IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 October 2023, accepted 28 December 2023, date of publication 11 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3353138

TOPICAL REVIEW

Audio- and Video-Based Human Activity Recognition Systems in Healthcare

STEFANIA CRISTINA[1], (Member, IEEE), VLADIMIR DESPOTOVIC[2], (Member, IEEE), RODRIGO PÉREZ-RODRÍGUEZ[3], AND SLAVISA ALEKSIC[4], (Senior Member, IEEE)

[1]Department of Systems and Control Engineering, Faculty of Engineering, University of Malta, MSD 2080 Msida, Malta
[2]Bioinformatics Platform, Data Integration and Analysis Unit, Luxembourg Institute of Health, 1445 Strassen, Luxembourg
[3]Teoría de la Señal y Comunicaciones y Sistemas Telemáticos y Computación, Universidad Rey Juan Carlos, 28943 Fuenlabrada, Madrid, Spain
[4]HTWK Leipzig, Fakultät Digitale Transformation, 04229 Leipzig, Germany

Corresponding author: Slavisa Aleksic (slavisa.aleksic@htwk-leipzig.de)

This work was supported by the COST Action CA19121—GoodBrother, Network on Privacy-Aware Audio- and Video-Based Applications for Active and Assisted Living through the European Cooperation in Science and Technology (COST).

ABSTRACT In recent years, human activity recognition (HAR) has gained importance in several domains such as surveillance, recognizing indoor and outdoor activities, and providing active and assisted living environments in smart homes and healthcare services. In all these scenarios, audio-, video- and image-based processing algorithms have been applied as well as systems using wearable sensors. This scoping review focuses on audio- and video-based activity recognition systems for healthcare applications. We provide a comprehensive overview of these systems and technologies and discuss their complexity, performance, robustness, stage of development, scalability as well as achievable privacy and security levels. Additionally, we present and discuss datasets that are designed for the evaluation of these activity recognition systems. Although a number of robust approaches have already been proposed, they still pose challenges when it comes to integrating them in larger systems or into clinical practice. We identify challenges for application of audio- and video-based HAR systems in real-word healthcare scenarios, draw recommendations and conclusions based on comparisons of existing approaches, and analyze future trends.

INDEX TERMS Human activity recognition (HAR), healthcare, ambient and assisted living (AAL).

I. INTRODUCTION

Human activity recognition (HAR) has received attention in a number of applications in the past decades, including surveillance, human-human interaction (HHI) and human-computer interaction (HCI). The fast development of novel technologies and smart sensors has led to numerous new components and systems that either have already found wide application, or are still in the development phase. However, application in healthcare still remains limited, even though there is a huge potential in monitoring of chronic diseases that require active lifestyle and regular physical activity, such as diabetes, cardiovascular diseases, or obesity [1]. HAR can be further used to detect abnormal activities or sedentary behavior in patients suffering from dementia, as well as for fall detection in elderly population [2]. Therefore, HAR can improve patients' quality of life, but also have positive effects for the healthcare system, by reducing hospital stays and enabling more accurate and timely diagnosis. However, several challenges still remain unsolved, mostly related to adoption of such HAR systems by the end users, and privacy and security issues when they are implemented in real-world scenarios.

Audio and video-based HAR plays a significant role in active and assisted living (AAL) applications. Vision-based activity recognition has gained particular attention. However, a number of factors such as the interference with the environmental light, shadowing, different angles, and privacy protection narrow its application. Vision-based activity recognition can be combined with other modalities

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

(e.g. audio, WiFi, wearables and smart sensors) in order to achieve better robustness and performance, as well as to extend the fields of application.

In this scoping review, we provide an overview and classification of various video- and audio-based approaches for activity/behaviour recognition with application in healthcare, addressing the possibilities of combining these methods with other technologies and approaches, particularly in the scope of AAL systems. A typical AAL scenario for HAR is shown in Figure 1), where different modalities (audio, video, motion sensors and wearable devices) are used for a wide range of healthcare applications in different settings. Furthermore, available datasets and methods for improving the performance and automating the process of human activity recognition (HAR) have also been reviewed.

Although there are a number of reviews focusing on HAR [3], [4], [5], [6], [7], [8], [9], they all address general approaches, disregarding specific requirements for implementation in healthcare. We address this gap in this review, by providing a systematic overview of audio/video based technologies for HAR in healthcare, with taxonomy of possible applications, comprehensive overview of datasets, and analysis of performance, robustness, scalability as well as privacy and security. Finally, we identify the main challenges that need to be addressed in the years ahead, and discern the prospective trends in development of HAR in the healthcare sector.

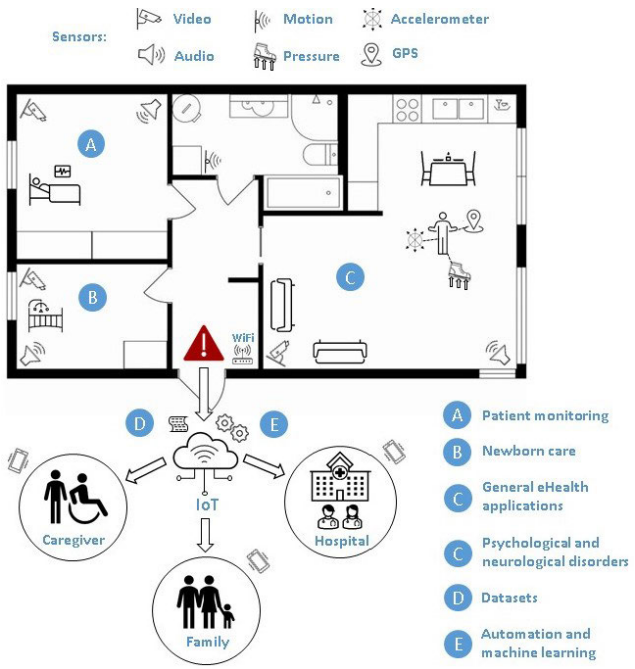

FIGURE 1. Typical AAL scenario for human activity recognition.

II. METHODS

In this study, we present results of a literature review on human activity recognition for healthcare applications published from 2000 to 2022. For this purpose, we carried out a search in two main scientific literature databases, namely Scopus and PubMed. For both the search and the assessment of published studies as well as synthesizing of research evidence, we followed the formal and explicit method for systematic reviews in healthcare known as the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) [10]. The PRISMA workflow including the number of identified, scanned, and selected articles during the review process is depicted in Figure 2.

In particular, we applied the following criteria to search both Scopus and PubMed databases:

> TITLE-ABS-KEY ((“activity recognition” OR “behavior recognition” OR “behaviour recognition”) AND (“video” OR “audio” OR “multimodal” OR “multi modal” OR “multi-modal”) AND (“healthcare” OR (“health” AND “care”) OR “health” OR “disease”))

The first search retrieved 353 results in Scopus and 53 results in PubMed, respectively. After identifying and removing duplicates, 396 articles remained for the subsequent screening step. The screening process focused on the eligibility of the records by examining the paper titles and abstracts, leaving only papers that have a clear link to HAR and at least one application in healthcare. 94 articles remained after the initial screening process. The next step was the eligibility proof with respect to entire articles, which led to the removal of 54 additional articles. Finally, the remaining 35 articles were selected for the inclusion in the quantitative and qualitative analysis.

Analyzing the keywords extracted from the selected articles in Figure 3, it can be clearly seen that “activity recognition” dominates the keyword landscape, followed by “deep learning” in combination with “video recording”. “Wearable sensors”, “intelligent buildings” and “ubiquitous computing” are keywords that play a significant role in multimodal approaches. The keywords “large datasets”, “daily life activity” and “computer-assisted image processing” also occur frequently in many publications. All the mentioned keywords are mainly found in recent research works published since 2017, which are indicated by the green and yellow colors.

III. AUDIO-BASED ACTIVITY/BEHAVIOR RECOGNITION IN HEALTHCARE

Audio is not extensively studied for human activity and/or behaviour recognition in healthcare, with only several applications where it was used as a standalone source of information, including emergency event detection [11], [12], and recognition of activities of daily living [13] in the context of elderly care, as well as recognition of suicidal behaviour [14]. In some cases audio was utilised in combination with other information sources (modalities), such as GPS, proximity and activity data for activity/behaviour monitoring of adolescent and young mothers with postpartum depression [15]; or combined with video, ultrasound, temperature, light and infrared sensors for detecting activities of people with mild

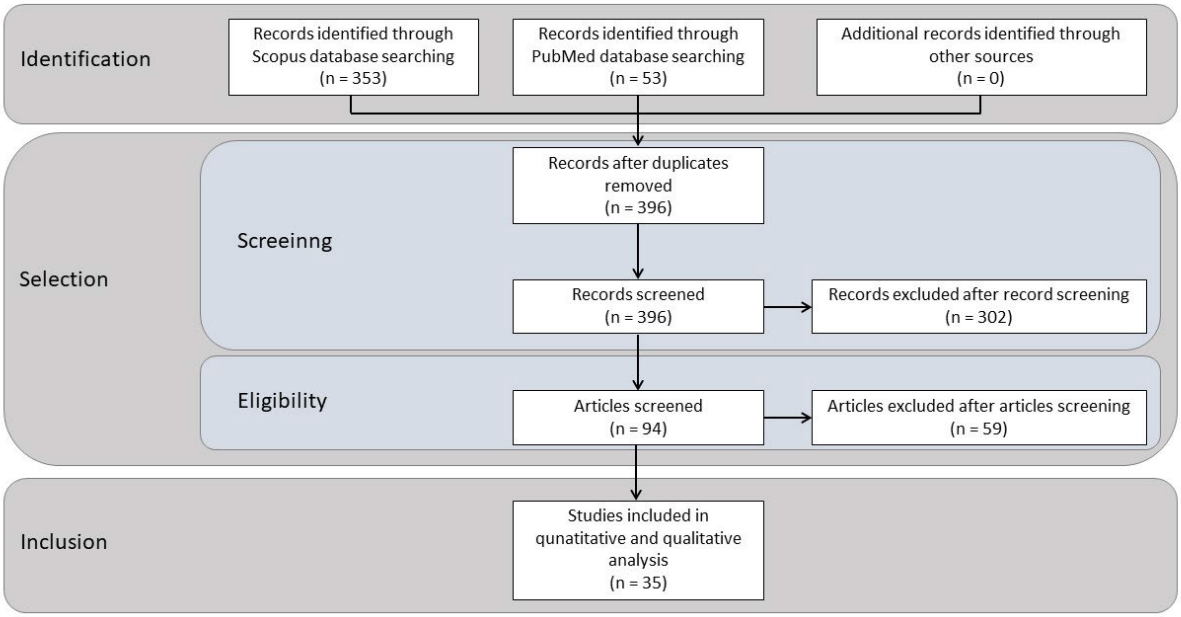

FIGURE 2. Review workflow according to the PRISMA framework.

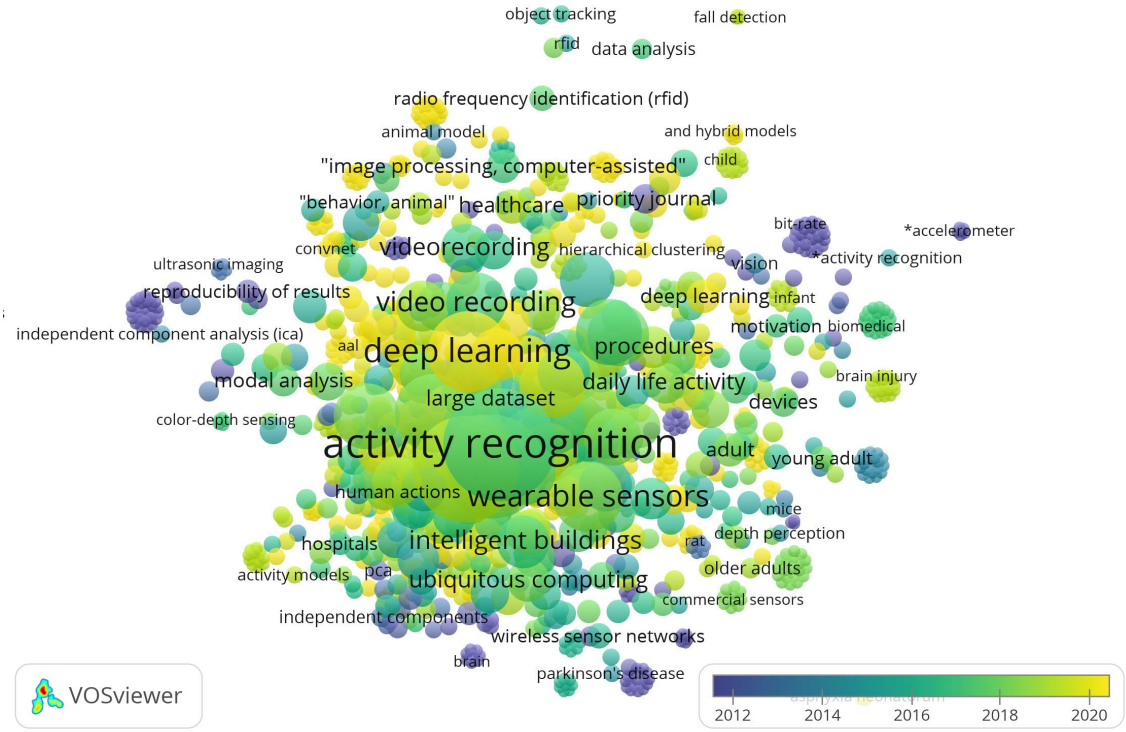

FIGURE 3. Keyword analysis of the literature search results (using both Scopus and PubMed data bases). Different colors represent different years of publication. The larger the bubbles, the more often the particular keyword occurs. Yellow color corresponds to the most recent articles, whereas purple corresponds to the oldest ones. The visualisation was made using the tool VOSviewer (www.vosviewer.com).

dementia [16]. More information about these multimodal approaches is provided in section V.

Kim et al. proposed an audio event recognition model for monitoring of elderly in single-person household environments and detection of emergency events potentially dangerous for their life [11]. Beside standard activities of daily living, such as cleaning, taking shower, doing laundry, sleeping, talking and listening to music, additional

emergency events were identified, including screaming, vomiting, burning, glass breaking, explosion and impact. The dataset is a collection of sounds acquired from five single-person household environments, and augmented by sounds from the online databases Freesound and Youtube-8M. Convolutional neural networks (CNN) and recurrent neural networks (RNN) were used for modeling, demonstrating capability to discriminate emergency events from the standard activities of daily living.

A similar approach for the detection of critical emergency events for people with disabilities or elderly people was studied in [12]. The dataset is a combination of 20 types of human and environmental sounds collected from 58 participants (47 male and 11 female) that may indicate potentially dangerous situations in a household or hospital. An eight-layer CNN with batch normalization and ReLu activation was used for classification, reaching approximately 97% of accuracy on independent test data.

Recognition of regular household activities that are relevant for remote health monitoring and assistive living of elderly people is analysed in [13]. The prototype device was implemented on Raspberry PI and enables real-time audio feature extraction and classification of activities of daily living. A graphical user interface was developed for Android mobile phones that acted as a remote controller for the Raspberry Pi device, and could be used both for data acquisition and training/classification using Support Vector Machines (SVM).

The tendency towards suicidal behaviour from speech data was investigated in [14]. The Butler-Brown Read Speech (BBRS) dataset was used that includes audio recordings of 226 psychiatric patients with a suicidal ideation or a previous suicide attempt, as well as 20 healthy controls with no history of suicidal behaviour. The system was composed of the automatic speech recognition module augmented with manual speech disfluency transcripts, and annotations for voice quality. On the other hand, the back-end of the system was fully automated using statistical machine learning methods. Disfluency features distinguish patients with suicidal behaviour from the healthy controls with 80% accuracy, outperforming voice quality features by 7%.

IV. VIDEO-BASED ACTIVITY/BEHAVIOR RECOGNITION IN HEALTHCARE

Video data is extensively exploited for the task of activity/behaviour recognition in healthcare, with applications ranging from monitoring of the elderly, to the resuscitation of newborns and human-robot interaction, among others. Videos are typically captured by cameras placed in the environment or embedded within wearable devices, and recognition of the human activity/behaviour is then carried out based on information that is extracted from their image frames. In recent years, methods based on machine learning and/or deep learning have been increasingly applied to the problem of activity/behaviour recognition from video. We present in the subsections below the most common applications of the video-based activity/behaviour recognition in healthcare.

A. PATIENT AND ELDERLY CARE

Monitoring of patients and the elderly are two related challenges that have been widely addressed through the application of different machine learning approaches. Anitha and Baghavathi [17], for instance, investigate the use of the R-transform, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for image feature extraction, and subsequently employ a Gaussian mixture model for activity classification based on the extracted features. Their method is aimed for the recognition of abnormal activities of the elderly people. They report an average accuracy of 82.2% on a self-collected dataset featuring 12 male participants and five activities (backward falling, chest pain, forward falling, headache, and vomiting). Siddiqi et al. [18], [19] investigate the use of a Hidden Markov Model (HMM) in monitoring the daily activities of patients. In their earlier work [19] they acquire patient images from a depth camera, and propose a segmentation algorithm that combines the Chan-Vese and the Battacharya energy functions to reduce the similarities within the human body parts, at the same time increasing the separation between the human body and the background. Following the segmentation process, salient features are extracted and used as input to an HMM. Their more recent work [19] extracts the feature vector that is fed into the HMM by matching a binary template image, representing an activity of interest, to an input image frame. They report accuracy measures ranging between 98.5% and 98.8% after testing their method on several datasets that include activities such as walking, jogging and lifting [19]. Jalal et al. [20] also make use of an HMM for human activity recognition. Their method utilises the depth silhouettes captured by a depth camera, which are processed to produce a set of 15 skeleton joints, representing the head, arms, torso, hip and legs, from which a feature vector containing the joint positions and angles is extracted. After representing the feature vector into a codebook generated by a k-means clustering algorithm, this information is fed into several HMMs for training and recognition of every activity of interest. Similarly, Khan and Sohn [21] generate a codebook using k-means clustering for training and testing of HMMs. However, their feature extraction and dimensionality reduction process is carried out using the R-transform and a posterior Kernel Discriminant Analysis (KDA) method, as applied to person silhouettes extracted from single camera images. Wang et al. [22] target the specific application of monitoring patients during sleep. Their approach requires the use of pyjamas with near-infrared (NIR) sensitive fabric that facilitates detection of the human body joint positions from NIR videos. Detection of the human body joints is performed by SIFT-based local features, while occlusion issues are addressed by applying a Bayesian inference algorithm to recognise poses.

Deep learning-based approaches have also been extensively applied to the challenge of monitoring patients and the elderly. For instance, Su et al. [23] investigate the use of several neural network architectures that include a Convolutional Neural Network (CNN), as well as single and multi-layer fully-connected neural networks, in detecting ten common types of activities such as standing, bending and eating. Their results indicate that the optimal outcome of a 95% accuracy rate among all tested neural networks was achieved by a deep neural network consisting of five hidden layers, with each layer consisting of 30 neurons. Subramanian [24] feed the acquired image frames to a YOLO neural network, which detects image frames in which a human subject is present. Following this, salient regions of significant change in the selected image frames are detected to narrow down the search space, based on which a GoogleNet neural network performs the final activity recognition. The proposed method is deployed onto a fog computing framework with the aim of increasing its computational efficiency.

Furthermore, the fact that videos also contain a temporal dimension is often exploited in different ways, e.g. by employing the use of a 3D CNN [25], [26], cascading the CNN or 3D CNN with a Long Short-Term Memory (LSTM) network [25], [26], [27], or encoding the spatio-temporal information of skeleton sequences into an image to be fed into a CNN [28]. Manocha et al. [25], for instance, initially aggregate the incoming image frames into frame segments. Each frame segment is, then, fed into a 3D CNN for feature extraction, subsequently passed on to an LSTM network, and finally to two fully-connected (FC) layers for activity classification. Similarly, Gao et al. [26] propose a 3D CNN that is based upon a simplified VGGNet architecture. The output of the 3D CNN is then fed into an LSTM network that, in turn, extracts the long-range features of the human actions from the video data. Singh and Vishwakarma [27] also cascade a CNN with a bidirectional-LSTM in one branch of their proposed network, while in a second branch, the proposed architecture feeds dynamic motion images into a fine-tuned CNN. Each branch feeds into a fully-connected layer and a softmax layer, and the probabilistic outputs of both branches are finally fused to produce an activity prediction. Chen et al. [28], on the other hand, encode skeleton sequences, extracted from RGB images, into spatio-temporal images and feed them to a CNN for classification. Liu et al. [29] propose a one-shot action recognition method in conjunction with their dataset, NTU RGB+D 120, that among its 120 different action classes also includes health-related activities. Their neural network adopts a two-dimensional spatio-temporal LSTM architecture, which is modified to receive three-dimensional coordinates of skeletal joints and produces features representing the corresponding body part. Action recognition is then carried out based on the features of each body part and the semantic relevance of the action classes. The method proposed by Andrade-Ambriz et al. [30] is targeted towards the provision of health care, where the care provider is a robot. The proposed architecture exploits the temporal information in videos by employing a 3D convolutional layer followed by a 2D LSTM layer and a 3D max pooling layer, in order to extract the spatio-temporal features. The extracted features are then fed into two fully-connected layers with ReLU activation, and a final dense layer with softmax activation to produce the output. Online experiments in a real-world environment were also conducted with a NAO humanoid robot connected to a personal computer, where the latter runs the neural network. During the experiment, the NAO robot reacted with a sentence upon receiving video data of an activity performed by the user.

B. MEDICATION INTAKE MONITORING

Closely related to the monitoring of patients is the application of activity recognition for monitoring and managing medication intake. Lee and Youm [31] propose a method in which videos of patients taking medications are recorded using a camera integrated into a smartwatch. Their objective was to monitor adherence to medication prescriptions to reduce the tax waste due to unconsumed drugs, optimise clinical trials that get extended due to lack of adherence, and address health outcomes and side effects related to poor medication management. A region-based CNN (R-CNN) and an SVM were applied for object detection and medication recognition, followed by an LSTM for action classification, achieving an accuracy of 92.7%. It is worth highlighting that this approach is not constrained to the administration of only a single type of medication. Rather, it has the potential of being applied for the recognition of other behaviours, such as the use and acceptance of various medical products (e.g. inhaler and glucometer usage), or to analyse the food and beverage consumption behaviour of patients with high blood pressure, obesity, and special diet requirements.

C. MONITORING PATIENTS WITH PSYCHOLOGICAL AND NEUROLOGICAL DISORDERS

Another area of interest for activity recognition relates to supporting persons with cognitive disorders, such as dementia, Alzheimer's and amnesia [32], [33], [34], [35], [36], and neuro-developmental disorders, such as autism [37].

The ontology-driven approaches are frequently used in monitoring of cognitive disorders to increase the interpretability of system decisions by incorporating domain knowledge and extracting semantic meaning of videos for the activity recognition task. Crispim-Junior et al. [32] proposed an ontology-driven approach to formalise knowledge about the persons and objects, as well as events happening in the scene. Their proposed pipeline performs ground plane estimation, followed by person detection and tracking, and finally event recognition based on their constraint-based ontology. The method proposed by Karakostas et al. [33] also exploits a constraint-based ontology to model activities of daily living, based on information related to motion,

posture and scene location patterns extracted from images taken by a three-dimensional camera. They employed a cross-comparison study at clinical sites situated in France and Greece to evaluate the autonomy of dementia patients.

Deep learning-based approaches have also been applied for a similar task, where, for instance, Lee at al. [34] propose a CNN-based method to facilitate machine-assisted human tagging of videos, for the purpose of logging activities carried out by older persons affected by Alzheimer's disease in a care facility. Their method starts with an object detection phase that uses a YOLO-v3 model to detect objects related to activities of interest. Subsequently, features are extracted, which contain information about the detected objects and the persons in the scene, and fed into a multi-dense layer network using a sliding window filter to reduce noise. Negin and Bremond [35] also target their method towards the monitoring of cognitive disorders, such as Alzheimer's disease, through an unsupervised human activity recognition method that combines global trajectory information with local dynamics of the human body. In their approach, they consider the use of either hand-crafted techniques, or CNNs for automatic feature extraction of the discovered activities, based on which a hierarchical activity model is later constructed via the integration of all extracted features. Washington et al. [37] propose a method to recognise autism-related behaviours, headbanging in particular, from unstable camera sources, and datasets that are relatively small and of low quality. In their approach, a CNN learns to extract visual features from each video frame, which are further passed to an LSTM network for classification of the behaviour of interest.

Zuo et al. [36] take a different approach towards supporting persons with memory problems, such as those suffering from dementia or amnesia, by exploiting the user's eye-gaze acquired by a wearable eye-gaze tracker. In their approach, Zuo et al. argue that the typical approaches of using regions-of-interest (ROI) or the image contents in their entirety, can lead to issues related to mismatches between the estimated and ground truth ROI, and moving backgrounds and irrelevant foregrounds. They, alternatively, propose to extract features for action recognition from gaze-informed regions (GROI) only, based on the user's gaze points, followed by feature extraction from the GROI, feature encoding into an activity identifier, and finally action classification. Several local feature descriptors have been tested, namely Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), Motion Boundary Histogram (MBH), and Histogram of Motion Gradients (HMG), while action classification is carried out by a back-propagation neural network, which takes the encoded and normalised features as input.

D. NEWBORN CARE

Newborn care is another related area of interest, where Meinich-Bache et al. [38] focus on recognising therapeutic resuscitation activities for birth asphyxia, such as stimulation, ventilation, suction, and the attachment, adjustment and removal of an ECG sensor. Their method consists of several steps that include object detection and tracking using CNNs, proposal of activity regions, followed by activity recognition through the use of 3D CNNs. Typical movements of particular activities during newborn resuscitation are recognised by exploiting the temporal information contained in sequential frames, instead of utilising only individual frames. This approach allows for recognition in noisy and low-quality data, where it might be difficult to detect activities from individual frames due to motion blurring, wth a mean accuracy of 92.40% in recognising activities such as stimulation, ventilation and suction [39].

V. MULTIMODAL ACTIVITY/BEHAVIOR RECOGNITION IN HEALTHCARE

Multimodal approaches for activity recognition are those using more than one sensing technique, but always including either audio or video information, or both. Different approaches can be found in the scientific literature using heterogeneous sensors. For instance, Aung Aung et al. [40] evaluate several approaches for activity recognition and monitoring in and around the bed. They use different sensing modalities that include pressure, video, ultrasound, and passive infrared (PIR), that are combined in a set of scenarios to compare their performance recognizing fine-grained primitives (lying postures) and coarse-grained primitives (dangerous situations in and around the bed). Authors obtained a maximum classification accuracy of 90% for fine-grained primitives using pressure and video information, and of 93% for coarse-grained ones with a combination of pressure, ultrasound and PIR sensors. Although these methods could potentially be applied in real settings, the related experiments are limited to a small population. Additionally, extensive training is required to correctly recognize different contexts. Furthermore, uncertainties and erroneous readings resulting from different sensor modalities are not sufficiently considered.

Feki et al. [16] propose a theoretical study of a system to assist people affected with mild dementia (an early stage of Alzheimer's Disease) performing their activities of daily living, and also continuously monitor them. In this work, the authors use multiple modalities for data acquisition, including video, audio, ultrasound, temperature, light and infrared sensors with to analyze the cognitive processes underlying executing the actions, detecting errors or inappropriate actions, and providing cues to the user when necessary. Dynamic Bayesian Networks, Reinforcement Q Learning and Fuzzy Partitioning are used for classification, but no quantitative performance evaluation is presented.

Byanjankar et al. [15] proposed an approach that combines audio recordings with passively collected proximity, activity and GPS data to identify and monitor activities and behaviour of adolescent and young mothers with postpartum depression. Their aim was to provide counseling beyond formal sessions. Data were processed in a cloud environment, and feedback about the behaviour patterns was automatically generated and

sent to the mothers. Additionally, the collected GPS data were visualized as heat maps, while proximity data together with the mothers' activities were visualized as charts within the so called StandStrong application. The generated information can be used during counselling sessions to discuss behavioral patterns as well as clinical progress. As authors state, the StandStrong platform has the potential to improve the quality and effectiveness of psychological services delivered by non-specialists in diverse global settings.

Khattak et al. [41] presented a human activity recognition engine (HARE) for people affected by Alzheimer's disease. To recognize activities, video is processed by applying independent component analysis (ICA) over the sequence of registered images (segmented body silhouettes), combined with k-means clustering and HMMs. Accelerometry data is processed using semi-Markov conditional random fields, which models the duration of the activity. Finally, to estimate location, an ANN calculates the distance to the beacons together with push-pull estimation [42]. The approach has been tested to recognize nine activities (bend, jump-jack, jump, run, gallop sideways, skip, walk, wave with one hand and wave with two hands) from a public database to assess the performance of the video processing in terms of action recognition. An average accuracy of 91.4% was obtained. Accelerometry was tested against four activities (dinner, commuting, lunch and office) from another open database, obtaining an average accuracy and recall of approximately 85%. Finally, localisation tracking does not perform so well, obtaining a RMSE of 3.12m.

Ning et al. [43] propose using multimodal inputs (accelerometry, image and sound) to improve classification accuracy with a very low proportion of labelled data compared to traditional supervised approaches. This method comprises three phases: (1) collaborative data collection from the heterogeneous data sources; (2) co-training based on HMMs; and (3) collaborative classifier combination using the product rule. Co-training appears to improve over the traditional supervised approaches using less than 20% of the labelled data. Although the authors claim that their method aims for healthcare applications, limited information is provided regarding potential use cases.

Bedri et al. introduced EarBit [44], a wearable device that includes Inertial Measurement Units (IMUs), a proximity sensor, and a microphone, aiming for the detection of chewing instances and eating episodes in unconstrained environments. The system was implemented and analysed in both a semi-controlled lab environment and outside-the-lab. The EarBit system accurately recognised all but one recorded eating episodes, which ranged from two minute snacks to 30 minute meals. The recognised eating activities include hand-to-mouth gestures, chewing, swallowing and drinking. For recognition they used Random Forests, while Sequential Forward Floating Selection (SFFS) was used for feature selection. Within a lab environment, the system achieved an accuracy of 90.1%, an F1 score of 90.9%, a precision of 86.2%, and a recall of 96.1%. Outside of the lab, the system achieved an accuracy of 93%, an F1 score of 80.%, a precision of 81.2%, and a recall of 79%.

VI. DISCUSSION

This section provides a summary and discussion of the presented approaches and an analysis of their main characteristics and performance parameters. Although there exist a relatively large number of studies focusing on human activity and/or behaviour recognition, only a small portion of them address concrete applications of audio, video or multimodal methods in healthcare, and provide an experimental validation or implementation in a real setting. The application of audio- and video-based HAR systems in healthcare has become a topic of interest within the last twelve years, especially since 2017, as indicated in Figure 4.

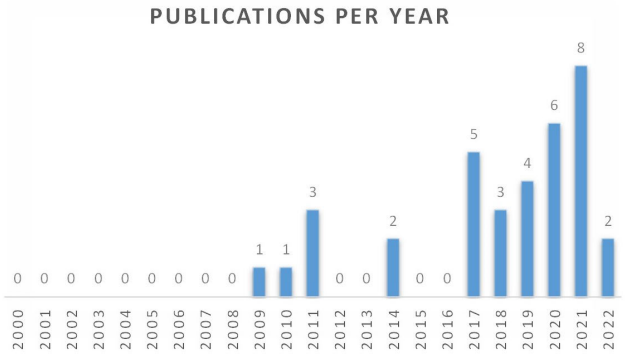

FIGURE 4. Distribution of selected relevant publications over the last twenty-two years.

A. CHARACTERISTICS OF APPROACHES

A detailed comparison of the selected research works for audio- and video-based HAR systems in healthcare is presented in Table 1, indicating the characteristics of the studies relating to the modality, intended application(s), stage of development, and whether privacy and security issues are addressed.

The analysis of the selected literature has shown that the majority of the recent works focus on video-based methods (71.5%), mostly using RGB cameras (see Figure 5). Multimodal approaches are represented by approximately 17% and approaches using audio processing by almost 11.5% of papers that have been considered in this review.

Most of the studies present experimental results (77%), where only 14% of the presented results are obtained in a real-world setting. The portion of publications that present results of a theoretical nature is less than 9%.

Table 2 presents details related to features and models used in recognising, processing and classifying patients' activities and behaviours, and provides an indication of the achieved performance.

We have found that a large variety of audio and video features has been exploited so far, but the mostly utilised

TABLE 1. Research works with applications related to human activity recognition in AAL for healthcare: Intended application(s), stage of development and privacy/security.

Modality	Ref.	Year	Intended application(s)	Stage of development	Privacy and security
Multimodal	[41]	2011	Dementia/Alzheimer's disease	Lab environment	Not assessed
Multimodal	[16]	2009	Dementia/Alzheimer's disease	Theoretical work	Not assessed
Multimodal	[43]	2011	General eHealth	Lab environment	Not assessed
Multimodal	[15]	2021	Psychological treatment	Real setting	Two-layer of consent: written and verbal
Multimodal	[40]	2010	Patient monitoring	Lab environment	Not assessed
Multimodal	[44]	2017	Recognition of eating episodes	Real setting	Not assessed
Audio	[11]	2020	Emergency event detection	Lab environment	Not assessed
Audio	[12]	2017	Emergency event detection	Lab environment	Not assessed
Audio	[13]	2017	Remote monitoring	Lab environment	Not assessed
Audio	[14]	2021	Recognition of suicidal behavior	Real setting	Not assessed
Video (RGB)	[39]	2020	Neonatal resuscitation	Lab environment	Not assessed
Video (RGB)	[26]	2018	Remote patient monitoring	Lab environment	Not assessed
Video (RGB, silhouette)	[21]	2011	Recognition of abnormal activities	Lab environment	Privacy through silhouette
Video (RGB)	[30]	2011	Human-robot interaction	Lab environment	Not assessed
Video (RGB)	[31]	2021	Medication monitoring	Lab environment	Not assessed
Video (RGB, skeleton)	[23]	2020	Elderly care	Lab environment	Not assessed
Video (RGB)	[34]	2020	Dementia/Alzheimer's disease	Real setting	Not assessed
Video (RGB)	[27]	2021	Elderly care	Lab environment	Not assessed
Video (RGB)	[28]	2018	Elderly care	Lab environment	Not assessed
Video (RGB-D)	[18]	2021	General eHealth	Lab environment	Privacy through using depth cameras
Video (RGB)	[19]	2022	Remote monitoring	Lab environment	Not assessed
Video (RGB-D)	[32]	2017	Dementia/Alzheimer's disease	Real setting	Not assessed
Video (RGB-D)	[33]	2020	Dementia/Alzheimer's disease	Lab environment	Not assessed
Video (RGB)	[37]	2021	Autism	Theoretical work	Not assessed
Video (RGB)	[35]	2019	Dementia/Alzheimer's disease	Lab environment	Not assessed
Video (RGB-D)	[29]	2020	General eHealth	Lab environment	Not assessed
Video (RGB and RGB-D)	[45]	2014	Elderly care	Lab environment	Privacy-by-context
Video (RGB)	[36]	2018	Dementia/Alzheimer's disease	Lab environment	Not assessed
Video (RGB)	[24]	2021	Elderly and children care	Lab environment	Not assessed
Video (RGB, silhouette)	[17]	2019	Patient monitoring	Lab environment	Privacy through silhouette
Video (RGB-D)	[25]	2020	General eHealth	Lab environment	Not assessed
Video (IR camera)	[22]	2019	Sleep monitoring	Lab environment	Not assessed

features are spectrograms in audio-based approaches, and spatiotemporal features when working with video data. A few video-based approaches use skeletons or silhouettes as an abstract representation of a person or an object. Several other video-based approaches make use of trajectory shapes, SIFT-based local features or demographic features. Multimodal approaches augment their feature set with time and frequency domain features, motion features, acceleration, gravity and energy levels.

As can be seen from Table 2, CNNs and RNNs have been dominant prediction models in most audio-based studies, followed by SVMs which were also extensively used. Although similar trend can be observed for video-based studies, a variety of algorithms is bigger, with HMMs, BPNN, LDA, KDA, and Bayesian networks, appearing among others. The choice of prediction algorithms for multimodal approaches depends on the features and sensors in use, as well as the requirements that are set on the data processing, classification and performance parameters. These models and algorithms include HMMs, ANNs, PPE, Dynamic Bayesian Networks, ICA, k-means clustering, Reinforcement Q learning and Fuzzy Partitioning (see Table 2).

TABLE 2. Research works with applications related to human activity recognition in AAL for healthcare: Features, models and performance.

ICA: Independent Component Analysis; CRF: Conditional Random Field; ANN: Artificial Neural Network; PPE: Push-Pull Estimation; HMM: Hidden Markov Model; GPS: Global Positioning System; SVM: Support Vector Machine; SFFS: Sequential Forward Feature Selection; CNN: Convolutional Neural Network; RNN: Recurrent Neural Network; LSTM: Long Short-Term Memory; kNN: k-Nearest Neighbor Algorithm; R-CNN: Region-Based Convolutional Neural Network; YOLO: You Only Look Once; SSD: Single-Shot Detector; DNN: Deep Neural Network; RBF: Radial Basis Function; ANOVA: ANalysis Of VAriance; MBH: Motion Boundaries Histogram; HOF: Histogram of Optical Flow; HOG: Histogram of Oriented Gradients; TDD: Trajectory-Pooled Deep-Convolutional Descriptors, HMG: Histogram of Motion Gradients; HAM: Hierarchical Activity Mode; DTW: Dynamic Time Warping; BPNN: Backpropagation neural network; PCS: Principal Component Analysis; ICA:Independent Component Analysis; SIFT: Scale invariant feature transform.

Modality	Ref.	Features	Type of model used	Performance
Multimodal	[41]	Motion features from sequences of images	Videos: ICA, K-means clustering and HMM Sensors: semi-CRF Location: ANN and PPE	Average recognition accuracy: 91.4% Average precision: 0.759 Recall: 0.636
Multimodal	[16]	Features from raw accelerometer data	Dynamic Bayesian networks Reinforcement Q learning Fuzzy Partitioning	Good recognition accuracy (no quantitative analysis presented)
Multimodal	[43]	Acceleration, image and audio features	Collaborative data collection Co-training based on HMMs Collaborative classifier combination using the product rule	Verification prediction: 99.4%
Multimodal	[15]	Automatic feedback on behavior patterns	Post-processing and visualizing sensor data Proximity chart Speech count percentage Mother's activities GPS heat map	Not provided (no quantitative analysis presented)
Multimodal	[40]	Spatiotemporal statistics, centre of gravity (CoG), distance and energy level	Multi-class SVM Weighted voting method Bayesian networks	Classification accuracy: 92% for pressure/video 93% for pressure/ultrasound/passive IR
Multimodal	[44]	Thirteen time and frequency domain features	Random forests algorithm SFFS	Precision, generalizability, and realism In Lab: accuracy 90.1%, F1-score 90.9%, precision 86.2%, recall 96.1% Outside of the lab: accuracy 93%, F1-score 80.1%, precision 81.2%, recall 79%
Audio	[11]	Mel-spectrogram	CNN and RNN (LSTM)	Precision: 78%, recall: 90.8%
Audio	[12]	Spectrogram	CNN, RNN (LSTM)	Accuracy: 96.8% Mean average precision: 0.991
Audio	[13]	Six audio features	SVM, kNN, Extremely Randomized Trees Random Forest, Gradient Boosting	F1 score: 84%- 92% (depending on a machine learning model)
Audio	[14]	GRBASI voice quality	Linear SVM and kNN	Accuracy: 73%, F1-score: 0.69
Video (RGB)	[39]	Object detection and tracking	CNN, Faster R-CNN, RetinaNet SSD MulitBox, YOLOv3	Mean precision: 77.67% Mean recall: 77,64% Max. mean accuracy: 92.40%
Video (RGB)	[26]	Spatiotemporal features extracted from videos	CNN, LSTM and SVM classifier	Accuracy: 86.8% at 386 fps
Video (RGB, silhouette)	[21]	Invariant silhouette features (periodic, scale, translation)	R-transform, LDA, KDA, k-means, HMM	Accuracy: 79,3% (KDA), 95,8% (LDA)
Video (RGB)	[30]	Spatiotemporal features extracted from videos	CNN and LSTM	Precision: 100%, recall: 100% (KARD and CAS-60 datasets) Accuracy: 95.6% (MSR daily activity 3D dataset)
Video (RGB)	[31]	Spatiotemporal features extracted from videos	R-CNN, SVM and LSTM	Accuracy: 92.7%
Video (RGB, skeleton)	[23]	Angles between the joints, changes in lengths	DNN	Accuracy: 95%
Video (RGB)	[34]	Person type, distances, change of distances	YOLO-v3, CNN	Precision: 98.4% (theoretical) Accuracy: 81.4% (live data)
Video (RGB)	[27]	Spatiotemporal features extracted from videos	CNN (Inception-v3) + bi-LSTM	Accuracy: 98.70% (SBU dataset) 99.41% (MIVIA dataset) 98.30% (MSR Action Pair dataset) 94.37% (MSR Daily Activity 3D dataset)

TABLE 2. *(Continued.)* **Research works with applications related to human activity recognition in AAL for healthcare: Features, models and performance.**

Video (RGB, skeleton)	[28]	Spatiotemporal features of skeleton	CNN	Accuracy: 89.4% on grey images Accuracy: 100% on color images
Video (RGB-D)	[18]	Confined features from the series of frames	HMM	Classification rate: 97.3 Standard deviation of +/- 2.7
Video (RGB)	[19]	Feature vector generated after matching a binary template to an underlying image	Template matching, HMM	Accuracy: 98.8% (Weizmann dataset) 98.5% (KTH action dataset) 98.6% (UCF dataset) 98.6% (IXMAS dataset)
Video (RGB-D)	[32]	Bag-of-visual-word embeddings over descriptors of trajectories features	SVM with RBF kernel	F1 score, precision Recall: 79.3 ± 13.0 (CHUN dataset), 76.4 ± 21.0 (GAADRD dataset)
Video (RGB-D)	[33]	Demographic features and activity duration	one-way ANOVA and correlation analyses	Statistically significant differences between human event annotation and automatic event detection (ANOVA, $\rho < 0.05$ and $\rho < 0.01$)
Video (RGB)	[37]	Visual features (e. g. head pose keypoints)	CNN and LSTM	F1-score: 90.77%
Video (RGB)	[35]	Trajectory shape, motion descriptors (HOF, MBH), HOG, TDD	HAM, SVM	Recall, F1 score Precision: 77% - 94%
Video (RGB-D)	[29]	2D spatiotemporal features	LSTM, Soft RNN, FSNet, Multi-task CNN, etc.	Cross-subject accuracy: 64.0% Cross-setup accuracy: 66.1%
Video (RGB and RGB-D)	[45]	Single-view and multi-view features of binary silhouettes	Weighted feature fusion scheme for multi-view action recognition DTW for the comparison of sequences	LOSO cross-validation: 95,2% for DHA dataset 91,45% for IXMAS dataset 94,4% for multi-view depth dataset
Video (RGB)	[36]	HOG, HOF, MBH, HMG	BPNN	Accuracy: 100% (HOG features)
Video (RGB)	[24]	Features extracted by a GoogleNet neural network	YOLO for human detection GoogleNet for activity recognition	Hollywood2: Precision 0.8802 Recall 0.8997
Video (RGB, silhouette)	[17]	PCA and ICA from binary silhouette images	Gaussian mixture model for activity classification	Average accuracy: 82.2%
Video (RGB-D)	[25]	Spatiotemporal features from frame segments	3D-CNN for feature extraction LSTM for activity classification	Recall: 90.33% Precision: 93.45% Specificity: 94.28% F-measure: 90.87%
Video (IR camera)	[22]	SIFT-based local features	Statistical inference using a Bayesian network	Positive predictive value: 80% Negative predictive value: 71% Specificity: 63% Sensitivity: 86%.

The reported performance of the analysed approaches is mostly good, with accuracy over 90%, often close to 100%. The precision values have often been reported to be between 70% and 90%, whereas sensitivity (recall) and specificity range between 60% and 100%. An interesting observation is that the performance of systems deployed in a laboratory environment and in a real setting differs only slightly. This is an indication that many proposed approaches have potential to provide a high performance once implemented in real environments. However, given the relatively small datasets the models were trained and tested on in most of the cases, the generalisability of the proposed methods remains questionable, requiring external validation.

An interesting observation is that only a few of the selected studies (14.3%) mention privacy and security, with only one of them addressing these important issues appropriately, by proposing concrete steps to ensure that the privacy of patients is protected and the sensitive data obtained by the HAR systems are well-handled (recorded, processed, transmitted and stored in a secure and safe manner) [45]. Chaaraoui et al. [45] suggest a privacy-by-context scheme that focuses on different visualisation levels to protect the user's identity. Their objective is to modify the raw image data in such a way as to retain the usefulness of the information, without jeopardizing the user's privacy, by using image blurring and replacement with a 3D avatar.

B. APPLICATIONS

Audio- and video-based HAR systems can be used for various applications in healthcare. They can be divided

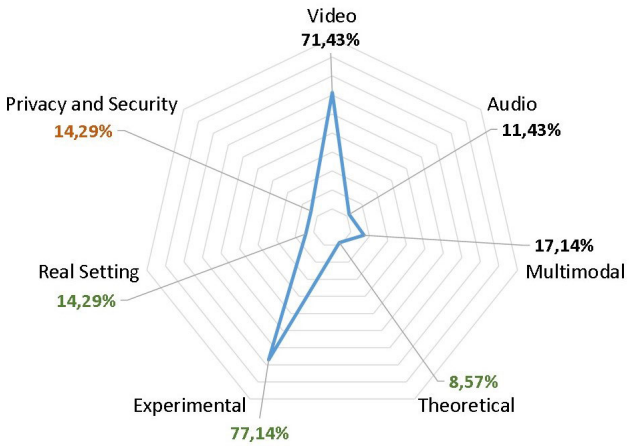

FIGURE 5. Characteristics of selected research works.

into four main groups as depicted in Figure 6, including: (1) patient monitoring, (2) psychological and neurological disorders, (3) newborn care, and (4) general health applications.

Patient monitoring is assumed to happen at a distance, i.e. the patients are monitored remotely either from another building, or from another room within the same building. Patient monitoring applications related to the detection of emergency events and abnormal activities can be used to identify risks of falls, prevent injuries or accidents, and provide timely assistance to patients in need. The recognition of eating episodes enables monitoring of patients' nutritional intake and evaluation of dietary habits, which is vital for patients with eating disorders, and further helps synchronizing medication administration with patients' mealtimes; thus, reducing the risks of adverse drug interactions or side effects. Monitoring of sleep phases is essential for diagnosing sleep disorders (e.g insomnia, narcolepsy, sleep apnea), but also for management of chronic conditions, such as mental health disorders, cardiovascular diseases, diabetes or obesity (see Table 1).

The largest research effort on HAR systems for treating psychological and neurological disorders is devoted to monitoring dementia/Alzheimer disease, with a smaller portion of the analysed works paying attention to depression and suicidal behaviour, as well as autism. HAR can detect subtle changes in behaviour and walking patterns, daily routine or social interactions, which could indicate symptoms of cognitive decline, dementia, depression, or even suicidal behaviour.

General eHealth applications mostly focus on elderly care and human-robot interaction, where HAR can support longer independent living of elderly individuals by monitoring their health status, facilitating caregiver support, and helping them to age gracefully.

Finally, newborn care is a niche application that mainly focuses on neonatal resuscitation and birth asphyxia.

C. DATASETS

The available audio datasets for human activity/behaviour recognition in healthcare are used either for emergency event detection, or for suicidal behaviour monitoring. In the case of emergency event detection, datasets (TUC, Intenta) are collected by simulating critical acoustic events for elderly people, e.g. calls for help, screaming, crying, whimpering, collapsing on the floor, striking the wall, trampling on carpet, or dislocating furniture [12]. In the case of evaluating suicidal behaviour using voice quality and speech disfluency attributes, the BBRS dataset contains speech recordings of psychiatric patients with suicidal ideation or a previous suicide attempt, and healthy control subjects with no history of suicidal behaviour [14]. All audio datasets are limited in size ranging from only 4 minutes (Intenta) up to 7 hours of recordings (BBRS).

Video datasets for HAR in healthcare may contain a single video modality (RGB or depth), or a combination of multiple modalities (RGB, depth, skeleton). The applications covered by such datasets range from remote monitoring [46], [47], [48], [49], [50] and elderly care [28], [51], [52], [53], [54], [55], to monitoring of people with dementia/Alzheimer's disease [32], [36] and autism [56]. In addition to datasets designed for a particular healthcare activity recognition task, general purpose activity recognition video datasets are also utilised that contain a subset of healthcare related activities, such as UCF Sports Action [49], UCF101 [46], Weizmann Human Action [47], KTH Action [48], IXMAS [50], KARD [57], CAD−60 [52], MSR Daily activity 3D [58] or NTU RGB+D 120 [29]. However, they might not always be well matched to the healthcare application domain, taking into account the observed activities. Excluding the general purpose video datasets, most of the video datasets specialised for healthcare are limited in size. Even when the dataset size is not a limitation, as in Nursing Home which is 72 hours long [32], the fact that it is collected from only one participant diagnosed with Alzheimer's disease constraints its generalisability.

VII. CHALLENGES AND FUTURE DIRECTIONS

Although audio- and video-based approaches for HAR have already found application in healthcare systems, the potential of such techniques has not been sufficiently exploited. While video-based approaches were utilised for screening, diagnosing and supporting the treatment of various diseases, the audio and multimodal techniques have so far been used primarily in the context of monitoring of daily activities, elderly care, and recognition of suicidal behaviour. Systems based solely on audio features are rarely considered. More often, audio is combined with other modalities in a multimodal setup. Although multimodal approaches are generally able to cover a large variety of applications for healthcare, its relatively complex implementation, feature extraction and processing, as well as the lack of suitable,

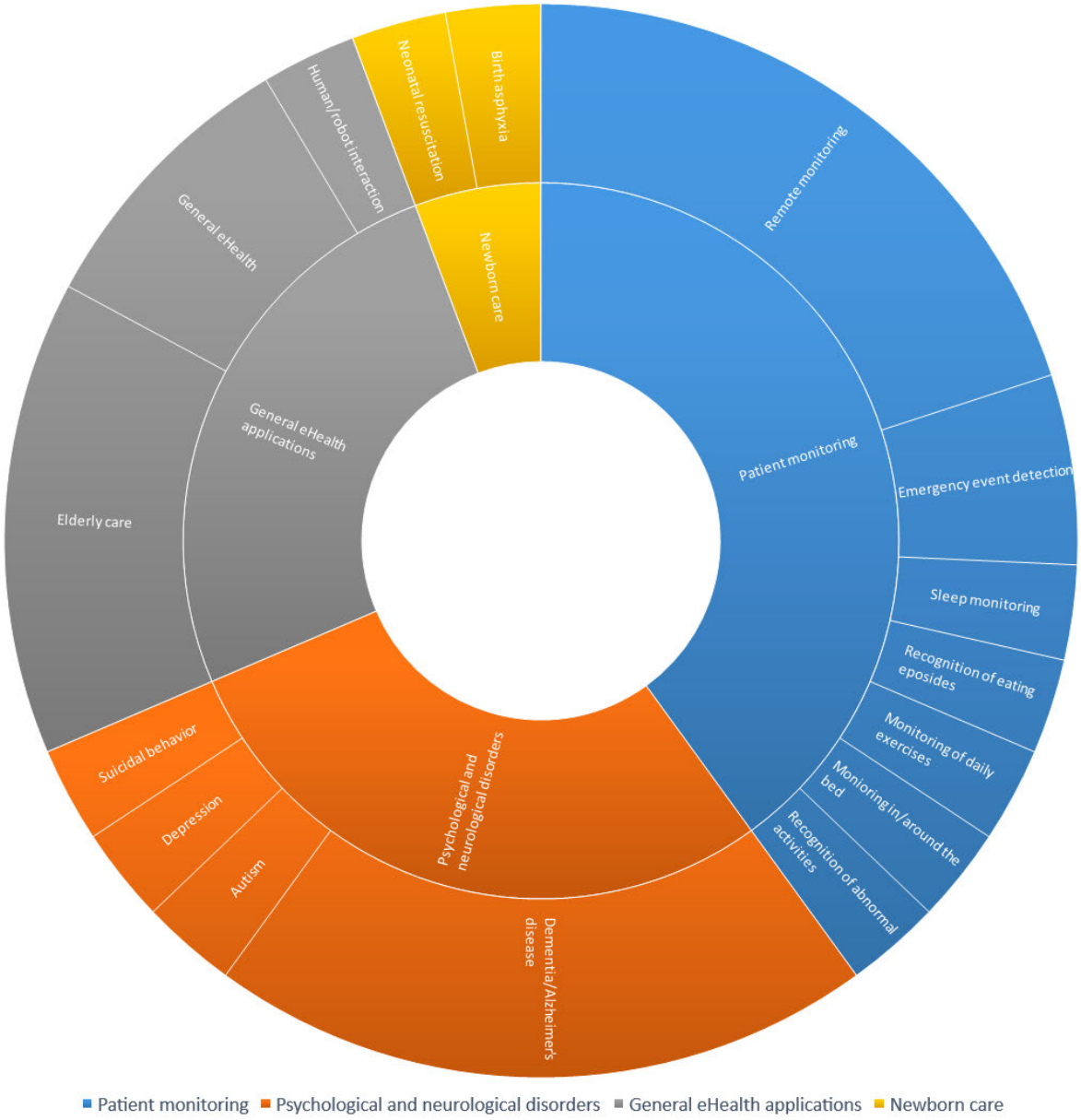

FIGURE 6. **Summary of applications of audio- and video-based HAR systems in healthcare.**

large and high-quality datasets, make this approach more difficult to design, test and implement in real settings. However, multimodal approaches could potentially provide higher flexibility and improved performance than pure audio- or video-based ones.

Another aspect that needs to be addressed in future studies is the implementation and evaluation of HAR systems in real-world conditions, for example by applying large-scale studies in various healthcare settings. Experiments in controlled laboratory environments and theoretical studies are mostly reported in the analysed literature, which may lead to poor generalisation when these systems are applied in practice. Limited available data, often only mimicking real-world environments, prevents from a wider adoption of HAR systems in healthcare. Given that they primarily cater to the elderly or frail population, data collection remains challenging. A possible solution to address this issue involves the utilisation of models that can be trained effectively even with limited available data, leveraging techniques such as unsupervised, semi-supervised, or self-supervised learning. Another approach entails employing transfer learning strategies, where the model is pre-trained on a different domain where data acquisition is easier, and subsequently fine-tuning the model using a restricted amount of real-world data for a particular health outcome. Availability of large-scale general purpose activity data sets,

TABLE 3. Datasets for human activity/behavior recognition in AAL for healthcare.

Modality	Ref.	Application	Dataset	Participants	Size	Data instances
Audio	[12]	Emergency event detection	TUC	58	54 min	1612
Audio	[12]	Emergency event detection	Intenta	NA	4 min	704
Audio	[14]	Suicidal behavior	BBRS	246	7h	5166
Video (RGB)	[46]	Remote monitoring	UCF101	NA	27h	13320
Video (RGB)	[28]	Elderly care	NAD	4	12 min	84
Video (RGB)	[47]	Remote monitoring	Weizmann Human Action	9	NA	90
Video (RGB)	[48]	Remote monitoring	KTH Action	NA	NA	2391
Video (RGB)	[49]	Remote monitoring	UCF Sports Action	NA	16 min	150
Video (RGB)	[50]	Remote monitoring	IXMAS Action	10	NA	1148
Video (RGB)	[36]	Dementia	UNN-GazeEAR	NA	5 min	50
Video (RGB)	[56]	Autism	SSBD	NA	112 min	75
Video (depth)	[32]	Alzheimer's disease	Nursing Home	1	72 h	NA
Video (RGB-D)	[57]	Human-robot interaction	KARD	10	1h	2160
Video (RGB-D)	[51]	Elderly care	MIVIA Action	14	NA	500
Video (RGB-D)	[32]	Alzheimer's disease	CHUN	27	NA	27
Video (RGB-D)	[32]	Alzheimer's disease	GAADRD	25	NA	25
Video (RGB-D, skeleton)	[52]	Human-robot interaction, Elderly care	CAD-60	4	36 min	60
Video (RGB-D, skeleton)	[58]	Human-robot interaction	MSR Daily Activity 3D	10	NA	320
Video (RGB-D, skeleton)	[53]	Elderly care	SBU Kinect Interaction	7	NA	300
Video (RGB-D, skeleton)	[54]	Elderly care	MSR Action Pairs 3D	NA	21 min	360
Video (RGB-D, skeleton)	[55]	Elderly care	Florence 3D	10	NA	215
Video (RGB-D, skeleton, IR)	[29]	Physical irregularity recognition	NTU RGB+D 120	40	NA	56880

such as Kinetics [59], [60] or NTU RGB+D [29], makes this task feasible.

Our results show that privacy and security issues are barely considered, both in audio and video-based HAR, with only few research works that ensure the involvement of all stakeholders in order to increase the privacy and security level of the proposed HAR approach. Even though these works propose concrete steps to ensure a certain level of privacy such as employing privacy-by-context methods [45], written and verbal consent [15] or depth cameras and silhouettes [17], [18], [21], the effectiveness of these methods is not verified in real settings, and policies and legal issues are not sufficiently considered. Addressing the privacy and confidentiality related concerns of study participants by providing them with instructions to remove sensitive audio recordings from their devices, as in [15], may not be sufficient. Processing audio data on-the-fly, thereby avoiding the need for storing them on user devices, such as smartphones and smartwatches, would be a more promising approach, but unfortunately not yet implemented [15]. Audio carries sensitive information that may reveal a person's identity, either related to the speaker's voice characteristics, or to the content of speech. More generally, paralinguistic audio events as stuttering or the style of laughter, may also be considered sensitive, whereas background noise can reveal the current location [61]. Voice masking techniques to cover or degrade the speech, making it unintelligible, while preserving audio information relevant for HAR should be preferred [62]. This is especially important for applications in healthcare, where speech can disclose not just the person's identity, but also their health status.

In video-based HAR, privacy and security are mostly addressed by technical approaches, e.g. by employing depth cameras and silhouettes [17], [18], [21]. In a privacy-by-context setting, users are able to decide how, when and by whom they are watched [45], and may have their visual identity protected by image blurring and replacement with a 3D avatar. Nonetheless, such techniques focus mainly on the user, but leave the remaining image information visible and unprotected. The remaining parts of the image that are left unprotected can not only contain private information about a person's belongings and living environment, but may also contain cues that can be used to infer the person's location. For this reason, within the context of video-based HAR, it is important that the preservation of visual privacy is broadened to also consider image cues that do not necessarily belong to the person's identity, but may still leak private information. Furthermore, the awareness of being monitored

by cameras in one's personal environment may result in modifications of the natural behaviour, leading a person to perform activities in a different manner when they know they are being observed. Protecting a person's privacy may help in easing one's concerns regarding the sensitivity of the data that is being captured, while preserving their natural behaviour.

Through a number of technical innovations for AAL applications, more powerful and effective automated solutions are being developed and experimented, such as those based on conversational agents and robotic assistants. Consequently, the increasing power and pervasiveness of AAL technologies calls for a more holistic approach for privacy preservation that involves various stakeholders such as researchers, policy makers, component and system vendors, service providers and funding bodies. Thus, new approaches to ensure that such solutions respond to the highest ethical, legal and privacy standards and requirements are needed. An extensive survey of different methods and current initiatives in this area has recently been published by the project GoodBrother [63].

VIII. CONCLUSION

In this scoping review, we provided an overview and classification of various video and audio-based approaches for activity/behaviour recognition as applied to healthcare. Our literature search revealed that, while a large number of studies focusing on activity/behaviour recognition exist, only a small portion are applied to healthcare, with the majority of these being video-based methods. Furthermore, our search also revealed an increasing shift towards the adoption of deep learning methods for HAR over the recent years.

Despite the research efforts for human activity recognition (HAR) in healthcare, we have identified open challenges that call for further attention from the HAR community. One of the most pressing issues relates to an under-consideration of privacy and security concerns in many of the proposed HAR approaches; an issue that may have significant consequences in determining the uptake of such HAR approaches in the healthcare industry. Additionally, there is an urgent need for adoption and testing the proposed approaches in realistic settings using appropriate datasets, especially those suitable for multimodal approaches and involving large population and complex situations. Moreover, various other emerging technologies with possible applications in human activity recognition such as, for example, ultra-wide band (UWB) radar may also be a suitable topic for a future work.

REFERENCES

[1] G. Ogbuabor and R. La, "Human activity recognition for healthcare using smartphones," in *Proc. 10th Int. Conf. Mach. Learn. Comput.*, Feb. 2018, pp. 41–46.

[2] E. C. Dinarevic, J. B. Husic, and S. Barakovic, "Issues of human activity recognition in healthcare," in *Proc. 18th Int. Symp. Infoteh-Jahorina (INFOTEH)*, Mar. 2019, pp. 1–6.

[3] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers Robot. AI*, vol. 2, pp. 1–28, Nov. 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2015.00028

[4] M. H. Arshad, M. Bilal, and A. Gani, "Human activity recognition: Review, taxonomy and open challenges," *Sensors*, vol. 22, no. 17, p. 6463, Aug. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/17/6463

[5] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, Jun. 2013. [Online]. Available: https://www.mdpi.com/2073-431X/2/2/88

[6] O. C. Ann and L. B. Theng, "Human activity recognition: A review," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng. (ICCSCE)*, Nov. 2014, pp. 389–393.

[7] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: A literature review," *Pattern Recognit.*, vol. 48, no. 8, pp. 2329–2345, Aug. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320315000953

[8] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: A survey," *Proc. Comput. Sci.*, vol. 155, pp. 698–703, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919310166

[9] N. C. Tay, T. Connie, T. S. Ong, A. B. J. Teoh, and P. S. Teh, "A review of abnormal behavior detection in activities of daily living," *IEEE Access*, vol. 11, pp. 5069–5088, 2023.

[10] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, p. 264, Aug. 2009. [Online]. Available: https://www.bmj.com/content/339/bmj.b2535

[11] J. Kim, K. Min, M. Jung, and S. Chi, "Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition," *Building Environ.*, vol. 181, Aug. 2020, Art. no. 107092.

[12] S. Kahl, H. Hussein, E. Fabian, J. Schlohauer, E. Thangaraju, D. Kowerko, and M. Eibl, "Acoustic event classification using convolutional neural networks," in *Proc. INFORMATIK*, M. Eibl and M. Gaedke, Eds. Gesellschaft fr Informatik, Bonn, 2017, pp. 2177–2188.

[13] T. Giannakopoulos and S. Konstantopoulos, *Daily Activity Recognition Based on Meta-Classification of Lowlevel Audio Events*. Setúbal, Portugal: ScitePress, May 2017, doi: 10.5220/0006372502200227.

[14] B. Stasak, J. Epps, H. T. Schatten, I. W. Miller, E. M. Provost, and M. F. Armey, "Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt," *Speech Commun.*, vol. 132, pp. 10–20, Sep. 2021.

[15] P. Byanjankar, A. Poudyal, B. A. Kohrt, S. M. Maharjan, A. Hagaman, and A. van Heerden, "Utilizing passive sensing data to provide personalized psychological care in low-resource settings," *Gates Open Res.*, vol. 4, p. 118, Aug. 2020.

[16] M. A. Feki, J. Biswas, and A. Tolstikov, "Model and algorithmic framework for detection and correction of cognitive errors," *Technol. Health Care*, vol. 17, no. 3, pp. 203–219, Jul. 2009.

[17] G. Anitha and S. Baghavathi Priya, "Posture based health monitoring and unusual behavior recognition system for elderly using dynamic Bayesian network," *Cluster Comput.*, vol. 22, no. S6, pp. 13583–13590, Nov. 2019.

[18] M. H. Siddiqi, N. Almashfi, A. Ali, M. Alruwaili, Y. Alhwaiti, S. Alanazi, and M. M. Kamruzzaman, "A unified approach for patient activity recognition in healthcare using depth camera," *IEEE Access*, vol. 9, pp. 92300–92317, 2021.

[19] M. H. Siddiqi, H. Alshammari, A. Ali, M. Alruwaili, Y. Alhwaiti, S. Alanazi, and M. M. Kamruzzaman, "A template matching based feature extraction for activity recognition," *Comput., Mater. Continua*, vol. 72, no. 1, pp. 611–634, 2022.

[20] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, no. 7, pp. 11735–11759, Jul. 2014.

[21] Z. A. Khan and W. Sohn, "Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1843–1850, Nov. 2011.

[22] Y.-K. Wang, H.-Y. Chen, and J.-R. Chen, "Unobtrusive sleep monitoring using movement activity by video analysis," *Electronics*, vol. 8, no. 7, p. 812, Jul. 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/7/812

[23] M. Su, D. W. Hayati, S. Tseng, J. Chen, and H. Wei, "Smart care using a DNN-based approach for activities of daily living (ADL) recognition," *Appl. Sci.*, vol. 11, no. 1, p. 10, Dec. 2020.

[24] R. R. Subramanian and V. Vasudevan, "A deep genetic algorithm for human activity recognition leveraging fog computing frameworks," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103132.

[25] A. Manocha, G. Kumar, M. Bhatia, and A. Sharma, "Video-assisted smart health monitoring for affliction determination based on fog analytics," *J. Biomed. Informat.*, vol. 109, Sep. 2020, Art. no. 103513.

[26] Y. Gao, X. Xiang, N. Xiong, B. Huang, H. J. Lee, R. Alrifai, X. Jiang, and Z. Fang, "Human action monitoring for healthcare based on deep learning," *IEEE Access*, vol. 6, pp. 52277–52285, 2018.

[27] T. Singh and D. K. Vishwakarma, "A deeply coupled ConvNet for human activity recognition using dynamic and RGB images," *Neural Comput. Appl.*, vol. 33, no. 1, pp. 469–485, Jan. 2021.

[28] Y. Chen, L. Yu, K. Ota, and M. Dong, "Robust activity recognition for aging society," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 6, pp. 1754–1764, Nov. 2018.

[29] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[30] Y. A. Andrade-Ambriz, S. Ledesma, M.-A. Ibarra-Manzano, M. I. Oros-Flores, and D.-L. Almanza-Ojeda, "Human activity recognition using temporal convolutional neural network architecture," *Expert Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116287, doi: 10.1016/j.eswa.2021.116287.

[31] H. Lee and S. Youm, "Development of a wearable camera and AI algorithm for medication behavior recognition," *Sensors*, vol. 21, no. 11, p. 3594, May 2021.

[32] C. Crispim-Junior, A. Gómez Uría, C. Strumia, M. Koperski, A. König, F. Negin, S. Cosar, A. Nghiem, D. Chau, G. Charpiat, and F. Bremond, "Online recognition of daily activities by color-depth sensing and knowledge models," *Sensors*, vol. 17, no. 7, p. 1528, Jun. 2017.

[33] A. Karakostas, A. König, C. F. Crispim-Junior, F. Bremond, A. Derreumaux, I. Lazarou, I. Kompatsiaris, M. Tsolaki, and P. Robert, "A French–Greek cross-site comparison study of the use of automatic video analyses for the assessment of autonomy in dementia patients," *Biosensors*, vol. 10, no. 9, p. 103, Aug. 2020.

[34] C. Lee, H. Choi, S. Muralidharan, H. Ko, B. Yoo, and G. J. Kim, "Machine assisted video tagging of elderly activities in K-log centre," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2020, pp. 237–242.

[35] F. Negin and F. Brémond, "An unsupervised framework for online spatiotemporal detection of activities of daily living by hierarchical activity models," *Sensors*, vol. 19, no. 19, p. 4237, Sep. 2019.

[36] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu, "Gaze-informed egocentric action recognition for memory aid systems," *IEEE Access*, vol. 6, pp. 12894–12904, 2018.

[37] P. Washington, A. Kline, O. C. Mutlu, E. Leblanc, C. Hou, N. Stockham, K. Paskov, B. Chrisman, and D. Wall, "Activity recognition with moving cameras and few training examples: Applications for detection of autism-related headbanging," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–7.

[38] Ø. Meinich-Bache, S. L. Austnes, K. Engan, I. Austvoll, T. Eftestøl, H. Myklebust, S. Kusulla, H. Kidanto, and H. Ersdal, "Activity recognition from newborn resuscitation videos," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3258–3267, Nov. 2020.

[39] Ø. Meinich-Bache, K. Engan, I. Austvoll, T. Eftestøl, H. Myklebust, L. B. Yarrot, H. Kidanto, and H. Ersdal, "Object detection during newborn resuscitation activities," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 3, pp. 796–803, Mar. 2020.

[40] P. W. Aung Aung, S. F. V. Foo, W. Huang, J. Biswas, J. E. Phua, K. Liou, and C.-C. Hsia, "Evaluation and analysis of multimodal sensors for developing in and around the bed patient monitoring system," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 2159–2162.

[41] A. M. Khattak, P. T. H. Truc, L. X. Hung, L. T. Vinh, V.-H. Dang, D. Guan, Z. Pervez, M. Han, S. Lee, and Y.-K. Lee, "Towards smart homes using low level sensory data," *Sensors*, vol. 11, no. 12, pp. 11581–11604, Dec. 2011.

[42] V.-D. Le, V.-H. Dang, S. Lee, and S.-H. Lee, "Distributed localization in wireless sensor networks based on force-vectors," in *Proc. Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process.*, Dec. 2008, pp. 31–36.

[43] Q. Ning, Y. Chen, J. Liu, and H. Zhang, "Heterogeneous multimodal sensors based activity recognition system," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–4.

[44] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, "EarBit: Using wearable sensors to detect eating episodes in unconstrained environments," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–20, Sep. 2017, doi: 10.1145/3130902.

[45] A. Chaaraoui, J. Padilla-López, F. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context," *Sensors*, vol. 14, no. 5, pp. 8895–8925, May 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/5/8895

[46] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[47] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2005, pp. 1395–1402.

[48] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, 2004, pp. 32–36.

[49] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer Vision in Sports*. Cham, Switzerland: Springer, 2014.

[50] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 249–257, Nov. 2006.

[51] V. Carletti, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "Recognition of human actions from RGB-D videos using a reject option," in *Proc. Int. Workshop Social Behav. Anal.*, 2013, pp. 1–15.

[52] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 842–849.

[53] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 28–35.

[54] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 716–723.

[55] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 479–485.

[56] S. S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 755–761.

[57] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-D posture data," *IEEE Trans. Hum.-Mach. Syst.*, vol. 45, no. 5, pp. 586–597, Oct. 2015.

[58] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1290–1297.

[59] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[60] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the Kinetics-700–2020 human action dataset," 2020, *arXiv:2010.10864*.

[61] J. Williams, K. Pizzi, S. Das, and P.-G. Noé, "New challenges for content privacy in speech and audio," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, Sep. 2022, p. 16.

[62] D. Liang, W. Song, and E. Thomaz, "Characterizing the effect of audio degradation on privacy perception and inference performance in audio-based human activity recognition," in *Proc. 22nd Int. Conf. Hum.-Comput. Interact. Mobile Devices Services*, Oct. 2020, pp. 1–12.

[63] S. Aleksic, L. Colonna, C. Dantas, A. Fedosov, F. Florez-Revuelta, E. Fosch-Villaronga, H. G. Jevremovic, A. Msaknic, S. Ravi, B. Rexha, and A. Tamö-Larrieux, "State of the art in privacy preservation in video data," Cost Action GoodBrother, Netw. Privacy-Aware Audio-Video-Based Appl. Active Assist. Living, Tech. Rep. WG02, 2022, doi: 10.5281/zenodo.6806206.

STEFANIA CRISTINA (Member, IEEE) received the Ph.D. degree in engineering from the University of Malta.

She is currently a Lecturer. Her research work combines computer vision with machine learning, with particular research interest includes video-based eye-gaze tracking as an assistive human–computer interaction tool for persons with mobility impairments. She is also the Chair of the Vision & Imaging Technical Network with the Institution of Engineering and Technology (IET).

VLADIMIR DESPOTOVIC (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Nis.

Currently, he is a Research Scientist with the Bioinformatics Platform, Luxembourg Institute of Health. Previously, he was an Adjunct Professor with the University of Belgrade and a Postdoctoral Researcher with the University of Luxembourg and the University of Paderborn. His research interests include the intersection of machine learning and audio/image signal processing, with an expertise in development of medical assistive technologies and digital health solutions for people with speech and motor disabilities.

RODRIGO PÉREZ-RODRÍGUEZ received the Ph.D. degree in biomedical engineering from Universidad Politécnica de Madrid.

He is currently a Senior Researcher and a Higher Education Professional. He is also a Professor with Universidad Rey Juan Carlos, he actively collaborates with the Intelligent Robotics Laboratory. He has been an eHealth Coordinator with the Biomedical Research Foundation—Getafe University Hospital and an Adjunct Professor with Universidad Carlos III de Madrid (UC3M). His research interests include robotics, cognitive models, and monitoring technologies, always applied to improve the wellness of the older population. He has participated in many research project both as a Collaborator and a Principal Investigator.

SLAVISA ALEKSIC (Senior Member, IEEE) received the Ph.D. degree in electrical engineering and the Habilitation (Venia Docendi Teaching Authorisation) degree in communication networks from the Vienna University of Technology.

He spent more than 15 years with the Vienna University of Technology in different positions and with several institutes. He was a Visiting Researcher with INTEC-IBCN, Ghent University, and a Professor with Hochschule für Telekommunikation (HfTL), Leipzig. Currently, he is a Professor with Hochschule für Technik, Wirtschaft und Kultur (HTWK), Leipzig, where he is also responsible for teaching and research within the broad area of network technologies and network management. He has authored or coauthored more than 140 scientific publications and acted as a collaborator and a principal investigator in a number of research projects funded by various international and national research funding organizations and in collaboration with industry. His research interests include communication technologies, networks, energy efficiency, network management, and smart environments. He received several international and national awards and grants, such as five best paper awards, the Biggest Austrian Business Plan Award i2b (second level), and the grant of the Austrian Federal Ministry of Education, Science and Culture.