sensors

MDPI

Perspective

RGB-D Cameras and Brain–Computer Interfaces for Human Activity Recognition: An Overview

Grazia Iadarola [1,*], Alessandro Mengarelli [1], Sabrina Iarlori [2], Andrea Monteriù [1] and Susanna Spinsante [1]

[1] Department of Information Engineering, Polytechnic University of Marche, 60131 Ancona, Italy; a.mengarelli@staff.univpm.it (A.M.); a.monteriu@staff.univpm.it (A.M.); s.spinsante@staff.univpm.it (S.S.)
[2] Department of Theoretical and Applied Sciences (DiSTA), Università degli Studi eCampus, 22060 Novedrate, Italy; sabrina.iarlori@uniecampus.it
* Correspondence: g.iadarola@staff.univpm.it

Abstract

This paper provides a perspective on the use of RGB-D cameras and non-invasive brain–computer interfaces (BCIs) for human activity recognition (HAR). Then, it explores the potential of integrating both the technologies for active and assisted living. RGB-D cameras can offer monitoring of users in their living environments, preserving their privacy in human activity recognition through depth images and skeleton tracking. Concurrently, non-invasive BCIs can provide access to intent and control of users by decoding neural signals. The synergy between these technologies may allow holistic understanding of both physical context and cognitive state of users, to enhance personalized assistance inside smart homes. The successful deployment in integrating the two technologies needs addressing critical technical hurdles, including computational demands for real-time multi-modal data processing, and user acceptance challenges related to data privacy, security, and BCI illiteracy. Continued interdisciplinary research is essential to realize the full potential of RGB-D cameras and BCIs as AAL solutions, in order to improve the quality of life for independent or impaired people.

Keywords: human activity recognition; assisted living; RGB-D cameras; brain–computer interfaces; wearable devices

Academic Editor: Sylvain Girard

Received: 18 July 2025
Revised: 26 September 2025
Accepted: 5 October 2025
Published: 10 October 2025

Citation: Iadarola, G.; Mengarelli, A.; Iarlori, S.; Monteriù, A.; Spinsante, S. RGB-D Cameras and Brain–Computer Interfaces for Human Activity Recognition: An Overview. *Sensors* **2025**, *25*, 6286. https://doi.org/10.3390/s25206286

1. Introduction

Born from the joint use of sensors, communication protocols and artificial intelligence techniques, active and assisted living (AAL) plays a fundamental role in supporting independent and disabled people. AAL is aimed at increasing the autonomy of persons at home, by improving their quality of life or caregiving activities for elderly and impaired people. In such a way, AAL solutions can be beneficial to healthcare in general, by concurring in saving resources through the cost reduction related to institutionalized treatments, and also by easing prevention strategies for chronic diseases. Indeed, pathologies affecting musculoskeletal and neurological systems primarily impact elderly independent life and require resource-intensive care [1]. The potential benefits carried out by AAL systems appear nowadays essential considering the perspectives of developed countries from a demographic point of view. For example, in Europe, the population with 65 years or older will expand significantly, rising from 90.5 million at the start of 2019 to 129.8 million by 2050. Specifically, the cohort aged 75–84 years is projected to increase by 56.1%, while the cohort aged 65–74 years is projected to increase by 16.6% [2]. Furthermore, the social

change due to population aging has entailed the demand for assistance to elderly with disabilities to cope with a reduction in social activities and consequent frustration, sadness and depression [3,4].

The need for innovative solutions to AAL challenges to help people pushes towards the development of smart homes that can provide valuable support to maintain an active lifestyle. With the development of more affordable and sophisticated technologies, human activity recognition (HAR) continues to raise great interest among the research community. HAR systems exploit sensors, user-generated datasets and learning algorithms to identify and discriminate everyday activities recurring in people [5–8], especially elderly or impaired [9,10]. Within this context, AAL technologies found their application in easing the independence of impaired people through the support of daily living activities [11]. Furthermore, they should allow relatives and caregivers to carry out continuous monitoring on the daily habits and to receive prompt warnings in case of falls or possibly risky situations [12]. AAL solutions should also be able to monitor the emotional state and the cognitive deterioration of the person [13]. An additional important aspect is the decoding of the user intention so that the interaction with assistive technologies would be safer and more effective [14]. Finally, the connection between people at home and the health and social services is also of paramount importance [15].

HAR is the object of many research efforts [16], being fundamental for the development of smart home environments [12,17]. In the HAR field, the approach employing vision-based devices is the most common, involving technologies external to body location, such as RGB, video, depth, or thermal cameras. The other possible approach is based on wireless wearable devices, such as sensors for acquisition of bioelectric signals, inertial sensors—inertial measurement units (IMU) and magnetic, angular rate and gravity (MARG) sensors—or brain–computer interfaces (BCI). Both approaches enable non-invasive monitoring of elderly or disabled people, at home or in care facilities [18]. Specifically, in the last fifteen years, among vision-based technologies, the most adopted is represented by RGB-D cameras, while, among wireless wearable technologies, the use of non-invasive BCIs has seen a rapid growth [19]. By collecting neural signals from the user, which can be converted into commands, BCIs enable new approaches to human–device interaction in smart living environments [20]. Their usage can be further advanced by the fusion with information on the performed activity provided by environmental sensors, such as RGB-D cameras.

Generally speaking, given the complexity of human activities, a uni-modal (single) sensing solution may not be enough for their recognition based on machine learning (ML), while multi-modal sensor fusion, enhancing the available features [21], emerges to face critical gaps in AAL technology [22]. Indeed, with multi-modal sensor fusion, combining various sensor types can overcome individual shortcomings, leading to more robust and accurate monitoring and assistance. AAL applications could be designed through RGB-D cameras and non-invasive BCI integration, specifically to improve the environmental perception of actions performed by users.

Differently from existing literature that highlights advantages and limitations of technologies for uni-modal sensing solutions [4,6,10,17,23–29], this overview is intended to propose a perspective on RGB-D cameras and non-invasive BCIs for HAR, exploring the rationale for the integration of the two technologies and opening a discussion on possible applications that could benefit from such integration. The rest of the paper is organized as follows. Section 2 offers an introduction to current approaches to HAR, while Sections 3 and 4 provide an overview on RGB-D cameras and non-invasive BCIs, respectively, also focusing on specific techniques based on the use of these two sensing

technologies. Potential innovations, but also barriers, stemming from integrated solutions are presented in Section 5. Finally, Section 6 draws the main conclusions.

2. Background: Approaches for HAR

Figure 1 reports the increasing trend of papers related to HAR in the time range from 2010 to 2024. Among human activities of interest, monitoring of body movements (such as walking or running) represents one of the most investigated aspects for the evaluation of effort and muscle fatigue [8], but applications can include a wider spectrum of physical tasks [25] and activities of daily living (ADLs), namely, hand gestures, sleeping, standing or falling.

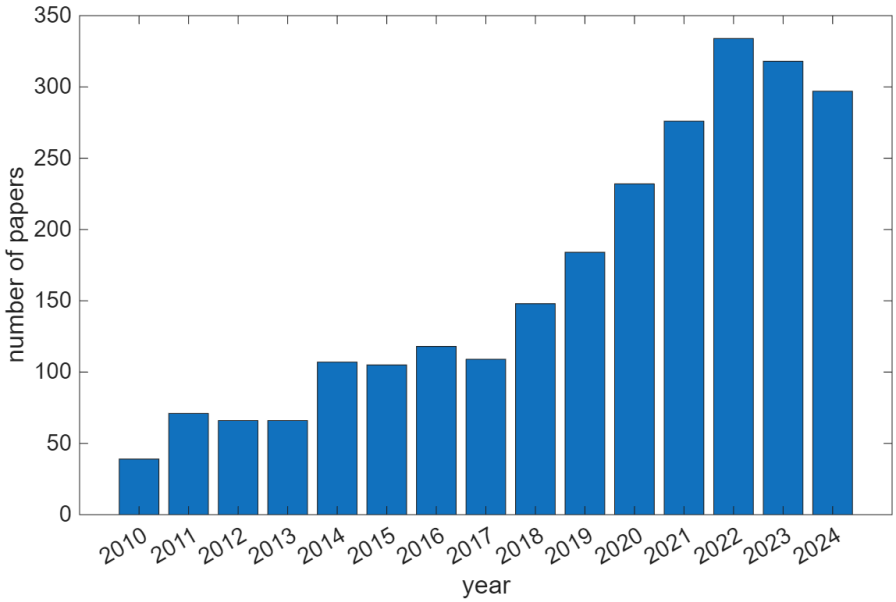

Figure 1. Number of papers related to HAR since 2010. Source: PubMed, June 2025, search query in title/abstract: (activity recognition) OR (HAR).

Although several human behaviors and activities have been investigated in the field of smart sensing, such as rehabilitation and even human emotional condition by affective computing [24], the outbreak of the COVID-19 pandemic has contributed to stress on two different aspects related to HAR. On the one hand, it put attention on the importance of correctly identifying a series of activities related to human well-being and health, such as those involved with personal hygiene [30]. On the other hand, the pandemic highlighted the value of technical solutions capable of remotely monitoring user activities, habits, and behaviors within a domestic scenario.

As mentioned in the Introduction, two common approaches can be leveraged for HAR within a closed environment, i.e., vision-based devices, such as cameras, and wearable devices. Although vision-based devices represent one of the most common approaches in research, they can mainly lead to two potential drawbacks that hamper their widespread diffusion for HAR purposes. The former is the need for instrumented homes with video-capture sensors to install within each room where the user is supposed to perform the activities of interest. The latter is that vision-based devices record images not only of the living environment but also of subjects, generating several issues regarding privacy and acceptability. Instead, wearable devices can provide information related to the performed activities not directly linked to a specific user and can be processed with a small computational delay, thus potentially allowing also an almost real-time recognition of specific activities performed by the subject. Anyway, also wearable devices, despite their mobility benefits, necessitate consistent user compliance, which can be a significant barrier

to long-term monitoring. A review specifically focused on the use of wearable devices for AAL has been proposed in [10]. Table 1 lists both vision-based devices and wearable devices for the main ADLs.

Table 1. Common devices for main ADLs.

Activity	Vision-Based Devices	Wearable Devices
Body movements	RGB-D cameras, video cameras, thermal cameras	IMUs, electromyography sensors, photoplethysmogram sensors, skin conductance sensors
Hand gestures (including eating and drinking)	RGB-D cameras, video cameras, thermal cameras	IMUs, electromyography sensors
Sleeping	RGB-D cameras	photoplethysmogram sensors, skin conductance sensors
Standing/falling	RGB-D cameras, video cameras	IMUs, electromyography sensors
Sitting with mental occupation (including studying and working)	RGB-D cameras, video cameras	BCIs, photoplethysmogram sensors, skin conductance sensors

When designing a measurement system that leverages both vision-based and wearable devices for HAR, the aim should be to develop a solution able to correctly identify ADLs with a specific and clear impact on subject well-being and health. Specifically, two main points merit attention. The first point regards the human activities to be recognized by the system and those activities that should be included in the whole set. As outlined above, a wide spectrum of possibilities is available in this case, ranging from activities involving full body movements, such as walking and running, to activities that mainly rely on upper limb motion, including interaction with objects [31]. In fact, the main activities to take into account should be those commonly performed during daily living with upper limbs, such as eating gestures. The rationale for this choice is based on their importance for many aspects related to health and well-being. For instance, a correct identification of the drinking gesture can be useful for monitoring fluid intake and hydration levels [32], and the recognition of bringing the hand to the mouth with a certain grasping shape can be linked to pill intake, useful for medical adherence assessment [33]. Thus, specific attention should be given to gestures, including also other daily living activities that can have similar execution mechanics, in order to test the robustness of the recognition methodologies to potentially confounding patterns. However, it should be noted that the inclusion of additional activities that involve whole-body movement can be considered in order to enlarge the spectrum of detectable activities through the specific sensor configuration chosen for this application. The second point is instead related to the reproducibility of the results [34]. Reproducibility is, in fact, a relevant property that should be observed, especially with calibrated instruments, but is typically overlooked when dealing with sensor data. Finally, users highly value their privacy and hence require a high degree of protection for their personal data, but they also accept to provide complete data transparency in case of perceived risks to their health and potential countermeasures offered by the monitoring technology [35]. Likewise, users exhibit willingness to pay for the implementation of these technologies in living environments if perceived beneficial to improve safety and health.

Originally focused on simple activity detection, HAR has evolved to more challenging tasks of identifying complex and multiple simultaneous activities, a change driven by technological advancements [36]. In this context, multi-modal sensor fusion, which involves the intelligent combination of data from different sensor types, becomes crucial to attain a more comprehensive, accurate, and robust understanding of both performed activities

and the environment, and even user intention. The latter point can nowadays be addressed by wearable BCIs. They provide direct access to user intent, enabling control and communication through neural signals, but they offer limited or no awareness of the external environment. In the following sections, the analysis of the current state of adoption of each sensing modality, i.e., RGB-D cameras and BCIs, is presented to lay the foundations for the proposal of their integration, discussing opportunities and open challenges to address.

3. HAR Based on RGB-D Cameras

Vision-based HAR still represents one of the major challenges in the field of computer vision [37], and RGB-D cameras remain a common solution to implement AAL systems and applications. Indeed, a large portion of HAR systems, especially for fall detection, is represented by depth cameras [23], such as the commercially available Orbbec [38] (Orbbec, Troy, MI, USA), Intel RealSense [39] (Intel Corp., Santa Clara, CA, USA), and Microsoft Kinect [40,41] (Microsoft Corp., Redmond, WA, USA). RGB-D cameras can actually gather information in a number of modalities and are typically devised to capture vision-based information under infrared (IR) or near-infrared (nIR) channels, skeletal and RGB. However, depth frames provided by RGB-D cameras are more commonly considered suitable for handling changes in room illumination and protecting the privacy of the users.

Table 2 reports advantages and disadvantages of RGB-D cameras. They may operate based on different techniques, i.e., stereo vision, IR speckle pattern projection and Time of Flight (ToF) at IR or nIR. In stereo vision, the depth map of the scene is reconstructed from a couple of RGB images acquired from two different directions by extracting distance information. In IR speckle pattern projection, an IR or nIR dot pattern is projected by the sensor onto the scene, and then the dots are detected by the IR camera. The depth information is extracted by analyzing the changes between the projected and detected dot patterns. In Time of Flight, the object-to-sensor distance is derived by measuring the time taken by a light pulse (emitted by an IR/nIR laser or an LED) to travel from a source onboard the sensor to the object and back to the sensor receiver. The depth-sensing technique implemented influences not only the cost of the corresponding camera but also the performance in terms of range of operation, depth frame resolution, frame rate, and accuracy [42].

Table 2. Advantages and disadvantages of RGB-D cameras.

Advantages	Disadvantages
reduced size	sensitivity to thermal noise and lighting variations
non-invasive monitoring	high storage resources for onboard data processing
costs and performance trade-off depending on sensing principle	wide occupied bandwidth for remote data processing

Although advances in manufacturing technologies have led to a reduction in the physical dimensions of RGB-D cameras, there are still some implementation challenges to face. In fact, regardless of the specific depth-sensing technique, factors of temperature and ambient light may influence the attainable performance and hinder the effectiveness of the adopted device in uncontrolled conditions, such as outside laboratories and in AAL home scenarios. Such influencing factors lead to noisy or incomplete depth images, which, in their turn, do not enable correct identification of skeleton joints or other body-tracking metrics. As shown in [43], not only noisy conditions but also processing platforms and algorithms may generate noticeable differences among the output results. Another

important factor to take into account regards the amount of data generated by RGB-D cameras. Indeed, depth frames, typically acquired at a rate of 15 frames per second (or even more), need adequate data storage capacity and power for processing, especially for real-time applications (e.g., fall detection). Two options are usually considered, namely, local data processing, which requires high storage resources onboard the sensor, or remote data processing, which removes the burden of computation from the sensor but necessitates enough wired or wireless data transfer rate. The latter approach can face implementation issues in situations where a network infrastructure is not already in place, such as in old buildings, or does not meet the application requirements, as highlighted in real-life evaluation studies.

Table 3 provides a comparison among notable commercial RGB-D cameras regarding the operating principle used. Cameras of the RealSense D400 series by Intel (Intel Corp., Santa Clara, CA, USA) are popular for HAR applications in the AAL domain for their versatility and relatively compact size, combined with affordable cost. They use stereo vision to calculate depth, while the L515 uses Light Detection and Ranging (LiDAR) technology for depth sensing, providing high-accuracy depth frames and low power consumption. The Kinect v1 and v2 cameras (Microsoft Corp., Redmond, WA, USA), using structured light and the ToF operating principle, respectively, are nowadays considered legacy, but they are historically significant for having made RGB-D imaging popular and affordable, first in the gaming space, then in the AAL domain. The more recent Kinect Azure camera (Microsoft Corp., Redmond, WA, USA) employing ToF mostly targeted developers and industrial applications, offering higher resolution and more advanced features than the consumer versions. In August 2023, Microsoft discontinued Kinect production, including Azure. Orbbec is nowadays a global leader in the design and manufacturing of 3D cameras, with a diverse offering of products tailored to different applications. The affordable Astra Pro series (Orbbec, Troy, MI, USA), among those more frequently used in AAL and HAR applications, exploits ToF in the nIR range, providing good performance for gesture and action recognition. Other solutions are provided by Asus, with the Xtion Pro device [44] (AsusTeK Computer Inc., Taipei, Taiwan) offered for more than ten years, and Structure IO (Structure, Boulder, CO, USA), both using structured light 3D sensing technology [45], but their usage is mostly limited to research and development activities.

Table 3. Comparison of sensing principle among commercial RGB-D cameras.

Manufacturer	Commercial Name	Depth-Sensing Principle
Intel RealSense	D400 series	stereo vision
	L515	LiDAR technology
Microsoft	Kinect v1	structured light (IR projection)
	Kinect v2	ToF
Orbbec	Astra Pro	ToF (nIR)
Others	Structure IO	structured light (IR projection)
	Asus Xtion Pro	structured light (IR projection)

Selecting an appropriate RGB-D camera requires a comprehensive evaluation of several technical and practical specifications to match them with the demands of the intended application. From the technical perspective, for applications demanding precise 3D data (detailed object reconstruction, accurate spatial mapping), higher resolution and superior accuracy are paramount, as they directly impact the fidelity and reliability of the captured depth information. Furthermore, the maximum operating distance of the camera is a critical factor, particularly when the application involves larger environments or requires

capturing objects or moving subjects at varying distances. An adequate range ensures comprehensive data acquisition across the desired operational volume. Finally, a wider field of view (FoV) is advantageous for applications capturing larger areas or for the simultaneous observation of multiple subjects within a scene. A broader FoV can reduce the need for multiple camera placements (thus reducing also complexity and costs) or extensive frame stitching in post-processing steps.

Among the practical issues to consider, the physical dimensions and form factor of the camera are crucial for seamless integration into diverse systems and environments and for compact experimental setups. Miniaturization and specific form factors can significantly influence overall system design and acceptability from the subjects being monitored [35]. Alongside this, the financial investment associated with RGB-D cameras varies considerably across different models and manufacturers. Budgetary constraints often play a significant role in the selection process, necessitating a balance between desired performance and economic viability. Lastly, the availability of robust software development kits (SDKs) and comprehensive libraries for depth frame processing is fundamental for efficient system development and application integration: strong software support facilitates data interpretation, algorithm implementation, and overall system functionality.

Beyond HAR, RGB-D cameras are being adopted across a multitude of sectors within consumer electronics, robotics, machine vision, and the automotive industry. The increasing number of applications using RGB-D cameras supports projections of market expansion at a compound annual growth rate (CAGR) of over 10% until 2031 [46] and the expectation that these devices will be available at a more affordable price and with improved performance in the future. This is a relevant aspect to consider for the viability and sustainability of HAR-based AAL applications and systems in the long term.

3.1. ML Techniques

RGB-D cameras capture both traditional color (RGB) images and depth (D) information for each pixel, providing rich datasets that significantly improve ML and deep learning (DL) approaches for HAR. Consequently, RGB-D-based HAR methods can directly exploit acquired images (RGB and/or D ones) or skeletal information obtained from them. This combination of visual appearance and 3D geometric data allows algorithms to perceive and understand human activities with greater accuracy and robustness. For example, in [47], a recurrent neural network (RNN) is proposed for recognizing human activities from depth camera images. The learning model, however, is not fed directly by images acquired at each frame; instead, skeletal joint positions are extracted from each image, and then joint angle variations are computed at each recorded frame in order to avoid dependencies from specific positions of the joints. The full body kinematics is taken into account. For encoding time sequential information, an RNN with long-short-term-memory (LSTM) is chosen, where the number of LSTM matches the length of the activity video frame. A total of 12 activities were considered for testing the proposed architecture: lift arms, duck, push right, goggles, wind it up, shoot, bow, throw, had enough, change weapon, beat both, and kick. The proposed methodology was compared with other state-of-the-art classification schemes, i.e., hidden Markov model (HMM) and deep belief networks (DBN). RNN outperformed both HMM and DBN, with an average accuracy above 99%, whereas the DBN offered about 97% in correctly classifying the 12 different activities.

In general, a major role in RGB-D-based HAR is played by artificial intelligence architectures employed to process images in order to extract human silhouettes, identify skeletons, or compute features. Their fast progress in the last few years depends also upon the growing availability of large amounts of data and large-scale datasets required for training, like those reported in Table 4 and discussed later in this section. In this view,

convolutional neural networks (CNN) represent the state-of-the-art for image processing and computer vision-based applications [48]. CNNs also allow learning hierarchical characteristics from unprocessed data, avoiding the handcrafted feature extraction procedure commonly employed when more traditional machine learning approaches are adopted for pattern recognition purposes. Some studies also used multi-dimensional CNN to further exploit the potential of such algorithms for image processing. The use of 2D CNN allows us to enhance the spatial information that can be retrieved from images, but in some cases they showed good performances only in poor, challenging scenarios [49]. For addressing this issue, 3DCNN found large employment, since they are able to learn spatial and temporal information by convolution within and across the video frames [50]. However, 3DCNNs become highly computationally demanding when long-range temporal dependencies across the frames have to be captured, imposing non-negligible problems when used for activities occurring for extended periods of time [51]. More recently, hybrid methods have been proposed to leverage the strengths of different models and improve the learning of discriminative features from the video frames. Spatial features can be extracted by pretrained nets and then concatenated for feeding LSTM as a sequence [52]. Multi-stream procedures were also proposed for learning spatial and temporal features for recognition of human activities [53,54].

However, due to the high number of environmental characteristics and challenging scenarios that can be encountered, e.g., texture variations, low resolution, scaling, and temporal flow, the recent literature in the field of HAR from RGB-D camera images focused its efforts on the development of novel architectures aimed at improving recognition performances in a variety of different contexts, pushing toward the generalization of such architectures. Hussain et al. [55] presented a two-module architecture. The former was devoted to feature extraction and based on a dynamic attention fusion unit. The latter was made by a temporal-spatial fusion network that encoded complex patterns and learned the most discriminative features for the final HAR in challenging, real-world conditions. Indeed, HAR from images requires understanding the interactions between the human body and the environment over time. Both spatial and temporal features are important because activities to recognize involve sequences of motion over space and time. The proposed methodology was evaluated on four available datasets, with various classes of human activities, that resemble real-world scenarios in terms of motion, appearance, viewpoint, and lighting conditions. Classification accuracy was well above 98% for three out of the four tested datasets, whereas for the last one about 80% accuracy was achieved. Although a wide range of different activities were taken into account, this study highlights that the input video data affect crucially the performance of the recognition model. Indeed, differences in image brightness and background between training and inference data are a well-known problem, referred to as distributional shift. It leads to a performance degradation, also when considering that an action is dependent on the performer and the same activity made by two subjects can be significantly different from each other. Thus, subject-specificity of movement data often lowers the capability of the model in correctly classifying actions [56].

One of the most promising solutions proposed to deal with this issue is domain adaptation, consisting of the training of a domain-invariant learning model, using domain labels as a clue [57]. A phase randomization approach was developed in [56], suitable for working on skeleton data extracted from RGB-D images. In this approach, a novel data augmentation procedure was developed to highlight subject features by decomposing motion data into amplitude and phase components. The phase randomization works by randomizing the phase of the frequency components, keeping untouched the relative amplitude, which most likely includes the individuality of the subjects. A more general framework was proposed in [58], with a two-stage pipeline made of a skeleton sequence

generation followed by a stacked ensemble classification of human activities. The 2D skeletal key points are extracted from image frames by using MoveNet [59], and then converted into 3D by a Gaussian RBF kernel. Then, spatial and temporal features, reflecting specific motion dynamics, are computed on 3D reconstructed skeleton data and used for the eventual classification of human activities. The latter includes a convolutional LSTM block, a spatial bidirectional gated temporal graph convolutional network, and a convolutional eXtreme gradient boosting. This general framework was tested on several datasets encompassing images recorded by different devices, namely the NTU-RGB+D60 dataset [60], where Kinect v2 was employed and 60 activities are included; the NTU-RGB+D120 dataset [61], where RGB-D cameras are used; the Kinetics-700-2020 dataset [62], which includes video clips taken from the YouTube website for over 700 activity classes; and the MA-52 [63], which includes 52 micro-actions (see also Table 4 for details). The framework provided above 95% accuracy for each dataset, proving to be reliable in several different working conditions and with different sources of information.

Interaction between humans and objects represents an additional field where image processing has been employed. Li et al. [64] introduced a novel paradigm through a model that leverages natural language processing for supervising visual feature learning, enhancing their expression capability. The proposed methodology was tested on two human-object interaction datasets, namely the CAD-120 dataset [65] and the Something-Else dataset [66]. The former contains 120 RGB-D videos for 10 complex activities, with labeled sub-activities, whereas the latter includes more than 10^5 videos for 174 interaction activities. Results showed that the inclusion of natural language supervision in the learning process improves the recognition performances on both datasets but is limited to indoor settings and interaction scenarios involving only a single human.

Rather than focusing exclusively on learning architecture and modules, some works also emphasize the role of feature extraction from RGB-D images. For instance, Elnady and Abdelmunim [67] combined an LSTM network with You Only Look Once (YOLO) for activity recognition in video sequences. In addition, a tracking model was included for maintaining temporal consistency of the detected objects across the video sequence. The proposed methodology was tested on four publicly available HAR datasets: the UFC101 dataset [68], the KTH dataset [69], the WEIZMANN dataset [70], and the IXMAS dataset [71]. The inclusion of YOLO for feature extraction dramatically enhanced the accuracy of object recognition within complex environments, paving also the way for a possible real-time usage of the method. Leveraging skeleton data represents an attractive way for dealing with HAR from images, since skeleton data are computationally efficient and more robust with respect to some environmental characteristics, such as changes in illumination, background noise, and camera view. For instance, Chen et al. [72] proposed a spatio-temporal graph CNN specifically designed for improving activity recognition from skeleton data by taking into account indirect connections between skeletal key points. The proposed network demonstrated the capability of extracting spatio-temporal features on a local and global scale by a graph CNN and a spatio-temporal Transformer. These two distinct streams allowed us to extract topological and motion-related structures, together with relationships between different skeleton joints, thus leveraging two different representations of the same object. Classification architecture was tested on the NTU-RGB+D60 dataset [60], the NTU-RGB+D120 dataset [61], and the Kinetic-Skeleton dataset, where skeleton representation was derived from videos of the Kinetics 400 dataset by using the OpenPose algorithm [73]. Outcomes showed that the proposed architecture was able to improve state-of-the-art solutions when tested on the same datasets. A short summary of the most used HAR datasets publicly available is reported in Table 4, with details on the type of collected data and activities included. Figure 2 provides an exemplary (though

not exhaustive) graphical representation of machine learning- and deep learning-based methods for HAR using RGB-D cameras.

Table 4. Publicly available datasets used for testing computational and learning frameworks for HAR from video recordings.

Dataset	Data	Activities
MSRC-12 [47]	6244 video samples with skeletal joints position	12 activities from 30 subjects
NTU-RGB+ D60 [60]	56,880 video samples, including RGB, infrared, depth, and skeleton data	60 ADLs
NTU-RGB+ D120 [61]	Expanded version of NTU-RGB+ D60, with 114,480 video samples	120 ADLs
MA-52 [63]	22,422 video samples	52 micro-actions from 205 subjects
CAD-120 [65]	120 video samples	10 high-level activities, 10 sub-activities, and 12 object affordance from 4 subjects
PKU-MMD [74]	1076 video sequences, including RGB, depth, infrared, and skeleton data	51 ADLs by 66 subjects
Berkeley-MHAD [75]	660 action sequences recorded from 4 cameras, 2 Kinect cameras, 6 accelerometers, and 4 microphones	11 ADLs by 12 subjects with 5 repetitions for each ADL
HWU-USP [76]	Recordings from ambient sensors, inertial units, and RGB videos for a total of 144 recording sessions	9 ADLs by 16 subjects
Northwestern-UCLA Multiview [77]	RBG, depth, and skeleton data recorded by 3 Kinect	10 ADLs by 10 subjects
UTD-MHAD [78]	RGB, depth, skeleton data recorded from 1 Kinect camera, and inertial data from a single inertial sensor	27 ADLs by 8 subjects, with 4 repetitions for each ADL
Toyota Smarthome [79]	16,115 video samples, including RGB, depth, and skeleton data recorded from 7 Kinect cameras	31 ADLs from 18 subjects
UFC101 [68]	13,320 video samples	101 ADLs
KTH [69]	2391 video samples from 25 people	6 ADLs in four different contexts
WEIZMANN [70]	90 low resolution video sequences	10 ADLs by 9 subjects
IXMAS [71]	2880 RGB video sequences	15 ADLs by 5 subjects, with each ADL repeated 3 times
Kinetics-700-2020 [62]	Collection of a series of datasets containing up to 700 video samples for each ADL	Up to 700 ADLs, depending on the considered specific datasets
Kinetic-Skeleteon [80]	Derived from Kinetics dataset, using OpenPose to extract skeleton key points, and made by 300,000 videos	400 ADLs
HMDB51 [81]	6776 annotated video samples from various sources as movies and YouTube	51 ADLs
STH-STH V1 [82]	108,499 video samples	174 ADLs from 1133 crowdsource workers

Table 4. *Cont.*

Dataset	Data	Activities
UCF50 [83]	6681 video samples	50 ADLs
YouTube Action [84]	1600 video samples from YouTube	11 ADLs
MSR-3D [85]	320 video frames, including skeleton data, RGB, and depth images	16 ADLs
JHMDB [86]	928 RGB video, for a total of 31,838 frames	21 ADLs

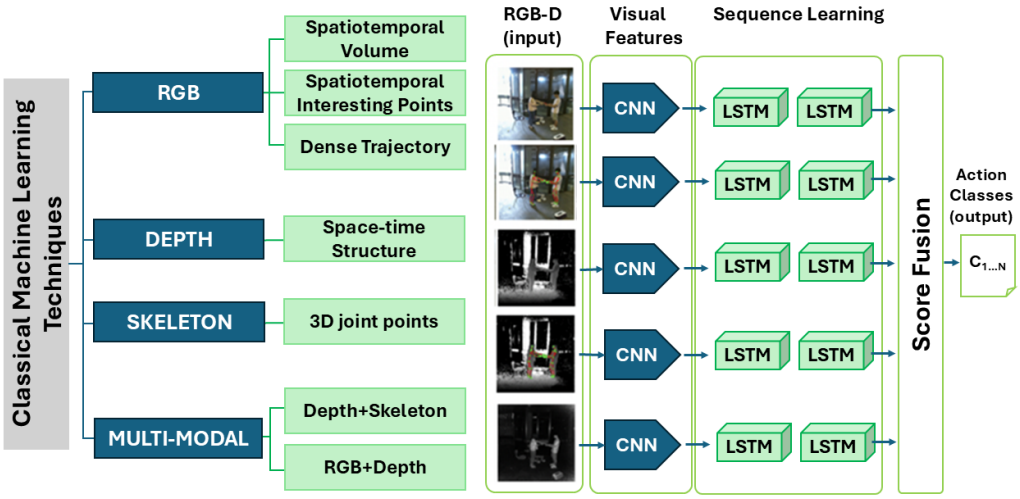

Figure 2. Exemplary graphical representation of (**a**) machine learning- and (**b**) deep learning-based methods, for HAR using RGB-D cameras, from [23].

In general, it is fair to say that skeleton data and RGB modality represent uni-modal approaches (Table 5) based on two different strategies for achieving reliable results in the field of computer vision-based HAR [87]. However, some limitations have been outlined throughout the years. Only to mention a few, the absence of a 3D structure as input is recognized as one of the major limitations of HAR from RGB-D cameras, together with the absence of environmental features [88]. On the other hand, multi-modal approaches, even though less investigated, proved to be particularly suitable for RGB-D-based HAR. In essence, multi-modal methods rely on data fusion on multiple levels, i.e., on a data, feature, or decision-level fusion (Table 6). The problem of multi-modal fusion was tackled in [88] by developing a multi-modal network integrating RGB images and skeletons relying on a model-based approach, where a graph CNN was used for learning weights to be transferred to another network devoted to RGB processing. Also in this case, extensive testing was performed on five datasets: the NTU RGB+D60 dataset [60], the NTU RGB+D120 dataset [61], the PKU-MMD dataset [74], the Northwestern-UCLA Multiview dataset [77], and the Toyota Smarthome dataset [79]. A similar multi-modal approach was proposed in [89], where images of skeletons, motion history (MH), and depth motion maps (DMM) are firstly retrieved from RGB-D cameras. Then, a 5-stack CNN was trained separately on MH and DMM only, whereas for skeleton images the output of the CNN was used for feeding a BiLSTM. Then, an additional fusion step was applied at the level of the score values of the three previous networks, and the final decision is based on the fusion value. The framework was tested on the UTD-MHAD dataset, which contains 27 human activities collected by using a depth camera, with over 96% accuracy on the validation

set. Multi-modal HAR was leveraged also by Batool et al. [90], who introduced a HAR solution based on RGB, RGB-D, and inertial data. Silhouette extraction was performed from RGB-D images by applying Laplacian fitting [91], whereas inertial data underwent a Kaiser windowed filter [92]. Then, specific features were extracted from RGB-D, namely dynamic likelihood random field and angle along the sagittal plane, and from inertial data, i.e., lag regression [93] and gammatone cepstral coefficients [94]. Features were then selected and fused by a genetic algorithm and given as input to a CNN-gated recurrent unit that leveraged Kalman gain instead of a rectified linear unit layer. Results gathered from testing on the UTD-MHAD dataset [78], HWU-USP dataset [76], Berkeley-MHAD dataset [75], NTU-RGB+D60 dataset [60], and NTU-RGB+D120 dataset [61] showed that the proposed methodology outperformed existing similar solutions by achieving an accuracy not below 95% for each of the five considered datasets. A multi-modal approach was also recently proposed by Liu et al. [95], who presented a semantic-assisted multi-modal network specifically designed for overcoming inherent limitations of skeleton and RGB uni-modal approaches. The method works by leveraging adaptive learning among three modalities, where text was added to skeleton and RGB video, thus including detailed textual descriptions of the activities, enriching the category expression. Also in this case, the method was tested on several different datasets, i.e., the NTU-RGB+ D60 dataset [60], the PKU-MMD dataset [74], and the Northwestern-UCLA Multiview dataset [77], showing promising improvements with respect to other existing solutions. Song et al. [96] presented a modality compensation network to properly leverage the complementary information given by different activity representation modalities. In particular, the RGB flow was designed as the source modality, whose feature extraction was improved by including information from the auxiliary modality, i.e., the skeleton data. A two-stream architecture was composed of a CNN and an LSTM, where the former was used for encoding the spatial features, while the latter collects and integrates the information flow over time. Then, the final decision is obtained from the score fusion of each network block. Notably, skeleton data were needed only for training. The architecture was tested on the NTU-RGB+D60, MSR-3D, UCF101, and JHMDB datasets (Table 4), showing improvements in HAR for all the tested datasets. Skeleton and RGB-D modalities have been leveraged also in [97], where a multi-modal framework, enhanced by temporal cues, was proposed. The weighted skeleton joints, learned from a dedicated graph convolutional network (GCN), are used for improving the identification of the spatio-temporal region of interest (ROI) from the RGB video modality. In addition, a temporal cues enhancement module was introduced for the RGB modality, made by a two-stream architecture with a CNN for the spatial domain and an RNN for the temporal cues. The proposed architecture was tested on three publicly available datasets, namely NTU-RGB+D60, PKU-MMD, and the Northwestern UCLA dataset (Table 4). Outcomes showed that the proposed methodology outperformed several previously proposed methods in both cross-subject and cross-view modalities.

3.2. Uni-Modal vs. Multi-Modal Sensing

In light of the previous overview, it is fair to say that both uni-modal (Table 5) and multi-modal (Table 6) approaches present some challenges and issues that still have to be addressed. For what concerns RGB-based HAR, the background in videos can significantly influence action recognition, often neglecting the actual modeling of the actions themselves. To address this issue, attention mechanisms working on foreground motion have been proposed in order to focus on the human appearance and moving regions, minimizing at the same time the impact of the background objects [98]. However, focusing on moving objects only can lead to wrong conclusions when dealing with complex environmental contexts, highlighting the need for a tailored recognition of those objects that are truly

relevant for HAR, which, however, represents itself as a complex and challenging task to be accomplished. Another critical aspect of RGB-based HAR is related to the computational costs for image processing. Although in the last few years computing power underwent remarkable improvements, enabling the widespread application of deep learning across various domains, the increased accuracy of this kind of architecture is achieved at the cost of even higher computational complexity. The latter aspect currently prevents the widespread application of RGB-based HAR technology on mobile devices or systems with limited processing capabilities. However, some solutions have been proposed, such as reducing redundant video information for achieving a more lightweight recognition or optimizing feature extraction pipelines [99,100].

Table 5. Comparison of works using uni-modal approach for HAR from RGB-D images. Classification accuracy values are from original studies.

Work	Architecture	Input Data	Outcomes	Key Novelty Point
Park et al. [47]	LSTM	Skeleton data	99.5% accuracy on MSRC-12 [47]	Leveraging time sequential encoding of activity features
Mitsuzumi et al. [56]	GCN	Skeleton data	67.4% accuracy on NTU-RGB+ D60 [60], 57.7% on NTU-RGB+ D120 [61]	Introducing a subject-agnostic domain adaptation, randomizing the frequency phase of motion data, leaving amplitude unchanged
Karthika et al. [58]	A stacked ensemble model made by SBGTGCN, 2DCNN+2P-LSTM, and 3DCNN+XGBoost	Skeleton data	97.9% accuracy on NTU-RGB+ D60 [60], 97.2% on NTU-RGB+ D120 [61], 97.5% on Kinetics-700-2020 [62], 95.2% on MA-52 [63]	Deriving 3D skeletal points by Gaussian RBF, and designing a stacked ensemble model that integrates multiple base learners and a meta-learner
Chen et al. [72]	GCN and Transformer model	Skeleton data	92.7% accuracy on NTU-RGB+ D60 [60], 86.8% on NTU-RGB+ D120 [61], 39.0% on Kinetic-Skeleton [80]	Design of GCN and Transformer parallel stream for extracting local and global features as topological structures and inter-joints connections
Wu et al. [53]	2D CNN	Video frames	91.9% accuracy on NTU-RGB+ D60 [60], 92.2% on Kinetics-700-2020 [62], 96.9% on UFC101 [68], 73.7% on HMDB51 [81], 77.9% on STH-STH [82]	Design of a multi-level channel attention excitation module to retrieve highly discriminative video feature representation

Table 5. *Cont.*

Work	Architecture	Input Data	Outcomes	Key Novelty Point
Zong et al. [54]	Hybrid CNN-LSTM	Video frames	94.7% accuracy on UFC101 [68], 67.2% on HMDB51 [81], 68.7% on Kinetics-700-2020 [62]	Design of a four-stream network, based on spatial and temporal saliency detection
Hussain et al. [55]	Hybrid CNN-LSTM	Video frames	98.7% accuracy on UFC101 [68], 80.3% on HMDB51 [81], 98.9% on UCF50 [83], 98.9% on YouTube Action [84]	Design of a dynamic attention fusion unit and a temporal-spatial fusion network to extract human-centric features from temporal, spatial, and behavioral dependencies
Li et al. [64]	Spatial-temporal mixed module MLP	Video frames	93.6% accuracy on CAD-120 [65], 86.1% on STH-STH V1 [82]	Introduction of text features for human-object interaction and of supervised natural language learning for augmentation of visual feature representation
Elnady and Abdelmunim [67]	LSTM	Video frames	96.0% accuracy on UFC101 [68], 99.0% on KTH [69], 98.0% on IXMAS [71], 100% on WEIZMANN [70]	Combining YOLO and LSTM to integrate highly discriminative features from individual frames and sequential temporal dynamics of motion

2P-LSTM: two-part LSTM; GCN: graph convolutional network; MLPs: multilayer perceptrons; SBGTGCN: spatial bidirectional gated temporal graph convolutional network with attention; RBF: radial basis function

Table 6. Comparison of works using a multi-modal approach for HAR from RGB-D images. Classification accuracy values are from original studies.

Work	Architecture	Input Data	Outcomes	Key Novelty Point
Bruce et al. [88]	GCN and attention mechanism	Skeleton and RGB data	93.9% accuracy on NTU-RGB+ D60 [60], 90.5% on NTU-RGB+ D120 [61], 96.3% on PKU-MMD [74], 77.5% on Toyota Smarthome [79], 93.5% on Northwestern-UCLA Multiview [77]	Multi-modal network that fuses skeleton and RGB complementary information, using GCN to learn attention weights from skeleton that are then shared with the RGB network

Table 6. *Cont.*

Work	Architecture	Input Data	Outcomes	Key Novelty Point
Kumar and Kumar [89]	Hybrid CNN-BiLSTM	Depth, RGB, and skeleton data	96.2% accuracy on UTD-MHAD [78]	Design of a multi-view, multi-modal framework, where RGB, depth, and skeleton data are separately processed by 5S-CNN and BiLSTM networks, and the outputs are fused by a weighted product model
Batool et al. [90]	CNNGRU	RGB, depth, and inertial data	97.9% accuracy on HWU-USP [76], 97.9% on Berkeley-MHAD [75], 96.6% on NTU-RGB+ D60 [60], 95.9% on NTU-RGB+ D120 [61], 97.9% on UTD-MHAD [78]	Introducing novel features extracted from RGB, depth, and inertial data, where redundant information is profiled out by a genetic algorithm
Liu et al. [95]	Hybrid GCN-CNN	Skeleton and RGB-D video	94.8% accuracy on NTU-RGB+ D60 [60], 97.0% on PKU-MMD [74], 93.7% Northwestern-UCLA Multiview [77]	Introducing a semantic-assisted framework where text modality is added to RGB and skeleton, with a visual-language module with contrastive language-image pretraining
Song et al. [96]	Hybrid CNN-LSTM	Skeleton and RGB-D video	93.8% accuracy on NTU-RGB+ D60 [60], 76.9% on MSR-3D [85], 85.7% on UCF101 [68], 64.8% on JHMDB [86]	Introducing a modality compensation network for fusing multiple representation modalities, including adaptation schemes for narrowing the distance between different modalities distributions
Liu et al. [97]	Hybrid GCN-CNN-RNN	Skeleton and RGB-D video	94.3% accuracy on NTU-RGB+ D60 [60], 96.8% on PKU-MMD [74], 93.9% on Northwestern-UCLA Multiview [77]	Designing of a temporal cues enhancement module for improving temporal modeling from RGB modality

CNNGRU: convolutional neural network–gated recurrent unit; GCN: graph convolutional network; RNN: recurrent neural network

The previous challenges can be mitigated by shifting the recognition paradigm toward the use of skeleton data, since the latter are more robust with respect to changes in the appearance and, in general, require fewer computational resources by neglecting background and object information [87]. When using skeleton data only, a key aspect is the modeling of the changes in the dynamics of actions [56,72]. However, it has also been highlighted that appearance information can enhance recognition of activities with similar motion dynamics [101], thus leading to a significant amount of work where an RGB-skeleton multi-modal approach was used, in order to balance computational needs and identification performances, by fusing temporal movement and appearance features (Table 6). Therefore, environmental characteristics appear to be a critical type of information for HAR applications, capable of enhancing the final recognition rate. In addition, skeleton-based action recognition also neglects information related to the objects involved in some particular actions, which instead represents an important point for many applications. Although some attempts have been made to include object-related information in a uni-modal, skeleton-based HAR framework, as highlighted, for instance, in [102], where objects were treated as an additional joint, combining skeleton data and object information still remains an open research aspect in this field. However, successfully accomplishing this task would offer significant advantages, since it would allow us to avoid privacy-related issues and rely only to a minor extent on the background characteristics, thus keeping computational costs at a reasonable level.

Even though a multi-modal approach thus offers, in general, a more robust approach for HAR from images, there are still some challenges that arise in particular when dealing with real usage scenarios outside controlled environments used for framework development. One of the major issues is to ensure reliable activity recognition when multiple persons are present in the same scenario and perform different activities, representing a condition that cannot be a priori avoided in real-world contexts. For addressing this issue, several works proposed solutions mainly based on the usage of multiple cameras or on refined feature weighting procedures [103,104]. However, unlocking the full potential of action recognition technology for multi-person and multi-action tasks remains a significant hurdle for real-life applications in this field. Connected to the multi-person, multi-action issue, there is also the problem of occlusion, which is unavoidable in densely populated scenes. To address this issue, widely adopted solutions are the fusion of multiple cameras' information and the a posteriori reconstruction [105,106]. However, while these approaches have made significant progress in mitigating occlusion-related issues, challenges such as large-scale occlusion and the inability to capture multi-angle video data remain still the object of much research effort.

Finally, it is worth mentioning that an additional challenge when deep learning approaches are employed for HAR is to have enough data to achieve a robust training of the learning models. Although throughout the years many datasets have been recorded and made publicly available (Table 4), their dimension is often limited with respect to the adopted architectures, and few of them were collected in real-life scenarios. Further, also the lack of labels represents an additional challenge. To address these issues, data augmentation schemes offer a viable solution [107], together with the usage of fine-tuning strategies of pre-trained models and unsupervised learning architectures [108].

Overall, multi-modal approaches seem to be a convenient solution for addressing some problems arising with uni-modal strategies. However, as outlined above, also integrating RGB and skeleton modalities within a more comprehensive approach encompasses some drawbacks that are still far from being completely solved. Therefore, the latter studies highlighted the value of fusing different types of information to enhance HAR from images recorded by RGB-D cameras. This supports the opportunity of investigating the inclusion

of additional devices and data flow when dealing with the identification of human activities and gestures from images and video recordings. In this view, BCI could represent a viable solution, since it provides a completely different kind of information with respect to RGB and/or skeleton data, with the potential of limiting some of the problems given by a fully image-based scheme. At the same time, BCI provides information that is specific for the single user, with the possibility of designing more subject-centered strategies for HAR also within real-life scenarios.

4. HAR Based on BCIs

In AAL environments, wearable devices have been increasingly utilized to improve the quality of life of frail and elderly individuals in their living environments and to promote autonomy and safety. This innovative strategy enables residents to seamlessly control their living spaces, making everyday tasks easier through intuitive interactions with various devices and robots. Wearable technologies also allow for collect real-time data on user health status and physical environment, allowing personalized adjustments to improve comfort, safety, and efficiency. This integration fosters a symbiotic relationship between the user and the environment, moving away from traditional interaction and user interfaces, where human feedback and sensor data continuously inform and optimize each other responses, ensuring a supportive and adaptable living space that caters to the unique needs of its inhabitants in a more human-centric approach. Users can command and communicate with their surroundings in a more natural and effortless manner, using the capabilities of wearable technology, such as smartwatches, fitness bracelets, and even more advanced tools such as BCI.

Among the wearable technologies being deployed for user feedback in AAL environments, BCIs are particularly significant and promising, especially for people with severe impairments. BCIs have a wide range of applications, prominently in the field of assistive technologies. They have been used to restore communication to control prosthetic limbs or robotic arms, offering new possibilities for people with different frailties [109]. Another important application area is the rehabilitation of stroke patients, where BCIs can facilitate motor recovery by enhancing neuroplasticity. This is achieved through the use of BCIs to control virtual reality environments or robotic devices that provide patients with feedback and assistance in performing motor tasks [110]. Wearable BCIs, based on the acquisition of electroencephalography (EEG) signals from the user scalp, enable the direct interpretation of the user neural signals, which can be encoded into machine-readable commands, for a natural and personalized control of the devices located in the living setting. With this scope, BCI applications can support users as interfaces to monitor and record data from cognitive signals during interaction tasks or as control interfaces for human–robot interaction or cooperation activities. In fact, direct control over home devices and robots is thus facilitated, enabling users to manage their surroundings with simple gestures or voice commands.

BCIs offer a revolutionary means to control both the living environment and various devices within it, including essential services and personal robots that assist with daily activities [111]. Their technology enables direct brain-to-machine communication, bypassing traditional physical interaction modalities and offering a more inclusive and accessible way for users to interact with their surroundings. BCIs operate on the principle of detecting, decoding, and translating brain activity into external actions. The brain activity can be recorded using various methods, including non-invasive techniques such as electroencephalography (EEG) and more invasive approaches such as electrocorticography or intracortical recordings. EEG, due to its ease of use, safety, and cost-effectiveness, is the most commonly used method in BCI systems [112]. With respect to this approach, the most used BCIs in the scientific research are the non-invasive ones that detect brain

activity externally. In Table 7, some of the commercial BCI prototypes and products used in different applications and with various scopes in scientific studies are reported. As reported before, in the literature invasive approaches and, in particular, invasive BCIs that require surgical implantation are also presented. In particular, authors report some of the commercial products/companies as Neuralink, Synchron (Stentrode), Blackrock Neurotech, Paradromics and Precision Neuroscience, which are characterized by high-density brain implants, minimally invasive via blood vessels, and invasive brain interfaces. These high-density implantable BCI systems are developed with the aim of supporting neural decoding and rehabilitation approaches, particularly for paralyzed patients and people with severe disabilities and neurological conditions, to restore and improve motor control and communication.

Table 7. Primary applications of main non-invasive BCIs.

Company/Product	Brief Description	Applications
Emotiv (EPOC X, MN8 EEG)	Wireless EEG headsets	Neuroscience research, neurofeedback, gaming, cognitive wellness monitoring
OpenBCI (Ultracortex Mark IV, Cyton, Galea)	Open-source BCI hardware and software	Research, development, hobbyist projects, gaming, AR/VR/XR with integrated biosensing
g.tec (Unicorn Brain Interface, g.HIamp PRO)	Complete BCI systems and components (amplifiers, EEG headsets)	Research, rehabilitation, clinical applications
NeuroMaker BCI	BCI kit for educational purposes	Learning about neuroscience, visualizing brainwaves, mind-controlled games
PiEEG	Low-cost EEG devices and BCI kits	Learning, research, development
Neurable	BCI technology for controlling digital objects	Gaming, augmented/virtual reality, thought control
NextMind (by Snap)	BCI technology for decoding neural activity	Controlling digital objects (acquired by Snap, focus on AR/VR)

The integration of BCIs into AAL environments represents a leap forward in human-environment interaction, allowing for a level of autonomy and engagement previously unattainable for many users. With BCIs, individuals can control aspects of their environment and interact with service robots [113] through only thought patterns. This capability is particularly transformative for those who face mobility or communication challenges, as it provides them with a new avenue to interact with their environment effectively and independently.

From this point of view, BCIs represent a cutting-edge technological paradigm that facilitates direct communication pathways between the brain and external devices, bypassing conventional neuromuscular routes. This technology harnesses neural signals, interprets them through computational algorithms, and translates these signals into actionable commands for various applications, ranging from medical rehabilitative tools to augmentative communication devices and control systems for the environment or robotics. The decoding process involves sophisticated signal processing and machine learning algorithms [114] to interpret complex patterns of brain activity. These algorithms are designed to identify specific signal features associated with intentions or thought processes and translate them into commands [115]. Furthermore, the application of BCIs in these settings shifts towards more naturalistic and human-centric user interfaces, moving from the human-computer

interface to the human-environment interaction, where the focus is on creating smart environments that adapt to and understand the needs of their users through feedback mechanisms. In this context, Streitz proposed a transition from human–computer interaction to human-environment interaction, highlighting the future, where individuals interact with smart environments composed of various devices. These environments often feature integrated computing devices in everyday objects, making them less noticeable, which are referred to as disappearing computers [116]. This integration poses new challenges in providing suitable interaction possibilities, and it results in crucial importance to focus on supporting human interactions within these settings, considering the importance of user control, transparency, ethics, privacy, and other significant issues. In the realm of human-environment interaction, a key development is the need to effectively use human feedback to adjust environmental behavior. This feedback is crucial in cooperative tasks to manage and mitigate factors that can hinder performance, as highlighted in several studies. Human feedback is particularly vital to improve safety in various situations. Using BCIs, these environments can become more responsive and tailored to individual preferences and requirements, enhancing not only the usability but also the safety and well-being of their inhabitants.

BCIs, therefore, not only enrich the user control over their living space but also contribute to the ongoing dialog on the ethical, privacy, and transparency concerns inherent in these technologies. As BCIs become more integrated into AAL environments, it is essential to address these issues to ensure that the technology empowers users without compromising their rights or autonomy. The potential of BCIs in AAL environments underscores the importance of developing smart living spaces that prioritize user interaction and feedback, thereby fostering a more inclusive, safe, and responsive living environment for all.

Despite promising applications, BCIs face several challenges. The variability in individual brain signals, the need for personalized calibration, and the limited accuracy and reliability of signal decoding are significant hurdles, as reported also in Table 8, where advantages and disadvantages of the BCI applications are reported. Moreover, non-invasive BCIs often suffer from low signal resolution due to the interference of signals by the skull and scalp [117].

Table 8. Advantages and disadvantages of BCI interfaces.

Advantages	Disadvantages
customized properties for users	required training phase on users
invasive/non-invasive mode depending on acquired signals	sensitivity to non-linearity and noise
signals monitored by multiple channels	non-stationary acquisition process

Future research directions include improving the usability and robustness of BCI systems, enhancing signal processing algorithms for better accuracy, and developing adaptive systems that can learn and adjust to change user brain patterns over time. Moreover, integrating BCIs with other technologies, such as augmented reality, virtual reality and mixed reality, opens new avenues for immersive and interactive applications [118]. In this context, in [119] a study related to an innovative methodology to train an artificial neural network is proposed to identify and tag target visual objects in a given database. In particular, with the aim of improving the tag data phase, the study uses the advantages of human cognition and machine learning by combining BCI, human-in-the-loop, and deep learning. The users are equipped with a BCI system and images are shown using a rapid serial visual presentation. Some images are target objects, while others are not. Based on the activity

of the brain waves of the users, when the target objects are shown, the computer learns to identify and tag the target objects already in the learning stage.

The authors report, in this scenario, the experience of a pivotal work to realize a human-in-the-loop approach, where the human can provide feedback to a specific robot, namely, a smart wheelchair, to augment its artificial sensory set, extend and improve its capabilities to detect and avoid obstacles. In the implemented approach, input from humans is facilitated through a BCI. In pursuit of this goal, the study also encompasses the development of a protocol to trigger and capture event-related potentials in the human brain. The entire framework is preliminary and will be tested in simulated robotic environment, using electroencephalography signals to collect from different users [120].

The focus of the research is on a smart wheelchair capable of indoor navigation and autonomously avoiding obstacles using its in-built sensors and intelligence. Human feedback comes into play when the operator detects an obstacle not recognized by the sensory capabilities wheelchair. Upon receiving human input, the robot navigation system adjusts to incorporate this feedback, creating a virtual obstacle. This adjustment alters the robot local path planning to circumvent the detected obstacle. BCI is integrated as a feedback method, together with a protocol designed to invoke the event-related potential (ERP) signal when encountering obstacles. Initially, the efficacy of the system will be tested using keyboard inputs to provide feedback, establishing a performance baseline. Subsequently, a BCI classifier will be used, showcasing the system capabilities with this advanced feedback mechanism.

Human-In-The-Loop Approaches

Human-in-the-loop is an increasingly researched area that refers to the integration of human feedback into computation processes [119]. The human-in-the-loop approach can be applied to different scenarios in which human control can be involved in the designed task. To provide human feedback, it is well-established that brain activity and mental states are decoded to varying degrees of accuracy, depending on the paradigm, equipment, and experimental setup used [121]. The BCI should learn from the data and be tailored to a specific subject or session as more data becomes available.

Nowadays, different studies investigate how to implement the human-in-the-loop control as feedback for robots in different activities to support people with special needs. In the literature, the study proposed in [122] investigates a BCI system that allows interaction with the external environment by naturally bypassing the musculoskeletal system. The approach performed a hybrid EEG-based BCI training involving healthy volunteers enrolled in a reach-and-grasp action operated by a robotic arm. The results obtained showed that hand grasping motor imagery timing significantly affects the evolution of BCI accuracy, as well as the spatio-temporal brain dynamics.

Often, in different approaches, the EEG signals acquired by BCIs are evoked through visual stimuli. The work [123] proposes a close-loop BCI with contextual visual feedback through an augmented reality headset. In such a BCI, the EEG patterns from multiple voluntary eye blinks are considered as input, and the online detection algorithm is proposed, whose average accuracy can reach a high value. In this context, the objective of the study [124] is to explore the feasibility of decoding the EEG signature of visual recognition under experimental conditions, promoting the natural ocular behavior when interacting with a dynamic environment. The results show that visual recognition of sudden events can be decoded during active driving. Therefore, this study lays a foundation for assistive systems based on driver brain signals. In the article [125], a novel strategy was proposed that uses the human mind to choose the mode of ambulation of prosthetic legs. The article described methods for acquiring information about human brain activity and recognizing

intentions. In order to realize stable and flexible walking of prosthetic legs in different terrains according to human intention, a brain–computer interface based on motor imagery (MI) is developed.

Furthermore, [120] reports an approach with the goal of involving the user directly in the robot navigation process by using human feedback to improve the built-in robot sensory system, thus improving its ability to identify obstacles. For user feedback, it is crucial to accurately estimate the location of obstacles that the robot might not detect. This feedback is then used to generate a virtual obstacle within the virtual environment of the robot, based on its sensory data. Consequently, the robot path planning system, unable to differentiate between actual and virtual obstacles, adjusts its course to avoid the obstacle. This strategy remains valid as long as the virtual obstacle placement accurately reflects the real location of the obstacle (Figure 3).

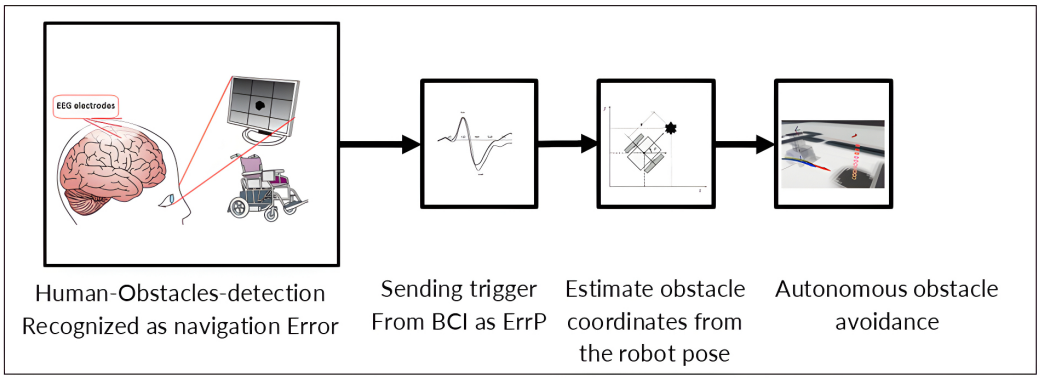

Figure 3. Passive BCI: Utilizing error-related potentials (ErrP) for error detection in the robot navigation path [126].In the first square participants watched the robot navigate through a simulated environment and instinctively responded to any navigational errors the robot makes. The generated errp is recorded by BCI and sent as trigger in order to change the path planning. In the last two squares the new obstacles coordinates are estimated by the algorithm and the obstacle avoidance is implemented in the simulated environment.

The involvement of human control and the implementation of feedback mechanism can be facilitated through the BCI system. In particular, it is important to design a protocol for EEG signal acquisition that can provide the useful trigger for the robotic system and for the effectiveness of the implemented approach. Focusing on control navigation, the EEG signals chosen to evoke as a trigger were the error-related potentials (ErrPs). The captured EEG signals revealed slow brain waves indicative of cognitive responses to perceived navigation risks, useful as navigation triggers. The test case was designed through evoked visual stimuli, leaving the user observing the screen displaying a smart wheelchair navigating around different holes in a virtual laboratory environment. Human feedback was generated using standard input devices, including keyboards and touchscreens. The task required the user to signal (via keyboard press) when a hole was noticed. Then, to respond to the same task, equipped with the BCI system, whenever they perceive the wheelchair crossing or evading a hole, they have to press the keyboard [113]. Repetition of the task with the BCI tests the viability of incorporating human feedback via the BCI into the navigation system, training the BCI classifier to recognize ERPs as navigation triggers. EEG signals were processed in MATLAB R2023b, with detected triggers sent to a Robot Operating System (ROS) node for publishing a Boolean trigger value. Within ROS, another node will receive the trigger and adjust the robot navigation strategy to approach and avoid obstacles, ensuring effective avoidance as long as there is proximity between the actual and virtual obstacles.

5. RGB-D and BCI Integration: Opportunities, Challenges, and Open Issues

In the previous sections, the current state of adoption of RGB-D camera sensing and non-invasive BCIs as single technologies in research dedicated to HAR in daily life environments has been examined. The presented analysis highlights that the two modalities offer complementary perspectives. RGB-D cameras capture the external context of the user, the physical environment, body movements and interactions, while non-invasive BCIs capture, through neural signals, the internal context as well as cognitive state, intentions, volitional commands or mental engagement of users during activities [20,127]. Neither RGB-D cameras nor non-invasive BCIs, though, when used alone, can ensure a comprehensive and reliable solution for the complex and dynamic needs of AAL. As highlighted by Diraco et al. in [128], the fusion of different sensing modalities enables the design of more comprehensive and robust HAR systems in smart living environments, with increased accuracy and reliability. This is already demonstrated in the literature for the case of inertial, RGB, and skeleton data, or for the case of RGB and radar sensor data: HAR performance may be dramatically improved by resorting to the integration of different sensing technologies, despite the challenges in developing effective models to capture spatio-temporal dependencies in complex settings. Based on these premises, the fusion of RGB-D camera sensing and wearable, non-invasive BCIs is envisioned here to obtain a dual perspective on user activity in AAL.

5.1. RGB-D and BCI Fusion Pipeline

A few attempts to fuse the information generated by integrated RGB-D and BCI sensing technologies. Pereira et al. [129] proposed an integrated framework in which data from an RGB-D camera is exploited to improve the BCI-enabled control (by a P300 speller [130]) of a smart wheelchair during navigation in indoor environments. The visual information provided as a background to the P300 graphical control interface helps the user to control the wheelchair with better precision and accuracy than using the BCI alone. In [131], Mezzina et al. proposed a smart sensor system aiming to realize a human-robot interface (HRI) for AAL. The acquired brain signals were wirelessly transmitted to a PEPPER personal care robot (SoftBank Robotics Group, Tokyo, Japan), which navigated the environment, showing the user the available goods and services through an onboard camera. Supported by video information, the user could better formalize unambiguous requests to the robot. Experiments showed that commands could be sent to the actuator in less than 900 ms and executed with an accuracy higher than 80%. A similar approach was presented in [132], again exploiting multi-modal sensing to improve the ability of the BCI user in controlling and refining a NAO robot (SoftBank Robotics Group, Tokyo, Japan) operation. In [133], the integration of a BCI and Kinect sensor to record motion capture information was used for designing an effective rehabilitation system based on serious games. The movements of the hands or the body were sensed by the depth camera for a more natural and dynamic interaction of the user with the serious game; while these works show the feasibility of the integration of RGB-D and BCI technologies with near real-time performance, the degree of multimodal fusion potentially achieved is far from being fully achieved, as this perspective paper aims to suggest.

A multi-modal approach can leverage the complementary nature of the data types generated by the two technologies: RGB-D provides rich environmental and physical information (visual and depth), while BCI provides direct user intent and cognitive state. To achieve complete integration based on fusion, as shown in Figure 4, a pipeline composed of several key steps needs to be developed, including data acquisition, preprocessing and synchronization, feature extraction, data fusion and classification, and finally application

and control. These same steps basically rule out the multimodal fusion of RGB, skeleton and inertial data, or RGB and radar sensor data, that are commonly found in the literature to maximize HAR performance, providing increased accuracy with respect to unimodal-based solutions and almost real-time recognition capabilities [134,135].

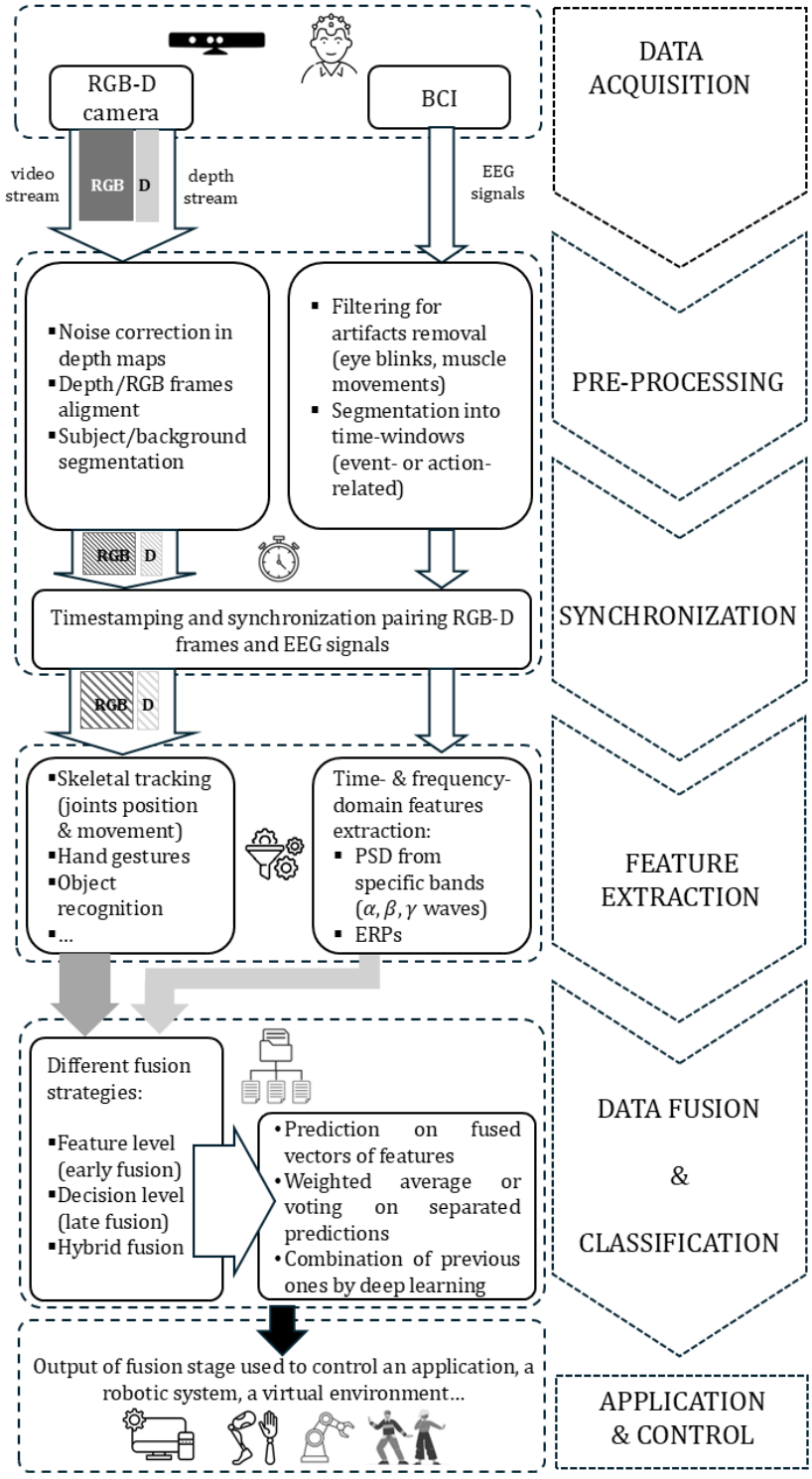

Figure 4. Pipeline of RGB-D and BCI integration steps.

The process starts with data acquisition, i.e., collecting raw data from both sensing modalities simultaneously. The video streams and depth maps captured by the RGB-D camera provide information on physical actions and gestures of users, and the 3D structure

of the environment. Raw EEG data, typically generated by non-invasive BCIs, reflect the user cognitive state or intention, which is the key information the BCI system needs to interpret. Raw data from both sources needs to be cleaned and aligned for effective fusion, thus requiring a pre-processing and synchronization step. Following the removal of noise, artifacts, and mismatches from the raw data streams with data-specific techniques as presented in [136] for BCIs, since the two sensors operate at different sampling rates and capture different aspects of the user and environment, their data streams must be precisely time-stamped and synchronized. This ensures that a given BCI signal is correctly paired with the corresponding RGB-D frame. After pre-processing, meaningful features are extracted from each data stream, which enables the core step of the pipeline: data streams are combined to create a more robust and accurate output by different possible fusion strategies at the feature or decision level, or by a hybrid approach [29]. Early fusion approaches working at the feature level are based on concatenating the extracted features from both the RGB-D and BCI streams into a single, combined feature vector. Then, a classifier is trained on the fused vector to make a prediction. Late fusion approaches apply fusion at the decision level. Each data stream is processed independently: an RGB-D classifier makes a prediction based on visual features, a BCI classifier makes a separate prediction based on brain signals, and then the final decision is made by a fusion algorithm combining the outputs of the individual classifiers. Other approaches, namely hybrid ones, exploit deep models that learn to fuse features at different levels of the network. At this point, different classifiers may be applied, depending on the specific target application or control task one aims to accomplish [137]. The fused information provides a more comprehensive understanding of the user intent and context, enabling more intuitive and precise control, as presented in [138] with the integration of audio-visual feedback into a BCI-based control of a humanoid robot. For example, a BCI command to control a prosthetic limb could be validated and refined by RGB-D data showing the current user orientation and body posture. A similar solution was presented by Zhang et al. in [139], where an intelligent BCI system switch, based on a deep learning object detection algorithm (YOLOv4), was designed to improve the level of user interaction. The performance of the brain-controlled prosthetic hand, and its practical effectiveness, were tested by experiments in real scenarios. Real time detection of the prosthetic hand with confidence higher than 96% was achieved by the YOLOv4 tiny model, and the system was able to support the execution of four types of daily life tasks in real conditions, with acceptable execution time.

RGB-D sensing and BCI integration would consequently represent a powerful application of multi-modal fusion, where the complementary strengths of each technology are leveraged for an advanced form of adaptive human–machine collaboration. A dynamic intelligence loop, where the system continuously infers user intent from BCI signals and contextualizes it with information from RGB-D data, would create a full user-in-context understanding, moving AAL systems from passive monitoring to truly intelligent, adaptive, and user-centric assistance. A recent paper by Bellicha et al. [140] describes an assist-as-needed sensor-based shared control (SC) method, relying on the blending of BCI and depth-sensor-based control, which is able to reduce the time needed for the user to perform tasks and to avoid unwanted actions of a robotic manipulator. However, in this case, though, a cortical implant featuring 64 electrodes has been employed. Table 9 summarizes the relevant studies mentioned above, with regard to their objective, methodology, test population and main results.

Table 9. Comparison of relevant studies adopting RGB-D cameras and BCI integration in AAL.

Study	Objective	Method	Population	Results
Pereira et al. [129]	A dynamic visual interface to navigate the indoor environment, which exploits RGB-D perception to improve BCI-based actuation of a wheelchair	RGB-D images and BCI signals, separately processed, are merged in a dynamic visual interface. User intent gathered by the P300 device is matched to bounding boxes detected on RGB-D images	5 participants (23–31 years old, 3 males, 2 females)	89.9% and 86.4% an average accuracy, respectively, for non-self-paced and self-paced selection of 30 predefined target events, with average effective symbol per minute (eSPM) of 4.8 and 4.7
Mezzina et al. [131]	A smart sensor system to implement a BCI-controlled human–robot interface for AAL	Direct communication path between the human brain and external actuator. BCI decodes user intention by a fast classifier; heterogeneous sensors (RGB-D cameras, sonar, and IR sensors) onboard a personal care robot are exploited to ensure correct actuation of user intent	4 participants (26 ± 1 years old)	84% accuracy in user intent classification—75% success rate in correct actuation execution
Ban et al. [132]	A multifunctional (10 actions) NAO v6 robot control system based on a multi-modal brain–machine interface (BMI) that fuses three signals: steady-state visual evoked potential (SSVEP), electrooculography (EOG), and gyroscope	Hybrid convolutional neural network—bidirectional—long short-term memory (CNN-Bi-LSTM) architecture based on attention modules, to extract temporal information from sequence data and enable classification	16 participants (19–25 years old, 8 males, 8 females)	93.78% accuracy in completion of complex tasks, by all participants—average response time of 3 s
Muñoz et al. [133]	A cost-effective rehabilitation system (named brain–kinect interface, BKI) based on videogames and multi-modal recordings of physiological signals (by a consumer-level EEG device + Kinect)	A gesture interaction module (using Kinect) and a BCI dynamically monitor physiological variables of patients while they are playing selected exergames for rehabilitation	Not specified for the proposed system validation—up to 700 patients involved in exergames only	Technical figures not reported—improvements in postural balance (+15% balance time increase) and range of motion (up to +18%)
Tidoni et al. [138]	Visual information and auditory feedback fused to improve the BCI-based remote control of a humanoid surrogate (HRP-2 robot) by people with sensorimotor disorders	An SSVEP BCI classifier decodes user intent that is associated with visually recognized objects captured by the embedded robot camera. Objects are paired with offline pre-defined tasks, triggered by the user SSVEP	14 healthy participants (25.8 ± 6.0 years old, 6 females) and 3 subjects who had suffered traumatic spinal cord injury (22–31 years old, 3 males)	Action-related feedback may improve subject information processing and decisions about when to start an action—no significant differences in task completion time and placing objects accuracy
Zhang et al. [139]	To improve the interaction and practical application of a prosthetic hand with a BCI system, by integrating augmented reality (AR) technology	An asynchronous pattern recognition algorithm, combining center extended canonical correlation analysis and support vector machine (Center-ECCA-SVM), is proposed, together with a deep learning object detection algorithm (YOLOv4) to improve the level of user interaction in 8 pre-defined control modes	12 participants	96.7% average stimulus pattern recognition accuracy—96.4% confidence in prosthetic hand real-time detection by YOLOv4 tiny model
Bellicha et al. [140]	An assist-as-needed sensor-based shared control (SC) method relying on the blending of BCI and depth-sensor-based control targeting mobility and manipulation needs in home settings	Shared control (SC): control shared between user command (retrieved from a control interface) and sensor-based control by features detected by robot-embedded sensors	A quadriplegic patient	Time to perform tasks and number of changes in mental tasks were reduced, unwanted actions avoided

5.2. Opportunities of RGB-D and BCI Fusion

Once integrated, RGB-D and BCI systems may operate under shared control or hybrid control paradigms [127]. In these models, the BCI-derived user intent, such as a high-level command or a desired direction of motion, is intelligently blended with sensor-based environmental awareness from RGB-D cameras. This would enable full adaptation of the assistance levels, ensuring that technology augments, rather than replaces, human action, which is crucial for user empowerment and long-term acceptance in AAL.

The synergistic integration of RGB-D cameras and non-intrusive BCIs could unlock a new generation of AAL applications, moving beyond conventional monitoring to provide more comprehensive, personalized, and proactive support. These novel applications would represent a significant leap in AAL, transforming solutions to empower people with greater independence, safety, and a higher quality of life. The fusion would enable systems that not only understand what is happening (RGB-D) but also why (user intent from BCI) and how to best intervene (environmental context joint user state), leading to highly personalized, proactive, and effective support. New possibilities for maintaining dignity and autonomy in aging would be enabled in domains such as adaptive robotic assistance (with assistive robots capable of performing activities of daily living with unprecedented precision and user-driven intent), personalized fall prevention and response (monitoring the user cognitive state through BCIs could enable fall prediction and proactive risk mitigation), and intuitive smart home control (with contextual information provided by RGB-D cameras and natural interaction supported by BCIs). Two possible examples are shown in Figure 5.

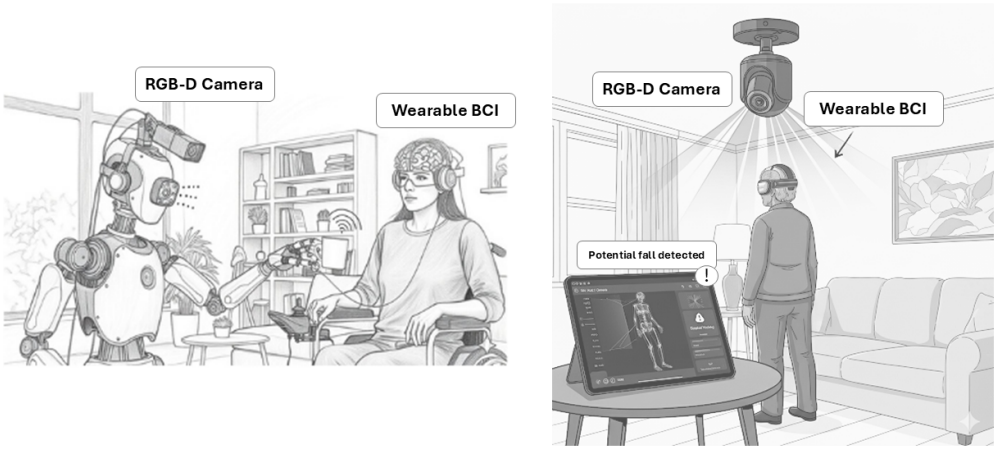

Figure 5. Examples of envisioned AAL scenarios supported by RGB-D camera and BCI integration: (**a**) adaptive robotic assistance, (**b**) personalized fall prevention.

5.3. Challenges in RGB-D and BCI Integration

Despite its promises, the integration of RGB-D cameras and BCIs also introduces significant challenges that must be addressed. They are similar to those pointed out by papers proposing multi-modal sensing applied to HAR, based on the integration of RGB-D cameras, wearable devices and inertial sensors [28], or RGB-D cameras and radar sensors [135], and several considerations are needed regarding multi-modal systems. First, real-time fusion of high-resolution video/depth data with high-dimensional EEG data is computationally demanding, and multi-modal sensor integration demands computational power and real-time processing capabilities. Indeed, RGB-D cameras generate

high-resolution RGB, depth and skeleton data, while non-invasive BCIs produce noisy real-time brain signals [141], especially in uncontrolled scenarios such as home living environments. As highlighted by Karim et al. [29], computational requirements of HAR systems vary significantly across diverse applications. Low latency and almost real-time processing are typically needed by systems designed for healthcare monitoring and smart homes. Modern approaches based on data mining and analytics may ensure robustness against noise and occlusions and computational efficiency as required by real-world AAL systems [142].

Additionally, there is a need for improved hardware and signal processing to make wearable BCIs more practical in daily living. Current EEG-based BCIs can be cumbersome or intrusive, and their signals are prone to artifacts from motion and muscle activity. On the one hand, miniaturized depth sensors have already been proposed and their performance evaluated [143]. On the other hand, innovations such as miniaturized, wireless BCI headsets with better signal quality are needed to ensure the system is both unobtrusive and reliable during routine activities. Progresses has been recently reported in the design and experimental validation of a minimally obtrusive platform for EEG recordings by a network of miniaturized wireless electrodes [144], which paves the way to future improvements and optimized design. Indeed, without reducing the burden of wearing and maintaining BCI devices, users may be reluctant to use them continuously, undermining the system effectiveness.

Privacy implications and data security issues cannot be overlooked, with both technologies (RGB-D cameras and BCIs) being deployed in intimate private spaces and used for monitoring sensitive activities. BCIs may raise even more profound privacy concerns than RGB-D cameras, as they collect exceptionally sensitive and personal neural data, capable of revealing intimate information about individual thoughts, emotions, and subconscious states [145]. While approaches to ensure security and privacy when using BCIs have been experimentally validated, as detailed in [146], new solutions are being proposed for the deployment of trusted BCI applications outside of controlled laboratory settings [147]. Mitigating privacy concerns through personalized, adaptive, and transparent design is essential to overcome psychological barriers, build long-term trust, and ensure the sustained engagement and effectiveness of these advanced assistive technologies.

User acceptance, comfort, and technology illiteracy are paramount factors for long-term successful deployment and widespread adoption of AAL systems [148,149]. Although the acceptance of camera-based monitoring has been analyzed in many studies in the literature [150], additional investigations would probably be needed for AAL systems that integrate BCIs, for which illiteracy refers to the inability of some individuals to effectively use BCIs despite their best efforts, thus preventing widespread adoption [151]. Variability in user capability means the system must be adaptable and cannot assume that one size fits all in terms of brain control performance. User-centric design and training will be necessary to mitigate these adoption barriers. Finally, properly structured ethical and legal frameworks for neural and visual (i.e., RGB-D) data will be of crucial importance to ensure responsible innovation and compliance, but also to build public trust, mitigate potential harms (e.g., discrimination, manipulation), and defend fundamental human rights, such as cognitive freedom [152,153]. In fact, current regulations for home monitoring and medical devices may not fully cover the nuanced risks coupled with the processing of neurophysiological data. Issues of consent, data ownership, and potential misuse (for instance, using brain data for purposes beyond the stated intent) are not yet fully resolved [154]. Without well-defined governance, there is a risk that user cognitive privacy or agency could be compromised. Developing guidelines

to protect user neural data and uphold rights like cognitive freedom is therefore essential as neurotechnology becomes more pervasive in the living environment.

5.4. Open Issues in RGB-D and BCI Integration

Looking ahead, future research should focus on several key directions to fully realize the envisioned benefits of RGB-D and BCI integration for HAR:

- *Lightweight multi-modal fusion algorithms:*There is a pressing need to develop more efficient data fusion and ML/DL algorithms to handle multi-stream RGB-D and EEG data in real time, without excessive computational load. In particular, this can include the design of optimized and efficient feature extraction methodologies [155], that, however, must take into account the heterogeneous nature of the input data, i.e., images and biosignal time series. In this case, adopting multi-domain features and advanced methodologies for feature selection and information fusion [156] represent attractive possibilities for the optimization of the feature extraction pipeline, reducing the computational burden of this processing step. However, it is worth noticing that RGB-D images require the major part of the computational resources in terms of data processing. Thus, other viable solutions can be found by considering techniques for compressing images and removing non-necessary or redundant information [99]. On the other hand, there is also the need for exploring computational solutions, optimized neural networks, and advanced signal processing techniques to ensure the system can run continuously in a home environment (potentially on embedded hardware) without sacrificing accuracy. This can be achieved by leveraging tiny learning models specifically designed for mobile or low-resource devices [157], and by relying on the growing computational power of newer devices. Another approach to deal with these issues is to transfer part of the computational demands to the cloud or shared services, or to adopt edge computing solutions within an optimized framework [158].
- *Robustness in unconstrained environments:* To be practical in daily living, HAR systems must be resilient to the messy, unpredictable nature of real homes. Future studies should emphasize improving the robustness and accuracy of activity recognition under real-world conditions, such as varying lighting, background clutter, the presence of multiple subjects, or user movement, as well as coping with EEG noise from muscle activity or electrical interference. The latter can be addressed by relying on lightweight but robust recently proposed algorithmic solutions [159]. This may involve creating large and diverse datasets for training, using adaptive algorithms that can learn from a user routine, or integrating additional context (e.g., time of day, habitual patterns) to reduce false detections. This point represents one of the major issues for the deployment of image-based, hybrid HAR solutions within real environments. Indeed, as outlined also in previous sections, in practice it is not possible to enforce the presence of a single user within a certain environment. In practical applications, it is also likely that different users would perform different activities to be recognized. To address the complexities imposed by a multi-person scenario, the usage of multiple cameras can be a viable solution [103], while the recognition of multiple activities performed at the same time would require the usage of tailored and specific processing procedures [104]. Real environments, with their complex backgrounds, can often lead to occlusion problems that can be faced by relying on multiple-angle views of the same scene [160]. Variable lighting conditions, which can heavily affect the outcome of HAR under consideration, could also be successfully handled through specific post-processing procedures, nowadays also tailored to limited computational resources, thus reducing their impact on the demands of the overall recognition architecture [161].

- *More user-friendly BCI devices:* Engineering advances are needed to design BCIs that are comfortable, unobtrusive, and easy to operate for non-expert users. This could mean wireless, miniaturized EEG systems with dry electrodes (avoiding lengthy setup), longer battery life, and auto-calibration features. Improving signal quality through better sensors or algorithms (to filter out artifacts) will also enhance reliability. By making the BCI hardware invisible and hassle-free, users are more likely to wear it regularly, which is crucial for continuous HAR in AAL. In particular, an attractive solution for the envisioned framework can be the usage of in-ear EEG-based BCI [162], which has been conceived in order to offer an alternative way for recording brain activity by using probes placed on the ear and in the ear canal. Even though the EEG acquired with this kind of sensing technology does not fully meet all the characteristics provided by scalp EEG, in-ear technology represents a suitable compromise for developing lightweight, unobtrusive, and comfortable BCI applications. Furthermore, in-ear EEG systems have also been integrated within large body-area networks for health monitoring [163], thus pointing out the combination of this kind of technology with RGB-D systems as an actual possible solution for RGB-D and BCI fusion.
- *User acceptance and usability studies:* Going forward, researchers should extensively study how target user groups (e.g., older adults, people with disabilities) perceive and interact with these integrated systems. Early involvement of end-users through participatory design can identify usability pain points and preferences, ensuring the solutions truly meet user needs. In fact, further investigations into user acceptance of BCI-augmented AAL systems are needed [164], as prior work has mainly examined acceptance of camera-based monitoring alone. Understanding factors that influence trust, such as system transparency, feedback provided to the user, and perceived benefits, will be vital. Strategies to improve acceptance might include personalized adaptation (tuning the system responses to individual comfort levels), training and onboarding programs to help users get comfortable with BCIs, and integrating privacy-respecting options (for example, allowing users to easily pause or control data collection). User acceptance and usability could be further enhanced by developing user-independent architectures for BCI applications in order to avoid the need for retraining the model when used on unseen, new individuals. For achieving this objective, some promising solutions are currently available, allowing also to address the cross-session issue [165]. By applying such kind of processing on EEG data, it would also be possible to lower computational demands and time needed for calibration, since data from other users or from different sessions could be used in the initialization phase, enhancing also the usability and acceptability of the BCI usage in a real-life scenario.
- *Privacy protection and ethical frameworks:* In tandem with technical improvements, there must be a concerted effort to establish comprehensive ethical and legal guidelines for deploying these technologies in private homes. Future work should engage ethicists, legal experts, and policymakers to develop standards for data handling that ensure strict privacy, security, and user consent. This includes determining how neural and video data should be stored and used and setting limits to prevent any form of data abuse [153]. Researchers have noted that mitigating privacy concerns through personalized, transparent system design is essential to overcome user barriers and build long-term trust, and that properly structured frameworks will be crucial for responsible innovation in this domain [154]. By embedding ethical considerations into the design (such as on-device data processing to keep raw data private or providing users with clear control over their information), developers can foster greater user confidence and societal acceptance of these advanced AAL solutions [166]. It is worth noticing that for addressing privacy issues, the usage of skeleton data ex-

tracted from RGB-D images appears convenient, since this kind of representation modality discards any user identity information, also making the background scene not identifiable. However, as discussed also in the previous sections, the usage of a multi-modal approach, where skeleton and RGB-D modalities are combined, provides significant advantages in terms of HAR results. Also, multi-modal schemes proved to be less vulnerable under adversarial attack, thus providing more robust applications in terms of safety and privacy concerns [87,167]. Therefore, balancing ethical concerns and technical outcomes still remains an open issue that should be evaluated depending on settings, environments, target users, and desired results for each specific HAR application.

6. Conclusions

Starting from the overview about the use of RGB-D cameras and non-invasive BCIs as single sensing modalities in HAR for AAL, this perspective shed light on the opportunities, challenges, and potential disruptive innovations that could stem from an integrated dual-perception framework, underpinning the transformative potential of RGB-D+BCI systems for AAL. Fusing environmental and cognitive data streams, such systems can infer not only what a user is performing, but also why and how they are performing it. This kind of enriched context is simply unattainable with vision or BCI alone. It is precisely the integration of these two dimensions that enables the development of innovative solutions to improve quality of life in AAL, aligning with the core goal of AAL technologies to support independent living and well-being. In essence, the system becomes both situation- and user-aware, which represents a paradigm shift in HAR: from reactive monitoring of observable actions to proactive understanding of user needs and states, enabling possible anticipatory actions. This approach promises to enhance user independence, safety, and quality of life, which are core goals of AAL, by allowing proactive and situation-specific interventions.

Integrating RGB-D environmental perception with non-invasive BCI-derived insight into the user state constitutes a powerful dual-perspective approach to HAR. This perspective paper has articulated how such synergy offers a more complete picture of daily activities, effectively bridging the gap between external actions and internal experiences of a person. By focusing on the combined analysis of what is happening around the user and what is happening within the user's mind, future HAR systems can become significantly more context-aware, personalized, and responsive. The concept of integrated dual perception highlighted here is not only a novel contribution to the field of HAR but also a promising foundation for next-generation AAL technologies that aim to improve safety, independence, and quality of life for aging or vulnerable populations.

Continued research and development along the lines discussed, from algorithmic innovations to privacy-preserving and human-centered design, will yield intelligent living environments capable of not just monitoring their inhabitants but truly understanding and anticipating their needs. This integration of external and internal sensing thus serves as a key enabler in achieving the full potential of active and assisted living. In fact, a key insight that arises from this envisioned dual-modality approach is that the fusion of these two dimensions, namely environmental context and user state, provides a far more holistic understanding of human activities than either modality alone. Their integration enables context-aware interpretations of behavior that were not previously possible.

Author Contributions: Conceptualization, G.I.; analysis methodology, G.I., A.M. (Alessandro Mengarelli) and S.I.; formal analysis on data, G.I.; investigation, G.I. and A.M. (Alessandro Mengarelli); resources, G.I.; writing—original draft preparation, G.I., A.M. (Alessandro Mengarelli), S.I., A.M. (Andrea Monteriù) and S.S.; writing—review and editing, G.I., A.M. (Alessandro Mengarelli), S.I., A.M. (Andrea Monteriù) and S.S.; supervision, G.I.; funding acquisition, A.M. (Andrea Monteriù) and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has received funding from the project Vitality—Project Code ECS00000041, CUP I33C22001330007; under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5—Creation and strengthening of "innovation ecosystems", construction of "territorial leaders in R&D"—Innovation Ecosystems—Project "Innovation, digitalization and sustainability for the diffused economy in Central Italy VITALITY" Call for tender No. 3277 of 30/12/2021, and Concession Decree No. 0001057.23-06-2022 of the Italian Ministry of University funded by the European Union—NextGenerationEU.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Colantonio, S.; Aleksic, S.; Calleja Agius, J.; Camilleri, K.P.; Čartolovni, A.; Climent-Pérez, P.; Cristina, S.; Despotovic, V.; Ekenel, H.K.; Erakin, M.E.; et al. A Historical View of Active Assisted Living. In *Privacy-Aware Monitoring for Assisted Living: Ethical, Legal, and Technological Aspects of Audio- and Video-Based AAL Solutions*; Salah, A.A., Colonna, L., Florez-Revuelta, F., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 3–44. [CrossRef]
2. Eurostat. Population Structure and Ageing. 2025. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing (accessed on 10 July 2025).
3. Periša, M.; Teskera, P.; Cvitić, I.; Grgurević, I. Empowering People with Disabilities in Smart Homes Using Predictive Informing. *Sensors* **2025**, *25*, 284. [CrossRef]
4. Cicirelli, G.; Marani, R.; Petitti, A.; Milella, A.; D'Orazio, T. Ambient Assisted Living: A Review of Technologies, Methodologies and Future Perspectives for Healthy Aging of Population. *Sensors* **2021**, *21*, 3549. [CrossRef] [PubMed]
5. Choudhury, N.A.; Soni, B. In-depth analysis of design & development for sensor-based human activity recognition system. *Multimed. Tools Appl.* **2023**, *83*, 73233–73272. [CrossRef]
6. Newaz, N.T.; Hanada, E. The Methods of Fall Detection: A Literature Review. *Sensors* **2023**, *23*, 5212. [CrossRef] [PubMed]
7. Seo, K.J.; Lee, J.; Cho, J.E.; Kim, H.; Kim, J.H. Gait Environment Recognition Using Biomechanical and Physiological Signals with Feed-Forward Neural Network: A Pilot Study. *Sensors* **2025**, *25*, 4302. [CrossRef]
8. Iadarola, G.; Mengarelli, A.; Spinsante, S. Classification of Physical Fatigue on Heart Rate by Wearable Devices. In Proceedings of the 2025 IEEE Medical Measurements & Applications (MeMeA), Chania, Greece, 28–30 May 2025; pp. 1–6. [CrossRef]
9. Pavaiyarkarasi, R.; Paulraj, D. A Concept to Reality: New Horizons in Alzheimer's Assistive Devices. In Proceedings of the 2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 21–23 May 2025; pp. 1264–1271. [CrossRef]
10. Iadarola, G.; Mengarelli, A.; Crippa, P.; Fioretti, S.; Spinsante, S. A Review on Assisted Living Using Wearable Devices. *Sensors* **2024**, *24*, 7439. [CrossRef]
11. Salem, Z.; Weiss, A.P. Improved Spatiotemporal Framework for Human Activity Recognition in Smart Environment. *Sensors* **2023**, *23*, 132. [CrossRef]
12. Ceron, J.D.; López, D.M.; Kluge, F.; Eskofier, B.M. Framework for Simultaneous Indoor Localization, Mapping, and Human Activity Recognition in Ambient Assisted Living Scenarios. *Sensors* **2022**, *22*, 3364. [CrossRef]
13. Chimamiwa, G.; Giaretta, A.; Alirezaie, M.; Pecora, F.; Loutfi, A. Are Smart Homes Adequate for Older Adults with Dementia? *Sensors* **2022**, *22*, 4254. [CrossRef]
14. Vasylkiv, Y.; Neshati, A.; Sakamoto, Y.; Gomez, R.; Nakamura, K.; Irani, P. Smart home interactions for people with reduced hand mobility using subtle EMG-signal gestures. In *Improving Usability, Safety and Patient Outcomes with Health Information Technology*; IOS Press: London, UK, 2019; pp. 436–443. [CrossRef]

15. De Venuto, D.; Annese, V.F.; Sangiovanni-Vincentelli, A.L. The ultimate IoT application: A cyber-physical system for ambient assisted living. In Proceedings of the 2016 IEEE International Symposium on Circuits and Systems (ISCAS), Montréal, QC, Canada, 22–25 May 2016; pp. 2042–2045. [CrossRef]
16. Scattolini, M.; Tigrini, A.; Verdini, F.; Iadarola, G.; Spinsante, S.; Fioretti, S.; Burattini, L.; Mengarelli, A. Leveraging inertial information from a single IMU for human daily activity recognition. In Proceedings of the 2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Eindhoven, The Netherlands, 26–28 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
17. Iadarola, G.; Scoccia, C.; Spinsante, S.; Rossi, L.; Monteriù, A. An Overview on Current Technologies for Assisted Living. In Proceedings of the 2024 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv), Chania, Greece, 12–14 June 2024; pp. 190–195. [CrossRef]
18. Villarroel F, M.J.; Villarroel G, C.H. Wireless smart environment in Ambient Assisted Living for people that suffer from cognitive disabilities. *Ingeniare. Rev. Chil. IngenierÃa* **2014**, *22*, 158–168. [CrossRef]
19. Caiado, F.; Ukolov, A. The history, current state and future possibilities of the non-invasive brain computer interfaces. *Med. Nov. Technol. Devices* **2025**, *25*, 100353. [CrossRef]
20. Kumar Gouda, S.; Choudhry, A.; Satpathy, S.P.; Shukla, K.M.; Dash, A.K.; Pasayat, A.K. Integration of EEG-based BCI technology in IoT enabled smart home environment: An in-depth comparative analysis on human-computer interaction techniques. *Expert Syst. Appl.* **2025**, *294*, 128730. [CrossRef]
21. Iadarola, G.; Cosoli, G.; Scalise, L.; Spinsante, S. Unsupervised Learning of Physical Effort: Proposal of a simple metric for wearable devices. In Proceedings of the 2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Chemnitz, Germany, 19–22 May 2025; pp. 1–6. [CrossRef]
22. Kang, H.; Lee, C.; Kang, S.J. A smart device for non-invasive ADL estimation through multi-environmental sensor fusion. *Sci. Rep.* **2023**, *13*, 17246. [CrossRef] [PubMed]
23. Shaikh, M.B.; Chai, D. Rgb-d data-based action recognition: A review. *Sensors* **2021**, *21*, 4246. [CrossRef] [PubMed]
24. Leelaarporn, P.; Wachiraphan, P.; Kaewlee, T.; Udsa, T.; Chaisaen, R.; Choksatchawathi, T.; Laosirirat, R.; Lakhan, P.; Natnithikarat, P.; Thanontip, K.; et al. Sensor-Driven Achieving of Smart Living: A Review. *IEEE Sens. J.* **2021**, *21*, 10369–10391. [CrossRef]
25. Kristoffersson, A.; Lindén, M. A systematic review of wearable sensors for monitoring physical activity. *Sensors* **2022**, *22*, 573. [CrossRef]
26. Oyibo, K.; Wang, K.; Morita, P.P. Using Smart Home Technologies to Promote Physical Activity Among the General and Aging Populations: Scoping Review. *J. Med. Internet Res.* **2023**, *25*, e41942. [CrossRef]
27. Al Farid, F.; Bari, A.; Miah, A.S.M.; Mansor, S.; Uddin, J.; Kumaresan, S.P. A Structured and Methodological Review on Multi-View Human Activity Recognition for Ambient Assisted Living. *J. Imaging* **2025**, *11*, 182. [CrossRef]
28. Qi, W.; Xu, X.; Qian, K.; Schuller, B.W.; Fortino, G.; Aliverti, A. A Review of AIoT-Based Human Activity Recognition: From Application to Technique. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 2425–2438. [CrossRef]
29. Karim, M.; Khalid, S.; Lee, S.; Almutairi, S.; Namoun, A.; Abohashrh, M. Next Generation Human Action Recognition: A Comprehensive Review of State-of-the-Art Signal Processing Techniques. *IEEE Access* **2025**, *13*, 135609–135633. [CrossRef]
30. Wang, C.; Jiang, W.; Yang, K.; Yu, D.; Newn, J.; Sarsenbayeva, Z.; Goncalves, J.; Kostakos, V. Electronic monitoring systems for hand hygiene: Systematic review of technology. *J. Med. Internet Res.* **2021**, *23*, e27880. [CrossRef]
31. Chun, K.S.; Sanders, A.B.; Adaimi, R.; Streeper, N.; Conroy, D.E.; Thomaz, E. Towards a generalizable method for detecting fluid intake with wrist-mounted sensors and adaptive segmentation. In Proceedings of the 24th International Conference on Intelligent User Interfaces, New York, NY, USA, 17–20 March 2019; pp. 80–85. [CrossRef]
32. Sabry, F.; Eltaras, T.; Labda, W.; Hamza, F.; Alzoubi, K.; Malluhi, Q. Towards on-device dehydration monitoring using machine learning from wearable device's data. *Sensors* **2022**, *22*, 1887. [CrossRef] [PubMed]
33. Moccia, S.; Solbiati, S.; Khornegah, M.; Bossi, F.F.; Caiani, E.G. Automated classification of hand gestures using a wristband and machine learning for possible application in pill intake monitoring. *Comput. Methods Programs Biomed.* **2022**, *219*, 106753. [CrossRef] [PubMed]
34. *JCGM 100:2008*; GUM 1995 with Minor Corrections. Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement. Sèvres International Bureau of Weights and Measures (BIPM): Sèvres, France, 2008.
35. Offermann, J.; Wilkowska, W.; Poli, A.; Spinsante, S.; Ziefle, M. Acceptance and Preferences of Using Ambient Sensor-Based Lifelogging Technologies in Home Environments. *Sensors* **2021**, *21*, 8297. [CrossRef] [PubMed]
36. Ankalaki, S. Simple to Complex, Single to Concurrent Sensor-Based Human Activity Recognition: Perception and Open Challenges. *IEEE Access* **2024**, *12*, 93450–93486. [CrossRef]
37. Vrigkas, M.; Nikou, C.; Kakadiaris, I.A. A review of human activity recognition methods. *Front. Robot. AI* **2015**, *2*, 28. [CrossRef]
38. Yeung, L.F.; Yang, Z.; Cheng, K.C.; Du, D.; Tong, R.K. Effects of camera viewing angles on tracking kinematic gait patterns using Azure Kinect, Kinect v2 and Orbbec Astra Pro v2. *Gait Posture* **2021**, *87*, 19 – 26. [CrossRef]

39. Carfagni, M.; Furferi, R.; Governi, L.; Santarelli, C.; Servi, M.; Uccheddu, F.; Volpe, Y. Metrological and Critical Characterization of the Intel D415 Stereo Depth Camera. *Sensors* **2019**, *19*, 489. [CrossRef]
40. Tölgyessy, M.; Dekan, M.; Chovanec, Ľ.; Hubinský, P. Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2. *Sensors* **2021**, *21*, 413. [CrossRef]
41. Kurillo, G.; Hemingway, E.; Cheng, M.L.; Cheng, L. Evaluating the Accuracy of the Azure Kinect and Kinect v2. *Sensors* **2022**, *22*, 2469. [CrossRef]
42. Burger, L.; Burger, L.; Sharan, L.; Karl, R.; Wang, C.; Karck, M.; De Simone, R.; Wolf, I.; Romano, G.; Engelhardt, S. Comparative evaluation of three commercially available markerless depth sensors for close-range use in surgical simulation. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 1109–1118. [CrossRef]
43. Büker, L.; Quinten, V.; Hackbarth, M.; Hellmers, S.; Diekmann, R.; Hein, A. How the Processing Mode Influences Azure Kinect Body Tracking Results. *Sensors* **2023**, *23*, 878. [CrossRef]
44. Gonzalez-Jorge, H.; Riveiro, B.; Vazquez-Fernandez, E.; Martínez-Sánchez, J.; Arias, P. Metrological evaluation of Microsoft Kinect and Asus Xtion sensors. *Measurement* **2013**, *46*, 1800–1806. [CrossRef]
45. Haider, A.; Hel-Or, H. What Can We Learn from Depth Camera Sensor Noise? *Sensors* **2022**, *22*, 5448. [CrossRef]
46. OpenPR—Worldwide Public Relations. Depth Camera Market Share, Trends Analysis 2031 by Key Vendors—Texas Instruments, STMicroelectronics, PMD Technologies, Infineon, PrimeSense (Apple). 2025. Available online: https://www.openpr.com/news/3903319/latest-size-depth-camera-market-share-trends-analysis-2031# (accessed on 10 July 2025).
47. Park, S.; Park, J.; Al-Masni, M.A.; Al-Antari, M.A.; Uddin, M.Z.; Kim, T.S. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. *Procedia Comput. Sci.* **2016**, *100*, 78–84. [CrossRef]
48. Raj, R.; Kos, A. An improved human activity recognition technique based on convolutional neural network. *Sci. Rep.* **2023**, *13*, 22581. [CrossRef] [PubMed]
49. Himeur, Y.; Al-Maadeed, S.; Kheddar, H.; Al-Maadeed, N.; Abualsaud, K.; Mohamed, A.; Khattab, T. Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105698. [CrossRef]
50. Huang, X.; Cai, Z. A review of video action recognition based on 3D convolution. *Comput. Electr. Eng.* **2023**, *108*, 108713. [CrossRef]
51. Maqsood, R.; Bajwa, U.I.; Saleem, G.; Raza, R.H.; Anwar, M.W. Anomaly recognition from surveillance videos using 3D convolution neural network. *Multimed. Tools Appl.* **2021**, *80*, 18693–18716. [CrossRef]
52. Muhammad, K.; Ullah, H.; Obaidat, M.S.; Ullah, A.; Munir, A.; Sajjad, M.; De Albuquerque, V.H.C. AI-driven salient soccer events recognition framework for next-generation IoT-enabled environments. *IEEE Internet Things J.* **2021**, *10*, 2202–2214. [CrossRef]
53. Wu, H.; Ma, X.; Li, Y. Multi-level channel attention excitation network for human action recognition in videos. *Signal Process. Image Commun.* **2023**, *114*, 116940. [CrossRef]
54. Zong, M.; Wang, R.; Ma, Y.; Ji, W. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition. *Appl. Soft Comput.* **2023**, *132*, 109884. [CrossRef]
55. Hussain, A.; Khan, S.U.; Khan, N.; Shabaz, M.; Baik, S.W. AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107218. [CrossRef]
56. Mitsuzumi, Y.; Irie, G.; Kimura, A.; Nakazawa, A. Phase randomization: A data augmentation for domain adaptation in human action recognition. *Pattern Recognit.* **2024**, *146*, 110051. [CrossRef]
57. Yin, Y.; Yang, Z.; Hu, H.; Wu, X. Universal multi-source domain adaptation for image classification. *Pattern Recognit.* **2022**, *121*, 108238. [CrossRef]
58. Karthika, S.; Jane, Y.N.; Nehemiah, H.K. Spatio temporal 3D skeleton kinematic joint point classification model for human activity recognition. *J. Vis. Commun. Image Represent.* **2025**, *110*, 104471. [CrossRef]
59. Jo, B.; Kim, S. Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices. *Trait. Signal* **2022**, *39*, 119. [CrossRef]
60. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019. [CrossRef]
61. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef]
62. Smaira, L.; Carreira, J.; Noland, E.; Clancy, E.; Wu, A.; Zisserman, A. A short note on the kinetics-700-2020 human action dataset. *arXiv* **2020**, arXiv:2010.10864. [CrossRef]
63. Guo, D.; Li, K.; Hu, B.; Zhang, Y.; Wang, M. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 6238–6252. [CrossRef]
64. Li, Q.; Xie, X.; Zhang, J.; Shi, G. Recognizing human-object interactions in videos with the supervision of natural language. *Neural Netw.* **2025**, *190*, 107606. [CrossRef]

65. Koppula, H.S.; Gupta, R.; Saxena, A. Learning human activities and object affordances from rgb-d videos. *Int. J. Robot. Res.* **2013**, *32*, 951–970. [CrossRef]
66. Materzynska, J.; Xiao, T.; Herzig, R.; Xu, H.; Wang, X.; Darrell, T. Something-else: Compositional action recognition with spatial-temporal interaction networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1049–1059. [CrossRef]
67. Elnady, M.; Abdelmunim, H.E. A novel YOLO LSTM approach for enhanced human action recognition in video sequences. *Sci. Rep.* **2025**, *15*, 17036. [CrossRef] [PubMed]
68. Soomro, K.; Zamir, A.R.; Shah, M. A dataset of 101 human action classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402. [CrossRef]
69. 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; IEEE: Piscataway, NJ, USA, 2004; Volume 3, pp. 32–36. [CrossRef]
70. 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]
71. Weinland, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257. [CrossRef]
72. Chen, D.; Chen, M.; Wu, P.; Wu, M.; Zhang, T.; Li, C. Two-stream spatio-temporal GCN-transformer networks for skeleton-based action recognition. *Sci. Rep.* **2025**, *15*, 4982. [CrossRef]
73. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef]
74. Liu, C.; Hu, Y.; Li, Y.; Song, S.; Liu, J. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, Mountain View, CA, USA, 23 October 2017; pp. 1–8. [CrossRef]
75. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Berkeley MHAD: A comprehensive multimodal human action database. In Proceedings of the 2013 IEEE workshop on applications of computer vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 53–60. [CrossRef]
76. Ranieri, C.M.; MacLeod, S.; Dragone, M.; Vargas, P.A.; Romero, R.A.F. Activity recognition for ambient assisted living with videos, inertial units and ambient sensors. *Sensors* **2021**, *21*, 768. [CrossRef]
77. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2649–2656. [CrossRef]
78. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International conference on image processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 168–172. [CrossRef]
79. Das, S.; Dai, R.; Koperski, M.; Minciullo, L.; Garattoni, L.; Bremond, F.; Francesca, G. Toyota smarthome: Real-world activities of daily living. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 833–842. [CrossRef]
80. 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. [CrossRef]
81. 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; IEEE: Piscataway, NJ, USA, 2011; pp. 2556–2563. [CrossRef]
82. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5842–5850. [CrossRef]
83. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [CrossRef]
84. 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; IEEE: Piscataway, NJ, USA, 2009; pp. 1996–2003. [CrossRef]
85. 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; IEEE: Piscataway, NJ, USA, 2012; pp. 1290–1297. [CrossRef]
86. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3192–3199. [CrossRef]
87. Yue, R.; Tian, Z.; Du, S. Action recognition based on RGB and skeleton data sets: A survey. *Neurocomputing* **2022**, *512*, 287–306. [CrossRef]

88. Bruce, X.; Liu, Y.; Zhang, X.; Zhong, S.h.; Chan, K.C. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3522–3538. [CrossRef]
89. Kumar, R.; Kumar, S. Multi-view multi-modal approach based on 5s-cnn and bilstm using skeleton, depth and rgb data for human activity recognition. *Wirel. Pers. Commun.* **2023**, *130*, 1141–1159. [CrossRef]
90. Batool, M.; Alotaibi, M.; Alotaibi, S.R.; AlHammadi, D.A.; Jamal, M.A.; Jalal, A.; Lee, B. Multimodal Human Action Recognition Framework using an Improved CNNGRU Classifier. *IEEE Access* **2024**, *12*, 158388–158406. [CrossRef]
91. Tian, Y.; Chen, W. MEMS-based human activity recognition using smartphone. In Proceedings of the 2016 35th Chinese Control Conference (CCC), Chengdu, China, 27–29 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 3984–3989. [CrossRef]
92. William, P.; Lanke, G.R.; Bordoloi, D.; Shrivastava, A.; Srivastavaa, A.P.; Deshmukh, S.V. Assessment of human activity recognition based on impact of feature extraction prediction accuracy. In Proceedings of the 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 9–11 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [CrossRef]
93. Mekruksavanich, S.; Jantawong, P.; Jitpattanakul, A. Deep learning approaches for har of daily living activities using imu sensors in smart glasses. 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), Phitsanulok, Thailand, 22–25 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 474–478. [CrossRef]
94. Jantawong, P.; Hnoohom, N.; Jitpattanakul, A.; Mekruksavanich, S. A lightweight deep learning network for sensor-based human activity recognition using imu sensors of a low-power wearable device. In Proceedings of the 2021 25th International Computer Science and Engineering Conference (ICSEC), Chiang Rai, Thailand, 18–20 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 459–463. [CrossRef]
95. Liu, D.; Meng, F.; Mi, J.; Ye, M.; Li, Q.; Zhang, J. SAM-Net: Semantic-assisted multimodal network for action recognition in RGB-D videos. *Pattern Recognit.* **2025**, *168*, 111725. [CrossRef]
96. Song, S.; Liu, J.; Li, Y.; Guo, Z. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Trans. Image Process.* **2020**, *29*, 3957–3969. [CrossRef]
97. Liu, D.; Meng, F.; Xia, Q.; Ma, Z.; Mi, J.; Gan, Y.; Ye, M.; Zhang, J. Temporal cues enhanced multimodal learning for action recognition in RGB-D videos. *Neurocomputing* **2024**, *594*, 127882. [CrossRef]
98. Demir, U.; Rawat, Y.S.; Shah, M. Tinyvirat: Low-resolution video action recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 7387–7394. [CrossRef]
99. Wu, C.Y.; Zaheer, M.; Hu, H.; Manmatha, R.; Smola, A.J.; Krähenbühl, P. Compressed video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6026–6035. [CrossRef]
100. Fan, L.; Buch, S.; Wang, G.; Cao, R.; Zhu, Y.; Niebles, J.C.; Fei-Fei, L. Rubiksnet: Learnable 3d-shift for efficient video action recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 505–521. [CrossRef]
101. Liu, G.; Qian, J.; Wen, F.; Zhu, X.; Ying, R.; Liu, P. Action recognition based on 3d skeleton and rgb frame fusion. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 258–264. [CrossRef]
102. Kim, S.; Yun, K.; Park, J.; Choi, J.Y. Skeleton-based action recognition of people handling objects. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 61–70. [CrossRef]
103. Phang, J.T.S.; Lim, K.H. Real-time multi-camera multi-person action recognition using pose estimation. In Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, Da Lat, Vietnam, 25–28 January 2019; pp. 175–180. [CrossRef]
104. Gilbert, A.; Illingworth, J.; Bowden, R. Fast realistic multi-action recognition using mined dense spatio-temporal features. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 925–931. [CrossRef]
105. Angelini, F.; Fu, Z.; Long, Y.; Shao, L.; Naqvi, S.M. 2D pose-based real-time human action recognition with occlusion-handling. *IEEE Trans. Multimed.* **2019**, *22*, 1433–1446. [CrossRef]
106. Papadopoulos, G.T.; Daras, P. Human action recognition using 3d reconstruction data. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 1807–1823. [CrossRef]
107. Huynh-The, T.; Hua, C.H.; Kim, D.S. Encoding pose features to images with data augmentation for 3-D action recognition. *IEEE Trans. Ind. Inform.* **2019**, *16*, 3100–3111. [CrossRef]
108. Su, K.; Liu, X.; Shlizerman, E. Predict & cluster: Unsupervised skeleton based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, 14–19 June 2020; pp. 9631–9640. [CrossRef]

109. Hochberg, L.; Bacher, D.; Jarosiewicz, B.; Masse, N.; Simeral, J.; Vogel, J.; Haddadin, S.; Liu, J.; Cash, S.; van der Smagt, P.; et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* **2012**, *485*, 372–375. [CrossRef]
110. Ang, K.; Chua, K.; Phua, K.S.; Wang, C.; Chin, Z.; Kuah, C.; Low, W.; Guan, C. A Randomized Controlled Trial of EEG-Based Motor Imagery Brain-Computer Interface Robotic Rehabilitation for Stroke. *Clin. EEG Neurosci. Off. J. EEG Clin. Neurosci. Soc. (ENCS)* **2015**, *46*, 310–320. [CrossRef] [PubMed]
111. Nchekwube, D.; Iarlori, S.; Monteriù, A. An assistive robot in healthcare scenarios requiring monitoring and rules regulation: Exploring Pepper use case. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkey, 5–8 December 2023; IEEE: Piscataway, NJ, USA, 2023; Volume 12, pp. 4812–4819. [CrossRef]
112. Wolpaw, J.R.; Birbaumer, N.; McFarland, D.J.; Pfurtscheller, G.; Vaughan, T.M. Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* **2002**, *113*, 767–791. [CrossRef] [PubMed]
113. Omer, K.; Vella, F.; Ferracuti, F.; Freddi, A.; Iarlori, S.; Monteriù, A. Mental Fatigue Evaluation for Passive and Active BCI Methods for Wheelchair-Robot During Human-in-the-Loop Control. In Proceedings of the 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), Milan, Italy, 25–27 December 2023; IEEE: Piscataway, NJ, USA, 2023; Volume 10, pp. 787–792. [CrossRef]
114. Moaveninejad, S.; D'Onofrio, V.; Tecchio, F.; Ferracuti, F.; Iarlori, S.; Monteriù, A.; Porcaro, C. Fractal Dimension as a discriminative feature for high accuracy classification in motor imagery EEG-based brain-computer interface. *Comput. Methods Programs Biomed.* **2024**, *244*, 107944. [CrossRef] [PubMed]
115. Wolpaw, J.; Wolpaw, E.W. *Brain Computer Interfaces: Principles and Practice*; Oxford University Press: Oxford, UK, 2012. [CrossRef]
116. Streitz, N.A. From Human–Computer Interaction to Human–Environment Interaction: Ambient Intelligence and the Disappearing Computer. In Proceedings of the Universal Access in Ambient Intelligence Environments, Austria, Salzburg, 28 January 2007; Stephanidis, C., Pieper, M., Eds., Springer: Berlin/Heidelberg, Germany, 2007; pp. 3–13.
117. Makeig, S.; Kothe, C.; Mullen, T.; Bigdely-Shamlo, N.; Zhang, Z.; Kreutz-Delgado, K. Evolving Signal Processing for Brain-Computer Interfaces. *Proc. IEEE* **2012**, *100*, 1567–1584. [CrossRef]
118. Rao, R.P.N. *Brain-Computer Interfacing: An Introduction*; Cambridge University Press: Cambridge, UK, 2013.
119. Netzer, E.; Geva, A. Human-in-the-loop active learning via brain computer interface. *Ann. Math. Artif. Intell.* **2020**, *88*, 1191–1205. [CrossRef]
120. Ferracuti, F.; Freddi, A.; Iarlori, S.; Monteriù, A.; Omer, K.I.M.; Porcaro, C. A human-in-the-loop approach for enhancing mobile robot navigation in presence of obstacles not detected by the sensory set. *Front. Robot. AI* **2022**, *9*, 2022. [CrossRef]
121. Gemborn Nilsson, M.; Tufvesson, P.; Heskebeck, F.; Johansson, M. An open-source human-in-the-loop BCI research framework: Method and design. *Front. Hum. Neurosci.* **2023**, *17*, 2023. [CrossRef]
122. Venot, T.; Desbois, A.; Corsi, M.C.; Hugueville, L.; Saint-Bauzel, L.; De Vico Fallani, F. Intentional binding for noninvasive BCI control. *J. Neural Eng.* **2024**, *21*, 046026. [CrossRef]
123. Ji, Z.; Liu, Q.; Xu, W.; Yao, B.; Liu, J.; Zhou, Z. A Closed-Loop Brain-Computer Interface with Augmented Reality Feedback for Industrial Human-Robot Collaboration. *Int. J. Adv. Manuf. Technol.* **2023**, *124*, 3083–3098. [CrossRef]
124. Aydarkhanov, R.; Uščumlić, M.; Chavarriaga, R.; Gheorghe, L.; del R Millán, J. Closed-loop EEG study on visual recognition during driving. *J. Neural Eng.* **2021**, *18*, 026010. [CrossRef]
125. Gao, H.; Luo, L.; Pi, M.; Li, Z.; Li, Q.; Zhao, K.; Huang, J. EEG-Based Volitional Control of Prosthetic Legs for Walking in Different Terrains. *IEEE Trans. Autom. Sci. Eng.* **2021**, *18*, 530–540. [CrossRef]
126. Omer, K.; Ferracuti, F.; Freddi, A.; Iarlori, S.; Vella, F.; Monteriù, A. Real-Time Mobile Robot Obstacles Detection and Avoidance Through EEG Signals. *Brain Sci.* **2025**, *15*, 359. [CrossRef] [PubMed]
127. Xu, B.; Li, W.; Liu, D.; Zhang, K.; Miao, M.; Xu, G.; Song, A. Continuous Hybrid BCI Control for Robotic Arm Using Noninvasive Electroencephalogram, Computer Vision, and Eye Tracking. *Mathematics* **2022**, *10*, 618. [CrossRef]
128. Diraco, G.; Rescio, G.; Siciliano, P.; Leone, A. Review on Human Action Recognition in Smart Living: Sensing Technology, Multimodality, Real-Time Processing, Interoperability, and Resource-Constrained Processing. *Sensors* **2023**, *23*, 5281. [CrossRef]
129. Pereira, R.; Cruz, A.; Garrote, L.; Pires, G.; Lopes, A.; Nunes, U.J. Dynamic environment-based visual user interface system for intuitive navigation target selection for brain-actuated wheelchairs. In Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Napoli, Italy, 29 August–2 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 198–204.
130. Sun, H.; Li, C.; Zhang, H. Image Segmentation-P300 Selector: A Brain–Computer Interface System for Target Selection. *Comput. Mater. Contin.* **2024**, *79*, 2505. [CrossRef]
131. Mezzina, G.; De Venuto, D. Smart Sensors HW/SW Interface based on Brain-actuated Personal Care Robot for Ambient Assisted Living. In Proceedings of the 2020 IEEE SENSORS, Rotterdam, The Netherlands, 25–28 October 2020; pp. 1–4. [CrossRef]
132. Ban, N.; Xie, S.; Qu, C.; Chen, X.; Pan, J. Multifunctional robot based on multimodal brain-machine interface. *Biomed. Signal Process. Control* **2024**, *91*, 106063. [CrossRef]

133. Muñoz, J.E.; Chavarriaga, R.; Villada, J.F.; SebastianLopez, D. BCI and motion capture technologies for rehabilitation based on videogames. In Proceedings of the IEEE Global Humanitarian Technology Conference (GHTC 2014), San Jose, CA, USA, 10–13 October 2014; pp. 396–401. [CrossRef]
134. Chen, C.; Jafari, R.; Kehtarnavaz, N. A Real-Time Human Action Recognition System Using Depth and Inertial Sensor Fusion. *IEEE Sens. J.* **2016**, *16*, 773–781. [CrossRef]
135. Feng, X.; Weng, Y.; Li, W.; Chen, P.; Zheng, H. DAMUN: A Domain Adaptive Human Activity Recognition Network Based on Multimodal Feature Fusion. *IEEE Sens. J.* **2023**, *23*, 22019–22030. [CrossRef]
136. Abbasi, H.F.; Ahmed Abbasi, M.; Jianbo, S.; Liping, X.; Yu, X. TriNet: A Hybrid Feature Integration Approach for Motor Imagery Classification in Brain-Computer Interface. *IEEE Access* **2025**, *13*, 115406–115418. [CrossRef]
137. Mir, A.A.; Khalid, A.S.; Musa, S.; Faizal Ahmad Fauzi, M.; Norfiza Abdul Razak, N.; Boon Tang, T. Machine Learning in Ambient Assisted Living for Enhanced Elderly Healthcare: A Systematic Literature Review. *IEEE Access* **2025**, *13*, 110508–110527. [CrossRef]
138. Tidoni, E.; Gergondet, P.; Fusco, G.; Kheddar, A.; Aglioti, S.M. The Role of Audio-Visual Feedback in a Thought-Based Control of a Humanoid Robot: A BCI Study in Healthy and Spinal Cord Injured People. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 772–781. [CrossRef] [PubMed]
139. Zhang, X.; Zhang, T.; Jiang, Y.; Zhang, W.; Lu, Z.; Wang, Y.; Tao, Q. A novel brain-controlled prosthetic hand method integrating AR-SSVEP augmentation, asynchronous control, and machine vision assistance. *Heliyon* **2024**, *10*, e26521. [CrossRef]
140. Bellicha, A.; Struber, L.; Pasteau, F.; Juillard, V.; Devigne, L.; Karakas, S.; Chabardes, S.; Babel, M.; Charvet, G. Depth-sensor-based shared control assistance for mobility and object manipulation: Toward long-term home-use of BCI-controlled assistive robotic devices. *J. Neural Eng.* **2025**, *22*, 016045. [CrossRef]
141. Li, S.; Wang, H.; Chen, X.; Wu, D. Multimodal Brain-Computer Interfaces: AI-powered Decoding Methodologies. *arXiv* **2025**, arXiv:2502.02830. [CrossRef]
142. Zakka, V.G.; Dai, Z.; Manso, L.J. Action Recognition in Real-World Ambient Assisted Living Environment. *Big Data Min. Anal.* **2025**, *8*, 914–932. [CrossRef]
143. Caroleo, G.; Albini, A.; Maiolino, P. Soft Robot Localization Using Distributed Miniaturized Time-of-Flight Sensors. In Proceedings of the 2025 IEEE 8th International Conference on Soft Robotics (RoboSoft), Lausanne, Switzerland, 22–26 April 2025; pp. 1–6. [CrossRef]
144. Ding, R.; Hovine, C.; Callemeyn, P.; Kraft, M.; Bertrand, A. A Wireless, Scalable, and Modular EEG Sensor Network Platform for Unobtrusive Brain Recordings. *IEEE Sens. J.* **2025**, *25*, 22580–22590. [CrossRef]
145. Dickey, J. The Rise of Neurotech and the Risks for Our Brain Data: Privacy and Security Challenges—Future Security. March 2025. Available online: https://www.newamerica.org/future-security/reports/the-rise-of-neurotech-and-the-risks-for-our-brain-data/privacy-and-security-challenges/ (accessed on 12 July 2025).
146. Xia, K.; Duch, W.; Sun, Y.; Xu, K.; Fang, W.; Luo, H.; Zhang, Y.; Sang, D.; Xu, X.; Wang, F.Y.; et al. Privacy-Preserving Brain–Computer Interfaces: A Systematic Review. *IEEE Trans. Comput. Soc. Syst.* **2023**, *10*, 2312–2324. [CrossRef]
147. Wu, H.; Ma, Z.; Guo, Z.; Wu, Y.; Zhang, J.; Zhou, G.; Long, J. Online Privacy-Preserving EEG Classification by Source-Free Transfer Learning. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2024**, *32*, 3059–3070. [CrossRef]
148. Bechtold, U.; Stauder, N.; Fieder, M. Attitudes towards Technology: Insights on Rarely Discussed Influences on Older Adults' Willingness to Adopt Active Assisted Living (AAL). *Int. J. Environ. Res. Public Health* **2024**, *21*, 628. [CrossRef]
149. Botchway, B.; Ghansah, F.A.; Edwards, D.J.; Kumi-Amoah, E.; Amo-Larbi, J. Critical Smart Functions for Smart Living Based on User Perspectives. *Buildings* **2025**, *15*, 2727. [CrossRef]
150. Bastardo, R.; Martins, A.I.; Pavão, J.; Silva, A.G.; Rocha, N.P. Methodological Quality of User-Centered Usability Evaluation of Ambient Assisted Living Solutions: A Systematic Literature Review. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11507. [CrossRef] [PubMed]
151. Padfield, N.; Anastasi, A.A.; Camilleri, T.; Fabri, S.; Bugeja, M.; Camilleri, K. BCI-controlled wheelchairs: End-users' perceptions, needs, and expectations, an interview-based study. *Disabil. Rehabil. Assist. Technol.* **2024**, *19*, 1539–1551. [CrossRef] [PubMed]
152. Kristen Mathews, T.S. Unlocking Neural Privacy: The Legal and Ethical Frontiers of Neural Data. March 2025. Available online: https://www.cooley.com/news/insight/2025/2025-03-13-unlocking-neural-privacy-the-legal-and-ethical-frontiers-of-neural-data (accessed on 12 July 2025).
153. Yang, H.; Jiang, L. Regulating neural data processing in the age of BCIs: Ethical concerns and legal approaches. *Digit. Health* **2025**, *11*, 20552076251326123. [CrossRef]
154. Ochang, P.; Eke, D.; Stahl, B.C. Perceptions on the Ethical and Legal Principles that Influence Global Brain Data Governance. *Neuroethics* **2024**, *17*, 23. [CrossRef]
155. Sudhakaran, S.; Escalera, S.; Lanz, O. Gate-shift-fuse for video action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10913–10928. [CrossRef]

156. Maruotto, I.; Ciliberti, F.K.; Gargiulo, P.; Recenti, M. Feature Selection in Healthcare Datasets: Towards a Generalizable Solution. *Comput. Biol. Med.* **2025**, *196*, 110812. [CrossRef]
157. Wu, J.; Leng, C.; Wang, Y.; Hu, Q.; Cheng, J. Quantized convolutional neural networks for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4820–4828. [CrossRef]
158. Qiao, D.; Wang, Z.; Liu, J.; Chen, X.; Zhang, D.; Zhang, M. EECF: An Edge-End Collaborative Framework with Optimized Lightweight Model. *Expert Syst. Appl.* **2025**, *297*, 129319. [CrossRef]
159. Hajhassani, D.; Barthélemy, Q.; Mattout, J.; Congedo, M. Improved Riemannian potato field: An Automatic Artifact Rejection Method for EEG. *Biomed. Signal Process. Control* **2026**, *112*, 108505. [CrossRef]
160. Iosifidis, A.; Tefas, A.; Pitas, I. Multi-view human action recognition under occlusion based on fuzzy distances and neural networks. In Proceedings of the 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1129–1133.
161. Bai, G.; Yan, H.; Liu, W.; Deng, Y.; Dong, E. Towards Lightest Low-Light Image Enhancement Architecture for Mobile Devices. *Expert Syst. Appl.* **2025**, *296*, 129125. [CrossRef]
162. Kaveh, R.; Doong, J.; Zhou, A.; Schwendeman, C.; Gopalan, K.; Burghardt, F.L.; Arias, A.C.; Maharbiz, M.M.; Muller, R. Wireless user-generic ear EEG. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 727–737. [CrossRef]
163. Paul, A.; Lee, M.S.; Xu, Y.; Deiss, S.R.; Cauwenberghs, G. A versatile in-ear biosensing system and body-area network for unobtrusive continuous health monitoring. *IEEE Trans. Biomed. Circuits Syst.* **2023**, *17*, 483–494. [CrossRef] [PubMed]
164. Lombardi, I.; Buono, M.; Giugliano, G.; Senese, V.P.; Capece, S. Usability and Acceptance Analysis of Wearable BCI Devices. *Appl. Sci.* **2025**, *15*, 3512. [CrossRef]
165. Zanini, P.; Congedo, M.; Jutten, C.; Said, S.; Berthoumieu, Y. Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces. *IEEE Trans. Biomed. Eng.* **2017**, *65*, 1107–1116. [CrossRef] [PubMed]
166. Carboni, A.; Russo, D.; Moroni, D.; Barsocchi, P. Privacy by design in systems for assisted living, personalised care, and wellbeing: A stakeholder analysis. *Front. Digit. Health* **2023**, *4*, 2022. [CrossRef]
167. Kumar, D.; Kumar, C.; Seah, C.W.; Xia, S.; Shao, M. Finding Achilles' Heel: Adversarial Attack on Multi-modal Action Recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3829–3837. [CrossRef]