machine learning & knowledge extraction

MDPI

Article

A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition

Michail Kaseris [1,2,*], Ioannis Kostavelis [1,2] and Sotiris Malassiotis [2]

[1] Department of Supply Chain Management, International Hellenic University, Kanellopoulou 2, 60132 Katerini, Greece; gkostave@ihu.gr
[2] Information Technologies Institute (ITI) Center of Research and Technology Hellas (CERTH), 57001 Thessaloniki, Greece; malasiot@iti.gr
* Correspondence: kasermich@ihu.gr

Abstract: Human activity recognition (HAR) remains an essential field of research with increasing real-world applications ranging from healthcare to industrial environments. As the volume of publications in this domain continues to grow, staying abreast of the most pertinent and innovative methodologies can be challenging. This survey provides a comprehensive overview of the state-of-the-art methods employed in HAR, embracing both classical machine learning techniques and their recent advancements. We investigate a plethora of approaches that leverage diverse input modalities including, but not limited to, accelerometer data, video sequences, and audio signals. Recognizing the challenge of navigating the vast and ever-growing HAR literature, we introduce a novel methodology that employs large language models to efficiently filter and pinpoint relevant academic papers. This not only reduces manual effort but also ensures the inclusion of the most influential works. We also provide a taxonomy of the examined literature to enable scholars to have rapid and organized access when studying HAR approaches. Through this survey, we aim to inform researchers and practitioners with a holistic understanding of the current HAR landscape, its evolution, and the promising avenues for future exploration.

Keywords: human activity recognition(HAR); survey; machine learning; wearable sensors; datasets; daily and industrial activities

Citation: Kaseris, M.; Kostavelis, I.; Malassiotis, S. A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 842–876. https://doi.org/10.3390/make6020040

Academic Editor: Luca Longo

Received: 22 February 2024
Revised: 3 April 2024
Accepted: 11 April 2024
Published: 18 April 2024

1. Introduction

Human activity recognition (HAR) pertains to the systematic identification and classification of activities undertaken by individuals based on diverse sensor-derived data [1]. This interdisciplinary research domain intersects computer science, engineering, and data science to decipher patterns in sensor readings and correlate them with specific human motions or actions. The importance of HAR is underscored by the rapid proliferation of wearable devices, mobile sensors, and the burgeoning Internet of Things (IoT) environment. Recognizing human activities accurately can not only improve user experience and automation but also assist in a myriad of applications where understanding human behavior is essential.

The applications of HAR are multifaceted, encompassing health monitoring, smart homes, security surveillance, sports analytics, and human–robot interaction, to name just a few. For instance, in the healthcare sector, HAR can facilitate the remote monitoring of elderly or patients with chronic diseases, enabling timely interventions and reducing hospital readmissions. Similarly, in smart homes, recognizing daily activities can lead to energy savings, enhanced comfort, and improved safety. In the realm of sports, HAR can aid athletes in refining their techniques and postures, providing feedback in real time. Moreover, in security and surveillance, anomalous activities can be promptly detected, thereby ensuring timely responses. In summary, HAR holds transformative potential across numerous sectors, shaping a more responsive and intuitive environment [2].

An example of HAR involves the use of wearable devices, such as smartwatches or fitness trackers, to monitor and classify user activities, as shown in Figure 1. These devices are often equipped with a range of sensors, the most prevalent being accelerometers and gyroscopes, which measure linear accelerations and angular velocities, respectively. Raw sensor data are captured at defined intervals, resulting in a time-series dataset [3–6]. For instance, an accelerometer would produce a triaxial dataset corresponding to acceleration values along the x, y, and z axes. Prior to feature extraction, several preprocessing steps are typically undertaken. In typical HAR applications, the proper predictor selection and data normalization are proven critical for the performance of machine learning algorithms in the field of HAR [7]. The raw time-series data are often noisy and may contain irrelevant fluctuations. Hence, they undergo filtering, commonly using a low-pass filter, to eliminate high-frequency noise [8]. Subsequently, the continuous data stream is segmented into overlapping windows, each representing a specific timeframe (e.g., 2.56 s with a 50% overlap). The windowing technique facilitates the extraction of local features that can encapsulate distinct activity patterns. For each window, multiple features are computed. Temporal domain features such as mean, variance, standard deviation, and correlation between axes are commonly extracted. Traditional algorithms such as support vector machines (SVMs), decision trees, random forests, and k-nearest neighbors (k-NN) have been extensively applied. However, with the advent of deep learning, methods like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have also found prominence due to their ability to model complex temporal relationships. Once trained, the model can classify incoming data into predefined activity classes such as walking, running, sitting, or standing. The granularity and accuracy of the classification are contingent on the quality of the data, the features extracted, and the efficacy of the chosen algorithm. In summation, HAR through wearable devices encompasses a systematic pipeline from raw sensor data acquisition to refined activity classification, leveraging advanced computational methods and algorithms.

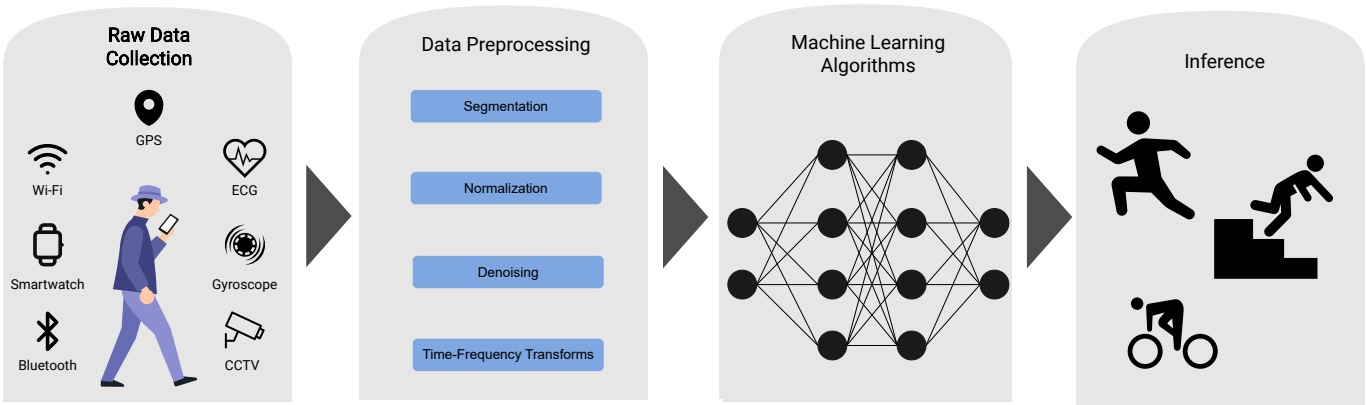

Figure 1. Generic flow of processes in human activity recognition tasks.

In a survey conducted in [9] about vision-based HAR, the authors point out their findings about related surveys in the topic of HAR. Specifically, the statistics of related surveys from 2010 until 2019 show that the majority of related surveys are focused on providing details about specific aspects of HAR [9–12], rather than giving a broader picture of the topic. We argue that in order to inform the reader about the recent trends in the field, more work needs to be performed on describing the wider spectrum of methodologies and settings adopted in every aspect of HAR. This article fills in the gap in the bibliography by giving an overview of various approaches of HAR, utilizing a large array of sensors and modalities, enabling the reader to identify the gaps in the literature.

Due to the expansive applications of human activity recognition (HAR) utilizing machine learning techniques, we structured our survey to distinctly address both sensor-based and vision-based methods. An additional salient contribution of this article is the

deployment of large language models (LLMs) to extract pertinent keywords and respond to questions, thereby facilitating the ranking and filtering of our comprehensive database of papers. This paper is organized as follows:

1. Introduction: This section delineates the problem of HAR, setting the context for the remainder of the paper.
2. Related Work: Here, we reference the seminal and recent literature on HAR, underscoring the importance of comprehensive literature reviews in the domain.
3. Methodology: In this section, we present our methodology, highlighting our data sources and the processes we employed to distill key information.
4. Taxonomy of Methods: This section presents a deeper categorization of HAR methods and distinctly partitions them into sensor-based and vision-based techniques.
5. Datasets: In this section, we catalog the most prevalent datasets employed in HAR research.
6. Conclusion and Future Directions: Finally, in this section we wrap up the paper by discussing potential avenues for future research in the realm of HAR.

2. Related Work

A plethora of surveys have emerged to record the advancements, methodologies, and applications shaping this dynamic domain. These comprehensive reviews offer critical insights into the evolution of HAR enabled by machine learning and deep learning techniques. Notably, the surveys by Ray et al. [12] and Singh et al. [13] underscore the pivotal role of transfer learning and the paradigm shift towards automatic feature extraction through deep neural networks, respectively, in enhancing HAR systems. Additionally, the study by Gu et al. [14] fills a crucial gap in the literature by providing an in-depth analysis of contemporary deep learning methods applied to HAR. Collectively, these works not only delineate the current state of HAR but also highlight the challenges and future directions, thereby offering a comprehensive backdrop for understanding the progress and potential within this field.

The field of human activity recognition (HAR) has seen significant research, particularly in its applications to eldercare and healthcare within assistive technologies, as highlighted by the comprehensive surveys conducted by various researchers. The survey by Hussain et al. [15] spans research from 2010 to 2018, focusing on device-free HAR solutions, which eliminate the need for subjects to carry devices by utilizing environmental sensors instead. This innovative approach is categorized into action-based, motion-based, and interaction-based research, offering a unique perspective across all sub-areas of HAR. The authors present a detailed analysis using ten critical metrics and discuss future research directions, underlining the importance of device-free solutions in the evolution of HAR.

Jobanputra et al. [16] explores a variety of state-of-the-art HAR methods that employ sensors, images, accelerometers, and gyroscopes placed in different settings to collect data. The paper delves into the effectiveness of various machine learning and deep neural network techniques, such as decision trees, k-nearest neighbors, support vector machines, and more advanced models like convolutional and recurrent neural networks, in interpreting the data collected for HAR. By comparing these techniques and reviewing their performance across different datasets, the paper complements Hussain et al.'s [15] taxonomy and analysis by providing a deeper understanding of the methodologies employed in HAR and their practical implications, especially in healthcare and eldercare.

Further extending the recording of the advancements in HAR, Lara et al. [10] provide a survey on HAR using wearable sensors, emphasizing the role of pervasive computing across sectors such as the medical, security, entertainment, and tactical fields. This paper introduces a general HAR system architecture and proposes a two-level taxonomy based on the learning approach and response time. By evaluating twenty-eight HAR systems qualitatively on various parameters like recognition performance and energy consumption, this survey highlights the critical challenges and suggests future research areas. The focus on wearable sensors offers a different yet complementary perspective to the device-free

approach discussed by Hussain et al. [15], showcasing the diversity in HAR research and its potential to revolutionize interactions between individuals and mobile devices.

The article by Ramamurthy et al. [17] investigates the effectiveness of machine learning in extracting and learning from activity datasets, transitioning from traditional algorithms that utilize hand-crafted features to advanced deep learning algorithms that evolve features hierarchically. The complexity of AR in uncontrolled environments, exacerbated by the volatile nature of activity data, underscores the ongoing challenges in the field. The paper provides a broad overview of the current machine learning and data mining techniques used in AR, identifying the challenges of existing systems and pointing towards future research directions.

Building on the discussion of machine learning's role in HAR, the survey by Dang et al. [18] offers a thorough review of HAR technologies that are crucial for the development of context-aware applications, especially within the Internet of Things (IoT) and healthcare sectors. The paper stands out by aiming for a comprehensive coverage of HAR, categorizing methodologies into sensor-based and vision-based HAR, and further dissecting these groups into subcategories related to data collection, preprocessing, feature engineering, and training. The conclusion highlights the existing challenges and suggests directions for future research, adding depth to the ongoing conversation about HAR technologies.

Furthermore, the work by Vrigkas et al. [19] focuses on the classification of human activities from videos and images, acknowledging the complications introduced by background clutter, occlusions, and variations in appearance. This review is particularly pertinent to applications in surveillance, human–computer interaction, and robotics. The authors categorize activity classification methodologies and analyze their pros and cons, dividing them based on the data modalities used and further into subcategories based on activity modeling and focus. The paper also assesses existing datasets for activity classification and discusses the criteria for an ideal dataset, concluding with future research directions in HAR. This review complements the broader discussions by Ramamurthy et al. [17] and Dang et al. [18], by providing specific insights into video and image-based HAR, thereby enriching the understanding of the challenges and potential advancements in the field.

Finally, Ke et al. [11] offer an in-depth survey focusing on three pivotal aspects: the core technologies behind HAR, the various systems developed for recognizing human activities, and the wide array of applications these systems serve. It thoroughly explores the processing stages essential for HAR, such as human object segmentation, feature extraction, and the algorithms for activity detection and classification. The review then categorizes HAR systems based on their functionality—from recognizing activities of individual persons to understanding interactions among multiple people and identifying abnormal behaviors in crowds. Special emphasis is placed on the application domains, particularly highlighting the roles of HAR in surveillance, entertainment, and healthcare, thereby illustrating the breadth of HAR's impact.

Saleem et al. [20] broadens the scope by presenting a comprehensive overview that encapsulates the multifaceted nature of HAR. This research introduces a detailed taxonomy that classifies HAR methods along several dimensions, including their operation mode (online/offline), data modality (multimodal/unimodal), feature type (handcrafted/learning-based), and more. By covering a wide range of application areas and methodological approaches, this study underscores the interdisciplinary nature of HAR, providing a comparative analysis of contemporary methods against various benchmarks such as activity complexity and recognition accuracy. This comparative lens not only deepens the understanding of state-of-the-art HAR techniques but also underscores the ongoing challenges and future avenues for research within the field.

Focusing on the computer vision aspect of HAR, Ref. [21] delves into the specific challenges and advancements in human action recognition within the broader spectrum of vision-based methods. This research highlights the paradigm shift towards feature learning-based representations driven by the adoption of deep learning methodologies. It offers a thorough examination of the latest developments in HAR, which are relevant

across diverse applications from augmented reality to surveillance. The paper delineates a taxonomy of techniques for feature acquisition, including those leveraging RGB and depth data, and discusses the interplay between cutting-edge deep learning approaches and traditional hand-crafted methods. Further delving into the realm of computer vision, the paper by Singh et al. [13] addresses the critical aspect of human action identification from video, which is a cornerstone in applications ranging from healthcare to security. Their study provides a meticulous evaluation and comparison of these methodologies, offering insights into their effectiveness based on accuracy, classifier types, and datasets, thereby presenting a holistic view of the current advancements in activity detection.

Complementing these discussions, the survey by Gu et al. [10,14] focuses on the advancements and challenges in HAR, particularly through the lens of deep learning. This is especially pertinent given the rapid evolution of deep learning techniques and their significant potential in enhancing HAR systems. By offering a detailed review of contemporary deep learning methods applied in HAR, this study addresses a critical need for an in-depth exploration of the field, highlighting the ongoing developments and future directions. The study by Ray et al. [12] emphasizes the role of transfer learning. Transfer learning is known for its ability to enhance accuracy, reduce data collection and labeling efforts, and adapt to evolving data distributions.

The summarized overview provided in Table 1 encapsulates the breadth and depth of research in the domain of human activity recognition (HAR), revealing a diverse array of focuses, from device-free solutions to the nuanced application of deep learning techniques. Notably, several surveys, such as those concentrating on device-free, wearable device, and sensor-based HAR, not only dissect the current methodologies and systems but also rigorously compare various techniques, underscoring the field's multifaceted nature. These comparisons offer valuable insights into the strengths and weaknesses of different approaches, guiding future research directions. Interestingly, while the majority of the surveys delve into comparative analyses, a select few, particularly those focusing on machine learning, video-based HAR, and deep learning applications, choose to emphasize the discussion of challenges, recent advances, and future directions without directly comparing techniques. This distinction might reflect the rapidly evolving landscape of HAR, where the emphasis is increasingly shifting towards understanding the underlying complexities and potential advancements rather than solely focusing on methodological comparisons. The recurring mention of future directions across all categories highlights a collective acknowledgment of the untapped potential and unresolved challenges within HAR. The emphasis on vision-based and video-based HAR underscores the growing importance of visual data in understanding human activities, which is a trend further amplified by the advent of deep learning and transfer learning. Specifically, the impact of transfer learning, as discussed in the surveys, signifies a transformative shift towards leveraging pre-existing knowledge, thereby enhancing efficiency and accuracy in HAR applications. This comprehensive table of surveys not only provides a snapshot of the current state of HAR research but also sets the stage for future explorations.

The surveys highlight several key future challenges, including integrating data from various sensors to recognize complex behaviors beyond simple actions like walking [14,17,18]. They note the high cost and effort in gathering labeled data and the limitations of generative models like AEs and GANs in human activity recognition (HAR) [14,17]. Additionally, they emphasize the need to account for simultaneous and overlapping activities, which are more common in daily life than singular tasks [10,18,21].

Table 1. Summary of survey papers on human activity recognition.

Categories	Main Focus	Future Directions Discussed	Comparison of Techniques
Device-free	Comprehensive survey of human activity recognition focusing on device-free solutions; taxonomy proposed.	Yes	Yes
Multiple	Survey of HAR methods in eldercare and healthcare using IoT; compares various data collection methods and machine learning techniques.	Yes	Yes
Wearable Device	Survey of HAR using wearable sensors; general architecture, taxonomy, key issues, challenges, and system evaluation.	Yes	Yes
Machine Learning	Overview of machine learning techniques in activity recognition; discusses challenges and recent advances.	Yes	No
Sensor-based and Vision-based	Comprehensive review of HAR technology; classification of methodologies and evaluation of advantages and weaknesses.	Yes	Yes
Vision-based	Detailed review of human activity classification from videos and images; categorization of methodologies and dataset analysis.	Yes	Yes
Video-based	Extensive survey of video-based human activity recognition; covers core technology, recognition systems, and applications.	Yes	No
Multiple	Overview of HAR categorizing methods; comparative analysis of state-of-the-art techniques.	Yes	Yes
Multiple	Analysis of human action recognition systems; focus on feature learning-based representations and deep learning.	Yes	Yes
Transfer Learning in HAR	Impact of transfer learning in HAR and other areas; reviews related research articles focusing on vision sensor-based HAR.	Yes	Yes
Video-based	Survey of human action identification from video; comparison of hand-crafted and automatic feature extraction approaches.	Yes	Yes
Deep Learning in HAR	Extensive survey on deep learning applications in HAR; detailed review of contemporary deep learning methods.	Yes	No

3. Methodology

The field of human activity recognition (HAR) has witnessed an exponential growth in published research, resulting in an immense corpus of literature. The recent emergence of large language models (LLMs) offers a promising avenue for processing such extensive corpora with remarkable precision in condensed time frames. In this section, we elucidate our methodology. We commence by outlining the origins of our paper sources, underscoring that our analysis, while primarily based on scraped data, is not limited to them. Subsequently, we delineate the techniques employed to mine keywords from textual content and to respond to specific queries. This section culminates with an account of our approach to filtering and refining the selection of papers, ensuring our review is both comprehensive and discerning. In this particular work, we cut off works prior to 2020 and cover works up to the year of authoring this paper, 2023.

3.1. Data Sources

To amass a comprehensive dataset for our survey, we mined data from prominent academic research repositories. These included IEEEXplore, from which we extracted 8005 papers; arXiv, contributing 271 papers; and MDPI, accounting for 1300 papers. All the data procured are publicly accessible. The specific attributes scraped for each paper, contingent upon availability, encompass the paper title, its publication year, the number of citations, and most notably, the abstract. An overview of the distribution of paper number can be represented graphically in the pie chart in Figure 2. Given the assumption that an abstract encapsulates a paper's content prior to a full perusal, our extraction efforts prioritized it. A succinct overview of the information available from each repository is furnished in Table 2. The data collection procedure is presented in Figure 3.

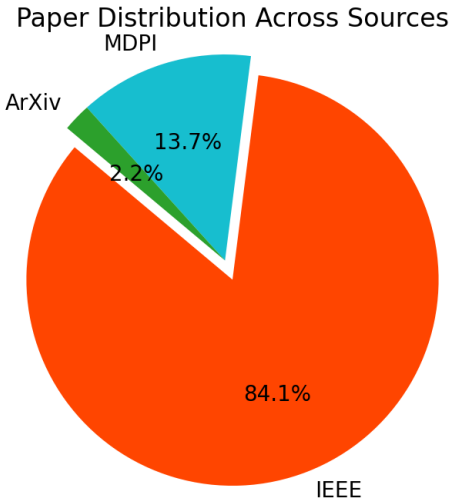

Figure 2. Pie chart representation of the number of papers scraped from our paper sources.

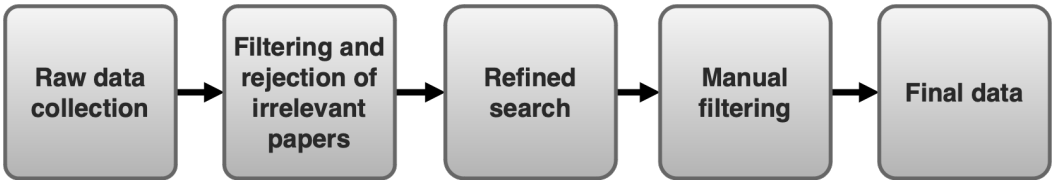

Figure 3. Overview of the data collection, filtering, and compilation pipeline.

Table 2. Available extracted information from the pool of academic research paper repositories.

Repository Name	Title	# Citations	Abstract	Year
IEEEXplore	Yes	Yes	Yes	Yes
arXiv	Yes	No	Yes	Yes
MDPI	Yes	No	Yes	Yes

The analysis encompassed three main sources: IEEE, MDPI, and ArXiv. The IEEE repository is the most extensive, comprising 8005 papers. The distribution of publications in IEEE (Figure 4b) demonstrates a consistent growth over the years, with citations ranging from none to a substantial 2659, showcasing the varied impact of these works. MDPI contributes 1300 articles, though detailed yearly trends were not discernible due to data constraints. ArXiv, while the smallest with 270 articles, offers unique insights, particularly from its submission year trends (Figure 4a). Collectively, this triad of sources offers a comprehensive glimpse into the academic landscape of the topics in question. The citation distribution is markedly right-skewed (Figure 4b), indicating that a large portion of the IEEE papers garner a relatively low number of citations. This pattern is characteristic of many academic publications, where only a fraction of works achieve broad recognition and accrue a high citation count. The presence of a few papers with exceptionally high citations suggests the existence of seminal works that have profoundly impacted the field. These outliers, albeit few, underscore the importance of breakthrough research and its far-reaching implications in the academic landscape. The majority of papers, however, reside within a modest citation bracket, emphasizing the collective contribution of incremental research to the broader knowledge base.

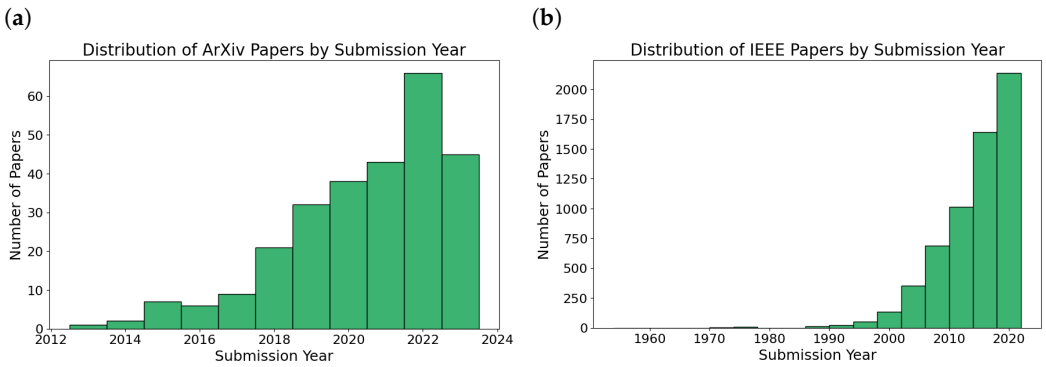

Figure 4. Statistics for the paper sources for IEEE and arXiv. (**a**) Distribution of ArXiv papers by submission year. (**b**) Average and Median number of citations per year for IEEE papers.

3.2. Use of Natural Language Processing

In our methodology, we harnessed the capabilities of natural language processing (NLP) to enhance the comprehensibility and assessment of scientific abstracts. Specifically, we employed the `Voicelab/vlt5-base-keywords`(https://huggingface.co/Voicelab/vlt5-base-keywords (accessed on 12 December 2023)) summarization transformer to extract salient keywords. Furthermore, to facilitate a deeper understanding of the content, we integrated a question answering transformer (https://huggingface.co/consciousAI/question-answering-roberta-base-s-v2 (accessed on 12 December 2023)). Our approach involved constructing an input context by amalgamating the paper's title and abstract. This composite text was then posed with a series of pertinent questions designed to assess the abstract's writing quality and information richness. Through this approach, we aimed to create an efficient tool for academic researchers to quickly ascertain the relevance and quality of a given scientific paper.

To further analyze our collected data, we proceed to the investigation of the keywords of our data. Concretely, we extract keywords from the title-abstract string concatenation

and then for each keyword, we extract their vector embeddings. Afterwards, we project them onto the 2D plane using the T-SNE method [22], and then cluster them using K-Means clustering. We can see a clear trend emerging in Figure 5, with the most dominant neighborhoods being occupied by the keywords "Human Action Recognition". We can also see that during the years 2020–2023, deep learning and transfer learning are obvious choices of approach for solving the problem of HAR. Other regions represent model proposal names as are seen on the far left side of the plane. The most repeated keywords that are close to the centroid that corresponds to "Sensor-Based" methods, are words that refer to accelerometers, IMUs, WiFi, and radars, hence we make the choice of focusing on those sensors. Another keyword that influenced our choice of encompassing vision-based methods is the centroid that corresponds to "Video and Image Processing" as well as the multiple occurrences of computer vision convolutional neural network-based methods. Finally, it should be noted that K-Means clustering was implemented after the projection of the embeddings onto the 2D plane and not vice-versa. We intentionally did this, so there is a clear distinction of the regions.

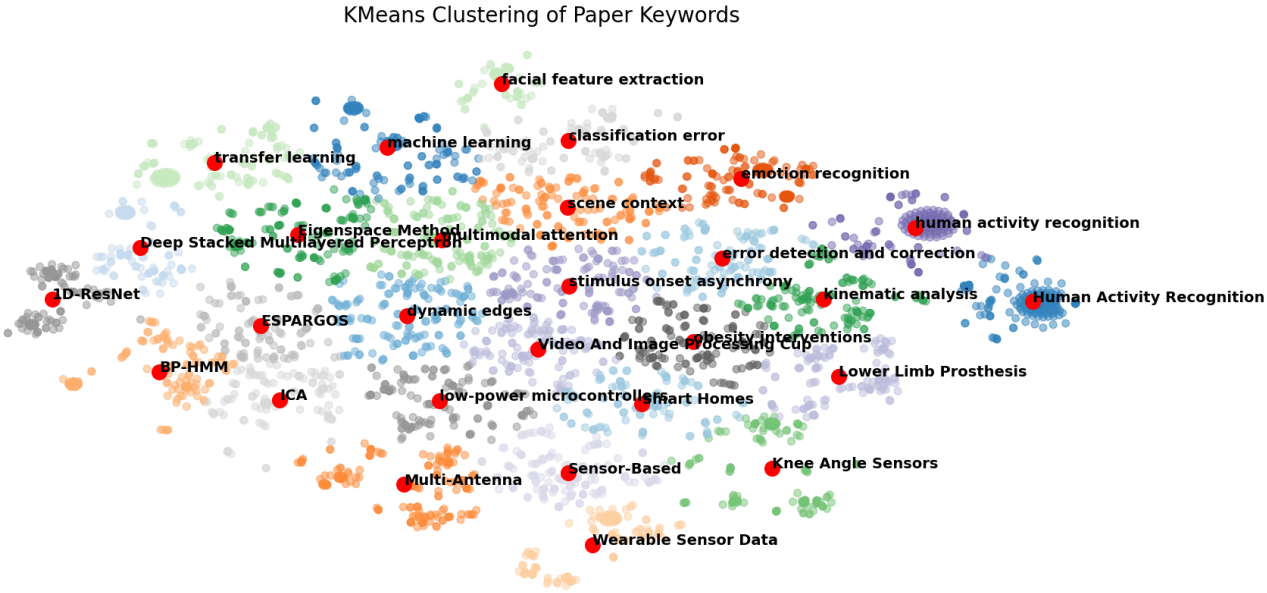

Figure 5. Projection on the 2D plane of the keyword embeddings using the t-SNE method [23]. The red dots represent the respective keyword closer to its corresponding centroid. Each point on the plot corresponds to a vector representation of the keyword embeddings, which is reduced to 2D for better visualization. The different colors differentiate the embeddings' clusters.

We filter our papers using the aforementioned models by asking the following template questions. We refer the reader to Appendix A for a detailed demonstration of the question answering module. The first criterion of filtering is the date of publication of the paper under examination; specifically, we are filtering out papers whose publication year is before 2020. Afterwards, we prompt the LLM to extract keywords and compare it to the column that corresponds to the scraped keywords, if available. Once the keywords are extracted, we proceed to ask a series of questions about the paper itself. As context we provide the LLM the string concatenation of the title and its abstract. We deliberately use these two components as input, as we assume that the title and the paper abstract contain just as much information as we need to infer whether it proposes a survey or a new methodology. The previous two steps have proven useful to discard the majority of irrelevant papers.

3.3. Data Compilation

In this subsection, we describe the methodology we adopted in order to obtain our data for our research. An overview of the pipeline is presented in Figure 3. First of all, we

initiate our raw data search with prompting our search with the keywords "human activity recognition" and "human-robot interaction" and we limit our search by constraining to results from 2020 until the present. Once the search is complete, we end up amassing a large volume of papers from the public sources mentioned in Table 2, where most of the retrieved papers may be irrelevant for our purposes. As a first stage of filtering, we introduce the natural language processing method, as described in the Section 3.2, resulting in a volume of papers of about 500 articles. From the remaining pool of articles, we consequently proceed to manually refine our search within the collected data and leave out the irrelevant by reading the papers' abstracts. The criterion was that we keep every article that introduces either a methodology, a dataset or a survey in the topic of HAR. Finally, after the conclusion of this step we end up with 159 papers.

3.4. Taxonomy Method

In this subsection, we outline the various comparison metrics that have been employed to taxonomize the literature we have collected. Our focus is on providing a comprehensive classification of the various HAR methods, ensuring that they encapsulate the breadth and depth of our research. This approach is crucial for establishing a clear understanding of the criteria and standards we have used to categorize and analyze the collected works. The inclusion of these metrics serves not only to enhance the clarity and rigor of our taxonomy but also aids in offering a structured and methodical overview of the literature for our readers.

- Cost Under this metric, we investigate the computational cost associated with training and deploying each method discussed in the literature. This metric is critically assessed based on the deep learning approach utilized in each method. Our empirical evaluation takes into account several key factors: the memory requirements essential for training and deploying, the complexity encountered during the inference process, and the usage of ensemble methods by the authors. By scrutinizing these elements, we aim to provide a thorough and nuanced understanding of the computational costs, offering insights into the practicality and efficiency of each method in real-world scenarios. This metric is of vital importance, as it is crucial for HAR methods to be cost-effective, as they are deployed in memory- and computing-constrained embedded systems.
- Approach We focus on identifying the machine learning models adopted by the authors to address the problem at hand. This aspect is crucial for understanding the benefits and shortcomings of each method. By identifying the types of deep learning models used, we enable readers to discern the benefits and drawbacks inherent to each approach. Such an understanding is pivotal for researchers and practitioners alike, as it not only provides a clear picture of the current state of the field but also aids in identifying potential areas for future exploration and development. We aim to offer a comprehensive overview that not only informs but also inspires readers to bridge gaps and contribute to the evolution of the future literature.
- Performance In this segment of our analysis, we turn our attention to the evaluation performance of the various methods within the datasets they were validated on. We categorize performance into three distinct tiers: low, medium, and high. A performance is deemed 'low' when scores fall below 75%, 'medium' for those ranging between 75% and 95%, and 'high' for scores exceeding 95%. However, it is crucial for our readers to understand that these performance degrees are indicative and not absolute. This is because different methods are often evaluated using diverse metrics, making direct comparisons challenging. Therefore, while these performance categories provide a helpful framework for initial assessment, they should be interpreted with an understanding of the varied and specific contexts in which each method is tested. Our intention is to offer a guide that aids in gauging performance, while also acknowledging the complexities and nuances inherent in methodological evaluations.

- Datasets This component of our taxonomy is essential, as it provides a clear insight into the environments and conditions under which each method was tested and refined. By presenting this information, we aim to give readers a comprehensive understanding of the types of data each method is best suited for, as well as the potential limitations or biases inherent in these datasets.
- Supervision In the 'Supervision' section, we report the nature of supervision employed in the training of the methods we have examined. This aspect is pivotal, as the type of supervision has a significant impact on several facets of the developmental process, most notably in the cost and effort associated with data labeling. Methods that utilize supervised learning often require large datasets, which in turn necessitate extensive input from human annotators, thereby increasing costs. Conversely, methods based on unsupervised learning, while alleviating the need for labeled data, often confront challenges in maintaining a consistent quality metric. Such methods are also more prone to collapsing during training. By outlining the supervision techniques used, we aim to provide insights into the trade-offs and considerations inherent in each approach, offering a comprehensive perspective on how the choice of supervision influences not just the method's development but also its potential applications and efficacy in real-world scenarios.

3.5. Budget

In this subsection, we analyze the costs for conducting our experiments. Since most of the queries were carried out through API calls, it is worth mentioning the overall financial cost. Given the cost (https://openai.com/pricing (accessed on 28 March 2024)) of using the OpenAI's platform for using the chat completion API, it costed us around USD 8.5. Most of the cost was dominated from the chat completion (USD $30/10^6$ tokens), whereas the embedding generation cost was nearly negligible (USD $0.02/10^6$ tokens). As far as the computational cost is concerned, the most critical aspect of our implementation is the clustering, which has a space complexity of $\mathcal{O}(N(M+K)$, where N is the number of samples in our available dataset, M is the embedding dimensionality, which for this use case, was set to 256, and K the number of clusters, which was set to 28.

4. HAR Devices and Processing Algorithms

In the domain of human activity recognition (HAR), a synergistic interplay between sophisticated algorithms and high-fidelity sensors is quintessential for achieving accurate and reliable recognition outcomes. This section meticulously delineates the pivotal constituents underpinning HAR, segregated into two focal subsections. The initial subsection expounds on the diverse array of sensors employed in HAR, shedding light on their operational principles and the distinct types of data they capture, which are indispensable for discerning human actions and postures. Following this, the subsequent subsection delves into the algorithms that are instrumental in processing the signals derived from these sensors. It elucidates the computational techniques and algorithmic frameworks that are adept at deciphering the intricate patterns embedded in the sensor data, thereby enabling the robust identification and classification of human activities. Through an in-depth examination of both the sensor technologies and the algorithmic methodologies, this section aims to furnish a comprehensive understanding of the technological underpinnings that propel the field of human activity recognition forward. Figure 6 illustrates the two main categories of HAR methods that we be discussed for the rest of the paper.

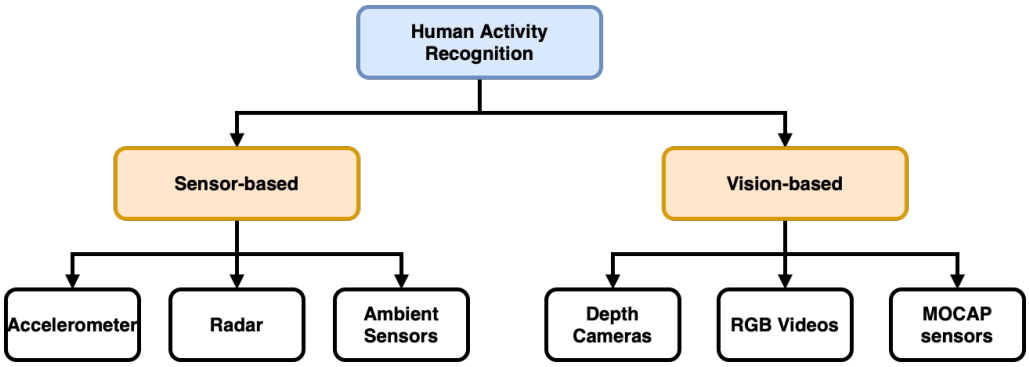

Figure 6. A general hierarchy of HAR methods. Typically, HAR is separated in sensor-based and vision-based approaches. The two major subcategories are divided into their respective sensors that are leveraged.

4.1. Devices

While some HAR approaches can apply to all sensor modalities, many are specific to certain types. The modalities can be categorized into three aspects: body-worn sensors, object sensors, and ambient sensors. Body-worn sensors, such as accelerometers, magnetometers, and gyroscopes, are commonly used in HAR. They capture human body movements and are found in devices like smartphones, watches, and helmets. Object sensors are placed on objects to detect their movement and infer human activities. For example, an accelerometer attached to a cup can detect drinking water activity. Radio frequency identifier (RFID) tags are often used as object sensors in smart home and medical applications. Ambient sensors capture the interaction between humans and the environment and are embedded in smart environments. They include radar, sound sensors, pressure sensors, and temperature sensors. Ambient sensors are challenging to deploy and are influenced by the environment. They are used to recognize activities and hand gestures in settings like smart homes. Some HAR approaches combine different sensor types, such as combining acceleration with acoustic information or using a combination of body-worn, object, and ambient sensors in a smart home environment. These hybrid sensor approaches enable the capture of rich information about human activities.

4.1.1. Body-Worn Sensors

Body-worn sensors, such as accelerometers, magnetometers, and gyroscopes, are commonly used in HAR. These sensors are typically worn by users on devices like smartphones, watches, and helmets. They capture changes in acceleration and angular velocity caused by human body movements, enabling the inference of human activities. Body-worn sensors, particularly accelerometers, have been extensively employed in deep learning-based HAR research. Gyroscopes and magnetometers are also commonly used in combination with accelerometers [24,25]. These sensors are primarily utilized to recognize activities of daily living (ADL) and sports [26]. Rather than extracting statistical or frequency-based features from movement data, the original sensor signals are directly used as inputs for the deep learning network.

4.1.2. Object Sensors

Object sensors are employed to detect the movement of specific objects, unlike body-worn sensors that capture human movements. These sensors enable the inference of human activities based on object movement [27,28]. For example, an accelerometer attached to a cup can detect the activity of drinking water [29]. Object sensors, typically using radio frequency identifier (RFID) tags, are commonly utilized in smart home environments and medical activities. RFID tags provide detailed information that facilitates complex activity recognition [30–33].

It is important to note that object sensors are less commonly used compared to body-worn sensors due to deployment challenges. However, there is a growing trend of combining object sensors with other sensor types to recognize higher-level activities.

4.1.3. Ambient Sensors

Ambient sensors are designed to capture the interactions between humans and their surrounding environment. These sensors are typically integrated into smart environments. Various types of ambient sensors are available, including radar, sound sensors, pressure sensors, and temperature sensors. Unlike object sensors that focus on measuring object movements, ambient sensors are used to monitor changes in the environment.

Several studies have utilized ambient sensors to recognize daily activities and hand gestures. Most of these studies were conducted in the context of smart home environments. Similar to object sensors, deploying ambient sensors can be challenging. Additionally, ambient sensors are susceptible to environmental influences, and only specific types of activities can be reliably inferred [34].

4.1.4. Hybrid Sensors

Some studies have explored the utilization of hybrid sensors in HAR by combining different types of sensors. For instance, research has shown that incorporating both acceleration and acoustic information can enhance the accuracy of HAR. Additionally, there are cases where ambient sensors are employed alongside object sensors, enabling the capture of both object movements and environmental conditions. An example is the development of a smart home environment called A-Wristocracy, where body-worn, object, and ambient sensors are utilized to recognize a wide range of intricate activities performed by multiple occupants. The combination of sensors demonstrates the potential to gather comprehensive information about human activities, which holds promise for future real-world smart home systems.

4.1.5. Vision Sensors

RGB data are widely recognized for their high availability, affordability, and the rich texture details they provide of subjects, making them a popular choice in various applications [18]. Despite these advantages, RGB sensors have limitations, such as a restricted range and sensitivity to calibration issues, and their performance is heavily influenced by environmental conditions like lighting, illumination, and the presence of cluttered backgrounds. The field of vision-based HAR has seen significant interest due to its real-world applications, and research in this area can be categorized based on the type of data used, such as RGB [35–37] and RGB-D data [38,39]. RGB-D data, which include depth information alongside traditional RGB data, provides additional layers of information for more accurate activity recognition. Furthermore, from depth data, it is possible to extract skeleton data, which offers a simplified yet effective representation of the human body's skeleton. These skeleton data occupy a lower-dimensional space, enabling HAR models to operate more efficiently and swiftly, which is an essential factor for real-time applications like surveillance. The most common sensor that can provide RGB, RGB-D, and skeleton information is the Kinect. An example of a both RGB and RGB-D dataset is the Human4D dataset [40] (Figure 7) where the authors provide samples with RGB and depth readings. A common way to represent a skeleton is illustrated in Figure 8 and typically is represented in the form of joint rotations in either exponential maps or quaternions [41,42].

Figure 7. An example of an RGB-D image. The source RGB image appears to the left and its corresponding depth map to the right. Objects that appear closer have a darker red color while the furthest objects appear blue. A notable dataset that is available for HAR problems and provides both RGB and LiDAR scanned depth maps is the Human4D dataset [40].

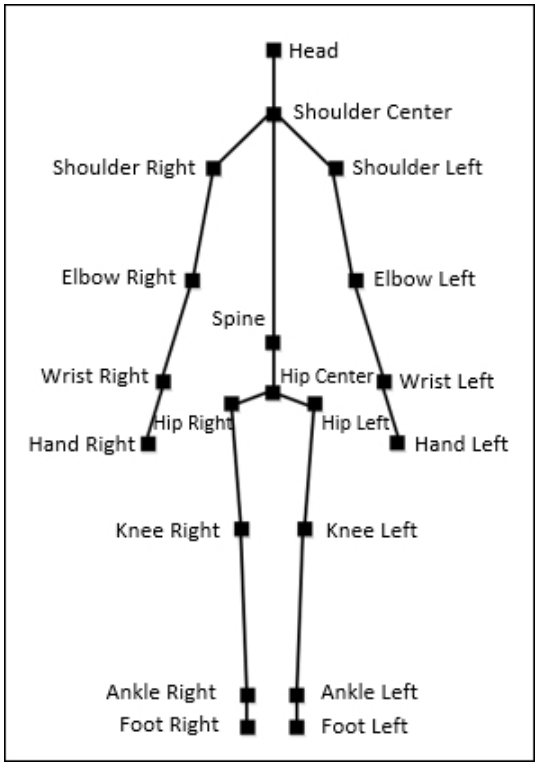

Figure 8. Example of a human body skeleton representation.

4.2. Algorithms

We outline the deep neural network (DNN)-based approaches employed in the literature for solving the problem of human activity recognition that utilize sensor inputs. The most common methodologies include convolutional neural networks (CNNs), autoencoders, and recurrent neural networks (RNNs).

4.2.1. Convolutional Neural Networks

This subsection discusses the application of convolutional neural networks (CNNs) in human activity recognition (HAR) and highlights its advantages and considerations. CNNs leverage sparse interactions, parameter sharing, and equivariant representations. After convolution, pooling and fully connected layers are typically used for classification or regression tasks.

CNNs have proven effective in extracting features from signals and have achieved promising results in various domains such as image classification, speech recognition, and text analysis. When applied to time-series classification like HAR, CNN offers two advantages over other models: local dependency and scale invariance. Local dependency

refers to the correlation between nearby signals in HAR, while scale invariance means the model can handle different paces or frequencies. Due to the effectiveness of CNNs, most of the surveyed work in HAR has focused on this area.

When applying CNNs to HAR, several aspects need to be considered: input adaptation, pooling, and weight-sharing.

4.2.2. Input Adaptation

HAR sensors typically produce time-series readings, such as acceleration signals, which are temporal multi-dimensional 1D readings. Input adaptation is necessary to transform these inputs into a suitable format for CNN. There are two main types of adaptation approaches:

4.2.3. Data-Driven Approach

Each dimension is treated as a channel, and 1D convolution is performed on them. The outputs of each channel are then flattened to unified deep neural network (DNN) layers. This approach treats the 1D sensor readings as a 1D image, and examples include treating accelerometer dimensions as separate RGB channels [43] and using shared weights in multi-sensor CNNs [44]. While this approach is simple, it ignores dependencies between dimensions and sensors, which can impact performance.

4.2.4. Model-Driven Approach

The inputs are resized to form a virtual 2D image, allowing for 2D convolution. This approach requires non-trivial input tuning techniques and domain knowledge. Examples include combining all dimensions into a single image [45] or using complex algorithms to transform time series into images [46]. This approach considers the temporal correlation of sensors but requires additional efforts in mapping time series to images.

4.2.5. Weight-Sharing

Weight-sharing is an efficient method to speed up the training process on new tasks. Different weight-sharing techniques have been explored, including relaxed partial weight-sharing [43] and CNN-pf and CNN-pff structures [47]. Partial weight-sharing has shown improvements in CNN performance.

In summary, this subsection provides an overview of the key concepts and considerations when applying CNNs to HAR, including input adaptation, pooling, and weight-sharing techniques.

4.2.6. Recurrent Neural Networks

RNN is a popular neural network architecture that leverages temporal correlations between neurons and is commonly used in speech recognition and natural language processing. In HAR, a RNN is often combined with long short-term memory (LSTM) cells, which act as memory units and help capture long-term dependencies through gradient descent.

There have been relatively few works that utilize RNNs for HAR tasks. In these works [48–51], the main concerns are the learning speed and resource consumption in HAR. One study [51] focused on investigating various model parameters and proposed a model that achieves high throughput for HAR. Another study [48] introduced a binarized-BLSTM-RNN model, where the weight parameters, inputs, and outputs of all hidden layers are binary values. The main focus of RNN-based HAR models is to address resource-constrained environments while still achieving good performance.

In summary, this subsection highlights the limited use of RNNs in HAR and emphasizes the importance of addressing learning speed and resource consumption in HAR applications. It mentions specific studies that have explored different approaches, including optimizing model parameters and introducing resource-efficient models like the binarized-BLSTM-RNN model.

5. Sensor-Based HAR

Human activity recognition (HAR) has witnessed transformative advancements with the incorporation of body-worn sensors, ushering in a new era of ubiquitous and continuous monitoring. Accelerometer and Inertial Measurement Unit (IMU) sensors serve as foundational elements, capturing motion-based data to deduce activities ranging from simple gestures to complex movements. In parallel, WiFi sensors have unlocked the potential for HAR even without direct contact with the human body, by leveraging the ambient wireless signals that reflect off our bodies. Advancements in radar sensors further expand the spectrum, providing a richer dataset with the capability to penetrate obstacles and discern minute motions, such as breathing patterns. However, the pinnacle of HAR's potential is realized through sensor fusion, where the combined prowesses of these varied sensors coalesce, offering enhanced accuracy, resilience to environmental noise, and the ability to operate in multifarious scenarios. This section delves into the multifaceted applications of these sensors, both in isolation and in harmony, illustrating their transformative impact on HAR's landscape. In the following subsections, we present applications of HAR that leverage accelerometer and IMU signals, WiFi, and radar imaging, as well as methods that combine all of the aforementioned modalities. The findings of this section are summarized in Tables 3–6, for the modalities of accelerometers, Wi-Fi, RADAR-based methods and finally for the various modalities, respectively.

Table 3. Accelerometer and IMU sensor-based methods.

Paper	Approach	Cost	Performance	Dataset	Supervised?
[3]	SVM + 1D CNN	Low	High	UCI-HAR	
[4]	MLP Ensemble	High	High	REALDISP	
[5]	1D CNN	Medium	Medium	SHO, MHealth	
[52]	1D CNN	High	High	UCI-HAR, WISDM, Skoda Dataset, self-prepared	
[6]	Self-attention + 1D CNN	High	High	UCI-HAR, MHealth	
[53]	Transformers	High	High	WISDM	
[54]	LSTM	Medium	Medium	WISDM	
[55]	MLP Ensembles	High	High	MotionSense (kaggle), self-prepared	
[56]	ConvLSTM	High	High	WISDM, UCI, PAMAP2, OPPORTUNITY	
[8]	Deformable CNN	High (4×3090)	Medium	OPPORTUNITY, UNIMIB-SHAR, WISDM	
[57]	LSTMs, Hierarchical Clustering	Medium	Medium	MHealth, UCI-HAR	
[58]	CNN	Medium	High	UCI-HAR, OPPORTUNITY, UNIMIB-SHAR, WISDM, PAMAP2	
[59]	Temporal CNN	High	Medium	PAMAP2, OPPORTUNITY, LISSI	Semi
[60]	GRU-ResNet	High	High	UCI-HAR	Supervised
[61]	MLP	Medium	High	N/A	Unsupervised
[62]	CNN	High	Medium	OPPORTUNITY, UNIMIB-SHAR, WISDM	Supervised
[63]	Linear Discriminant Analysis	Low	High	Self-prepared	Supervised

Table 3. *Cont.*

Paper	Approach	Cost	Performance	Dataset	Supervised?
[64]	GRU-CNN	High	High	UCI-HAR, OPPORTUNITY, MHealth	Supervised
[65]	CNN	Medium	High	PAMAP2	Supervised
[66]	CNN	Low	High	UCI-HAR, PAMAP2, WISDM, UNIMIB-SHAR	Supervised
[67]	MLP	Low	High	UCI ML Repository	Supervised

Table 4. WiFi sensor-based methods.

Paper	Approach	Cost	Performance	Dataset	Supervised?
[68]	CNN-RNN	Medium	Medium	Self collected	Unsupervised
[69]	SVM, MLP, CNN	Low	Medium	Self collected	Unsupervised
[70]	CNN	Low	High	Self collected	Supervised
[71]	ConvLSTM, PCA with STFT	Low	High	Self collected	Supervised
[72]	CNN	Low	High with Transfer Learning, Low w/o Transfer Learning	Self collected	Supervised
[73]	CNN Unet	Low	Medium	Self collected	Supervised
[74]	CNN	Low	High	Self collected	Unsupervised

Table 5. Classification of radar-based methods.

Paper	Approach	Cost	Performance	Dataset	Supervised?
[75]	2DPCA, 2DLDA, kNN	Low	High	University of Glasgow Dataset	Unsupervised
[76]	Transformer	High	High	4d Imaging Radar Dataset	Supervised
[77]	RNN	Medium	N/A	PARRad Dataset	Unsupervised

Table 6. Classification of methods that leverage a fusion of various modalities.

Paper	Approach	Cost	Performance	Dataset	Supervised?
[78]	Transformer	High	High	[79–81]	Supervised
[82]	GMM, HMM	Low	Medium	NGSIM (Next Generation Simulation (NGSIM) https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajector/8ect-6jqj (accessed on 12 December 2023))	Unsupervised
[83]	SVM	Low	High	UCI-HAR	Supervised
[84]	Attention based	High	Medium	OPPORTUNITY, UCI ML REPOSITORY, Daily life activities	Supervised
[85]	CNN	High	High	Self-supervised	Supervised

5.1. Accelerometer and IMU Modalities

A contribution to accelerometer-based HAR is presented in [3], where the authors introduce an adaptive HAR model that employs a two-stage learning process. This model utilizes data recorded from a waist-mounted accelerometer and gyroscope sensor to first distinguish between static and moving activities using a random forest classifier. Subsequent classification of static activities is performed by a Support Vector Machine (SVM), while moving activities are identified with the aid of a 1D convolutional neural network (CNN). Building on the idea of utilizing deep learning for HAR, the work in [4] proposes an ensemble deep learning approach that integrates data from sensor nodes located at various body sites, including the waist, chest, leg, and arm. The implementation involves training three distinct deep learning networks on a publicly available dataset, encompassing data from eight human actions, demonstrating the potential of ensemble methods in improving HAR accuracy. Furthermore, the research in [5] introduces a multi-modal deep convolutional neural network designed to leverage accelerometer data from multiple body positions for activity recognition. The study in [52] explores the application of a one-dimensional (1D) CNN model for HAR, utilizing tri-axis accelerometer data collected from a smartwatch. This research focuses on distinguishing between complex activities such as studying, playing games, and mobile scrolling.

Another development in this area is presented in [6], where the authors introduce a convolved self-attention neural network model optimized for gait detection and HAR tasks. Further addressing the challenges of capturing spatial and temporal relationships in time-series data from wearable devices, the work in [53] explores the application of a transformer-based deep learning architecture. Recognizing the limitations of traditional AI algorithms, where convolutional models focus on local features and recurrent networks overlook spatial aspects, the authors propose leveraging the transformer model's self-attention mechanism. In another approach, Ref. [54] employs Bidirectional long short-term memory (LSTM) networks for data generation using the WISDM dataset, a publicly available tri-axial accelerometer dataset. The study focuses on assessing the similarity between generated and original data and explores the impact of synthetic data on classifier performance. Addressing the variability in human subjects' physical attributes, which often leads to inconsistent model performance, the authors in [55] propose a physique-based HAR approach. By leveraging raw data from smartphone accelerometers and gyroscopes. Additionally, the study in [56] introduces an architecture called "ConvAE-LSTM", which synergizes the strengths of convolutional neural networks (CNNs), autoencoders (AEs), and Long Short-Term Memory (LSTM) networks. In parallel, the challenge of capturing salient activity features across varying sensor modalities and time intervals due to the fixed nature of traditional convolutional filters is addressed in [8]. This research presents a deformable convolutional network designed to enhance human activity recognition from complex sensory data. Ref. [57] contributes to the HAR field by proposing a hierarchical framework named HierHAR, which focuses on distinguishing between similar activities. This structure serves as the basis for a tree-based activity recognition model, complemented by a graph-based model to address potential compounding errors during the prediction process.

In [58], Teng et al. introduce a groundbreaking approach by implementing a CNN-based architecture with a local loss. The adoption of a local loss-based CNN represents a strategy in the domain that offers a more refined feature extraction capabilities and enhanced learning efficacy for activity recognition. Additionally, Ref. [59] explores the use of Generative Adversarial Networks (GANs) combined with temporal convolutions in a semisupervised learning framework for action recognition. This approach is particularly designed to tackle common challenges in HAR, such as the scarcity of annotated samples. Ref. [60] presents the BiGRUResNet model, which combines LSTM-CNN architectures with deep residual learning for improved HAR accuracy and reduced parameter count.

In [61], the authors propose XAI-BayesHAR, an integrated Bayesian framework by leveraging a Kalman filter to recursively track the feature embedding vector and its associ-

ated uncertainty. This feature is particularly valuable for practical deployment scenarios where understanding and quantifying predictive uncertainty are required. Furthermore, the framework's ability to function as an out-of-data distribution (OOD) detector adds an additional layer of practical utility by identifying inputs that significantly deviate from the trained data distribution. In contrast to traditional wrist-worn devices, Yen, Liao, and Huang in [86] explore the potential of a waist-worn wearable device specifically designed to accurately monitor six fundamental daily activities, particularly catering to the needs of patients with medical constraints such as artificial blood vessels in their arms. The hardware and software components of this wearable, including an ensemble of inertial sensors and a sophisticated activity recognition algorithm powered by a CNN-based model, highlight the importance of tailored hardware and algorithmic approaches in enhancing HAR accuracy in specialized healthcare applications. Furthermore, Ref. [87] delves into the utilization of smartphones and IoT devices for HAR. The work in [62] addresses the challenge of balancing HAR efficiency with the inference-time costs associated with deep convolutional networks, which is especially pertinent for resource-constrained environments.

The authors of [63] introduce a wearable prototype that combines an accelerometer with another launchpad device. This device is specifically designed to recognize and classify activities such as walking, running, and stationary states. The classification algorithm is based on analyzing statistical features from the accelerometer data, applying Linear Discriminant Analysis (LDA) for dimensionality reduction, and employing a support vector machine (SVM) for the final activity classification. Building on the integration of advanced neural network architectures for HAR, Ref. [64] presents a system that combines convolutional neural networks (CNNs) with Gated Recurrent Units (GRUs). This hybrid approach allows for enhanced feature extraction through the CNN layers, followed by the GRU layers capturing temporal dependencies. In response to the constraints imposed by resource-limited hardware, Ref. [65] introduces an innovative HAR system based on a partly binarized Hybrid Neural Network (HNN). This system is optimized for real-time activity recognition using data from a single tri-axial accelerometer, distinguishing among five key human activities. Finally, Tang et al. in [66] propose a CNN architecture that incorporates a hierarchical-split (HS) module, designed to enhance multiscale feature representation by capturing a broad range of receptive fields within a single feature layer.

Finally, in a similar direction, Ref. [67] explores the integration of cost-effective hardware focusing on leveraging the gyroscope and accelerometer to feed data into a deep neural network architecture. Their model, DS-MLP, represents a confluence of affordability and sophistication, aiming to make HAR more accessible and reliable.

5.2. Methods Leveraging WiFi Signals

The work presented in [68] introduces WiHARAN, a robust WiFi-based activity recognition system designed to perform effectively in various settings. The key to WiHARAN's success lies in its ability to learn environment-independent features from Channel State Information (CSI) traces, utilizing a base network adept at extracting temporal information from spectrograms. This is complemented by adversarial learning techniques that align the feature and label distributions across different environments, thus ensuring consistent performance despite changes in the operating conditions. Complementing this, the approach detailed in [69] explores a device-free methodology utilizing WiFi Received Signal Strength Indication (RSSI) for recognizing human activities within indoor spaces. By employing machine learning algorithms and collecting RSSI data from multiple access points and mobile phones, the system achieves remarkable accuracy, particularly in the 5 GHz frequency band. Further expanding the scope of WiFi-based activity recognition, Ref. [70] presents a comprehensive framework that integrates data collection and machine learning models for the simultaneous recognition of human orientation and activity. In a different vein, Ref. [71] focuses on the nuances of limb movement and its impact on WiFi signal propagation. The research identifies the challenges associated with variability in activity performance and individual-specific signal reflections, which complicate the recognition

process. To overcome these hurdles, a novel system is proposed that leverages diverse CSI transformation methods and deep learning models tailored for small datasets. The work by Ding et al. [72] showcases the application of a straightforward convolutional neural network (CNN) architecture, augmented with transfer learning techniques, with minimal or even no prior training. This approach not only demonstrates the effectiveness of transfer learning in overcoming the limitations of small training sets but also sets new benchmarks in location-independent, device-free human activity recognition. Amidst these technological advancements, the aspect of user privacy emerges as a pivotal consideration, particularly in the domain of HAR applications and, by extension, Assisted Daily Living (ADL). As the quality of service for IoT devices improves, users are increasingly faced with the balance between benefitting from enhanced services and risking the exposure of sensitive personal information, such as banking details, workout routines, and medical records. The participants of the study conducted in [88] highlight a nuanced stance towards this balance; they seem amenable to sharing sensitive information about their health and daily habits as long as the IoT services provided are tailored to their needs and yield better service performance. Nonetheless, their willingness to share personal data is conditional, underscored by the prerequisite of specific circumstances that safeguard their privacy while enabling the benefits of technology-enhanced living.

Wireless sensing technologies have seen remarkable advancements, particularly in the realms of indoor localization and human activity recognition (HAR), propelled by the nuanced capabilities of wireless signals to reflect human motions. This is vividly illustrated in the works of [73,74], where each study presents a unique approach to leveraging Channel State Information (CSI) for enhanced HAR and localization. In [73], the focus is on the dynamic nature of wireless signals and their interactions with human activities. The authors observe that the scattering of wireless signals, as captured in the CSI, varies with different human motions, such as walking or standing. They achieve this by leveraging the U-Net architecture, a specialized convolutional neural network (CNN). Ref. [74] integrates wireless sensing within the Internet of Things (IoT) ecosystem, emphasizing its importance for both HAR and precise location estimation. The authors in [74] introduce a hardware design, which simplifies CSI acquisition and enables the simultaneous execution of HAR and localization tasks using Siamese networks. This integrated approach is particularly advantageous in smart home environments, where it can facilitate gesture-based device control from specific locations without compromising user privacy, as it negates the need for wearable sensors or cameras.

5.3. Radar Signal HAR

In an effort to address challenges related to high data dimensions in Frequency Modulated Continuous Wave (FMCW) radar images, slow feature extraction, and complex recognition algorithms, Ref. [75] presents a method using two-dimensional feature extraction for FMCW radar. The approach begins by employing two-dimensional principal component analysis (2DPCA) to reduce the dimensionality of the radar Doppler–Time Map (DTM). The recognition task is accomplished using a k-nearest neighbor (KNN) classifier. Ref. [76] explores the application of three self-attention models, specifically Point Transformer models, in the classification of Activities of Daily Living (ADL). The experimental dataset, collected at TU Delft, serves as the foundation for investigating the optimal combination of various input features, assessing the impact of the proposed Adaptive Clutter Cancellation (ACC) method, and evaluating the model's robustness within a leave-one-subject-out scenario. In [77], the Bayesian Split Bidirectional recurrent neural network for human activity recognition is introduced, in order to compensate for the computational cost of deep neural networks. The proposed technique harnesses the computational capabilities of the off-premise device to quantify uncertainty, distinguishing between epistemic (uncertainty due to lack of training data) and aleatoric (inherent uncertainty in predictions) uncertainties. Radar signal modalities are used to predict human activities.

5.4. Various Modalities and Modality Fusion

One approach that leverages multiple modalities for HAR is the Data Efficient Separable Transformer (DeSepTr) framework, as introduced in [78]. This framework leverages the capabilities of transformer neural networks [89,90], specifically a Vision Transformer (ViT), trained using spectrograms generated from data collected by wearable sensors. In response to the limitations of model-based hidden Markov models, the switching Gaussian mixture model-based hidden Markov model (S-GMMHMM), is introduced in [82]. The S-GMMHMM utilizes a supervised algorithm for accurate parameter estimation and introduces a real-time recognition algorithm to compute activity posteriors recursively. Capitalizing on Enveloped Power Spectrum (EPS) to isolate impulse components from signals, Ref. [83] introduces Linear Discriminant Analysis (LDA), which is employed to reduce the feature dimensionality and extract discriminative features. The extracted features are then used to train a multiclass support vector machine (MCSVM). Building upon the need for efficient data handling and feature extraction in HAR, Ref. [84] introduces the HAP-DNN model, an advanced solution addressing the challenges posed by the extensive sensor data and multichannel signals typical in HAR tasks. Finally, in response to the complexities of recognizing human activities in varied real-world settings, Ref. [85] presents a method that combines data from Frequency Modulated Continuous Wave (FMCW) radar with image data, taking advantage of the complementary strengths of these modalities to boost recognition accuracy. This approach is further enhanced by incorporating domain adaptation techniques to address inconsistencies in data due to environmental variations and differing user behaviors.

6. Vision-Based HAR

Human activity recognition (HAR) represents a pivotal area of research in the realm of computer vision, aiming to identify and categorize human actions from a series of observational data. Primarily, the data feeding these models come from visual sources such as videos or sequences of images captured from a variety of devices like surveillance cameras, smartphones, or dedicated recording equipment. Traditional machine learning algorithms, having provided the foundation for HAR, have in recent years been complemented and even superseded by more advanced architectures. Particularly, convolutional neural networks (CNNs) have proven adept at extracting spatial hierarchies and patterns from visual data [91–93], while Vision Transformers [90,94] divide the input image into fixed-size patches and linearly embed them for attention-driven understanding. Furthermore, to capture the temporal dependencies inherent in video sequences, researchers often amalgamate CNNs with recurrent neural networks (RNNs) [95]. This combination exploits the spatial recognition capabilities of CNNs and the sequence understanding of RNNs, offering a comprehensive understanding of both the spatial and temporal aspects of human activities. As technology advances, the fusion of classical and modern algorithms offers promise in achieving more accurate and real-time HAR applications. The typical vision-based approaches re illustrated in Figure 9. Many approaches use a combination of primitive tasks, such as image segmentation [96] in order to extract semantic information about the scene and infer the ongoing activity [97], as shown in Figure 10. We summarize our findings in Table 7.

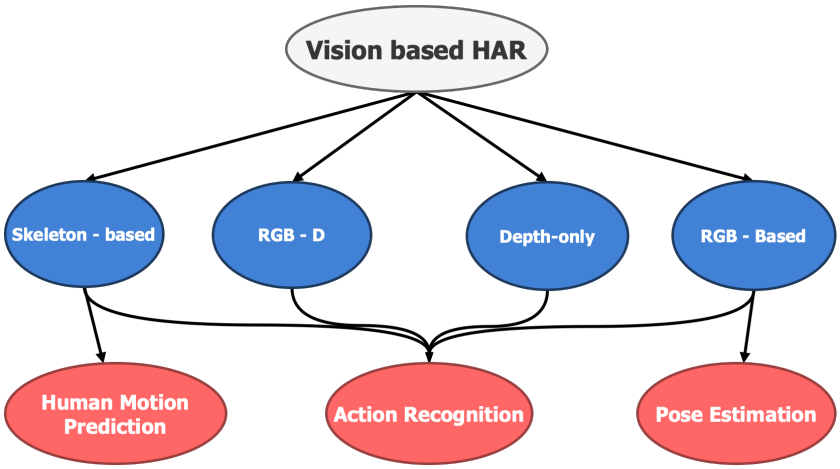

Figure 9. Typically used modalities in vision-based methods in HAR. Due to the wide range of applications of HAR, vision-based methods can be further categorized in subsequent tasks. Most notable are human motion prediction, action recognition and pose estimation.

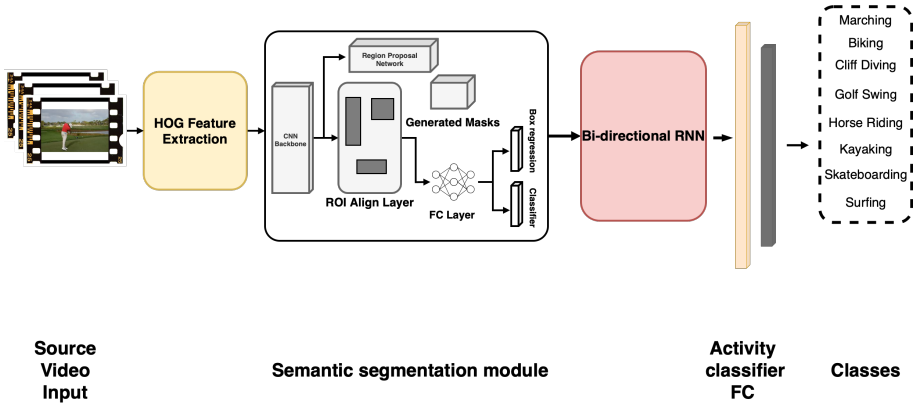

Figure 10. The framework proposed in [97]. First, the video is split into a sequence of RGB images and consequently fed into an instance segmentation module, typically a MaskRCNN [96]. The per-frame semantic representation extracted from the segmentation module are then processed by an RNN, before they are fed to a fully connected layer to finally infer the activity class.

Table 7. Classification of vision-based methods.

Paper	Approach	Cost	Performance	Dataset	Supervised?
[98]	Attention, LSTM	High	High	UTD-MHAD, UT-Kinect, UCSD MIT	Supervised
[99]	CNN + LSTM	High	High	Kinetic Activity Recognition Dataset	Supervised
[100]	3D CNNs	High	High	MSRDailyActivity3D, NTU RGB + D and UTD-MHAD, PRECIS HAR	Supervised
[101]	DBN	Low	High	HMDB51	Supervised
[102]	Decision Tree	Low	High	UT-Interaction	Unsupervised
[103]	CNN LSTM	Medium	High	iSPL, UCI-HAR	Supervised
[104]	LSTM	Medium	High	UCF-50	Supervised
[105]	HMM	Low	High	Depth Dataset Using Kinect Camera	Unsupervised

The work presented in HAMLET [98] is characterized by a hierarchical architecture designed for encoding spatio-temporal features from unimodal data. This is achieved through a multi-head self-attention mechanism that captures detailed temporal and spatial dynamics. In a similar fashion, the SDL-Net model introduced in [99] leverages the synergies of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to address HAR challenges. This model uniquely employs Part Affinity Fields (PAFs) to construct skeletal representations and skeletal gait energy images, which are instrumental in capturing the nuanced sequential patterns of human movement. Moreover, Popescu et al. [100] augment information from multiple channels within a 3D video framework, including RGB, depth data, skeletal information, and contextual objects. By processing each data stream through independent 2D convolutional neural networks and then integrating the outputs using sophisticated fusion mechanisms, their system achieves a comprehensive understanding of the video content. Ref. [101] proposes a novel approach that melds the strengths of Deep Belief Networks (DBNs) with a DBN-based Recursive Genetic Micro-Aggregation Approach (DBN-RGMAA) to ensure anonymity in HAR. The Hybrid Deep Fuzzy Hashing Algorithm (HDFHA) further enhances this approach by capturing complex dependencies between actions. Furthermore, the CNN-LSTM framework introduced in [103] by harmoniously integrating convolutional neural networks (CNNs) with long short-term memory networks (LSTMs), captures the spatio-temporal dependencies of human activities. The study by [104] introduces a pioneering hybrid approach that integrates the temporal learning capabilities of long short-term memory (LSTM) networks with the optimization prowess of Particle Swarm Optimization (PSO). This combination effectively analyzes both the temporal and spatial dimensions of video data. The introduction of the Maximum Entropy Markov Model (MEMM) by the authors of [105] treats sensor observations as inputs to the model and employs a customized Viterbi algorithm to infer the most likely sequence of activities.

7. Datasets

7.1. Sensor-Based Datasets

Ref. [106] offers an exhaustive survey on datasets tailored for human activity recognition (HAR) through the use of installed sensors. These datasets are pivotal for researchers aiming to develop efficient machine learning models for activity recognition, and they exclude datasets based on RGB or RGB-Depth video actions. The content of these datasets spans a broad spectrum of sensor-based activities. Firstly, there are datasets from the UCI Machine Learning Repository, which encompass a variety of domains, including activities of daily living, fall classification, and smartphone-based activity recognition. Another significant contributor is the Pervasive System Research Group from the University of Twente, offering datasets on smoking activity, complex human activities, and physical activity recognition. The Human Activity Sensing Consortium (HASC) provides large-scale databases, with datasets like HASC2010corpus and HASC2012corpus, focusing on human behavior monitoring using wearable sensors. Medical activities also find representation with datasets like the Daphnet Freezing of Gait (FoG) dataset, which monitors Parkinson's disease patients, and the Nursing Activity dataset that observes nursing activities in hospitals. For those interested in physical and sports activities, datasets such as the Body Attack Fitness and Swimmaster datasets provide insights into workout routines and swimming techniques, respectively. The rise of smart homes has led to the inclusion of household activities-related datasets like the MIT PlaceLab and CMU-MMAC, which utilize IoT to track daily household activities. Lastly, with the ubiquity of smart devices, the document also touches upon datasets related to device usage, capturing user behaviors and patterns while interacting with these devices.

The UCI Machine Learning Repository offers a diverse collection of datasets pivotal for human activity recognition (HAR) research. The HHAR dataset [107], for instance, encompasses data from nine subjects across six activities, with 16 attributes spanning about 44 million instances. Similarly, the UCIBWS dataset [108], derived from 14 subjects,

captures four activities and comprises nine attributes across 75,000 instances. The AReM dataset [109], although from a single subject, provides insights into six activities with six attributes, though the instance count remains unspecified. The HAR [110] and HAPT datasets [111], both sourced from 30 subjects, delve into 6 and 12 activities respectively, both having 10,000 instances. Single Chest [112] and OPPORTUNITY datasets [113] offer a broader activity spectrum, with the former capturing 7 activities from 15 subjects and the latter detailing 35 activities from 4 subjects. ADLs [114] and REALDISP datasets [115] provide a snapshot of daily activities, with the latter offering an extensive 33 activities and attributes from 17 subjects. UIFWA [112] and PAMAP2 [81] datasets, from 22 and 9 subjects, respectively, offer varied attributes, with the former's activity count unspecified. DSA and Wrist ADL datasets [116], both focusing on daily activities, provide data from 8 and 16 subjects respectively and proposed in the same publication. The RSS dataset [117], while not specifying subject count, offers insights into two activities with four attributes. Lastly, the MHEALTH [118], WISDM [119], and WESAD datasets [120], derived from 10, 51, and 15 subjects, respectively, provide a comprehensive view of various activities, with the WISDM dataset [119] detailing 18 activities across 15 million instances. Collectively, these datasets from the UCI Machine Learning Repository present a rich tapestry of information that is invaluable for researchers aiming to advance the domain of HAR. Table 8 summarizes all the datasets' attributes along with the measurement tools employed to create the samples.

Table 8. Summary of datasets from the UCI Machine Learning Repository for human activity recognition (HAR) research. The table provides details on the number of subjects, activities, instances, measurement sensors and the respective sources for each dataset.

Dataset	# Subjects	# Activities	Sensors	# Instances	# Source
HHAR	9	6	Accelerometer, gyroscope	44 million	[107]
UCIBWS	14	4	RFID	75k	[108]
AReM	1	6	IRIS Nodes	42k	[109]
HAR	30	6	Accelerometer, gyroscope	10k	[110]
HAPT	30	12	Accelerometer, gyroscope	10k	[111]
Single Chest	15	7	Accelerometer	N/A	[112]
OPPORTUNITY	4	35	Accelerometer, motion sensors, ambient sensors	2551	[113]
ADLs	2	10	PIR, magnetic, pressure and electric sensor	2747	[114]
REALDISP	17	33	Accelerometer, gyroscope	1419	[115]
UIFWA	22	2	Accelerometer	N/A	[112]
PAMAP2	9	19	IMU, ECG	3.8 million	[81]
DSA	8	19	Accelerometer, magnetometers, gyroscope	9120	[116]
Wrist ADL	16	14	Accelerometer	N/A	[116]
RSS	N/A	2	N/A	13,917	[117]

Table 8. *Cont.*

Dataset	# Subjects	# Activities	Sensors	# Instances	# Source
MHEALTH	10	12	Accelerometer, ECG	120	[118]
WISDM	51	18	Accelerometer, gyroscope	15 million	[119]
WESAD	15	3	N/A	63 million	[120]

7.2. Vision-Based Datasets

In terms of visual-based human activity recognition (HAR), several benchmark datasets have been established to evaluate and validate algorithmic approaches and learning methods. At the action level, the KTH Human Action Dataset, developed by the Royal Institute of Technology of Sweden in 2004, offers 2391 sequences spanning six distinct human actions [121]. Transitioning to the behavior level, the VISOR Dataset from the University of Modena and Reggio Emilia encompasses 130 video sequences tailored for human action and activity recognition [122]. Additionally, the Caviar Dataset (2004) provides videos capturing nine activities in varied settings, while the Multi-Camera Action Dataset (MCAD) [123] from the National University of Singapore focuses on 18 daily actions across five camera perspectives. On the interaction level, the MSR Daily Activity 3D Dataset by Microsoft Research Redmond [124] comprises 320 sequences across channels like depth maps, skeleton joint positions, and RGB video, capturing 10 subjects in 16 activities. Other notable datasets include the UCF50 [125] with 50 action categories from YouTube, the HMDB-51 Dataset from Brown University [126] with 6849 clips across 51 action categories, and the Hollywood Dataset by INRIA [127], featuring actions from 32 movies. We split our pool of the benchmark datasets into four major categories: (i) action-level datasets, (ii) behavioral-level datasets, (iii) interaction-level datasets, and (iv) group activities-level datasets. Table 9 summarizes the fundamental properties of all the datasets that will be described below.

Table 9. Taxonomy of the vision-based dataset based on their interaction properties.

Dataset	Action	Behavior	Human–Object Interaction	Human–Human Interaction	Group Activities
KTH [121]	✓				
Weizmann [128]	✓				
Stanford 40 [129]	✓				
IXMAS [130]	✓				
VISOR [122]		✓			
MCAD [123]		✓			
MSR Daily Activity 3D [124]	✓		✓	✓	
50 Salads [131]	✓		✓		
UCF50 [125]	✓		✓		
ETISEO [132]		✓		✓	
Olympic Sports [133]	✓		✓		
UT-Interaction [134]		✓		✓	
UT-Tower [135]		✓		✓	✓
ActivityNet [136]	✓	✓			✓
Kinetics [137]	✓			✓	✓
HMDB-51 [126]	✓			✓	✓
Hollywood [127]	✓			✓	
Hollywood2 [127]	✓			✓	
UCF-101 [138]	✓			✓	✓
YouTube Action [139]		✓		✓	

7.2.1. Action-Level Datasets

In the category of action-level datasets, the KTH Human Action Dataset, developed by the Royal Institute of Technology of Sweden in 2004, consists of 2391 sequences across six action classes, including actions such as walking, jogging, and boxing, recorded against consistent backgrounds with a stationary camera. In contrast, the Weizmann Human Action Dataset, introduced by the Weizmann Institute of Science in 2005, comprises 90 sequences with nine individuals performing 10 distinct actions, including jumping, skipping, and hand waving. The Stanford 40 Actions dataset, curated by the Stanford Vision Lab, encompasses 9532 images from 40 different action classes. Another notable dataset is the IXMAS dataset, established in 2006, offering a unique multi-view perspective on action recognition by capturing 11 actors executing 13 everyday actions using five cameras from varied angles. Lastly, the MSR Action 3D dataset, a brainchild of Wanqing Li from Microsoft Research Redmond, includes 567 depth map sequences, where 10 subjects perform 20 types of actions, captured using a Kinect device. Some characteristic dataset are shown in Figure 11.

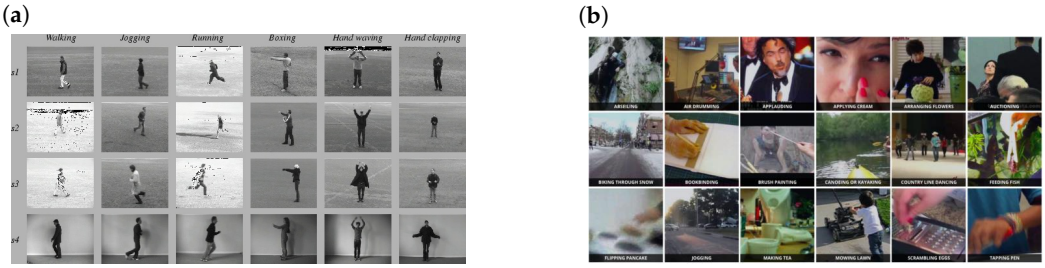

Figure 11. Comparison of activity snapshots from two renowned datasets, KTH [121] and Kinetics [137]. While both datasets provide a visual array of human actions, they vary in complexity and scene context. (**a**) Snapshots from the KTH dataset, showcasing various human activities such as walking, jogging, running, boxing, hand waving, and hand clapping. (**b**) A collection of snapshots from the Kinetics dataset, depicting diverse activities ranging from air drumming and applauding to making tea and mowing the lawn.

7.2.2. Behavioral-Level Datasets

Examining the behavioral-level datasets, we have the VISOR dataset, introduced by the Imagelab Laboratory of the University of Modena and Reggio Emilia in 2005, which includes a diverse range of videos categorized by type. One significant category, designated for human action recognition in video surveillance, encompasses 130 video sequences. The Caviar dataset, founded in 2004, splits into two distinct sets: one captured using a wide-angle lens in the INRIA Labs' lobby in Grenoble, France, and the other filmed in a shopping center in Lisbon. This dataset captures individuals engaging in nine activities across two contrasting settings. Lastly, the Multi-Camera Action Dataset (MCAD), a creation of the National University of Singapore, addresses the open-view classification challenge in surveillance contexts. It comprises recordings of 18 daily actions, sourced from other datasets like KTH and IXMAS, captured through five cameras and performed by 20 subjects. Each subject performs each action eight times, split evenly between daytime and evening sessions.

7.2.3. Interaction-Level Datasets

In the category of interaction-level datasets essential for vision-based human activity recognition, a large number of datasets can be encountered. Among these, the MSR Daily Activity 3D Dataset from Microsoft Research Redmond offers 320 sequences across channels including depth maps, skeleton joint positions, and RGB video, capturing 10 subjects engaging in 16 activities in both standing and sitting stances. The 50 Salads dataset from the University of Dundee chronicles 25 individuals preparing salads in videos summing up to 4 h. The MuHAVI dataset by Kingston University focuses on silhouette-based human action recognition with videos from eight different angles. Additionally, the University of Central Florida has contributed two significant datasets: UCF50, which extends from

the UCF11 dataset with 50 action categories, and the UCF Sports Action Dataset, offering 150 sequences from televised sports. The ETISEO dataset by the INRIA Institute aims at enhancing video surveillance techniques across varied settings. Meanwhile, the Olympic Sports Dataset by the Stanford Vision Lab comprises 50 videos of athletes from 16 sports disciplines. The University of Texas presented the UT-Interaction dataset, stemming from a research competition, spotlighting 20 video sequences of continuous human–human interactions across diverse attire conditions. Lastly, the UT-Tower dataset, also by the University of Texas, showcases 108 video sequences spanning two settings: a concrete square and a lawn.

7.2.4. Group Activity-Level Datasets

The study of group interaction in vision-based datasets has experienced significant expansion and diversification. A notable contribution is the ActivityNet Dataset, created in 2015, offering 849 video hours spanning 203 activity classes and three comparison scenarios, making it a robust choice for evaluating human activity understanding algorithms. The Kinetics Human Action Video Dataset, crafted by DeepMind in 2017, originally provided 400 human action classes (Kinetics 400) but later expanded to 600 (Kinetics 600), making it one of the most extensive datasets, comprising around 500,000 video clips sourced from YouTube. The HMDB-51 dataset by the Serre Lab of Brown University includes 6849 clips from 51 action categories, standing out as one of the largest datasets in human activity recognition. The Hollywood and Hollywood2 datasets, both hailing from INRIA, provide short sequences from 32 and 69 movies, respectively, highlighting multiple human actions in realistic settings. The UCF-101 Action Recognition Dataset from the University of Central Florida stands as an extension of the UCF50 dataset, featuring 13,320 videos of 101 action categories, making it a gold standard for diversity in terms of actions and realistic challenges. Finally, the YouTube Action Dataset from the same institution provides 11 action categories, offering challenges due to a multitude of factors like camera motion and illumination variations.

8. HAR In Robotics and Industry

Recognizing human activities and being able to predict human intent plays a crucial role in industrial environments and by extension in robotics. Specifically, in large production lines where humans have to closely cooperate with robotic agents, it is important for the robot to synergize efficiently by planning its path [140]. For this reason, complementary to HAR, it is also necessary to model human motion. Numerous approaches have been proposed to the task of modeling human motion and predicting it. Early methods incorporate RNNs in human skeleton models represented by their joint rotations [141,142]. Further extending this field of research, other methodologies combine RGB scene characteristics along with the skeleton representations [143]. The accelerating popularity of attention-based methods has also given the advantage of accurately modeling and predicting the human motion, and consequently, human intent. Notable works include [144,145].

The field of robotics is undoubtedly driven by the exploration of emulating human motion. In this context, Ref. [146] introduces a novel approach that combines deep learning and human motion imitation through the use of motion primitives. It further involves motion modeling using motion primitives and the replication of these motions in a simulated environment using the V-REP robotic simulator. The proposed framework represents an initial version of deep learning-powered video analytics for human motion imitation, employing motion primitive techniques.

9. Conclusions

The domain of human activity recognition (HAR) continues to grapple with several open problems that impede its advancement. One notable challenge lies in the limited availability of annotated data, which can potentially be mitigated by leveraging generative AI to autonomously generate data from text descriptions, thus alleviating data scarcity

issues. The diversification in data collection regarding the age, gender, and number of subjects, alongside handling postural transitions, is another area that demands attention. Moreover, despite the remarkable accuracies achieved by wearable sensor-based activity recognition methods, their adoption among elderly individuals remains low due to reluctance in wearing sensors. Moreover, the field faces hurdles in several core areas including data collection, data preprocessing, hardware and techniques, complex activity detection, and misalignment of activities. For instance, vision-based data, despite its larger size requiring more processing, is often more cost-effective compared to sensor-based data, the latter being more expensive. However, the trade-off between cost and data processing needs is an area of ongoing discussion and research within the community. Furthermore, the continuous evolution and expansion of HAR applications across various domains necessitate ongoing updates and reviews of the existing literature to keep pace with the emerging challenges and solutions. The complexity of these open problems underscores the necessity for a multidisciplinary approach, encompassing advancements in sensor technology, machine learning algorithms, and a deeper understanding of human behavior and activities, to propel the field of human activity recognition forward.

Author Contributions: Conceptualization, M.K. and I.K.; methodology, M.K.; formal analysis, M.K.; investigation, M.K.; writing—original draft preparation, M.K.; writing—review and editing, M.K., S.M. and I.K.; visualization, M.K.; supervision, I.K. All authors have read and agreed to the published version of the manuscript

Funding: This research was funded by the European Union's Horizon Europe Project "Sestosenso" (http://sestosenso.eu/ (accessed on 22 February 2024), HORIZON-CL4-Digital-Emerging) under Grant 101070310.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

We present the input to the LLM. As context we provide the paper's title concatenated with its abstract. The rows below context are the follow-up questions we provide to the LLM. If no answer is to be found, the model is prompted to return an N/A string as answer.

Table A1. Summary and query-response analysis of the paper on mobile convolutional neural networks for human action recognition

Item	Description
Context	"Human action recognition is a computer vision task that identifies how a person or a group acts on a video sequence. Various methods have been proposed to address this problem over the years; the most recent ones rely on deep-learning techniques. First, RGB video measurements and stacks of optical flows are fed as input, and then through two-dimensional convolutional neural networks (2D-CNNs), spatial and temporal analyses are performed. In another line of work, several applications use 2D-CNNs for feature extraction. This way, an image is represented via a feature vector employed to recurrent neural networks (RNNs) for temporal analysis. Motivated by the fact that most high complexity CNNs are utilized on human action recognition tasks and owing to the necessity for mobile implementations on platforms with restricted computational resources, this article evaluates the performance of four lightweight architectures. In particular, we examine how the models of certain mobile-oriented CNNs, viz., ShuffleNet-v2, EfficientNet-b0, MobileNet-v3, and GhostNet, execute in spatial analysis. To that end, we measure the classification accuracy on two human action datasets, the HMDB51, and the UCF101, when the presented models have been previously trained on ImageNet and BU101. The frameworks' evaluation is based on the average, max scores, and voting generated through the three and fifteen RGB frames of each video included in the test set. Finally, via the trained mobile 2D-CNNs extracted features, RNNs performance evaluation is also assessed where the temporal analysis is achieved."

Table A1. *Cont.*

Question	Answer
What is the paper's main contribution?	evaluates the performance of four lightweight architectures
What problem is being addressed?	Human action recognition
What is the input modality?	RGB video measurements and 3 stacks of optical flows
How is the method evaluated?	based on the average, 14 max scores, and voting
What is the methodology the authors approach the problem with?	deep-learning techniques

References

1. Gupta, S. Deep learning based human activity recognition (HAR) using wearable sensor data. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100046. [CrossRef]
2. Diraco, G.; Rescio, G.; Caroppo, A.; Manni, A.; Leone, A. Human Action Recognition in Smart Living Services and Applications: Context Awareness, Data Availability, Personalization, and Privacy. *Sensors* **2023**, *23*, 6040. [CrossRef] [PubMed]
3. Shuvo, M.M.H.; Ahmed, N.; Nouduri, K.; Palaniappan, K. A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network. In Proceedings of the 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, USA, 13–15 October 2020 ; pp. 1–5. [CrossRef]
4. Rojanavasu, P.; Jantawong, P.; Jitpattanakul, A.; Mekruksavanich, S. Improving Inertial Sensor-based Human Activity Recognition using Ensemble Deep Learning. In Proceedings of the 2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Phuket, Thailand, 22–25 March 2023; pp. 488–492. [CrossRef]
5. Muhoza, A.C.; Bergeret, E.; Brdys, C.; Gary, F. Multi-Position Human Activity Recognition using a Multi-Modal Deep Convolutional Neural Network. In Proceedings of the 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 20–23 June 2023; pp. 1–5.
6. Tao, S.; Goh, W.L.; Gao, Y. A Convolved Self-Attention Model for IMU-based Gait Detection and Human Activity Recognition. In Proceedings of the 2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS), Hangzhou, China, 11–13 June 2023; pp. 1–5.
7. Hassler, A.P.; Menasalvas, E.; García-García, F.J.; Rodríguez-Mañas, L.; Holzinger, A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 33. [CrossRef] [PubMed]
8. Xu, S.; Zhang, L.; Huang, W.; Wu, H.; Song, A. Deformable convolutional networks for multimodal human activity recognition using wearable sensors. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2505414. [CrossRef]
9. Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: A survey. *Multimed. Tools Appl.* **2020**, *79*, 30509–30555. [CrossRef]
10. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutorials* **2012**, *15*, 1192–1209. [CrossRef]
11. Ke, S.R.; Thuc, H.L.U.; Lee, Y.J.; Hwang, J.N.; Yoo, J.H.; Choi, K.H. A review on video-based human activity recognition. *Computers* **2013**, *2*, 88–131. [CrossRef]
12. Ray, A.; Kolekar, M.H.; Balasubramanian, R.; Hafiane, A. Transfer learning enhanced vision-based human activity recognition: A decade-long analysis. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100142. [CrossRef]
13. Singh, R.; Kushwaha, A.K.S.; Srivastava, R. Recent trends in human activity recognition–A comparative study. *Cogn. Syst. Res.* **2023**, *77*, 30–44. [CrossRef]
14. Gu, F.; Chung, M.H.; Chignell, M.; Valaee, S.; Zhou, B.; Liu, X. A survey on deep learning for human activity recognition. *Acm Comput. Surv. (CSUR)* **2021**, *54*, 1–34. [CrossRef]
15. Hussain, Z.; Sheng, M.; Zhang, W.E. Different approaches for human activity recognition: A survey. *arXiv* **2019**, arXiv:1906.05074.
16. Jobanputra, C.; Bavishi, J.; Doshi, N. Human activity recognition: A survey. *Procedia Comput. Sci.* **2019**, *155*, 698–703. [CrossRef]
17. Ramasamy Ramamurthy, S.; Roy, N. Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1254. [CrossRef]
18. Dang, L.M.; Min, K.; Wang, H.; Piran, M.J.; Lee, C.H.; Moon, H. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognit.* **2020**, *108*, 107561. [CrossRef]
19. Vrigkas, M.; Nikou, C.; Kakadiaris, I.A. A review of human activity recognition methods. *Front. Robot. AI* **2015**, *2*, 28. [CrossRef]
20. Saleem, G.; Bajwa, U.I.; Raza, R.H. Toward human activity recognition: A survey. *Neural Comput. Appl.* **2023**, *35*, 4145–4182. [CrossRef]

21. Morshed, M.G.; Sultana, T.; Alam, A.; Lee, Y.K. Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities. *Sensors* **2023**, *23*, 2182. [CrossRef]
22. Hinton, G.E.; Roweis, S. Stochastic neighbor embedding. *Adv. Neural Inf. Process. Syst.* **2002**, *15*.
23. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
24. Seyfioğlu, M.S.; Özbayoğlu, A.M.; Gürbüz, S.Z. Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. *IEEE Trans. Aerosp. Electron. Syst.* **2018**, *54*, 1709–1723. [CrossRef]
25. Ignatov , A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Appl. Soft Comput.* **2018**, *62*, 915–922. [CrossRef]
26. Hegde, N.; Bries, M.; Swibas, T.; Melanson, E.; Sazonov, E. Automatic recognition of activities of daily living utilizing insole-based and wrist-worn wearable sensors. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 979–988. [CrossRef] [PubMed]
27. Wang, W.; Liu, A.X.; Shahzad, M.; Ling, K.; Lu, S. Device-free human activity recognition using commercial WiFi devices. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1118–1131. [CrossRef]
28. Ruan, W.; Sheng, Q.Z.; Yao, L.; Li, X.; Falkner, N.J.; Yang, L. Device-free human localization and tracking with UHF passive RFID tags: A data-driven approach. *J. Netw. Comput. Appl.* **2018**, *104*, 78–96. [CrossRef]
29. Rol, L.; Lidauer, L.; Sattlecker, G.; Kickinger, F.; Auer, W.; Sturm, V.; Efrosinin, D.; Drillich, M.; Iwersen, M. Monitoring drinking behavior in bucket-fed dairy calves using an ear-attached tri-axial accelerometer: A pilot study. *Comput. Electron. Agric.* **2018**, *145*, 298–301.
30. Alsinglawi, B.; Nguyen, Q.V.; Gunawardana, U.; Maeder, A.; Simoff, S.J. RFID systems in healthcare settings and activity of daily living in smart homes: A review. E-Health Telecommun. *Syst. Netw.* **2017**, *6*, 1–17. [CrossRef]
31. Fan, X.; Wang, F.; Wang, F.; Gong, W.; Liu, J. When RFID meets deep learning: Exploring cognitive intelligence for activity identification. *IEEE Wirel. Commun.* **2019**, *26*, 19–25. [CrossRef]
32. Qi, J.; Yang, P.; Waraich, A.; Deng, Z.; Zhao, Y.; Yang, Y. Examining sensor-based physical activity recognition and monitoring for healthcare using inter- net of things: A systematic review. *J. Biomed. Inform.* **2018**, *87*, 138–153. [CrossRef] [PubMed]
33. Hao, J.; Bouzouane, A.; Gaboury, S. Recognizing multi-resident activities in non-intrusive sensor-based smart homes by formal concept analysis. *Neuro-Computing* **2018**, *318*, 75–89. [CrossRef]
34. Roy, N.; Misra, A.; Cook, D. Ambient and smartphone sensor assisted ADL recognition in multi-inhabitant smart environments. *J. Ambient Intell. Humaniz. Comput.* **2016**, *7*, 1–19. [CrossRef]
35. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [CrossRef]
36. Oyedotun, O.K.; Khashman, A. Deep learning in vision-based static hand gesture recognition. *Neural Comput. Appl.* **2017**, *28*, 3941–3951. [CrossRef]
37. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
38. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **2017**, *156*, 117–127. [CrossRef]
39. Zerrouki, N.; Harrou, F.; Sun, Y.; Houacine, A. Vision-based human action classification using adaptive boosting algorithm. *IEEE Sens. J.* **2018**, *18*, 5115–5121. [CrossRef]
40. Chatzitofis, A.; Saroglou, L.; Boutis, P.; Drakoulis, P.; Zioulis, N.; Subramanyam, S.; Kevelham, B.; Charbonnier, C.; Cesar, P.; Zarpalas, D.; et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access* **2020**, *8*, 176241–176262. [CrossRef]
41. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
42. Mahmood, N.; Ghorbani, N.; Troje, N.F.; Pons-Moll, G.; Black, M.J. AMASS: Archive of motion capture as surface shapes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5442–5451.
43. Zeng, M.; Nguyen, L.T.; Yu, B.; Mengshoel, O.J.; Zhu, J.; Wu, P.; Zhang, J. Convolutional neural networks for human activity recognition using mobile sensors. In Proceedings of the 6th International Conference on Mobile Computing, Applications and Services, Austin, TX, USA, 6–7 November 2014; pp. 197–205.
44. Yang, J.; Nguyen, M.N.; San, P.P.; Li, X.; Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In Proceedings of the Ijcai, Buenos Aires, Argentina, 25–31 July 2015; Volume 15, pp. 3995–4001.
45. Ha, S.; Yun, J.M.; Choi, S. Multi-modal convolutional neural networks for activity recognition. In Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon Tong, Hong Kong, China, 9–12 October 2015; pp. 3017–3022.
46. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
47. Ha, S.; Choi, S. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 381–388.

48. Edel, M.; Köppe, E. Binarized-blstm-rnn based human activity recognition. In Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Alcala de Henares, Spain, 4–7 October 2016; pp. 1–7.
49. Guan, Y.; Plötz, T. Ensembles of deep lstm learners for activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 1–28. [CrossRef]
50. Hammerla, N.Y.; Halloran, S.; Plötz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv* **2016**, arXiv:1604.08880.
51. Inoue, M.; Inoue, S.; Nishida, T. Deep recurrent neural network for mobile human activity recognition with high throughput. *Artif. Life Robot.* **2018**, *23*, 173–185. [CrossRef]
52. Maurya, R.; Teo, T.H.; Chua, S.H.; Chow, H.C.; Wey, I.C. Complex Human Activities Recognition Based on High Performance 1D CNN Model. In Proceedings of the 2022 IEEE 15th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoC), Penang, Malaysia, 19–22 December 2022; pp. 330–336.
53. Liang, Y.; Feng, K.; Ren, Z. Human Activity Recognition Based on Transformer via Smart-phone Sensors. In Proceedings of the 2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI), Taiyuan, China, 26–28 May 2023; pp. 267–271.
54. Aswal, V.; Sreeram, V.; Kuchik, A.; Ahuja, S.; Patel, H. Real-time human activity generation using bidirectional long short term memory networks. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; pp. 775–780.
55. Choudhury, N.A.; Moulik, S.; Roy, D.S. Physique-based human activity recognition using ensemble learning and smartphone sensors. *IEEE Sens. J.* **2021**, *21*, 16852–16860. [CrossRef]
56. Thakur, D.; Biswas, S.; Ho, E.S.; Chattopadhyay, S. Convae-lstm: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition. *IEEE Access* **2022**, *10*, 4137–4156. [CrossRef]
57. Dong, Y.; Zhou, R.; Zhu, C.; Cao, L.; Li, X. Hierarchical activity recognition based on belief functions theory in body sensor networks. *IEEE Sens. J.* **2022**, *22*, 15211–15221. [CrossRef]
58. Teng, Q.; Wang, K.; Zhang, L.; He, J. The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. *IEEE Sens. J.* **2020**, *20*, 7265–7274. [CrossRef]
59. Zilelioglu, H.; Khodabandelou, G.; Chibani, A.; Amirat, Y. Semi-Supervised Generative Adversarial Networks with Temporal Convolutions for Human Activity Recognition. *IEEE Sens. J.* **2023**, *23*, 12355–12369. [CrossRef]
60. Mekruksavanich, S.; Jantawong, P.; Hnoohom, N.; Jitpattanakul, A. A novel deep bigru-resnet model for human activity recognition using smartphone sensors. In Proceedings of the 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), Bangkok, Thailand, 22–25 June 2022; pp. 1–5.
61. Dubey, A.; Lyons, N.; Santra, A.; Pandey, A. XAI-BayesHAR: A novel Framework for Human Activity Recognition with Integrated Uncertainty and Shapely Values. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 12–14 December 2022; pp. 1281–1288.
62. Han, C.; Zhang, L.; Xu, S.; Wang, X.; Wu, H.; Song, A. An Efficient Diverse-branch Convolution Scheme for Sensor-Based Human Activity Recognition. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2509313. [CrossRef]
63. Stolovas, I.; Suárez, S.; Pereyra, D.; De Izaguirre, F.; Cabrera, V. Human activity recognition using machine learning techniques in a low-resource embedded system. In Proceedings of the 2021 IEEE URUCON, Montevideo, Uruguay, 24–26 November 2021; pp. 263–267.
64. Khatun, M.A.; Yousuf, M.A.; Moni, M.A. Deep CNN-GRU Based Human Activity Recognition with Automatic Feature Extraction Using Smartphone and Wearable Sensors. In Proceedings of the 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Kolkata, India, 20–21 January 2023; pp. 1–6.
65. De Vita, A.; Russo, A.; Pau, D.; Di Benedetto, L.; Rubino, A.; Licciardo, G.D. A partially binarized hybrid neural network system for low-power and resource constrained human activity recognition. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2020**, *67*, 3893–3904. [CrossRef]
66. Tang, Y.; Zhang, L.; Min, F.; He, J. Multiscale deep feature learning for human activity recognition using wearable sensors. *IEEE Trans. Ind. Electron.* **2022**, *70*, 2106–2116. [CrossRef]
67. Rustam, F.; Reshi, A.A.; Ashraf, I.; Mehmood, A.; Ullah, S.; Khan, D.M.; Choi, G.S. Sensor-based human activity recognition using deep stacked multilayered perceptron model. *IEEE Access* **2020**, *8*, 218898–218910. [CrossRef]
68. Wang, Z.; Chen, S.; Yang, W.; Xu, Y. Environment-independent wi-fi human activity recognition with adversarial network. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3330–3334.
69. Hsieh, C.F.; Chen, Y.C.; Hsieh, C.Y.; Ku, M.L. Device-free indoor human activity recognition using Wi-Fi RSSI: Machine learning approaches. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2.
70. Salehinejad, H.; Hasanzadeh, N.; Djogo, R.; Valaee, S. Joint Human Orientation-Activity Recognition Using WIFI Signals for Human-Machine Interaction. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
71. Zhang, J.; Wu, F.; Wei, B.; Zhang, Q.; Huang, H.; Shah, S.W.; Cheng, J. Data augmentation and dense-LSTM for human activity recognition using WiFi signal. *IEEE Internet Things J.* **2020**, *8*, 4628–4641. [CrossRef]

72. Ding, X.; Jiang, T.; Li, Y.; Xue, W.; Zhong, Y. Device-free location-independent human activity recognition using transfer learning based on CNN. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 7–11 June 2020; pp. 1–6.
73. Khan, D.; Ho, I.W.H. Deep learning of CSI for efficient device-free human activity recognition. In Proceedings of the 2021 IEEE 7th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 14 June–31 July 2021; pp. 19–24.
74. Zeeshan, M.; Pandey, A.; Kumar, S. CSI-based device-free joint activity recognition and localization using Siamese networks. In Proceedings of the 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 4–8 January 2022; pp. 260–264.
75. Xiang, F.; Nie, X.; Cui, C.; Nie, W.; Dong, X. Radar-based human activity recognition using two-dimensional feature extraction. In Proceedings of the 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 6–8 January 2023; pp. 267–271.
76. Guo, Z.; Guendel, R.G.; Yarovoy, A.; Fioranelli, F. Point Transformer-Based Human Activity Recognition Using High-Dimensional Radar Point Clouds. In Proceedings of the 2023 IEEE Radar Conference (RadarConf23), San Antonio, TX, USA, 1–5 May 2023; pp. 1–6.
77. Werthen-Brabants, L.; Bhavanasi, G.; Couckuyt, I.; Dhaene, T.; Deschrijver, D. Quantifying uncertainty in real time with split BiRNN for radar human activity recognition. In Proceedings of the 2022 19th European Radar Conference (EuRAD), Milan, Italy, 28–30 September 2022; pp. 173–176.
78. McQuire, J.; Watson, P.; Wright, N.; Hiden, H.; Catt, M. A Data Efficient Vision Transformer for Robust Human Activity Recognition from the Spectrograms of Wearable Sensor Data. In Proceedings of the 2023 IEEE Statistical Signal Processing Workshop (SSP), Hanoi, Vietnam, 2–5 July 2023; pp. 364–368. [CrossRef]
79. Luo, Y.; Coppola, S.M.; Dixon, P.C.; Li, S.; Dennerlein, J.T.; Hu, B. A database of human gait performance on irregular and uneven surfaces collected by wearable sensors. *Sci. Data* **2020**, *7*, 219. [CrossRef] [PubMed]
80. Reiss, A.; Stricker, D. Creating and benchmarking a new dataset for physical activity monitoring. In Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, Heraklion, Crete, Greece, 6–8 June 2012; pp. 1–8.
81. Reiss , A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In Proceedings of the 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 18–22 June 2012; pp. 108–109.
82. Qin, W.; Wu, H.N. Switching GMM-HMM for Complex Human Activity Modeling and Recognition. In Proceedings of the 2022 China Automation Congress (CAC), Xiamen, China, 25–27 November 2022; pp. 696–701.
83. Bhuiyan, R.A.; Amiruzzaman, M.; Ahmed, N.; Islam, M.R. Efficient frequency domain feature extraction model using EPS and LDA for human activity recognition. In Proceedings of the 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), Kaohsiung, Taiwan, 21–23 August 2020; pp. 344–347.
84. Zhou, Y.; Yang, Z.; Zhang, X.; Wang, Y. A hybrid attention-based deep neural network for simultaneous multi-sensor pruning and human activity recognition. *IEEE Internet Things J.* **2022**, *9*, 25363–25372. [CrossRef]
85. Li, W.; Feng, X.; He, Z.; Zheng, H. Human activity recognition based on data fusion of fmcw radar and image. In Proceedings of the 2021 7th International Conference on Computer and Communications (ICCC), Tianjin, China, 10–13 December 2021; pp. 943–947.
86. Yen, C.T.; Liao, J.X.; Huang, Y.K. Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms. *IEEE Access* **2020**, *8*, 174105–174114. [CrossRef]
87. Chowdhury, A.I.; Ashraf, M.; Islam, A.; Ahmed, E.; Jaman, M.S.; Rahman, M.M. hActNET: An improved neural network based method in recognizing human activities. In Proceedings of the 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 22–24 October 2020; pp. 1–6.
88. Psychoula, I.; Singh, D.; Chen, L.; Chen, F.; Holzinger, A.; Ning, H. Users' privacy concerns in IoT based applications. In Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 8–12 October 2018; pp. 1887–1894.
89. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advancements in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
90. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
91. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
92. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
93. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
94. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.

95. Keren, G.; Schuller, B. Convolutional RNN: An enhanced model for extracting features from sequential data. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 3412–3419.
96. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
97. Gundu, S.; Syed, H. Vision-Based HAR in UAV Videos Using Histograms and Deep Learning Techniques. *Sensors* **2023**, *23*, 2569. [CrossRef] [PubMed]
98. Islam, M.M.; Iqbal, T. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 10285–10292.
99. Gupta, D.; Singh, A.K.; Gupta, N.; Vishwakarma, D.K. SDL-Net: A Combined CNN & RNN Human Activity Recognition Model. In Proceedings of the 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), Bhubaneswar, India, 9–11 June 2023; pp. 1–5.
100. Popescu, A.C.; Mocanu, I.; Cramariuc, B. Fusion mechanisms for human activity recognition using automated machine learning. *IEEE Access* **2020**, *8*, 143996–144014. [CrossRef]
101. Kumar, K.V.; Harikiran, J.; Chandana, B.S. Human Activity Recognition with Privacy Preserving using Deep Learning Algorithms. In Proceedings of the 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP), Vijayawada, India, 12–14 February 2022; pp. 1–8.
102. Bukht, T.F.N.; Rahman, H.; Jalal, A. A Novel Framework for Human Action Recognition Based on Features Fusion and Decision Tree. In Proceedings of the 2023 4th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 20–22 February 2023; pp. 1–6.
103. Mutegeki, R.; Han, D.S. A CNN-LSTM approach to human activity recognition. In Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 19–21 February 2020; pp. 362–366.
104. Razmah, M.; Prabha, R.; Divya, B.; Sridevi, S.; Naveen, A. LSTM Method for Human Activity Recognition of Video Using PSO Algorithm. In Proceedings of the 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 8–9 December 2022; pp. 1–6.
105. Alrashdi, I.; Siddiqi, M.H.; Alhwaiti, Y.; Alruwaili, M.; Azad, M. Maximum entropy Markov model for human activity recognition using depth camera. *IEEE Access* **2021**, *9*, 160635–160645. [CrossRef]
106. Ahad, M.A.R.; Antar, A.D.; Ahmed, M.; Ahad, M.A.R.; Antar, A.D.; Ahmed, M. Sensor-based benchmark datasets: Comparison and analysis. In *IoT Sensor-Based Activity Recognition: Human Activity Recognition*; Springer: Cham, Switzerland, 2021; pp. 95–121.
107. Blunck, H.; Bhattacharya, S.; Stisen, A.; Prentow, T.S.; Kjærgaard, M.B.; Dey, A.; Jensen, M.M.; Sonne, T. Activity recognition on smart devices: Dealing with diversity in the wild. *Getmobile Mob. Comput. Commun.* **2016**, *20*, 34–38. [CrossRef]
108. Torres, R.L.S.; Ranasinghe, D.C.; Shi, Q.; Sample, A.P. Sensor enabled wearable RFID technology for mitigating the risk of falls near beds. In Proceedings of the 2013 IEEE International Conference on RFID (RFID), Orlando, FL, USA, 30 April–2 May 2013; pp. 191–198.
109. Palumbo, F.; Gallicchio, C.; Pucci, R.; Micheli, A. Human activity recognition using multisensor data fusion based on reservoir computing. *J. Ambient. Intell. Smart Environ.* **2016**, *8*, 87–107. [CrossRef]
110. Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A public domain dataset for human activity recognition using smartphones. In Proceedings of the Esann, Bruges, Belgium, 24–26 April 2013; Volume 3, p. 3.
111. Reyes-Ortiz, J.L.; Oneto, L.; Samà, A.; Parra, X.; Anguita, D. Transition-aware human activity recognition using smartphones. *Neurocomputing* **2016**, *171*, 754–767. [CrossRef]
112. Casale, P.; Pujol, O.; Radeva, P. Personalization and user verification in wearable systems using biometric walking patterns. *Pers. Ubiquitous Comput.* **2012**, *16*, 563–580. [CrossRef]
113. Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S.T.; Tröster, G.; Millán, J.D.R.; Roggen, D. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognit. Lett.* **2013**, *34*, 2033–2042. [CrossRef]
114. Ordónez, F.J.; De Toledo, P.; Sanchis, A. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors* **2013**, *13*, 5460–5477. [CrossRef] [PubMed]
115. Baños, O.; Damas, M.; Pomares, H.; Rojas, I.; Tóth, M.A.; Amft, O. A benchmark dataset to evaluate sensor displacement in activity recognition. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 1026–1035.
116. Altun, K.; Barshan, B.; Tunçel, O. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognit.* **2010**, *43*, 3605–3620. [CrossRef]
117. Bacciu, D.; Barsocchi, P.; Chessa, S.; Gallicchio, C.; Micheli, A. An experimental characterization of reservoir computing in ambient assisted living applications. *Neural Comput. Appl.* **2014**, *24*, 1451–1464. [CrossRef]
118. Banos, O.; Garcia, R.; Holgado-Terriza, J.A.; Damas, M.; Pomares, H.; Rojas, I.; Saez, A.; Villalonga, C. mHealthDroid: A novel framework for agile development of mobile health applications. In Proceedings of the Ambient Assisted Living and Daily Activities: 6th International Work-Conference, IWAAL 2014, Belfast, UK, 2–5 December 2014; Proceedings 6; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 91–98.
119. Weiss, G.M.; Yoneda, K.; Hayajneh, T. Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access* **2019**, *7*, 133190–133202. [CrossRef]

120. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM international conference on multimodal interaction, Boulder, CO, USA, 16–20 October 2018; pp. 400–408.
121. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, 26 August 2004; Volume 3, pp. 32–36.
122. Ballan, L.; Bertini, M.; Del Bimbo, A.; Seidenari, L.; Serra, G. Effective codebooks for human action categorization. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 27 September –4 October 2009; pp. 506–513.
123. Li, W.; Wong, Y.; Liu, A.A.; Li, Y.; Su, Y.T.; Kankanhalli, M. Multi-camera action dataset (MCAD): A dataset for studying non-overlapped cross-camera action recognition. *arXiv* **2016**, arXiv:1607.06408.
124. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1290–1297.
125. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [CrossRef]
126. Kuehne , H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
127. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2929–2936.
128. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef] [PubMed]
129. Yao, B.; Jiang, X.; Khosla, A.; Lin, A.L.; Guibas, L.; Fei-Fei, L. Human action recognition by learning bases of action attributes and parts. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1331–1338.
130. Weinl, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257.
131. Stein, S.; McKenna, S.J. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 8–12 September 2013; pp. 729–738.
132. Nghiem, A.T.; Bremond, F.; Thonnat, M.; Valentin, V. ETISEO, performance evaluation for video surveillance systems. In Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, London, UK, 5–7 September 2007; pp. 476–481.
133. Niebles, J.C.; Chen, C.W.; Fei-Fei, L. Modeling temporal structure of decomposable motion segments for activity classification. In Proceedings of the Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Proceedings, Part II 11; Springer: Berlin/Heidelberg, Germany, 2010; pp. 392–405.
134. Ryoo, M.S.; Aggarwal, J.K. UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). In Proceedings of the IEEE International Conference on Pattern Recognition Workshops, Istanbul, Turkey, 23–26 August 2010; Volume 2, p. 4.
135. Chen, C.-C.; Aggarwal, J.K. Recognizing human action from a far field of view. In Proceedings of the 2009 Workshop on Motion and Video Computing (WMVC), Snowbird, UT, USA, 8–9 December 2009; pp. 1–7.
136. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
137. Kay , W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
138. Soomro, K.; Zamir, A.R.; Shah, M. A dataset of 101 human action classes from videos in the wild. *Cent. Res. Comput. Vis.* **2012**, *2*, 1–7.
139. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos "in the wild". In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1996–2003.
140. Berenson, D.; Abbeel, P.; Goldberg, K. A robot path planning framework that learns from experience. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, St. Paul, MN, USA, 14–18 May 2012; pp. 3671–3678.
141. Martinez, J.; Black, M.J.; Romero, J. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2891–2900.
142. Wang, H.; Dong, J.; Cheng, B.; Feng, J. PVRED: A position-velocity recurrent encoder-decoder for human motion prediction. *IEEE Trans. Image Process.* **2021**, *30*, 6096–6106. [CrossRef]
143. Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.Z.; Vo M.; Malik, J. Long-term human motion prediction with scene context. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 387–404.
144. Aksan, E.; Kaufmann, M.; Cao, P.; Hilliges, O. A spatio-temporal transformer for 3d human motion prediction. In Proceedings of the 2021 International Conference on 3D Vision (3DV), Virtual Conference, 1–3 December 2021; pp. 565–574.

145. Medjaouri, O.; Desai, K. Hr-stan: High-resolution spatio-temporal attention network for 3d human motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LO, USA, 18–24 June 2022; pp. 2540–2549.
146. Tanberk, S.; Tükel, D.B.; Uysal, M. A Simple AI-Powered Video Analytics Framework for Human Motion Imitation. In Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–5.