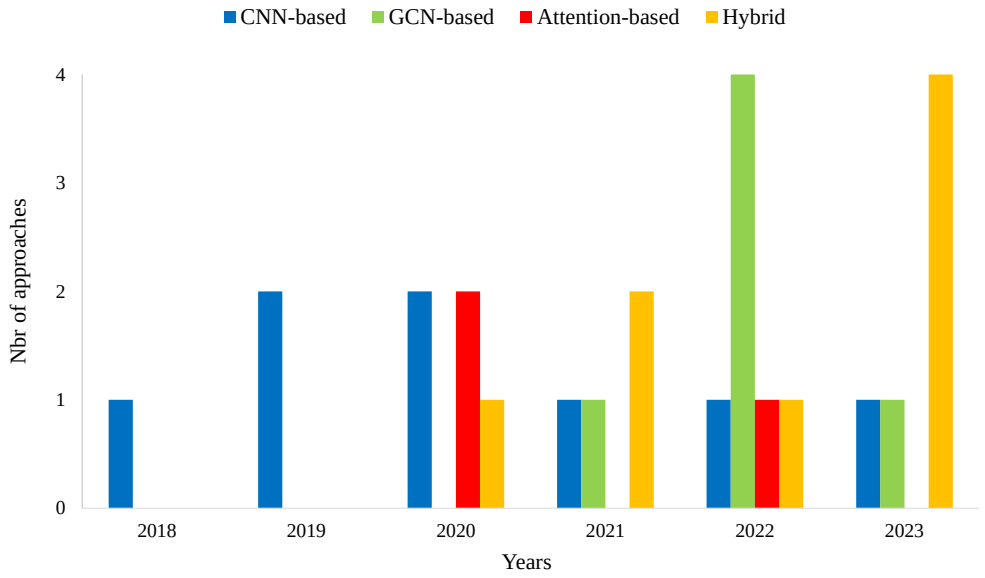


Figure 7: Distribution of the number of approaches in industry per year.

detection and recognition of objects and tools, as outlined in (Büsch et al., 2023), CNNs can significantly enhance the accuracy of action recognition. However, effectively modeling spatio-temporal information for skeleton-based HAR presents a challenge. Advanced approaches tackle this by representing skeleton sequences as pseudo-images for standard CNN processing (Sun et al., 2022). Pseudo-images encode spatial structure within frames and temporal dynamics between them. Extending 2D CNNs to 3D structures captures the crucial spatial and temporal context in videos. Robustness and accuracy make them widely adopted in computer vision tasks, especially action recognition. (Morshed et al., 2023).

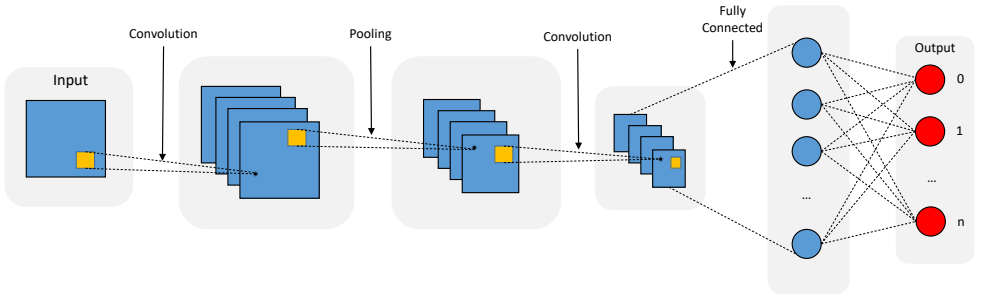


Figure 8: General CNN architecture.

The studied approaches in Table 8 vary in terms of the specific techniques used, datasets employed, model architectures and the problem domains they address. For example, Rana et al. (2023) proposed a framework to enhance the accuracy of HAR in

tricycle assembly tasks by leveraging semi-supervised learning techniques using a CNN classifier tuned based on a stepwise search method. The classification model is trained for activity recognition using the segmented data and the corresponding pseudo labels saving both time and resources by avoiding manually annotating data. Also, Deng et al. (2023) performed a thorough assessment of 2D CNN models for action recognition on a diverse set of 18 datasets. Notably, the TSM (Temporal Shift Module) model demonstrated superior performance on the InHARD dataset, while the TSN (Temporal Segment Networks) model exhibited superior performance on the MECCANO dataset. Another approach conducted by Wang et al. (2018) presented a modified version of the AlexNet deep convolutional neural network (DCNN) for improved recognition of human operator actions. They employed a transfer learning-enabled tuning method and achieved recognition accuracy of over 96% in an experimental case study on car engine assembly. Later, they proposed a method for real-time evaluation of operator actions (Wang et al., 2019). They utilized a CNN classifier for action recognition and applied hierarchical clustering to mitigate confusion among industrial actions. The method was effective in analyzing human actions in a reduced assembly line.

Numerous human action recognition and anticipation approaches were conducted for safety and ergonomic applications such as Tassi et al. (2022) who proposed a framework for mitigating musculoskeletal disorders. They employed action recognition models with surface classification to identify human actions and intentions. Surface classification classifies the sides of the object being handled during assembly tasks to adjust its orientation to match the operator's hand for better collaboration. They used a pre-trained SlowOnly neural network with ResNet50 as the underlying architecture of the action recognition model. The evaluation was performed on the HRI30 dataset, and the results demonstrated remarkable performance in recognizing human actions. In a different study, Kwon et al. (2021) proposed a deep learning-based method for collision identification on articulated robots, building upon the previous work of CollisionNet (Heo et al., 2019). The method detects collisions and accurately determines their locations. To improve accuracy, they employed uncertainty-aware knowledge distillation to transfer knowledge from a larger, more complex model (known as the teacher model) to a smaller, more efficient model (known as the student model). Following the same spirit, Mohammadi et al. (2020) introduced a safety-enhancing approach by combining human action recognition using visual perception and interpreting physical human-robot contact using tactile perception. Their system utilizes skeleton data as input for a 3D-CNN model for action recognition. Additionally, a 1D CNN is employed for contact detection to differentiate intentional and incidental physical contact.

5.2.2. Graph convolution-based methods

Graph-based learning models have gained attention for analyzing graph structures due to their expressive power (Bhatti et al., 2023). Skeleton data naturally takes the form of graphs, and representing it solely as vector sequences or 2D/3D maps fail to capture its complex spatiotemporal configurations and joint correlations (Sun et al., 2022). Therefore, employing topological graph representations is more suitable for effectively representing skeleton data (Ahmad et al., 2021). The GCN model processes an input that comprises an adjacency matrix and a node feature matrix. To capture spatial characteristics among the nodes within the graph, the GCN model utilizes a Fourier domain filter that considers their immediate neighbors. The model can be extended by stacking

multiple convolutional layers, allowing it to operate on the graph's nodes and employ the filter to effectively capture spatial information (Zhao et al., 2024). An architecture of GCN is illustrated in Fig. 9.

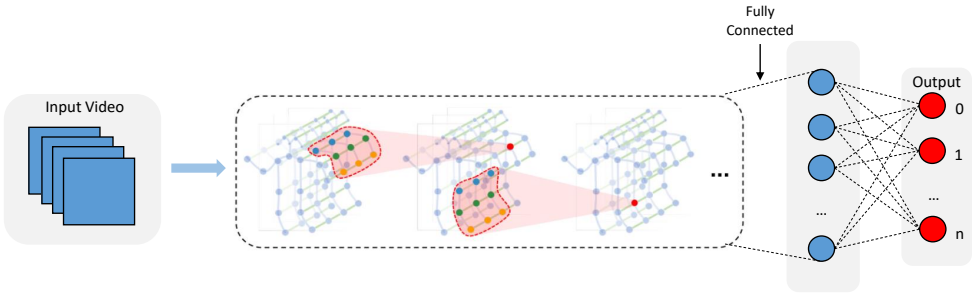


Figure 9: Spatio Temporal GCN architecture inspired from (Ahmad et al., 2021).

Several HAR methods based on GCNs have been proposed, treating the skeleton data as graph structures composed of nodes and edges (Zhu and Deng, 2023; Zhang et al., 2022c). For instance, Dallel et al. (2022) introduced an online HAR approach proposing a new method to the sliding window technique by adding a majority voting system. This skeleton-based method utilizes spatial-temporal graph convolutional networks (ST-GCNs) to automatically learn spatial and temporal information. The approach was evaluated on OAD (Li et al., 2016) and UOW (Tang et al., 2018) daily action datasets, demonstrating superior performance compared to state-of-the-art algorithms. Another study conducted by Koch et al. (2022), introduced a method utilizing generalized action primitives derived from time measurement analysis. These primitives are detected using a skeletal-based action recognition system. Then, a search algorithm combines information from HAR and methods of time measurement (MTM) to estimate the assembly process. Later, in a more recent study by Büsch et al. (2023), they considered the recognition of tools used during assembly tasks to enhance HAR. Additionally, Li et al. (2021) presented a deep transfer learning ST-GCN model designed to learn action representations that are invariant to domain variations between human body joints in the source and target domains. The authors utilized the maximum mean discrepancy (MMD) approach in the domain adaptation module to align the extracted features from the two domains. In general, GCN architectures are susceptible to novel motion patterns, for that, Liu et al. (2023a) introduced a mask graph convolutional network (Mask-GCN). This system prioritizes the learning of action-specific skeleton joints, which are crucial for conveying action information. Conversely, action-agnostic skeleton joints, which convey rare action information and are more susceptible to novel motion patterns, are masked. The utilization of skeleton data enables direct capture of human body structure while minimizing redundant information. Experimental results demonstrate that Mask-GCN outperforms the majority of GCN-based methods when confronted with diverse novel motion patterns.

5.2.3. Attention-based methods

Transformers have demonstrated success in various domains, including natural language processing (NLP) and vision tasks (Sun et al., 2022). Inspired by their effectiveness,

researchers have applied transformers to skeleton sequences for spatio-temporal modeling in human action recognition from videos (Ahn et al., 2023). Transformers architecture comprises encoder and decoder components with self-attention and encoder-decoder attention layers for effective long-term dependency modeling, multi-modal fusion and multi-task processing (Morshed et al., 2023). Fig 10 illustrates a general architecture of an encoder-decoder transformer.

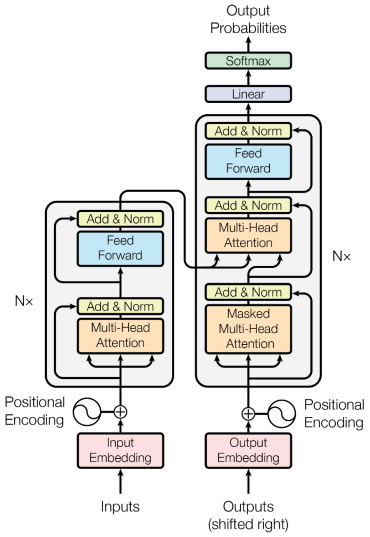


Figure 10: General Transformer architecture (Vaswani et al., 2017).

Recently, researchers have been exploiting these methods to apply them in the field of HAR. Upadhyay et al. introduced P3DAttnNet, a deep neural network designed to automatically generate assembly plans from video demonstrations. The network incorporates a spatiotemporal attention model to recognize actions within the video and utilizes a functional object-oriented network to model the assembly plan. The performance evaluation of the network was conducted on the IKEA ASM dataset. In another study conducted by Islam and Iqbal (2020), they introduced HAMLET, a hierarchical deep neural network algorithm for multimodal HAR. The algorithm employed a multi-head self-attention mechanism to encode spatiotemporal features and accurately identify human activities. Evaluation on three diverse human activity datasets(Chen et al., 2015; Xia et al., 2012; Kubota et al., 2019) demonstrated the superior performance of HAMLET compared to other baseline methods. Finally, Mahmud et al. (2020) introduced a self-attention-based deep learning framework that employs various attention mechanisms to generate higher-dimensional feature representations for human activity recognition. The model demonstrates impressive performance on widely used HAR datasets, namely PAMAP2 (2012), Opportunity (2010), Skoda (2008) and USC-HAD (2012).

5.2.4. Hybrid deep learning methods

Hybrid methods refer to combining two or more types of models to provide strong spatio-temporal modeling of human movements taking advantage of different deep net-

work architectures leading to enhanced efficiency (Morshed et al., 2023). An example of a hybrid network architecture is illustrated in Fig. 11.

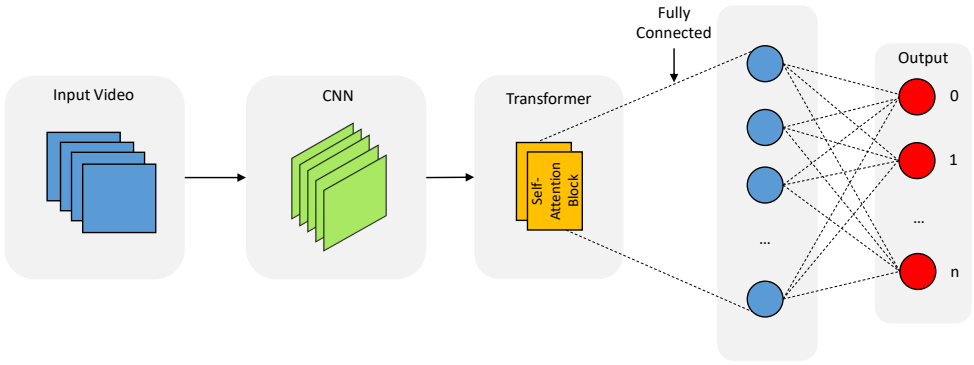


Figure 11: Example of Hybrid deep learning network architecture.

Hybrid approaches are gaining increasing traction as the prevailing methodology. For instance, Moutinho et al. (2023) proposed a residual convolutional neural network with 34 layers (ResNet-34) and an LSTM network to extract high-level context information for human action recognition of an industrial engine assembly operation. In the same fashion, Belay Tuli et al. (2022) proposed an activity recognition method using a combined CNN and LSTM technique. They evaluated 9 of 14 actions of the InHARD dataset including background actions and compared results with their private dataset captured in a lab environment. Also, Singhania et al. (2022) introduces C2F-TCN, an encoder-decoder architecture with a "coarse-to-fine" ensemble of decoders. They enhance the C2F-TCN framework with a novel model-agnostic temporal feature augmentation strategy using stochastic max-pooling of segments. This strategy improves accuracy and calibration in supervised action segmentation outperforming state-of-the-art approaches in coarse segmentation on the Assembly101 dataset.

Hybrid approaches incorporating transformer-based methods are becoming increasingly predominant. Zhou et al. (2023) presented an attention-based deep learning approach for inertial motion recognition and trajectory estimation. They incorporated a convolution module, residual module, attention module and Bi-LSTM to improve the predicted human motion trajectory accuracy so that the robot determines when and how to aid human workers. Following the same spirit, Zhang et al. (2021) presented a hybrid approach for context-aware human action recognition and prediction. They integrated a CNN and variable-length Markov modeling (VMM) to exploit spatial context from video images for action recognition. A bi-stream CNN structure is employed as the spatial context input, while the VMM analyzes dependencies embedded in the action sequences to determine the optimal consideration of current and past actions for accurate future action prediction. Later, they proposed IF-ConvTransformer (Zhang et al., 2022d), an HAR framework consisting of an IMU fusion block and a ConvTransformer subnet. The ConvTransformer network excels in capturing local features and modeling long-term dependencies. Extensive experiments conducted on five smartphone-based datasets and three wearable device-based datasets demonstrate the superior performance of the proposed framework. In another work conducted by Hernandez et al. (2021),

utilized an encoder-decoder-based classifier to recognize primitives and a continuous-time hidden Markov model for activity recognition using skeletal data and hand-centric features. Their proposed system provides valuable operational insights by computing productivity indicators such as worker availability, worker performance and overall labor effectiveness. Finally, Jiao et al. (2020) introduced a multi-deep-learning model, integrating spatial transformer graph convolutional network for action recognition in industrial workflows. Their approach combines CNNs, spatial transformer networks (STNs) and GCNs to extract spatial and temporal information from videos. STNs are employed to correct skeleton images and mitigate the impact of complex real-world environments. Additionally, an attention mechanism is incorporated to adjust the weight of key points, considering the workload disparity between the human upper and lower body in industrial operations, thereby enhancing the accurate identification of manual operations in industrial settings.

5.3. Self-Supervised and Zero-Shot Learning approaches

Self-supervised training is suitable when dealing with large amounts of data. It learns visual knowledge from massive unlabeled data which alleviates the annotation burden (Deng et al., 2023). However, most existing HMA approaches are supervised-learning based and therefore suffer from manual annotation of massive unlabeled data. As mentioned earlier, some approaches generated automatically synthetic data using digital twins to overcome this problem (Dallel et al., 2023; Vysocky et al., 2022).

In the context of human motion analysis in the industry, zero-shot learning offers a promising approach that aligns with the principles of frugality. Traditionally, motion analysis models require a large amount of labeled training data specific to each action class, which can be costly and time-consuming to collect (Dallel et al., 2023). Zero-shot learning aims to overcome this limitation by enabling the recognition and understanding of new, unseen actions without the need for explicit training examples. For instance, Deng et al. (2023) provided a zero-shot evaluation on several datasets which showed poor results, especially for challenging procedural datasets such as InHARD and MECCANO.

5.4. Frugality and embedded systems

In the pursuit of cost-effectiveness, researchers and practitioners strive to develop efficient algorithms, sensor configurations and data processing techniques that minimize hardware requirements and computational complexity.

Frugality metrics refer to a set of measures or criteria introduced to evaluate approaches and address complexity issues such as computational complexity and time complexity (Onsongo and Knorringa, 2020). The study conducted by Kwon et al. (2021) proposed a time processing metric that incorporates factors such as inference time, detection delay and post-processing time. In the same spirit, Wang et al. (2019) estimated the average computational cost for each recognition task to be below the recognition interval set to 1 second, Thereby fulfilling the criteria for real-time recognition. By emphasizing frugality, industry professionals can strike a balance between achieving reliable motion analysis outcomes and optimizing resource allocation. This approach ensures practical and sustainable solutions for various applications, including manufacturing, surveillance and human-robot collaboration.

5.5. Summary of findings

This review delves into deep learning approaches employed for analyzing, recognizing and predicting operator actions and gestures within industrial settings. Notably, there is a rising trend favoring GCN and Transformers over CNN. GCNs are used with human skeleton data, while attention-based models capture temporal aspects in continuous action sequences. It is noteworthy that hybrid approaches integrating multiple methods are increasingly being used to enhance the accuracy of recognition models. Most of the identified approaches primarily focus on recognizing human actions in trimmed videos achieving high performances. Recent studies have shifted their focus toward online recognition, driven by the opportunities it offers in industrial contexts. However, these online methods have shown lower performance. Furthermore, the study highlights a commonly overlooked aspect of existing approaches, which is frugality. Only a limited number of studies have addressed this dimension, employing methods such as zero-shot learning and unsupervised learning.

6. Discussion and Challenges

The study of different research works on human motion analysis across different applications in industry has yielded valuable observations and opportunities for improvement. We identified the following insights:

6.1. Human motion analysis evolution from industry 4.0 to industry 5.0:

In our literature review, we identified 36 articles discussing industry 4.0 and 9 articles mentioning industry 5.0 of which 5 are review papers. This disparity can be attributed to the fact that the majority of these articles primarily focus on improving productivity and efficiency, aligning with industry 4.0's objectives. Industry 4.0, is predominantly characterized by digitalization and automation, striving for high production and efficiency (Nguyen Ngoc et al., 2022) while it considers also safety and ergonomics (Bortolini et al., 2020). In our context, Industry 5.0 greatly enhances the significance of these two aspects, embracing a more pronounced human-centered orientation (Zizic et al., 2022).

In industry 5.0, HMA systems are primarily employed in applications that prioritize worker well-being by creating ergonomic work environments to reduce physical strain (Alves et al., 2023). These systems can analyze human pose to suggest better work postures and also evaluate safety and fatigue levels (Kim et al., 2022). Real-time monitoring of worker movements is used to prevent accidents and collisions in shared environments, enhancing safety in dynamic collaborative settings and enabling smoother HRI (Mohammadi et al., 2020). Furthermore, it allows personalization by adapting work environments to individual worker preferences and needs (Alves et al., 2023). This customization can lead to a more comfortable and accommodating workspace, and ultimately, improve the well-being and productivity of operators in industry.

To conclude, this transition from industry 4.0 to 5.0 regarding our context signifies mainly a move from process optimization towards a more comprehensive approach that takes into account the workforce requirements and the integration of autonomous systems (Moutinho et al., 2023).

6.2. Weakly supervised learning

In real-world industrial scenarios, datasets are generally specific to case studies. If new case studies or new activities are to be added, it would be hard to annotate new data as it is time-consuming and costly. However, most of the reviewed approaches in the literature were fully supervised, with a smaller proportion being semi-supervised (Rana et al., 2023; Singhania et al., 2022). To address this issue, more works using weak supervision learning and self-supervised learning methods should be encouraged (Vahdani and Tian, 2022). These approaches offer the advantage of leveraging unlabeled data, which is particularly beneficial in industrial scenarios.

Another problem is the lack of data especially for some actions that are difficult to reproduce which results in imbalanced datasets such as (Sener et al., 2022; ?; Ragusa et al., 2021), where some action classes are less represented than others affecting the accuracy of the recognition model. Exploring approaches that can perform well on a few data utilizing methods including zero-shot learning (Deng et al., 2023), Seames (Akremi et al., 2022) and domain adaptation (Zhuang et al., 2020) which is a branch of transfer learning seem interesting. These approaches enable the recognition model to generalize better and make predictions for classes with limited or no training samples.

Overall, weakly supervised learning in industrial HMA offers cost-effective, rapid implementation, utilization of contextual information, leveraging unlabeled data, scalability, flexibility and improved outcomes. These advantages make it a promising approach for analyzing human motion in real-world industrial applications. Future approaches can be inspired by the work presented in (Rana et al., 2023; Deng et al., 2023; Zhang et al., 2023) to incorporate these notions.

6.3. Action segmentation on untrimmed data

While offline action segmentation has demonstrated excellent performance such as in (Tassi et al., 2022), online segmentation has shown lower performance, highlighting the necessity for further advancements in untrimmed sequence segmentation (Singhania et al., 2022). With the growing availability of untrimmed datasets, there is an urgent requirement for better methodologies that can effectively and accurately segment sequences with various durations. Early research (Reining et al., 2019) mainly relied on object tracking and sliding windows to locate action positions for detecting the start and end frames of an action sequence. To address this challenge, one of the possible solutions is to investigate developing high-performing methods that can effectively perform sequence segmentation, regardless of the specific action labels involved. Such an approach can rely on the utilization of temporal convolutional networks (TCNs) (Ding et al., 2022), which have shown promise in this regard.

In addition, for sequence localization and better performance of continual action recognition, some work such in (Zhang et al., 2023) explored transfer learning techniques to adapt models trained on general datasets to the specific industrial context. They leverage pre-trained models on large general datasets and fine-tune them on smaller, industry-specific datasets which can help overcome data scarcity challenges and improve generalization performance. Notable examples include models such as TSM and TSN (Deng et al., 2023) pre-trained on the Kinetics-400 dataset, which have demonstrated promising results on the InHARD and MECCANO datasets, respectively.

6.4. The appropriate sensors

Camera-based sensors are susceptible to occlusions and can be affected by workspace conditions such as lighting conditions, reflections and temperature variations (Menolotto et al., 2020). For occlusion problems, employing multi-view cameras could be a viable solution. By strategically positioning them, the fusion of data from various sensing capabilities within the factory (such as onboard cameras, robots, augmented reality headsets, etc) could be leveraged. This provides a broader coverage of the workspace allowing tracking of the most important movements and human object interaction information. Another solution for resolving lightning factors is the utilization of digital twins to recreate industrial environments to generate data with almost fully controlled parameters capable of simulating various lightning positions or camera views. Furthermore, combining multiple sensors, such as IMUs and low-cost camera-based systems, can offer a comprehensive solution to compensate for the limitations of individual sensors.

Although capable of providing accurate data for body movement, IMUs present certain drawbacks such as time-consuming setup procedures and discomfort when worn during work. To address this issue, a potential solution involves utilizing IMUs more specifically during the training phase while relieving operators from the burden of wearing these sensors during the testing phase.

Images captured by RGB cameras have the potential to reveal the identity of operators, leading to ethical considerations regarding the anonymization of data, as the personal information of individuals is being recorded. One possible solution proposes that sensitive parts of the images, such as the face, can be blurred or masked during data processing steps. In Addition, it is proposed to consider safeguarding data through encryption techniques which can then be processed using an edge or cloud computing architecture (Parashar and Shekhawat, 2022). Another solution may be registering only pose estimation features and object information ensuring the protection of sensitive data while still enabling efficient analysis and extraction of valuable insights.

6.5. Data acquisition recommendations

The availability of benchmark datasets is important for advancing human motion analysis in industrial contexts for various reasons. Indeed, they provide a standardized foundation for method comparisons and accuracy validation. Moreover, the presence of accessible benchmarks accelerates research, alleviating researchers from the time and resource-consuming task of data collection. This, in turn, empowers them to concentrate on the development of innovative solutions and addressing specific challenges within industrial environments. Additionally, these datasets foster interdisciplinary collaboration and ensure that research aligns with industry requirements. Furthermore, it's worth noting that benchmark datasets often include predefined evaluation metrics, making them publicly available to facilitate objective performance measurement and contributing to the development of more effective solutions for the industry. However, it is noticed that most of the existing datasets are collected in controlled laboratory environments, which fail to accurately reflect the complexities encountered in actual industrial settings. To address these issues, the adoption of diverse and comprehensive datasets is proposed. These datasets should consider additional context information of tools and objects alongside human motion data (Niemann et al., 2021), different workstation backgrounds such in

(Ben-Shabat et al., 2021) and dynamic environments with the presence of moving entities (operators, objects, robots...) such in (Rudenko et al., 2020). Incorporating such datasets will provide robust solutions for better deployment in real-world applications

Another prominent issue is the lack of industrial benchmark datasets covering different aspects such as the simultaneous acquisition of different body parts (whole body, upper body, or partial body parts such as hands or arms), alongside different views (Ego-centric, Allocentric and multiple views). Incorporating most of these aspects is advised for creating industrial datasets. Acquisition protocol and data collection can be inspired by existing datasets that take into consideration these aspects such as the NTU-RGB+D120 dataset (Liu et al., 2019) representing an interesting protocol for massive data collection, Assembly101 dataset (Sener et al., 2022) with various viewpoints and NTU-X dataset (Trivedi et al., 2021) with dense representation of human face and full body.

Many researchers suggest using personalized and private datasets to evaluate their proposed approaches and experiments. While these datasets are valuable for the research community, they are often not accessible, preventing others from benefiting from the data or comparing their works. Moreover, these datasets could potentially enrich existing datasets by providing additional case studies. encouraging authors to make their data available is strongly advised, employing anonymization techniques if necessary to protect participant identities.

Another issue is that in industrial datasets, within each activity, some actions tend to be repeated more frequently than others. Consequently, this leads to imbalanced datasets, with certain actions being over-represented while others are under-represented. To address these limitations, the utilization of synthetic data generation techniques is proposed to provide more data for fair action class representation. Synthetic data generator such as digital twin (Dallel et al., 2023) and generative adversarial networks (GANs) (Ali et al., 2023) offers this opportunity to expand the available data enabling a more comprehensive representation of diverse industrial activities, and facilitating the training of robust and accurate HAR models.

Multimodal approaches demonstrate strong performance when utilizing data captured from diverse sensors (Sun et al., 2022; Ba et al., 2023), which provide information such as various angles of human posture or objects that operators interact with. These datasets should encompass different modalities such as RGB, depth, point cloud, IR, skeleton and even physiological signals, allowing a rich representation of the operator's movements and interaction by complementary information. Thus, it is encouraged to acquire more data following the example set by the HA4M dataset (Cicirelli et al., 2022).

6.6. Approaches

In industry, the objective is to deploy the systems developed for action recognition using the proposed approaches in embedded production settings. The system must meet specific criteria to be considered successful in such environments. Firstly, it should exhibit high performance, characterized by good accuracy in action recognition. Secondly, it needs to be fast, providing real-time processing and immediate results. Additionally, it should be able to work with untrimmed data and operate efficiently with limited training samples that have already been processed. Furthermore, the system should support continuous learning, allowing for the incorporation of new actions and updating the learning model accordingly. This particular application involves spatio-temporal modeling, with

a focus on time series analysis. Another important aspect is the interpretability of the system's output, enabling a clear understanding of why it works. In terms of existing techniques, it is worth considering different types of representations, such as graph-based approaches that provide a suitable representation of skeletal structures (Xing and Zhu, 2021). Additionally, transformers offer effective temporal data representation (Wen et al., 2022). Exploring the combination of graph transformers could be an interesting avenue of research, building upon previous work conducted by Jiao et al. (2020). Another aspect to address is reducing execution time through memory utilization optimization, which can be achieved through techniques such as continual graph-based learning (Hedegaard et al., 2022). These advancements are not limited to the industry domain but can also find applications in other fields. Additionally, the fusion of multimodal data within deep learning architectures can be achieved through various fusion strategies. These strategies encompass techniques such as early fusion, late fusion, or even attention mechanisms, which selectively focus on informative modalities. By effectively integrating deep learning-based approach and fusion techniques, the embedded production system can attain the desired performance, real-time processing and improved accuracy in action recognition tasks, making it a promising approach for practical implementation (Sun et al., 2022).

7. Conclusion

This study has presenedt a review of the human motion analysis topic from acquisition to application in the industrial context from 2018 to May 2023 (included). There has been a growing interest among researchers in applying human action recognition methods to industrial applications. This is due to the growing availability and variety of industrial datasets acquired using the newest technologies and presenting different case studies.

Various applications of HMA in the industry were reviewed following the proposed taxonomy that englobes safety, ergonomics and productivity, shedding light on their significance. We also reported different technologies employed in human motion data acquisition, with a particular focus on MoCap and signal-based sensors. Moreover, we identify and list relevant industrial datasets, emphasizing their distinct characteristics. Additionally, we conduct a comprehensive analysis and comparison of existing deep learning-based methods and online metrics used to evaluate these datasets with a specific focus on human action recognition. We conclude by discussing various challenges and by providing insights into future perspectives.

The findings of our research can significantly facilitate various aspects of future work in the field of HMA in the industry. Future researchers will be better equipped to discern which applications are most influenced by HMA, thereby directing more researchers toward these applications and accelerating the development of crucial areas in the industry. They can make informed decisions on selecting the most suitable sensors for acquiring movement data, thanks to the taxonomy of technologies we have provided. Additionally, researchers can identify datasets of interest provided in this paper, selecting the most appropriate one based on their needs, such as the type of interaction, the number of views and data modalities. Our research also provides a valuable starting point for researchers seeking guidance on approaches and methods for action and gesture recognition, along with insights into the relevant metrics for evaluation.

Our study reveals several scientific gaps related to online human activity recognition in industrial environments, which present their unique challenges. Further research is needed to enhance the recognition performance for practical applications, ensuring accurate analysis of human movement, in real-time deployed on embedded systems in industrial case studies. Among interesting strategies, the application of digital twins holds promising potential in producing balanced datasets and reducing the annotation burden. Furthermore, the integration of multimodal data into deep learning architectures has the potential to significantly enhance accuracy in action recognition tasks. Finally, special attention should also be given to frugality aspects for better optimization of approaches, emphasizing both simplicity in terms of complexity and optimal use of data.

References

Abu Al-Haija, Q., Al-Saraireh, J., 2022. Asymmetric identification model for human-robot contacts via supervised learning. Symmetry 14, 591.

Ahmad, T., Jin, L., Zhang, X., Lai, S., Tang, G., Lin, L., 2021. Graph convolutional neural network for human action recognition: A comprehensive survey. IEEE Transactions on Artificial Intelligence 2, 128–145. doi:10.1109/TAI.2021.3076974.

Ahn, D., Kim, S., Hong, H., Ko, B.C., 2023. Star-transformer: A spatio-temporal cross attention transformer for human action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3330–3339.

Akremi, M., Slama, R., Tabia, H., 2022. Spd siamese neural network for skeleton-based hand gesture recognition, in: 17th International Conference on Computer Vision Theory and Applications VISAPP 2022), SCITEPRESS-Science and Technology Publications. pp. 394–402.

Al-Amin, M., Tao, W., Doell, D., Lingard, R., Yin, Z., Leu, M.C., Qin, R., 2019. Action recognition in manufacturing assembly using multimodal sensor fusion. Procedia Manufacturing 39, 158–167.

Ali, H., Grönlund, C., Shah, Z., 2023. Leveraging gans for data scarcity of covid-19: Beyond the hype, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 659–667.

Alves, J., Lima, T.M., Gaspar, P.D., 2023. Is industry 5.0 a human-centred approach? a systematic review. Processes 11, 193.

Arshad, M.H., Bilal, M., Gani, A., 2022. Human activity recognition: Review, taxonomy and open challenges. Sensors 22, 6463.

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y., 2015. Physical human activity recognition using wearable sensors. Sensors 15, 31314–31338.

Ba, M., Ji, Z., Liu, Z., Yao, B., Xu, W., Zhong, Y., 2023. Human action detection based on multimodal feature fusion for human-robot collaborative assembly, in: 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), IEEE. pp. 1–6.

Barioul, R., Ghribi, S.F., Derbel, H.B.J., Kanoun, O., 2020. Four sensors bracelet for american sign language recognition based on wrist force myography, in: 2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), IEEE. pp. 1–5.

Bassani, G., Filippeschi, A., Avizzano, C.A., 2021. A dataset of human motion and muscular activities in manual material handling tasks for biomechanical and ergonomic analyses. IEEE Sensors Journal 21, 24731–24739.

Belay Tuli, T., Patel, V.M., Manns, M., 2022. Industrial human activity prediction and detection using sequential memory networks, in: Proceedings of the Conference on Production Systems and Logistics: CPSL 2022, Hannover: publish-Ing.. pp. 62–72.

Ben-Shabat, Y., Yu, X., Saleh, F., Campbell, D., Rodriguez-Opazo, C., Li, H., Gould, S., 2021. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 847–859.

Bhatti, U.A., Tang, H., Wu, G., Marjan, S., Hussain, A., 2023. Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence. International Journal of Intelligent Systems 2023, 1–28.

Birjandi, S.A.B., Kühn, J., Haddadin, S., 2020. Observer-extended direct method for collision monitoring in robot manipulators using proprioception and imu sensing. IEEE Robotics and Automation Letters 5, 954–961.

Bonci, A., Cen Cheng, P.D., Indri, M., Nabissi, G., Sibona, F., 2021. Human-robot perception in industrial environments: A survey. Sensors 21, 1571.

Borboni, A., Reddy, K.V.V., Elamvazuthi, I., AL-Quraishi, M.S., Natarajan, E., Azhar Ali, S.S., 2023. The expanding role of artificial intelligence in collaborative robots for industrial applications: A systematic review of recent works. Machines 11, 111.

Bortolini, M., Faccio, M., Gamberi, M., Pilati, F., 2020. Motion analysis system (mas) for production and ergonomics assessment in the manufacturing processes. Computers & Industrial Engineering 139, 105485.

Büsch, L., Koch, J., Schoepflin, D., Schulze, M., Schüppstuhl, T., 2023. Towards recognition of human actions in collaborative tasks with robots: Extending action recognition with tool recognition methods. Sensors 23, 5718.

Castro, A., Silva, F., Santos, V., 2021. Trends of human-robot collaboration in industry contexts: Handover, learning, and metrics. Sensors 21, 4113.

Chen, C., Jafari, R., Kehtarnavaz, N., 2015. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: 2015 IEEE International conference on image processing (ICIP), IEEE. pp. 168–172.

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

Chen, W., Yu, C., Tu, C., Lyu, Z., Tang, J., Ou, S., Fu, Y., Xue, Z., 2020b. A survey on hand pose estimation with wearable sensors and computer-vision-based methods. Sensors 20, 1074.

Chiabert, P., Aliev, K., 2020. Analyses and study of human operator monotonous tasks in small enterprises in the era of industry 4.0, in: Product Lifecycle Management Enabling Smart X: 17th IFIP WG 5.1 International Conference, PLM 2020, Rapperswil, Switzerland, July 5–8, 2020, Revised Selected Papers 17, Springer. pp. 83–97.

Cicirelli, G., Marani, R., Romeo, L., Domínguez, M.G., Heras, J., Perri, A.G., D'Orazio, T., 2022. The ha4m dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. Scientific Data 9, 745.

Colyer, S.L., Evans, M., Cosker, D.P., Salo, A.I., 2018. 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, 1–15.

Dallel, M., Havard, V., Dupuis, Y., Baudry, D., 2022. A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks, in: 2022 7th International Conference on Machine Learning Technologies (ICMLT), pp. 155–163.

Dallel, M., Havard, V., Dupuis, Y., Baudry, D., 2023. Digital twin of an industrial workstation: A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of human–robot collaboration. Engineering Applications of Artificial Intelligence 118, 105655.

Darvish, K., Wanderlingh, F., Bruno, B., Simetti, E., Mastrogiovanni, F., Casalino, G., 2018. Flexible human–robot cooperation models for assisted shop-floor tasks. Mechatronics 51, 97–114.

Delamare, M., Duval, F., Boutteau, R., 2020. A new dataset of people flow in an industrial site with uwb and motion capture systems. Sensors 20, 4511.

Delva, M.L., Lajoie, K., Khoshnam, M., Menon, C., 2020. Wrist-worn wearables based on force myography: on the significance of user anthropometry. BioMedical Engineering OnLine 19, 1–18.

Deng, A., Yang, T., Chen, C., 2023. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. arXiv preprint arXiv:2303.13505 .

Ding, G., Sener, F., Yao, A., 2022. Temporal action segmentation: An analysis of modern technique. arXiv preprint arXiv:2210.10352 .

Donisi, L., Cesarelli, G., Pisani, N., Ponsiglione, A.M., Ricciardi, C., Capodaglio, E., 2022. Wearable sensors and artificial intelligence for physical ergonomics: A systematic review of literature. Diagnostics 12, 3048.

Eisenbach, M., Aganian, D., Köhler, M., Stephan, B., Schröter, C., Groß, H.M., 2022. Visual scene understanding for enabling situation-aware cobots. Universitätsbibliothek.

Fan, J., Zheng, P., Li, S., 2022. Vision-based holistic scene understanding towards proactive human–robot collaboration. Robotics and Computer-Integrated Manufacturing 75, 102304.

Grushko, S., Vysockỳ, A., Heczko, D., Bobovskỳ, Z., 2021. Intuitive spatial tactile feedback for better awareness about robot trajectory during human–robot collaboration. Sensors 21, 5748.

Havard, V., Jeanne, B., Lacomblez, M., Baudry, D., 2019. Digital twin and virtual reality: a co-

simulation environment for design and assessment of industrial workstations. Production & Manufacturing Research 7, 472–489.
Hedegaard, L., Heidari, N., Iosifidis, A., 2022. Online skeleton-based action recognition with continual spatio-temporal graph convolutional networks. arXiv preprint arXiv:2203.11009 .
Heo, Y.J., Kim, D., Lee, W., Kim, H., Park, J., Chung, W.K., 2019. Collision detection for industrial collaborative robots: A deep learning approach. IEEE Robotics and Automation Letters 4, 740–746. doi:10.1109/LRA.2019.2893400.
Hernandez, J., Valarezo, G., Cobos, R., Kim, J.W., Palacios, R., Abad, A.G., 2021. Hierarchical human action recognition to measure the performance of manual labor. IEEE Access 9, 103110–103119.
Hjorth, S., Chrysostomou, D., 2022. Human–robot collaboration in industrial environments: A literature review on non-destructive disassembly. Robotics and Computer-Integrated Manufacturing 73, 102208.
Inkulu, A.K., Bahubalendruni, M.R., Dara, A., 2022. Challenges and opportunities in human robot collaboration context of industry 4.0-a state of the art review. Industrial Robot: the international journal of robotics research and application 49, 226–239.
Iodice, F., De Momi, E., Ajoudani, A., 2022. Hri30: An action recognition dataset for industrial human-robot interaction, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE. pp. 4941–4947.
Islam, M.M., Iqbal, T., 2020. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 10285–10292.
Jiang, S., Gao, Q., Liu, H., Shull, P.B., 2020. A novel, co-located emg-fmg-sensing wearable armband for hand gesture recognition. Sensors and Actuators A: Physical 301, 111738.
Jiao, Z., Jia, G., Cai, Y., 2020. Ensuring computers understand manual operations in production: Deep-learning-based action recognition in industrial workflows. Applied Sciences 10, 966.
Joshi, M., Deshpande, V., 2019. A systematic review of comparative studies on ergonomic assessment techniques. International Journal of Industrial Ergonomics 74, 102865.
Ke, A., Huang, J., Chen, L., Gao, Z., He, J., 2020. An ultra-sensitive modular hybrid emg–fmg sensor with floating electrodes. Sensors 20, 4775.
Khan, M.H., Zöller, M., Farid, M.S., Grzegorzek, M., 2020. Marker-based movement analysis of human body parts in therapeutic procedure. Sensors 20, 3312.
Kim, G.Y., Kim, D., Do Noh, S., Han, H.K., Kim, N.G., Kang, Y.S., Choi, S.H., Go, D.H., Song, J., Lee, D.Y., et al., 2022. Human digital twin system for operator safety and work management, in: IFIP International Conference on Advances in Production Management Systems, Springer. pp. 529–536.
Kim, W., Peternel, L., Lorenzini, M., Babič, J., Ajoudani, A., 2021. A human-robot collaboration framework for improving ergonomics during dexterous operation of power tools. Robotics and Computer-Integrated Manufacturing 68, 102084.
Koch, J., Büsch, L., Gomse, M., Schüppstuhl, T., 2022. A methods-time-measurement based approach to enable action recognition for multi-variant assembly in human-robot collaboration. Procedia CIRP 106, 233–238.
Kubota, A., Iqbal, T., Shah, J.A., Riek, L.D., 2019. Activity recognition in manufacturing: The roles of motion capture and semg+ inertial wearables in detecting fine vs. gross motion, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE. pp. 6533–6539.
Kwon, W., Jin, Y., Lee, S.J., 2021. Uncertainty-aware knowledge distillation for collision identification of collaborative robots. Sensors 21, 6674.
Lagamtzis, D., Schmidt, F., Seyler, J.R., Dang, T., 2022. Coax: Collaborative action dataset for human motion forecasting in an industrial workspace., in: ICAART (3), pp. 98–105.
Le, V.T., Tran-Trung, K., Hoang, V.T., 2022. A comprehensive review of recent deep learning techniques for human activity recognition. Computational Intelligence and Neuroscience 2022.
Li, S., Fan, J., Zheng, P., Wang, L., 2021. Transfer learning-enabled action recognition for human-robot collaborative assembly. Procedia CIRP 104, 1795–1800.
Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C., Liu, J., 2016. Online human action detection using joint classification-regression recurrent neural networks. European Conference on Computer Vision .
Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C., 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence 42, 2684–2701.
Liu, M., Meng, F., Chen, C., Wu, S., 2023a. Novel motion patterns matter for practical skeleton-based action recognition, in: AAAI Conference on Artificial Intelligence (AAAI).
Liu, Z., Liu, Q., Xu, W., Wang, L., Ji, Z., 2023b. Adaptive real-time similar repetitive manual procedure prediction and robotic procedure generation for human-robot collaboration. Advanced Engineering

Informatics 58, 102129.
Maddikunta, P.K.R., Pham, Q.V., Prabadevi, B., Deepa, N., Dev, K., Gadekallu, T.R., Ruby, R., Liyanage, M., 2022. Industry 5.0: A survey on enabling technologies and potential applications. Journal of Industrial Information Integration 26, 100257.
Mahmud, S., Tonmoy, M., Bhaumik, K.K., Rahman, A.M., Amin, M.A., Shoyaib, M., Khan, M.A.H., Ali, A.A., 2020. Human activity recognition from wearable sensor data using self-attention. arXiv preprint arXiv:2003.09018 .
Manghisi, V.M., Evangelista, A., Uva, A.E., 2022. A virtual reality approach for assisting sustainable human-centered ergonomic design: The ergovr tool. Procedia Computer Science 200, 1338–1346.
Manoj, T., Thyagaraju, G., 2018. Active and assisted living: A comprehensive review of enabling technologies and scenarios. International Journal of Advanced Research in Computer Science 9.
Matheson, E., Minto, R., Zampieri, E.G., Faccio, M., Rosati, G., 2019. Human–robot collaboration in manufacturing applications: A review. Robotics 8, 100.
Maurice, P., Malaisé, A., Amiot, C., Paris, N., Richard, G.J., Rochel, O., Ivaldi, S., 2019. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. The International Journal of Robotics Research 38, 1529–1537.
McGinnis, R.S., 2013. Advancing Applications of IMUs in Sports Training and Biomechanics. Ph.D. thesis. University of Michigan.
Menolotto, M., Komaris, D.S., Tedesco, S., O'Flynn, B., Walsh, M., 2020. Motion capture technology in industrial applications: A systematic review. Sensors 20, 5687.
Mohammadi, A.F., Rezayati, M., van de Venn, H.W., Karimpour, H., 2020. A mixed-perception approach for safe human–robot collaboration in industrial automation. Sensors 20, 6347.
Morshed, M.G., Sultana, T., Alam, A., Lee, Y.K., 2023. Human action recognition: A taxonomy-based survey, updates, and opportunities. Sensors 23, 2182.
Moutinho, D., Rocha, L.F., Costa, C.M., Teixeira, L.F., Veiga, G., 2023. Deep learning-based human action recognition to leverage context awareness in collaborative assembly. Robotics and Computer-Integrated Manufacturing 80, 102449.
Munasinghe, C., Amin, F.M., Scaramuzza, D., van de Venn, H.W., 2022. Covered, collaborative robot environment dataset for 3d semantic segmentation, in: 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4. doi:10.1109/ETFA52439.2022.9921525.
Nazmus Sakib, A., Basak, P., Doha Uddin, S., Mustavi Tasin, S., Ahad, M.A.R., 2022. Can ensemble of classifiers provide better recognition results in packaging activity?, in: Sensor-and Video-Based Activity and Behavior Computing: Proceedings of 3rd International Conference on Activity and Behavior Computing (ABC 2021), Springer. pp. 167–180.
Nguyen Ngoc, H., Lasa, G., Iriarte, I., 2022. Human-centred design in industry 4.0: case study review and opportunities for future research. Journal of Intelligent Manufacturing 33, 35–76.
Niemann, F., Bas, H., Steffens, J., Nair, N., ten Hompel, M., 2021. Context-aware activity recognition in logistics (caarl)–a optical marker-based motion capture dataset.
Niemann, F., Reining, C., Moya Rueda, F., Nair, N.R., Steffens, J.A., Fink, G.A., Ten Hompel, M., 2020. Lara: Creating a dataset for human activity recognition in logistics using semantic attributes. Sensors 20, 4083.
Onsongo, E.K., Knorringa, P., 2020. Comparing frugality and inclusion in innovation for development: Logic, process and outcome. Innovation and Development , 1–21.
Orsag, L., Stipancic, T., Koren, L., 2023. Towards a safe human–robot collaboration using information on human worker activity. Sensors 23, 1283.
Parashar, A., Shekhawat, R.S., 2022. Protection of gait data set for preserving its privacy in deep learning pipeline. IET Biometrics 11, 557–569.
Pastor, F., Lin-Yang, D.h., Gómez-de Gabriel, J.M., García-Cerezo, A.J., 2022. Dataset with tactile and kinesthetic information from a human forearm and its application to deep learning. Sensors 22, 8752.
Prakash, A., Sharma, N., Sharma, S., 2021. An affordable transradial prosthesis based on force myography sensor. Sensors and Actuators A: Physical 325, 112699.
Prunet, T., Absi, N., Borodin, V., Cattaruzza, D., 2022. Optimization of human-aware manufacturing and logistics systems: A survey on the modeling frameworks of human aspects .
Ragusa, F., Furnari, A., Livatino, S., Farinella, G.M., 2021. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1569–1578.
Rana, M., Rahman, A., Smith, D., 2023. A semi-supervised approach for activity recognition from indoor trajectory data. arXiv preprint arXiv:2301.03134 .

Reining, C., Niemann, F., Moya Rueda, F., Fink, G.A., ten Hompel, M., 2019. Human activity recognition for production and logistics—a systematic literature review. Information 10, 245.

Reiss, A., Stricker, D., 2012. Introducing a new benchmarked dataset for activity monitoring, in: 2012 16th international symposium on wearable computers, IEEE. pp. 108–109.

Ribeiro, P.M.S., Matos, A.C., Santos, P.H., Cardoso, J.S., 2020. Machine learning improvements to human motion tracking with imus. Sensors 20, 6383.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., et al., 2010. Collecting complex activity datasets in highly rich networked sensor environments, in: 2010 Seventh international conference on networked sensing systems (INSS), IEEE. pp. 233–240.

Rudenko, A., Kucner, T.P., Swaminathan, C.S., Chadalavada, R.T., Arras, K.O., Lilienthal, A.J., 2020. Thör: Human-robot navigation data collection and accurate motion trajectories dataset. IEEE Robotics and Automation Letters 5, 676–682.

Schreiter, T., de Almeida, T.R., Zhu, Y., Maestro, E.G., Morillo-Mendez, L., Rudenko, A., Kucner, T.P., Mozos, O.M., Magnusson, M., Palmieri, L., et al., 2022. The magni human motion dataset: Accurate, complex, multi-modal, natural, semantically-rich and contextualized. arXiv preprint arXiv:2208.14925 .

Sedlar, J., Stepanova, K., Skoviera, R., Behrens, J.K., Tuna, M., Sejnova, G., Sivic, J., Babuska, R., 2023. Imitrob: Imitation learning dataset for training and evaluating 6d object pose estimators. IEEE Robotics and Automation Letters .

Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A., 2022. 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, pp. 21096–21106.

da Silva, A.G., Mendes Gomes, M.V., Winkler, I., 2022. Virtual reality and digital human modeling for ergonomic assessment in industrial product development: a patent and literature review. Applied Sciences 12, 1084.

Singhania, D., Rahaman, R., Yao, A., 2022. C2f-tcn: A framework for semi and fully supervised temporal action segmentation. arXiv preprint arXiv:2212.11078 .

Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P., Tröster, G., 2008. Wearable activity tracking in car manufacturing. IEEE Pervasive Computing 7, 42–50.

Sturm, F., Hergenroether, E., ReC2Ft, J., Vojnovikj, P.S., Siegel, M., 2023. Challenges of the creation of a dataset for vision based human hand action recognition in industrial assembly. arXiv preprint arXiv:2303.03716 .

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J., 2022. Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence .

Tamantini, C., Cordella, F., Lauretti, C., Zollo, L., 2021. The wgd—a dataset of assembly line working gestures for ergonomic analysis and work-related injuries prevention. Sensors 21, 7600.

Tang, C., Li, W., Wang, P., Wang, L., 2018. Online human action recognition based on incremental learning of weighted covariance descriptors. Information Sciences 467, 219–237.

Tassi, F., Iodice, F., De Momi, E., Ajoudani, A., 2022. Sociable and ergonomic human-robot collaboration through action recognition and augmented hierarchical quadratic programming, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 10712–10719.

Toichoa Eyam, A., Mohammed, W.M., Martinez Lastra, J.L., 2021. Emotion-driven analysis and control of human-robot interactions in collaborative applications. Sensors 21, 4626.

Trivedi, N., Thatipelli, A., Sarvadevabhatla, R.K., 2021. Ntu-x: An enhanced large-scale dataset for improving pose-based recognition of subtle human actions, in: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1–9.

Upadhyay, A., Barua, P.A., Dubey, A., Sengupta, S., Kuriakose, S.M., Goenka, P., . P3dattnnet: Automated assembly plan generation from video demonstration .

Vahdani, E., Tian, Y., 2022. Deep learning-based action detection in untrimmed videos: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Vicentini, F., 2021. Collaborative robotics: a survey. Journal of Mechanical Design 143.

Vysocky, A., Grushko, S., Spurny, T., Pastor, R., Kot, T., 2022. Generating synthetic depth image dataset for industrial applications of hand localization. IEEE Access 10, 99734–99744.

Wang, P., Liu, H., Wang, L., Gao, R.X., 2018. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. CIRP annals 67, 17–20.

Wang, Z., Qin, R., Yan, J., Guo, C., 2019. Vision sensor based action recognition for improving efficiency

and quality under the environment of industry 4.0. Procedia CIRP 80, 711–716.
Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., Sun, L., 2022. Transformers in time series: A survey. arXiv preprint arXiv:2202.07125 .
Wu, Y.T., Gomes, M.K., da Silva, W.H., Lazari, P.M., Fujiwara, E., 2020. Integrated optical fiber force myography sensor as pervasive predictor of hand postures. Biomedical engineering and computational biology 11, 1179597220912825.
Xia, L., Chen, C.C., Aggarwal, J.K., 2012. View invariant human action recognition using histograms of 3d joints, in: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE. pp. 20–27.
Xiao, Z.G., Menon, C., 2019. A review of force myography research and development. Sensors 19, 4557.
Xing, Y., Zhu, J., 2021. Deep learning-based action recognition with 3d skeleton: a survey.
Xu, M., Xiong, Y., Chen, H., Li, X., Xia, W., Tu, Z., Soatto, S., 2021. Long short-term transformer for online action detection. Advances in Neural Information Processing Systems 34, 1086–1099.
Xu, W., Feng, S., Yao, B., Ji, Z., Liu, Z., 2023. Turn-taking prediction for human–robot collaborative assembly considering human uncertainty. Journal of Manufacturing Science and Engineering 145, 121007.
Yao, R., Lin, G., Shi, Q., Ranasinghe, D.C., 2018. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. Pattern Recognition 78, 252–266.
Yonga Chuengwa, T., Swanepoel, J.A., Kurien, A.M., Kanakana-Katumba, M.G., Djouani, K., 2023. Research perspectives in collaborative assembly: A review. Robotics 12, 37.
Yoshimura, N., Morales, J., Maekawa, T., Hara, T., 2022. Openpack: A large-scale dataset for recognizing packaging works in iot-enabled logistic environments. arXiv preprint arXiv:2212.11152 .
Yun, Y., Agarwal, P., Deshpande, A.D., 2013. Accurate, robust, and real-time estimation of finger pose with a motion capture system, in: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE. pp. 1626–1631.
Zakia, U., Menon, C., 2022. Dataset on force myography for human–robot interactions. Data 7, 154.
Zamora, H., Mauricio-Andrés, Castro-Vargas, J.A., Azorin-Lopez, J., Garcia-Rodriguez, J., 2021. Deep learning-based visual control assistant for assembly in industry 4.0. Computers in Industry 131, 103485.
Zhang, J., Liu, H., Chang, Q., Wang, L., Gao, R.X., 2020. Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly. CIRP annals 69, 9–12.
Zhang, J., Wang, P., Gao, R.X., 2021. Hybrid machine learning for human action recognition and prediction in assembly. Robotics and Computer-Integrated Manufacturing 72, 102184.
Zhang, M., Sawchuk, A.A., 2012. Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors, in: Proceedings of the 2012 ACM conference on ubiquitous computing, pp. 1036–1043.
Zhang, R., Li, J., Zheng, P., Lu, Y., Bao, J., Sun, X., 2022a. A fusion-based spiking neural network approach for predicting collaboration request in human-robot collaboration. Robotics and Computer-Integrated Manufacturing 78, 102383.
Zhang, R., Lv, J., Li, J., Bao, J., Zheng, P., Peng, T., 2022b. A graph-based reinforcement learning-enabled approach for adaptive human-robot collaborative assembly operations. Journal of Manufacturing Systems 63, 491–503.
Zhang, X.Y., Li, C., Shi, H., Zhu, X., Li, P., Dong, J., 2023. Adapnet: Adaptability decomposing encoder–decoder network for weakly supervised action recognition and localization. IEEE Transactions on Neural Networks and Learning Systems 34, 1852–1863. doi:10.1109/TNNLS.2019.2962815.
Zhang, Y., Ding, K., Hui, J., Lv, J., Zhou, X., Zheng, P., 2022c. Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly. Advanced Engineering Informatics 54, 101792.
Zhang, Y., Wang, L., Chen, H., Tian, A., Zhou, S., Guo, Y., 2022d. If-convtransformer: A framework for human activity recognition using imu fusion and convtransformer. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6, 1–26.
Zhao, J., Liu, P., Li, Z., 2024. Exploring the impact of trip patterns on spatially aggregated crashes using floating vehicle trajectory data and graph convolutional networks. Accident Analysis & Prevention 194, 107340.
Zheng, Z., Wu, Z., Zhao, R., Ni, Y., Jing, X., Gao, S., 2022. A review of emg-, fmg-, and eit-based biosensors and relevant human–machine interactivities and biomedical applications. Biosensors 12, 516.
Zhou, H., Yang, G., Wang, B., Li, X., Wang, R., Huang, X., Wu, H., Wang, X.V., 2023. An attention-based deep learning approach for inertial motion recognition and estimation in human-robot collabo-

ration. Journal of Manufacturing Systems 67, 97–110.
Zhu, Q., Deng, H., 2023. Spatial adaptive graph convolutional network for skeleton-based action recognition. Applied Intelligence , 1–13.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proceedings of the IEEE 109, 43–76.
Zizic, M.C., Mladineo, M., Gjeldum, N., Celent, L., 2022. From industry 4.0 towards industry 5.0: A review and analysis of paradigm shift for the people, organization and technology. Energies 15, 5221.