Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07665-9

REVIEW

A review of machine learning-based human activity recognition for diverse applications

Farzana Kulsoom[2] · Sanam Narejo[3] · Zahid Mehmood[1] · Hassan Nazeer Chaudhry[4] · butt Aisha[3] · Ali Kashif Bashir[5]

Received: 3 September 2021 / Accepted: 20 July 2022

Abstract
Human activity recognition (HAR) is a very active yet challenging and demanding area of computer science. Due to the articulated nature of human motion, it is not trivial to detect human activity with high accuracy for all applications. Generally, activities are recognized from a series of actions performed by the human through vision-based sensors or non-vision-based sensors. HAR's application areas span from health, sports, smart home-based, and other diverse areas. Moreover, detecting human activity is also needed to automate systems to monitor ambient and detect suspicious activity while performing surveillance. Besides, providing appropriate information about individuals is a necessary task in pervasive computing. However, identifying human activities and actions is challenging due to the complexity of activities, speed of action, dynamic recording, and diverse application areas. Besides that, all the actions and activities are performed in distinct situations and backgrounds. There is a lot of work done in HAR; finding a suitable algorithm and sensors for a certain application area is still challenging. While some surveys are already conducted in HAR, the comprehensive survey to investigate algorithms and sensors concerning diverse applications is not done yet. This survey investigates the best and optimal machine learning algorithms and techniques to recognize human activities in the field of HAR. It provides an in-depth analysis of which algorithms might be suitable for a certain application area. It also investigates which vision-based or non-vision-based acquisition devices are mostly employed in the literature and are suitable for a specific HAR application.

Keywords Machine learning · Human activity recognition · Activities of daily living · Sensors · Videos

1 Introduction

Human activity recognition (HAR) aims to recognize the actions carried out by a particular person based on certain information about that person and his ambiance. Accordingly, HAR is a study of the interpretation of human body gestures or motion with sensors, images, and video sequences [11]. It has been actively investigated for a wide range of applications and real-world problems, including healthcare [139, 144], sports training [73], abnormal behavior detection [77, 145], content-based video analysis [193], robotics, human–computer interaction [210], visual surveillance [25, 31, 97, 193], video indexing [122, 177], smart homes [99, 151, 181, 231] ambient intelligence [166, 177], and several other areas [134]. In ambient intelligence, ambient sensors are installed in the human habitat; these sensors are sensitive to the presence of humans and can respond to human activity. The sensors categories include a wide variety of sensors, such as motion detectors, door sensors, object sensors, pressure sensors, and temperature sensors. These types of sensors are deployed in the environment to monitor and record the actions [51, 74, 90]. Video indexing permits automating the recognition and isolation of videos efficiently based on their scenarios and contents. For example, it can identify and index videos based on different activities and conditions like sports-based videos, shopping malls videos, home videos, etc.

The interpretation of activity may vary as per the application area and domains; however, specific activity is generally a collection of a particular set of actions. For example, the activity of washing clothes may consist of

Extended author information available on the last page of the article

Published online: 07 August 2022

pre-soaking, rinsing, washing, and drying actions. Generally, these activities are performed in a specific time window and may be performed in different forms. Subsequently, activities can be categorized into four major categories, as shown in Fig. 1. **Composite activities** are composed of a set of complex and overlapping activities. Composite activity is made up of more complex behaviors, such as cooking or cleaning. It can be seen in Fig. 1; cooking involves turning on the stove, the addition of pasta, cooking pasta, and turning the stove off. Similarly, playing tennis is made up of volleys, smash, service, dropping the ball, running and so on. On the other hand, the **concurrent activity** involves the number of tasks performed simultaneously or concurrently. For example, a person might eat a snack while watching his favorite movie. A logical order or sequence is followed in the involved steps in an operational plan for execution in another type called **sequential activities**. For instance, drinking water from the refrigerator requires opening action before water can be consumed and logically followed by closing the fridge. Finally, **interleaved activities** are linked with each other and can be switched back and forth. For example, a person might read the novel, suspend it for a while, write its summary, and switch back to reading. Figure 1 shows composite, concurrent, sequential, and interleaved activities with examples.

The information or data in the HAR is indexed over time dimension. Thus, the time intervals are consecutive, non-overlapping, and non-empty. Generally, the activities are not simultaneous, i.e., a subject cannot "sit" and "stand," "run," and "walk" in a single time frame. Noticeably, the HAR problem is not feasible to be solved deterministically. It is also possible that the number of combination of input attributes and activities become very large or even infinite in some rare cases, and finding the transition points becomes challenging as the exact duration of each activity is generally unknown. Subsequently, before feature extraction and selection, a relaxed version of the problem is then introduced. In this step, the time series sequential data is divided into fixed-length time windows and thereby, filtering the relevant information from the raw signal or video sequences.

HAR consists of several steps; a typical flow for HAR is shown in Fig. 2. Initially, actions are recorded using data acquisition devices such as sensors, and cameras, further explained in Sects. in 1.1 and 1.2, respectively. The data obtained from these devices is mostly acquired in raw form with redundant information; therefore, a preprocessing is required. Besides, the data may not be in the required shape for the other steps in the pipeline. The preprocessing involves different types of filters, transformation, reductions, and other techniques, further explained in 1.3. Once data is preprocessed, machine learning techniques are applied to it to identify or classify different human activities in the next step. In the following paragraph, data acquisition using a diverse variety of sensors is discussed in detail.

1.1 Non-vision-based HAR

If the activity has to be monitored for a brief period, wearable sensors are preferred. For long-term monitoring of human activity, implanted and external sensors are employed. In the case of wearable sensors, the device is attached to the human body. Additionally, this category also includes smart devices, for example, smartwatch, smart glasses for the visual and hearing disabled, and smart shoes. In some implants, the devices monitor the body's internal activity; one particular example could be implanted EMG sensors. Another possible way is external sensors, where the devices are fixed in predetermined points of interest. These types of sensors are widely used in traffic control and management systems. Resulting in involuntary interaction between the users and sensors. It also includes objects that constitute the activity environment, namely **dense sensing.**

Wearable sensors often utilize inertial measurement units and radio frequency identification device (RFID) tags to gather an actor's behavioral information. This approach is effective for recognizing physical movements such as physical exercises. In contrast, dense sensing infers activities by monitoring human-object interactions through the usage of multiple multi-modal miniaturized sensors. Smartphone-based wearable sensors are popular alternative methods of inferring human activity details. It can be used to connect a wide range of sensors, i.e., Wi-Fi, Bluetooth, microphones, accelerometer, gyroscope, magnetometer, light sensors, and cellular radio sensors. These sensors are employed to infer human activity details for diverse applications. Sensors such as accelerometer, gyroscope, magnetometer, implanted sensors [15, 194] and, global position system (GPS) can be deployed for coarse grain and context activity recognition, user location, and social interaction between users. Motion sensors (Accelerometers, gyroscopes, magnetometers) provide significant information that facilitates recognition and monitoring of users' movements such as walking, standing, or running. Similarly, proximity and light sensors that are generally embedded in mobile devices to enhance user experiences can also be deployed to determine whether the user is in light or dark. Other sensors such as barometers, thermometers, air humidity, and pedometers have also been employed to monitor the healthy status of elderly citizens and for assisted living. For instance, the pedometer found in the Samsung Galaxy smartphones and exercises tracking wearable devices is essential for step counts, heart rate, and

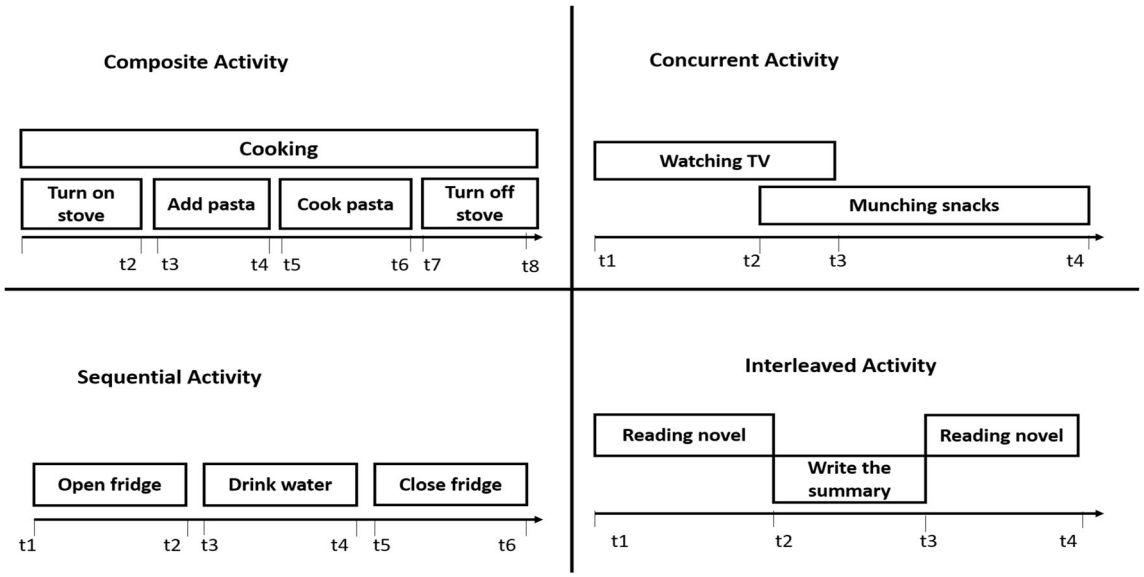


Fig. 1 Categorization of an Activity based on sequence of operations

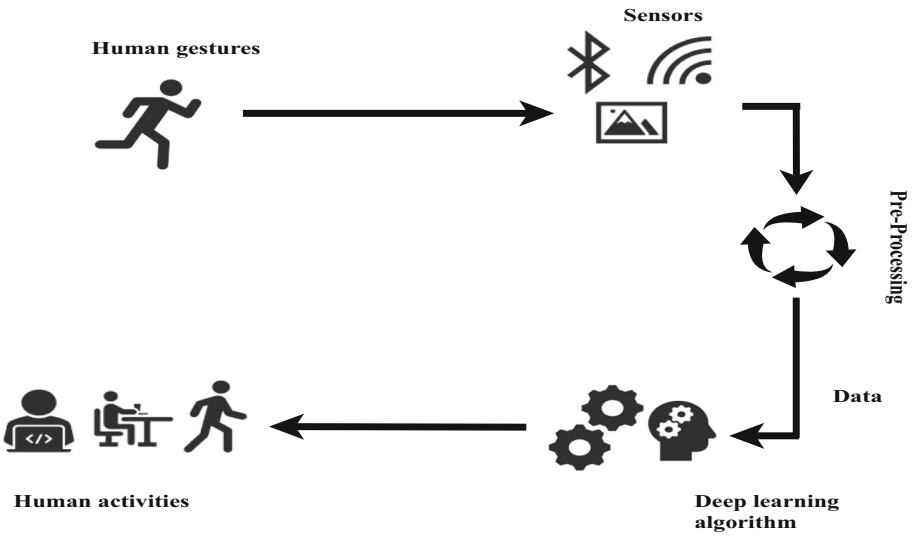


Fig. 2 Human activity recognition's Flow diagram

pulse monitoring. These sensors' measurements could be noisy because of displacement or other unsuitable conditions. To eliminate noise from the data different types of thresholding or filtering techniques could be applied [192].

Table 1 summarizes various dataset containing different types of activities. The activities include Ambient Assisted Living (AAL), Paroxysmal atrial fibrillation (PAF) detection, and activity of daily living (ADLs). The activities range from 2 to 35 activities with different machine learning techniques applied to them. Moreover, it also explains how different activities are captured using numerous sensors, including an accelerometer, gyroscope, magnetometer, and electrocardiogram (ECG). Each activity is recorded at a particular sampling rate; a higher sampling rate means that the activity has more data. The higher sampling rate translates into finer granularity and real-time detection of human activities. On the other hand, low sampling rates mean that fewer snaps shots are available per minute. It results in faster processing, less storage, and bandwidth consumption; however, it follows event recognition limitations and lower resolution. In some cases, the activities are not immediately trivial and are required to be extracted from data. For this purpose temporal pattern extraction and recognition, algorithms are applied to the data to extract activities [37].

1.2 Vision-based HAR

Vision-based activity recognition is one of the pioneering approaches. It has been a research focus for a long time due

to its significant role in intelligent video surveillance, health care, smart home, AAL, human–computer interaction, robot learning, emotion recognition [85] and video labeling. The primary aim of vision-based HAR is to investigate and interpret activities from a video (i.e., a sequence of image frames) or directly from images. Therefore, vision-based methods utilize video cameras to identify human actions and gestures from video sequences or visual data. Due to the prevailing development in technology, camera devices are upgrading. In response to this, novel approaches for vision-based HAR are constantly emerging.

In the recent past, an ample amount of valuable information, for instance, three-dimensional structures, can be obtained using 3D depth cameras compared to traditional cameras. Literature suggests that a wide variety of modalities, such as a single or multi-camera, stereo, and infra-red, are applied to understand and investigate various HAR applications. Vision-based methods employ cameras to detect and recognize activities using several different computer vision techniques, such as object segmentation, feature extraction, and feature representation. The appropriate cameras for capturing the activity greatly impact the overall functionality of the recognition system. As discussed earlier, vision-based HAR is a more challenging problem due to motion and variation in human shape, occlusions, cluttered backgrounds, stationary or moving cameras, different illumination conditions, light intensity, and viewpoint variations. However, the severeness of these challenges depends upon the kind of HAR application. Table 2 summarizes many datasets of video having data related to many human activities. As far as the time domain decomposition of activity is concerned, the variety of HAR applications results in a considerably extensive range.

The recognition of video sequences' actions involves complex steps, including pre-processing images or space-

Table 1 A comparison of literature on the activity's dataset for Non-Vision-based human activity recognition

Refs.	Year	Dataset	Types of activities	Num of activities	Sensors	Method learning	Sampling rate (Hz)
[121]	2020	HAR	ADL	3	Single triaxial A, triaxial G	SVM and Logistic Regression	50
[124]	2020	Opportunity	AAL	7	Wearable sensors, object sensors, and ambient sensors	DBN	50
[153]	2018	PAF	PAF	2	ECG	CNN	128
[127]	2018	PAMAP2	ADL	18	A,G,M	CNN	100
[5]	2017	UCI Smartphone	ADL	6	A,G	SAE	50
[62]	2016	Daphnet Gait	Gait	2	A	CNN	64
[7]	2016	WISDM	ADL	6	A	RBM	20
[81]	2016	Self	ADL	9		Bagged tree model	
[142]	2016	Opportunity	ADL	16	A, G, M, AM	CNN, RNN	32
[159]	2016	ActiveMiles	ADL	7	A	CNN	50–200
[215]	2015	ActRecTut	Gestures	12	A,G	CNN	32
[82]	2015	USC-HAD	ADL	12	A,G	CNN	100
		SHO	ADL	7	A, G, M	CNN	50
[61]	2015	MHEALTH	ADL	12	A, C, G	CNN	50
[65]	2015	HASC	ADL	13	A	RBM	200
[227]	2015	DSADS	ADL	19	A, G, M	DBN	25
[230]	2014	BIDMC	Heart failure	2	ECG	MC-DCNN	125
[221]	2014	Actitracker	ADL	6	A	CNN	20
[102]	2012	Self	ADL	6	A	Additive Logistic Regression	50
[151]	2011	Ambient kitchen	Food preparation	2	A	RBM	40
	2011	Darmstadt Daily	ADL, Food	35	A	RBM	100
		Routines	preparation, FACTORY				

time volume video data, feature extraction concerning actions, and action modeling based on the extracted features. Therefore, to acquire accurate and meaningful representations as input features for the classifier, some well-defined ways are categorized as global, local, and depth-based representations. It is evident from the literature that initially, studies attempted to model the whole images or silhouettes and represent human activities globally, where space-time shapes are generated as the image descriptors. Subsequently, significant attention was diverted towards the new local representation view, the evolution of space-time interest points (STIPs), which focuses on the informative interest points. Apart from this, other local descriptors, for example, a histogram of optical flow (HOF) and histogram of oriented gradients (HOG) from the domain of object recognition, are widely adopted to 3D in the HAR area. With the latest camera devices' latest advancements, specifically the evolution of RGB-D cameras, currently, depth image-based representations are used.

1.3 Pre-processing of data

After acquiring data from sensors, i.e., images, videos, and sensors, it is further processed to prepare it for upcoming blocks in the pipeline. The primary steps are performed to remove noise from data, extract salient and discriminative features, remove background or isolation of certain areas of interest, and resample data to meet specific requirements. The most primary and commonly used operation in pre-processing is the removal of unwanted noise. Therefore, various approaches can be utilized, such as nonlinear filtering, Laplacian, and Gaussian filters. Another frequently used operation is segmentation; it involves dividing the signal into small window sizes to extract prominent features. The next step is to extract features to reduce computational time and enhance classification accuracy. Additionally, if these features are still very huge, they are further reduced by utilizing the dimensionality reduction method or selecting the most discriminative features to identify human activity. There are two types of feature vectors for human activity recognition; the first one involves statistical features, and the other one is based on structural features. Common statistical features are mean, median, standard deviation, time, and frequency domain representation. These features are based on the qualitative properties of the acquired data. On the other hand, the structural features are based on the relationship between the mobile sensor's data. To reduce the computational complexity dimensionality reduction algorithms like principal component analysis (PCA), linear discriminate analysis (LDA), and empirical cumulative distribution functions (ECDF) are used.

While doing preprocessing on images and videos, the features can be represented in image space. With videos, these features represent the pose of human action in image space and represent the change in the state of that particular action. Hence with videos-based HAR, the feature representation is extended from 2D space to 3D space. In recent years several methods have been adopted to represent actions, including local and global features based on temporal and spatial changes [165], trajectory features based on keypoint tracking [9, 126, 207], motion changes based on depth information [23, 24, 217] and features based on human action and pose changes [46, 220]. Deep learning had been prevalent for image classification and object detection; many researchers have also applied deep learning to human action recognition. This approach enables to automatically generate action features from sensed data [142, 161, 230]. Human activity recognition is one of the popular research areas; therefore, several surveys are already published in this field as shown in Fig. 6 with the timeline. Then Table 3 demonstrates highlights of exiting surveys in terms of activities and algorithms discussed.

The works can be broadly classified into surveys related to vision-based [89, 193, 20, 177, 97] and non-vision-based HAR [13, 101, 203]. Due to the increasing application and

Table 2 A comparison of literature on the activity's dataset for Vision-based human activity recognition

Refs.	Year	Dataset	Type	Video clip	Classes
[96]	2011	HMDB51	Action recognition	7000	51
[175]	2012	UCF101	Sports	13,320	101
[87]	2014	Sports-1M	Sports	1,100,000	487
[18]	2015	ActivityNet	Human Activities	28,000	203
[172]	2016	Charades	Human activities	9848	157
[231]	2018	YouCook2	Cooking videos	2000	15,400
[163]	2018	How2	Instructions videos	13,168	1,84,949
[125]	2019	Moments in Time	Action recognition	1,000,000	339
[181]	2019	COIN	Instructions videos	11,827	180
[120]	2019	HowTo100M	Captioning	1,200,000	120
[45]	2019	Oops	Classification	20,723	–

popularity of deep learning recently, some surveys provided an in-depth deep learning perspective for HAR [199]. Similarly, other related surveys presented different machine learning techniques for HAR [157]. Some special surveys covered narrow areas like group activity recognition [97, 206], use of context, and domain knowledge [141], middleware [141], online activity recognition using mobile phones [171], and use of 3D data [2]. HAR is applied in various domains, and the existing literature does not cover application-based surveys for HAR. Though some works cover specific applications domain such as health care [203] and sports-based application [13]. However, to the best of our knowledge, no recent survey covers datasets, machine learning algorithms, and techniques for diverse application domains in depth. The literature indicates that State-of-the-art machine learning and deep learning algorithms are outperforming and providing excellent results in HAR's domain.

Although online activity recognition is very beneficial though challenging, in most of the literature only offline recognition of activities is covered. Moreover, it has been analyzed that decision trees, support vector machine (SVM), Hidden Markov Model (HMM), and K-Nearest-Neighbor (KNN) are mostly used classifiers for HAR. As per our analysis, this work not only covered the latest literature related to machine learning, some advanced learning-based techniques like reinforcement learning are only covered in detail in this paper. The primary focus of this survey is to investigate the best-suited algorithms and techniques for human activity recognition for diversified application domains. In the beginning, this paper provides a brief introduction to HAR with sensors, images, and videos. It provides an organized review of HAR's main techniques and solutions, including various Machine learning approaches. Moreover, the paper also provides a comprehensive survey and comparative analysis of HAR's applications. Additionally, this study indicates the current trends, challenges, and applications for HAR.

The remainder of the paper is organized as follows. Section 2 overviews the main concept of HAR with sensors, images, and videos and categorizes the different applications. Section 3 refers to a brief description of the traditional machine learning approaches in terms of discriminative and generative models and their implementation in HAR. Sections 4 and 5 deal with deep learning architectures and transfer learning. Section 6 presents Reinforcement learning. Subsequently, Sect. 7 deals with a few more machine learning-related techniques. Section 8 provides a discussion on the performance analysis of various HAR models by comparing a variety of research work that is recently used by different authors. Besides that future directions and limitation of HAR-based system is also presented in the aforementioned section. Section 9 deals with the conclusion of the study.

2 Applications of HAR

HAR finds applications in a wide spectrum of domains including health-care [139, 144], abnormal behavior and fall detection [77, 145], exercise and sports training assistance systems [73], smart homes [181, 231], crowd surveillance and video content analysis [193] are few examples. Each area has several modalities where HAR is applied in numerous subareas, for example, health-care HAR includes patient monitoring of ICU patients [28, 142]. Similarly, smart homes, it is used to assist elderly people, activity monitoring of children, and help dementia patients. The recent research on AI has made humans more inclined to identify objects, actions, and time series analysis. This section investigates which kind of sensors and videos-based acquisition devices are mostly used in the literature and suitable for a specific HAR application, as shown in Table 4. We summarize many recent works and present a new research survey on human action recognition techniques, including classic machine learning algorithms and advanced deep learning architectures over sensor-based, vision-based HAR and audio-based HAR [33]. For classification, SVM, neural network (NN), Gaussian Mixture Model (GMM), HMM, and Kernel extreme learning machine (KELM) classifier are considered the most popular in activity recognition. KELM classifier enhanced the capability of an extreme learning machine (ELM) by transforming linearly non-separable data in a low dimensional space into a linearly separable one. While GMM is mostly used in unsupervised learning, where Gaussian distributed sub-groups are formed within data based on a specific feature. On the other hand, HMM-based classification is still restricted to supervised learning. HMM has been proven very successful in classifying sequential events. Therefore if some activities require to get benefited from sequential information of events like in the online activity recognition, HMM is very robust.

The number of inertial sensors and their location on the human body has a significant effect on the type of human activity to be monitored and classified [12]. Several types of indoor movement, such as standing, walking, or climbing ascending and descending stairways, are determined in [84] using support vector machines in conjunction with an Inertial Measurement Unit (IMU). An IMU is a device that uses gyroscopes and accelerometers to measure and report angular rate and specific force. It is demonstrated in [131] that walking, running, and jogging share similar properties in terms of angular movement. This could be highly beneficial for discovering irregularities in human actions and

Table 3 Comparison of activities and algorithm covered in other surveys

Problem considered	The existing survey											
	[13]	[20]	[177]	[97]	[101]	[206]	[141]	[171]	[2]	[203]	[89]	Our work
Online recognition	✓				✓			✓				✓
Offline recognition	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Vision based		✓	✓	✓		✓	✓		✓	✓	✓	✓
Non-vision based	✓				✓			✓		✓		✓
Group activity			✓	✓	✓	✓					✓	✓
Daily activity				✓	✓			✓	✓	✓	✓	✓
Sports activity	✓	✓	✓	✓		✓	✓		✓	✓		✓
Surveillance		✓	✓	✓		✓	✓					✓
Health care	✓				✓			✓		✓	✓	✓
SVM	✓		✓	✓	✓			✓		✓	✓	✓
GMM	✓										✓	✓
HMM	✓			✓		✓	✓	✓	✓	✓	✓	✓
LSTM						✓				✓		✓
CNN						✓				✓		✓
RNN												✓
ANN	✓				✓						✓	✓
Reinforcement learning												✓
Bagging					✓							✓
Decision tree	✓							✓		✓		✓

identifying any outliers. The research conducted in [185] estimates the body joint angles features by co-registering a 3-D body model to the stereo information from the time-sequential activity video frames. The aforementioned study indicates that with 3-D joint angle information, substantially stronger features and attributes may be formed than depth and binary features, this could significantly enhance the HAR.

Joint movement recognition is mostly recognized in sports-related applications; mostly, depth maps are used. Depth maps are image channel that provides information about the distance of the targeted object from a viewpoint. However, the joint activity perception with a depth map will increase the processing time. To reduce the computational complexity and to increase speed, data dimension reduction is essential. For that purpose, data must be reduced efficiently like it must contain the depth information by completely preserving the depth map sequence. For example, [24] uses PCA for dimensionality reduction of features, and then classification is performed. In [105] authors introduces 3D human recognition method from offline to online. Methods use skeletal sequences [9, 39, 44, 46, 57, 113, 189, 220], depth maps [23, 24, 106, 202, 209, 217], both of skeletal sequences and depth maps [140, 198, 219], or RGB-D sequences [48, 205] as motion data for action recognition method. Heart rate monitoring is helpful not only for one's well-being [192] but also has some relation with physical activities [182]. Real-time daily and sports activities have been recognized in [182] with partial information from heart monitoring. It has been observed that only heart rate monitoring activities cannot be recognized accurately since it is influenced by environmental and emotional factors. However, heart rate has some relation with the energy consumption during various activities.

For health-based applications, irregular activity can be recognized by determining motion recognition. Motion recognition is very challenging, particularly if it contains the repetition of actions and abnormal activity. The authors in [187] have utilized smartphone sensors like accelerometer (A), gyroscope (G) proximity, light (L), and magnetometer (M) sensors to detect complex joint movements. It was observed that static states in which a person is in a steady-state concerning sensors, like lying, sitting, and standing, are easy to identify. In contrast to that, the dynamic states in which the person is in constant movement concerning sensors, like fast turn, U-turn, moving forward and backward, are challenging and difficult to recognize.

Further, in [139], experimental studies show that these sensors can be used individually to recognize human activity. The accelerometer sensor gives better

Table 4 Application wise categorization of HAR

Application	Refs.	Dataset	Type (vision/ non-vision)	ML algorithm	Objective
Sports	[24]	MSRAction3D and MSRGesture3D	Video	PCA, DMM-LBP-DF and KELM	Fast game postures
	[217]	MSR Action3D	Video	DMM-HOG	Track human body joints
	[23]	MSRAction3D and MSRGesture3D	Video	KELM	Game postures
	[106]	MSR Action3D	Video	GMM	Track human body joints
	[232]	KTH Actions, UCF Sports, Youtube Actions	Video	HMM	Sports activities
	[9]	MSR Action-3D, MSR Daily Activity and 3D Action	Video	SVM	Sports activities
	[182]	self	Sensor (A) and heart rate monitor	NB classifier	Gymnasium activities
Smart Home	[99]	off-self shore	Sensors based	ANN	Daily activities recognition
	[181]	COIN	Video	Fully and weakly supervised approach	Multi-domain daily activity analysis
	[151]	Ambient kitchen	Sensor (A)	RBM	Food preparation
	[231]	YouCook2.	Video	ProcNets	Food preparation
	[151]	Darmstadt Daily Activities	Sensor (A)	RBM	Food preparation
	[198]	CMU MoCap dataset, MSR-Action3D dataset , MSR-DailyActivity3D, Cornell Activity dataset, and Multiview 3D Event	Video	SVM	Human-object interactions
	[153]	Skoda checkpost	Sensor(A)	RBM	Factory, food preparation
	[205]	RGB-D	Video	Unsupervised Learning (HMM)	Composite action recognition
	[81]	self	Mobile sensors	ensembled-bagged tree	Detect ADL
Health care	[144]	self	Sensor	Junction Tree Algorithm	Automated health care
	[139]	UCI	Sensor (A, G)	MLP	Chronic diseases monitoring
	[187]	self-collected and public	Sensors(A, G, M)	KNN and K-mean	Automated healthcare
	[230]	BIDMC	ECG	CNN	Heart failure
	[61]	MHEALTH	Sensors (A, C, G)	CNN	ADL
	[185]	self	Video	HMM	Detect joint angle features
	[62]	Daphnet Gait	Sensor (A)	CNN	Abnormal gait detection
	[77]	self	Video	SIFT and HMM	Abnormal activity
	[218]	self	Sensor	SVM	Abnormal activity
	[145]	DLR German Aerospace	Sensor (A,G)	Multi-class SVM,	Abnormal activity
	[153]	self	Sensor (ECG)	CNN	PAF detection

Table 4 (continued)

Application	Refs.	Dataset	Type (vision/ non-vision)	ML algorithm	Objective
Video Indexing	[140]	MSR-Action3D	Video	HOG2	Detect joint angle features
	[105]	MSR 3D Online and MSR Daily Activity 3D	LBP	SVM	Online action recognition
	[125]	Moments in Time	Video	SVM	Dynamic events unfolding
	[219]	ORGBD	Video	Boosting and SVM	Orderlets
	[88]	IXMAS	Video	Multi-view activity recognition	Detect pixels-based motion information
	[209]	MSRAction3D, MSRDailyActivity3D	Depth cameras	Filtering and DCSF	Interest point detection
	[126]	CROSS	Video	GMM and HMM	Surveillance subjects
	[202]	MSRAction3D, MSRAction3DExt, UTKinect-Action and MSRDailyActivity3D	Video	HDMM and 3ConvNets	Object segmentation
	[57]	RGB-D	3D depth sensor	DMW	Motion similarity
	[135]	KTH dataset	Video	pLSA and LDA	Ontological Activity
Ambient	[34]	MIT-pedestrian	Images	SVM	Pedestrian activities
	[207]	ARG and APHill	Video	Particle trajectories and SVM,	Object motion tracking
	[213]	FPV activity, coupled ego-motion and eye-motion	Wearable cameras	K-mean and kernel K-mean SVM,	Multi-task clustering
	[46]	ChalearnLAP-2014	Video	GMM and multi-class SVM,	Continuous gesture recognition
	[79]	J-HMBD	Video	SVM and RBM	Detect annotated shapes of human
	[220]	MSR-Action3D dataset , MSR-DailyActivity3D	Video	KNN	Moving pose descriptor
	[69]	Self video recording	Video	TDNN	Pedestrian activities

performance than a gyroscope sensor. Although a combination of the aforementioned sensors gives better performance than individually used sensors but at the cost of high battery consumption.

With the constant use of cameras everywhere nowadays for surveillance, it becomes challenging and time-consuming to manually monitor human activity, especially the 'activity of interest' manually. There is sufficient research [48, 57, 88, 125, 126, 135, 140, 202, 209, 219] present on video comprehension and indexing, which is quite helpful for surveillance and to detect some suspicious activity. Group activity recognition is also very challenging and advantageous. Since it could be helpful in many applications like counting people, understanding crowd behavior, and group tracking. In [91] author has considered headcount in the high-density crowd and utilized the end-to-end scale-invariant method for headcount. Recognizing group activities can aid in understanding abnormal crowd behavior. Although recognizing abnormal activity is quite challenging in itself because of many reasons. For example, an activity may be considered normal in one scenario and abnormal in another. Secondly, discriminative feature extraction of such abnormal activity is also not an easy task. In [154] two convolution layers-based convolution neural networks (CNN) model has been employed for detecting abnormal crowd behaviors. To identify an individual or group-based behavior, events are recognized from videos, and then 'activity of interest' is extracted from

these events. Based on this, annotations are provided, which can be utilized for search indexing [126]. Mainly there are two ways of finding 'action of interest', offline [9, 23, 24, 39, 44, 106, 113, 140, 189, 198, 202, 217] and online [46, 48, 57, 205, 216, 219, 220]. In offline evaluation, the processing is done on static and stored data. The weight changes depending on the complete dataset and thus defining a global cost function. Contrary to that, in the case of the online evaluation, all data is not collected a-prior, data is acquired incessantly and evaluation is done, as the data is sensed.

The focus of most of the researchers is offline recognition, which works on segmented sequences. Although, with offline evaluation, a high level of accuracy can be obtained if robust classification algorithms are used. Mostly, SVM and HMM-based classification algorithms are used in the literature for offline evaluation. On the other hand, online evaluation is very challenging and practical not only for detecting suspicious activities but also for sports and health-based applications. In online evaluation, low latency and high accuracy are desired [105], but there is always a trade-off between them, a lot of research is required to mature online evaluation. Online methods are usually frame-based or sub-sequence-based, with a short duration frame in vision-based. As a matter of fact, human actions always have a temporal correlation. Exploiting such correlation can help recognize human activity accurately, especially in the online evaluation of the activity.

For temporal pattern recognition, different techniques such as HMM, DTW, CRF, Fourier Temporal Pyramid, and actionlet ensembled have been used in literature. Temporal smoothness aids in online evaluation to enforce consistency among sub-sequences. Figure 5 explained the training, testing, and online evaluation of vision and non-vision-based HAR systems. The available labeled data set is divided into training and testing data sets. The ten-fold cross-validation is performed over data to select the appropriate batch for testing as well as training. For vision-based data, frames are extracted, while subsequences are obtained from the given dataset.

After essential application-specific pre-processing and removing noise, the data is segmented to train the activity model. Since ground truth is available for the training dataset, this is used to find the optimal model for the machine learning algorithm as shown in Fig. 5. Once the model is obtained, sequence detection is done from the test dataset. After performing pre-processing, features are extracted from the test dataset. Finally, machine learning is performed on the test dataset by employing the model learned from the training dataset. Once the test dataset attains the required level of accuracy, online evaluation is performed over the trained model as shown in Fig. 5.

Smart home-based HAR includes applications like automatic food preparation and controlling home remotely by detecting human activity. Sensors are attached to kitchen utensils, and home objects [151] to determine some activity.

3 Machine learning approaches

Machine learning-based algorithms used for HAR, depending on the application, could be classified as discriminative and generative models. The generative models work on joint probabilities $p(x, y)$; for example, in HAR, each action collects different poses. Therefore, recognition of action depends on the joint probability of all poses. On the other hand, the discriminative models work on conditional probabilities $p(y|x)$. They work on labeled data and compare it with the action at hand. In general, discriminative models outperform their generative counterparts but require extensive training, which is difficult in some cases [58]. Further details of discriminative models are given in Sect. 3.1 and generative models are provided in Sect. 3.2.

3.1 Discriminative models

Discriminative Methods estimate the posterior probabilities directly without attempting to model the related probability distributions. SVM and KNN are well-known algorithms of the discriminative model. SVM is supervised while KNN is an unsupervised learning algorithm. In SVM each data point is represented in space using already extracted features, with a particular value in the coordinates [26]. Then different features are classified by building a hyper-plane to differentiate them, as shown in Fig. 3. Hence, the more likely features are labeled in each class.

KNN is based on premise that similar things (data points) are often in proximity. Therefore, it calculates the distance between the example in question and the current example from the data. After sorting, assign them a class based on distance similarity, as shown in Fig. 4. The literature indicates that aforementioned algorithms have been extensively used in HAR [9, 14, 34, 46, 66, 67, 187, 198, 213, 218, 220].

In [67], Discrete Cosine Transform (DCT) was used to extract the characteristics from accelerometer data. Subsequently, PCA was applied to reduce the feature dimension. Finally, the Multi-class SVM was selected and applied to classify distinct human activities. The researchers in [34] elaborate that utilizing a locally normalized histogram of gradient orientation features in a dense overlapping grid provides a perfect result for person detection. Moreover, it helps in reducing false-positive rates by more than an order of magnitude. Research work

[34] have trained linear SVM with SVM light by utilizing Gaussian kernel SVM. The improved performance is about 3% in their case.

The researchers in [81] have identified HAR by using data collected via mobile phone sensors. In their research, several classifiers such as Decision Tree, SVM, and KNN were trained. It was found that the Decision tree outperformed the rest models bearing the lowest error rate. Also, SVM was attempted with Linear, Polynomial, and RBF (Gaussian) Kernel, using L1-regulation with various box constrains where the performance rate of linear SVM kernel was found better than the other two. The authors have used a hierarchical approach for analyzing feature descriptors from videos, where classification was performed by applying a multiclass SVM classifier [79, 195]. They further suggested improving the optical flow and human detection algorithms by refining the underlying mid and low-level features. The authors in [117] have demonstrated that in the case of HAR with less number of instances, the SVM classifier performs marginally inferior to the existing results. However, the main focus of their research was computational time. Thus, in their research, it was demonstrated that SVM trained on an existing spatiotemporal feature descriptor is computationally cost-effective in comparison with metric learning. The researchers in [147] have examined the performance of the KNN classification algorithm, particularly for an online activity identification that enables online training and classification using just accelerometer data. Their study further revealed that on mobile platforms with limited resources, the clustered KNN technique performed considerably better than the KNN classifier in terms of accuracy. Research online approaches are used to reduce the number of training instances stored in the KNN search space. Even though KNN is amongst the most examined classifiers for HAR systems or other applications [17, 112, 148, 164], its storage and computation needs grow as the number of data and training examples increase, thus resulting in additional prototype problems. Consequently, the research in HAR also introduces basic, computationally intensive, energy-efficient, and viable economic strategies for keeping a maximum number of training examples stored by KNN at runtime to endure the issues related to time and memory restrictions in the online mode as well [50].

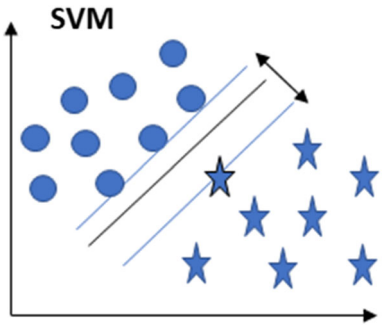


Fig. 3 Support Vector Machine graphical representation

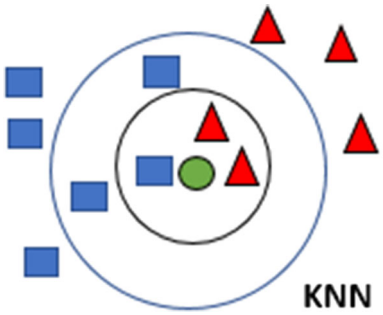


Fig. 4 Graphical representation of KNN

3.2 Generative models

In machine learning, generative modeling is unsupervised learning which detects and learns regularities or patterns automatically from the input data distribution. Henceforth, the model may be used to produce or output new instances that might have been taken from the original dataset. By modeling the underlying distribution of classes from the given feature space, generative techniques increase generalization ability. Although the parameters are not optimized, generative models are flexible because they learn the structure and relationships between classes by utilizing previous information, such as Markov assumptions, prior distributions, and probabilistic reasoning. Generative models are the preferred approach in case there is any ambiguity or uncertainty in the data; Nevertheless, these models require a vast quantity of data for providing accurate estimates [47]. In these models, initially joint probabilities are learned. Then it estimates the conditional probability using Bayes Theorem [22]. The two most popular algorithms of the generative model are the HMM and GMM.

3.2.1 Hidden Markov model

HMM are generative models which follow the Markov Chain process or rule. The mechanism refers to a series of potential occurrences in which the likelihood of each event is determined by the conditions of previously occurring events. A Markov process is a random process that follows a property that the probability of the next state depends on the current state and not on all previous states, $P(Future|Present) = P(Future|Present, Past)$. It could be mathematically formulated as in Eq. 1.

$$P(x_t + 1|x_t) = P(x_t + 1|x_{1\rightarrow t}) \quad (1)$$

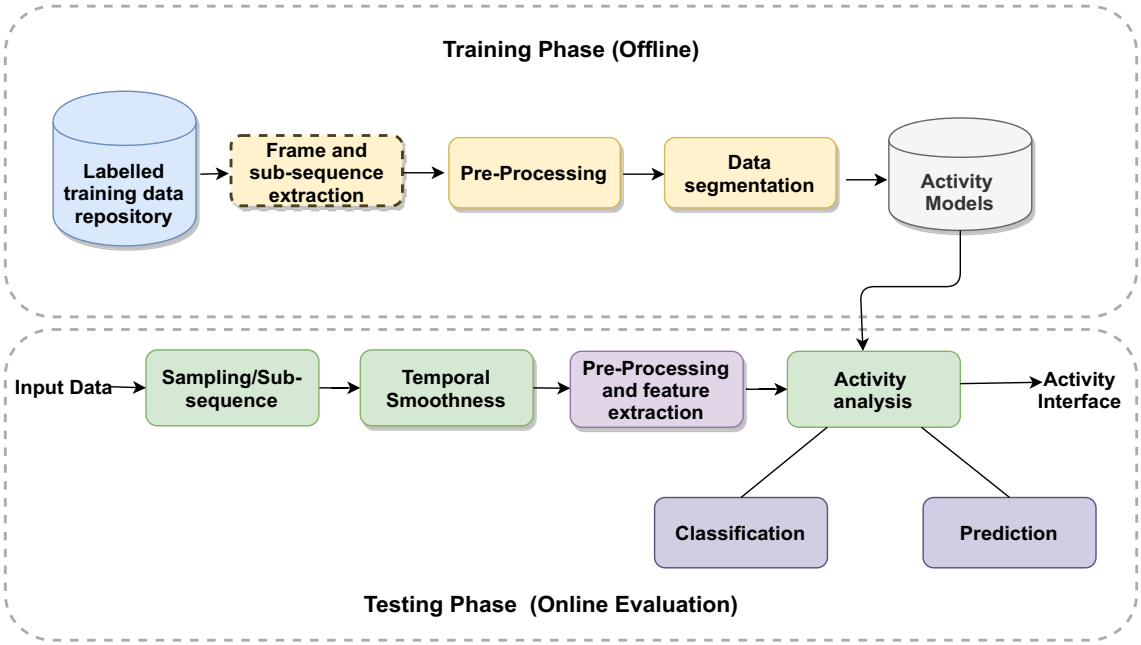


Fig. 5 General framework of Online Evaluation for Vision/Non-Vision-based HAR

A discrete variant of the Markov process also known as discrete-time Markov chain (DTMC) has a discrete set of times. HMM is a particular case of DTMC consisting of hidden variables, also called states, and a sequence of emitted observations. For any given measurement, x_k, the hidden states $Z_k^{(N)}$ are not directly measurable; however, their emitted observations can y_k could be observed as shown in Fig. 7. Any HMM is represented as a tuple of $\lambda = (\pi, \Phi, E)$, where π is initial state probabilities as shown in Eq. 2, Φ are state transition probabilities Eq. 3, and E are e Emission Probability Matrix Eq. 4, all symbols are described in table 5.

$$\pi = P(x_1 == i) \tag{2}$$

$$\Phi_{i,j} = P(x_{t+1} = i | x_t = j) \tag{3}$$

$$E_{i,j} = P(y_t = j | x_t = i) \tag{4}$$

There are three basic problems in HMM.

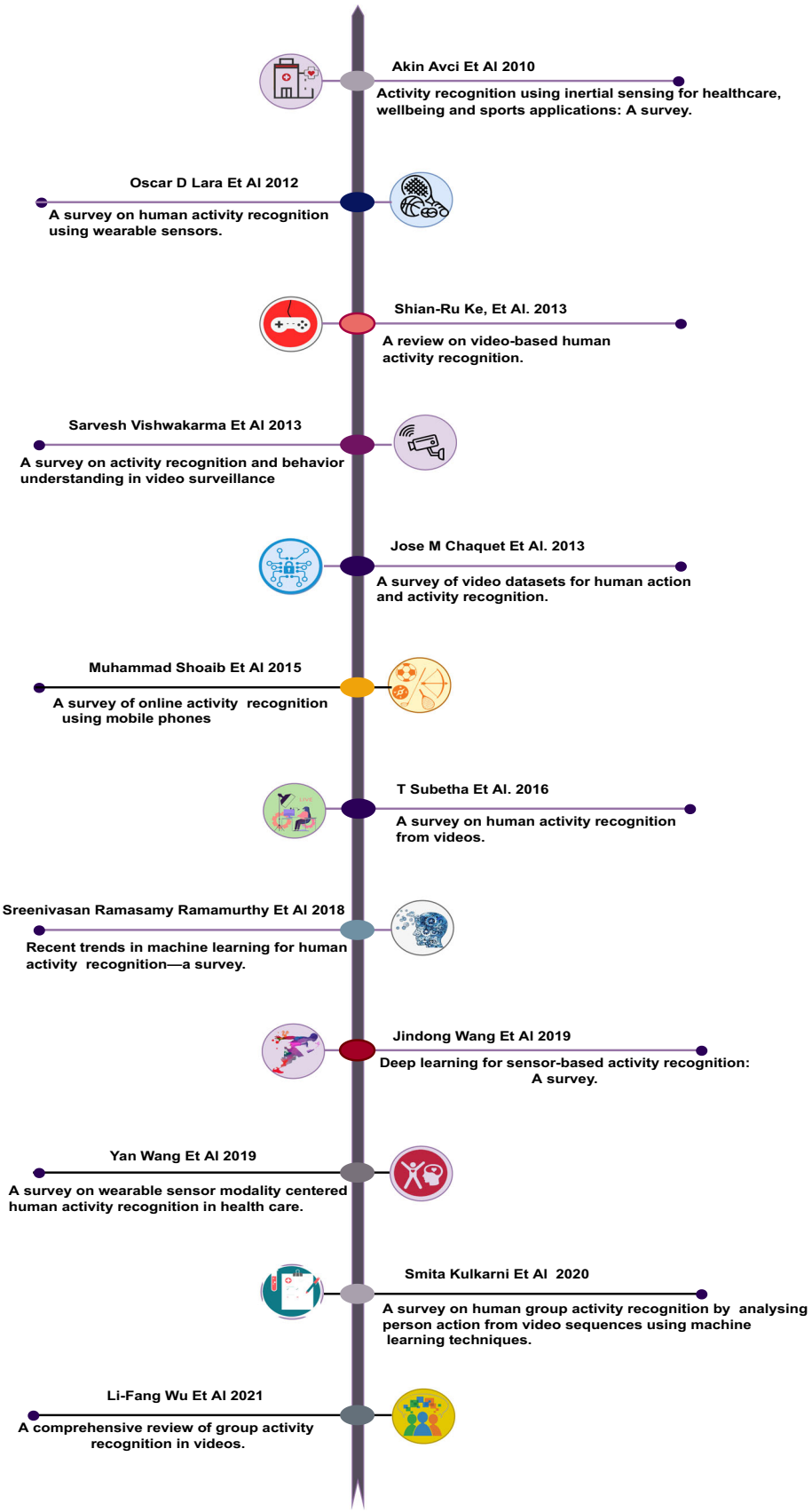
- *Likelihood:* Given the HMM $\lambda = (\Phi, E)$ and observed sequence Y, calculating the likelihood $P(Y|\lambda)$.
- *Decoding:* Having observations Y and $\lambda = (\Phi, E)$, find the hidden state sequence Z.
- *Learning:* Having observation sequence Y and states Z determined parameters Φ and E.

HMM are important in HAR since it can encode a sequence of events which is the fundamental concept in activity recognition. There is a large volume of published research describing the role of HMM in HAR [64, 75, 111] and [226]. The researchers in [162] proposed a user adaptation technique for improving the HAR system using HMM. Their system consists of a feature extractor to extract the significant properties from inertial signals, and a training module based on six HMMs, i.e., one for each human activity. Finally, a segmentation module that uses those models to segment activity sequences. Several researchers have also proposed the combination of HMM with discriminative model SVM for HAR [38, 53, 214]. A multi-layer HMM is proposed in [43] to recognize different levels of abstract Group Activities. Moreover, The research conducted in [154] demonstrates the use of the Hierarchical Hidden Markov Model (HHMM) for HAR. HHMM is an extension of HMM that works with hierarchical and complex data dependencies. The variants of HMM have also received a lot of attention in the realm of HAR, some examples are [205, 232, 126, 202] and [205]. Mostly HMM-based HAR lies in the area of decoding and learning problems of HMM. For example, [83] used Baum-Welch (BW) to learn the parameters of HMM. The Markovian property implicit in the traditional HMM presupposes that the present state is only a function of the former state. However, in practice, this assumption frequently fails to satisfy expectations. Furthermore, the generative property of HMM, as well as the assumption of independence between observations and states, limit its performance [100].

3.2.2 Gaussian Mixture Model (GMM)

As the name GMM implies, it is a mixture of several Gaussian distributions [156]. A Gaussian distribution is a

Fig. 6 Timeline for HAR survey



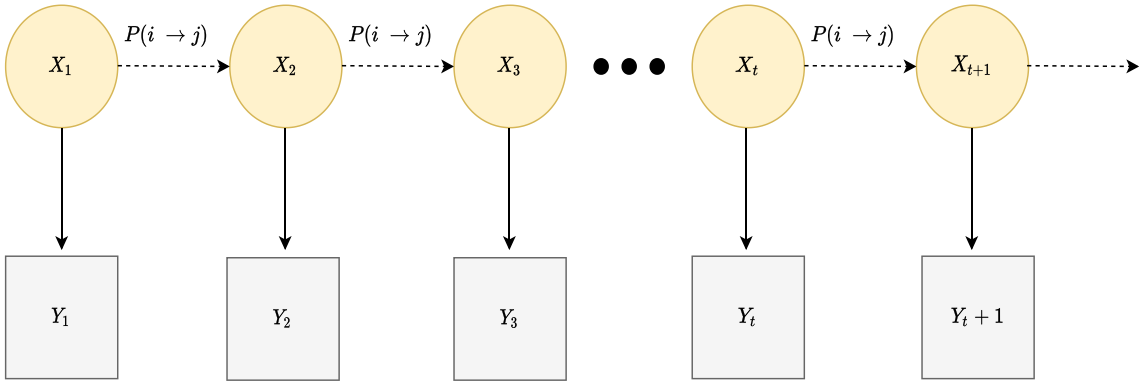


Fig. 7 Temporal evolution of a hidden Markov model

Table 5 Terminologies used in HMM

Symbol	Definition
Y	A set of observed states.
Z	A set of Hidden states which cannot be directly observed.
π	Initial States of HMM
Φ	States transition probabilities
E	Emission probability matrix /observation likelihoods.

symmetrical bell-shaped continuous probability distribution. Each Gaussian is identified by $k \in 1, ...K$ as presented below in Eq. (5).

$$\sum_{k=1}^{K} \pi_K = 1 \tag{5}$$

Where a specific weight π_k represents the probability of the k_{th} component. Mathematically, a univariate Gaussian distribution is expressed as in Eq. (6). Whereas, μ and σ are scalars representing the mean and standard deviation of the distribution. Correspondingly, Eq. (7) indicates the multivariate Gaussian distribution.

$$p(x \mid \mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{6}$$

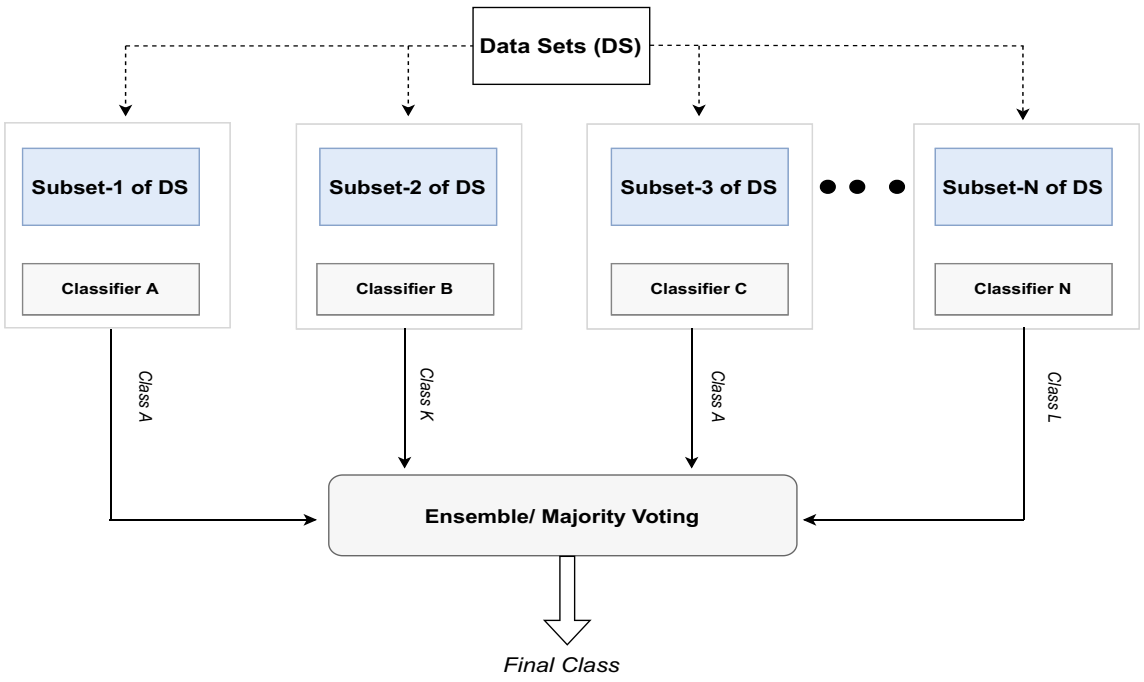


Fig. 8 Ensemble algorithms bagging

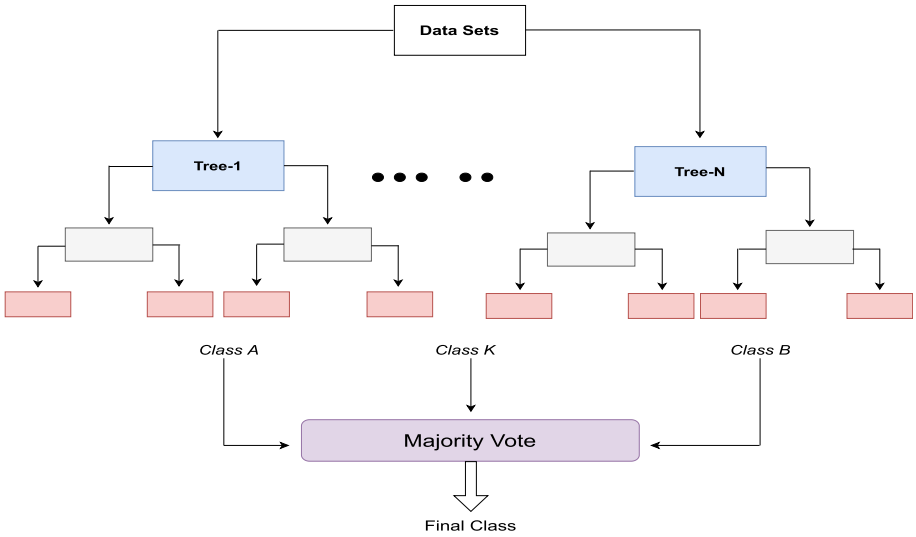


Fig. 9 Ensemble algorithms random forest

$$(p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{7}$$

Where Σ is a covariance matrix of X. The likelihood $p(x|\theta)$ is obtained through the marginalization of latent variable z. It consists on summation of the latent variables from the joint distribution $p(x, z)$ as shown in Eq. (8). Where θ is a vector of Gaussian parameters.

$$p(x \mid \boldsymbol{\theta}) = \sum_{z} p(x \mid \boldsymbol{\theta}, z)p(z \mid \boldsymbol{\theta}) \tag{8}$$

This marginalisation may now be linked to the GMM by considering that $p(x|\theta, z_k)$ is a Gaussian distribution, i.e, $\mathcal{N}(x|\mu_k, \sigma_k)$ with z comprising of K components as shown in Eq. 9. A specific weight π_k represents the probability of the k_{th} component so that $p(z_k = 1|\theta)$.

$$\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

In theory, the GMM is capable of approximating any probability density function with reasonable precision. GMM has proven to be an effective algorithm in time series analysis and modeling. GMM usually works on frame-based classification, while HMM is mostly focused on sequence-based classification. The research conducted in [176] is based on hierarchical recognition which consists of two phases, initially, activities are classified into two broad clusters, static and dynamic activity. Subsequently, within the identified class, activity recognition is carried out. It is evident from the literature that several researchers have proposed joint models based on HMM and GMM for HAR, for example, [29, 150] and [126]. In [126] GMM is proposed with expectation-maximization to find the point of interest (POI) in human activity while the evolution of activities is learned by employing HMM. The researchers in [149] developed a probabilistic graphical model-based human daily activity detection system by using an RGB-D camera. Using only skeleton characteristics provided by an RGB-D camera, they implemented a GMM-based HMM for human activity detection. As a collection of multinomial Gaussian distributions, Gaussian Mixtures can cluster data into multiple categories. Human actions are a collection of how various human body stances transmit consecutively at various periods. As a result, each body position can be modeled as a set of multinomial distributions, with HMM modeling the intra-slice dependencies between time periods.

The Bayesian network is based on a graphical model which establishes probabilistic relationships among variables of interest [68]. These graphical models work very well for HAR data analysis, especially when combined with statistical techniques [42] and [10]. The main reason for its good performance is its ability to establish dependencies among all variables. Therefore it can immediately estimate missing data entries as in [200]. In their work, State-based learning architectures were presented, namely HMMs and CHMMs. The objective was to model human behavior and its interaction with others. HMM was particularly used to model and classify human behavior while CHMMs (Coupled- Hidden Markov Model) purpose was to model interaction and coupled generative process.

3.3 Ensemble learning

Ensemble learning is a machine learning paradigm that combines multiple weak learners to improve their

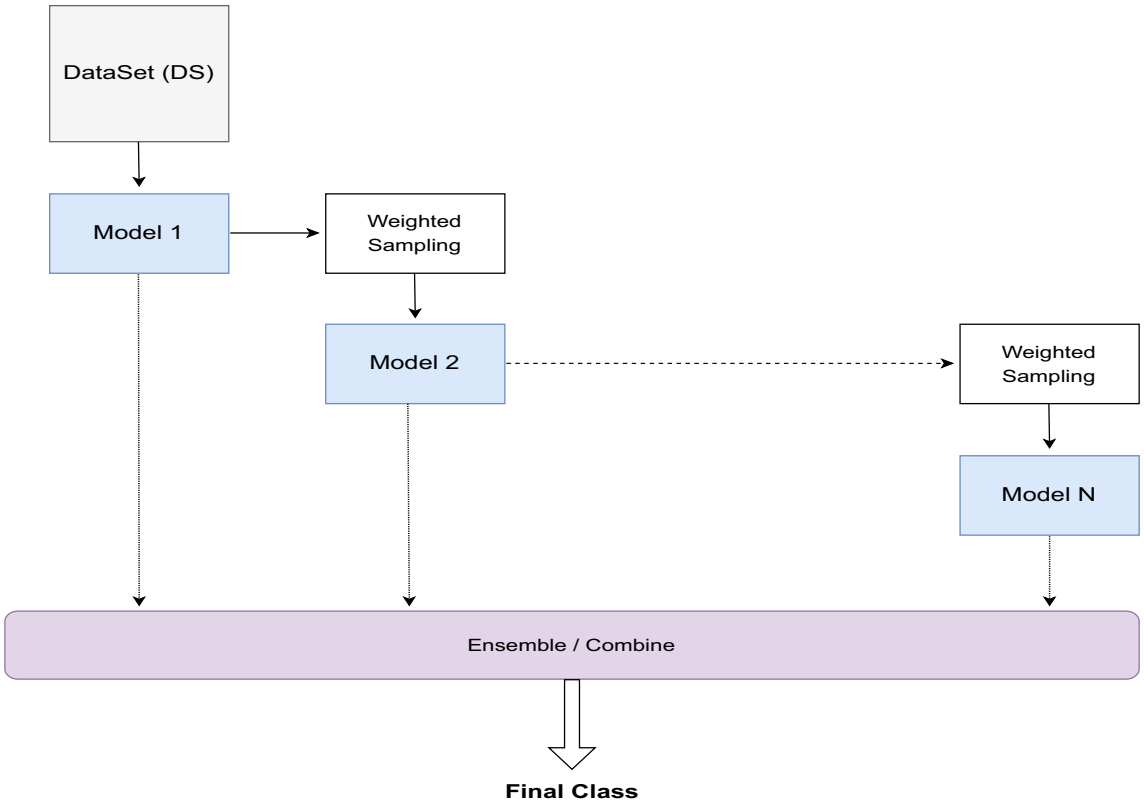
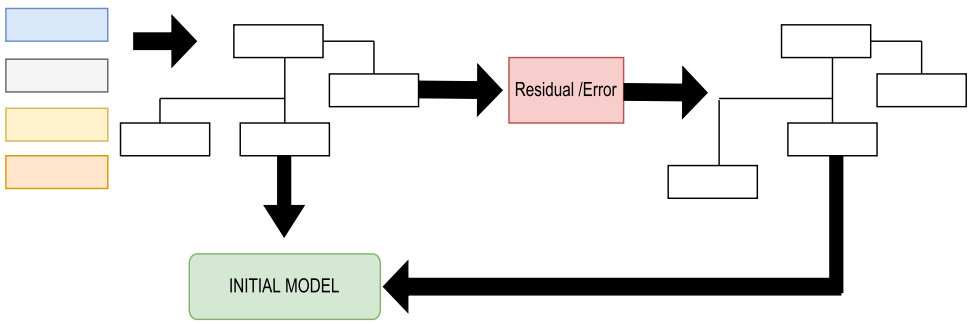


Fig. 10 Ensemble learning boosting

Fig. 11 Gradient boosting algorithm



performance. During the training phase, the model might tend to over or underfit and suffer a problem of high bias or variance. The ensemble learning methods combine multiple weak learners to achieve better performance. There are three major techniques used in ensemble learning.

- **Bagging**: In bagging similar weak learners are trained in parallel. Each one of them either classifies or predicts independently from other models. The result of all weak learners is combined using a majority vote or averaging process.
- **Boosting**: In boosting weak learners are trained sequentially while learning from the loss of the previous stage in each case.
- **Stacking**: Stacking uses different weak learners and trains them in parallel. The models are combined to train a metamodel which is used to predict the output based on the outputs of multiple predictors.

3.3.1 Bagging

Bagging stands for bootstrap aggregation. In this technique *N* homogenous weak learners are trained in parallel as shown in Fig. 8. Each classifier is tested on a subset of the dataset and their outputs are combined by using majority voting or averaging. The dataset is created through random sampling with replacement over the training dataset. For

any given dataset of size N, the bagging can be summarized as in algorithm 1.

Algorithm 1 Bagging Algorithm

Step-1: Repeat for K times

(a) Create data set of size N from the original dataset using sampling by replacement policy.
(b) Feed each dataset into the classifier and store the result of each stage

Step-2: Perform either majority voting on outcomes of K stages or averaging to get the final result.

Another popular variation of bagging is known as random forest (RF), in RF besides sampling the dataset, the features are also randomly sampled. The homogenous classifier in the case of RF is a forest tree. Each forest tree has a subset of the dataset as well as a subset of features as shown in Fig. 9.

It has been shown in [136, 137] that the RF outscored other decision tree techniques and machine learning classifiers in recognizing human activities utilizing the characteristics such as acceleration and jerk. The RF offers improved activity detection ability because it generates numerous decision trees and combines them to produce a more accurate and stable outcome [224, 225]. The research performed in [32] proposed an ensemble architecture, i.e., WiAReS. This integrates a multilayer perception (MLP), a random forest (RF), and SVM to enhance the recognition performance of human activities using the features extracted from convolutional neural networks.

3.3.2 Boosting

Unlike bagging where N classifier operates in parallel the boosting is a sequential algorithm. Boosting starts with a weak classifier employing sampling of the input dataset. Once the classifier is trained it is tested using the dataset. The points correctly and incorrectly predicted are assigned lower and higher weights respectively. The weighted sample points are now assigned to the next version of the model. The process is repeated for N stages as shown in Fig. 10. To summarize boosting improves each successive model by correcting the errors of the previous model. There are two major types of boosting i.e. Adaboost and gradient boosting. Adaboost or adaptive boosting adjust the weights based on the performance of the current iteration. This means that weights are adaptively recomputed in each iteration, as shown in algorithm 2.

Algorithm 2 AdaBoosting Algorithm

Step-1: Initialize weights with uniform random numbers. **Step-2: For Each** base learner do:

(a) Train base learner with weights.
(b) Test base learner with the given data.
(c) Assign errors to the learner weight.
(d) Set example weights based on ensemble predictions.

End For

Gradient boosting is a combination of gradient descent and boosting. It works in the same way as AdaBoost except the weights are updated on a residual error from the previous estimator as shown in Fig. 11. The step-by-step procedure of gradient boosting is given in algorithm 3. Due to better results of gradient boosting algorithm it is widely used in human activity literature, [60, 70], and [167] are some examples of gradient boosting-based human activity recognition.

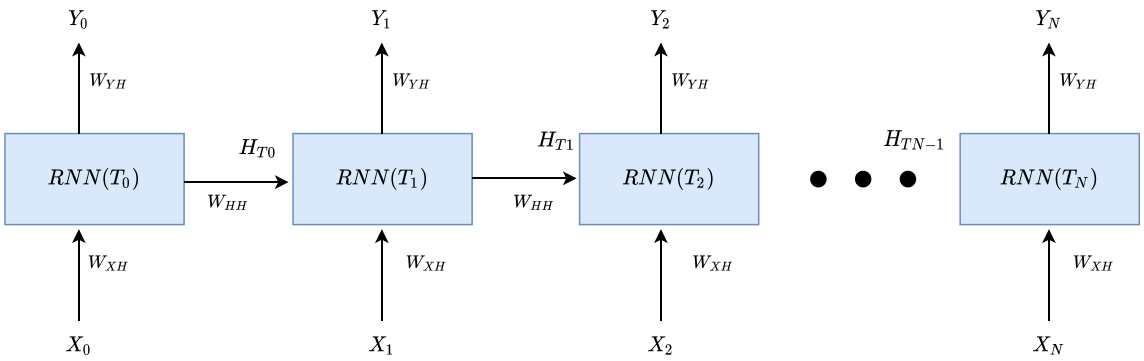


Fig. 12 Illustration of Recurrent neural networks (RNN)

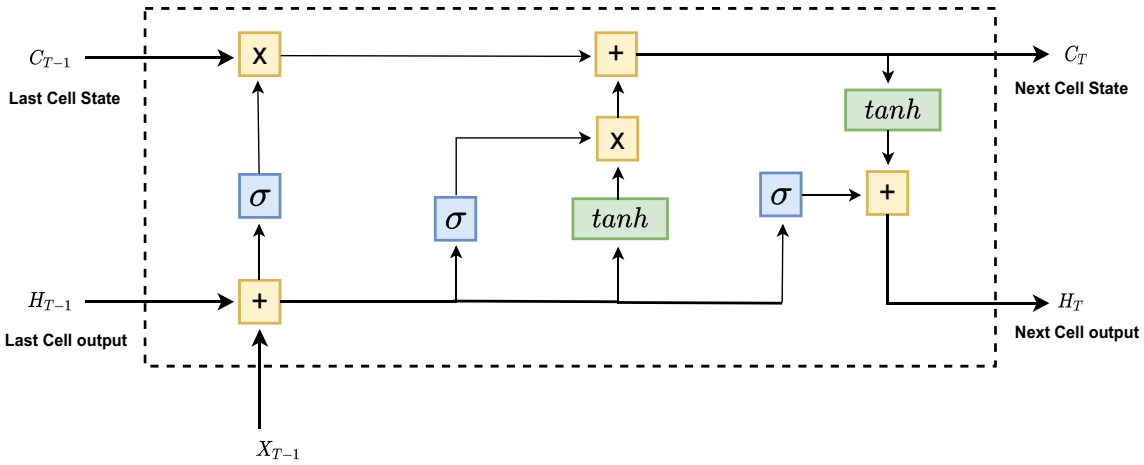


Fig. 13 Illustration of Long short term memory (LSTM)

Algorithm 3 Gradient boosting Algorithm

Step-1: Calculate error residuals e. The error can be calculated as difference of actual target value and predicted target value $e = y - \hat{y}$.

Step-2: Compute new model using residual error as target variable, known as $\hat{e}$.

Step-3: Combine previous predictions with the predicted residuals $\bar{y} = \hat{y} - \hat{e}$.

Step-4: Find another model to find the remaining error left $e_{left} = y - \bar{y}$ and repeat steps from 1 to 4 till sum of residual becomes constant.

3.3.3 Stacking

Stacking is another type of ensemble learning algorithm. It uses heterogeneous weak learners as compared to homogeneous learners in boosting or bagging. Stacking addresses the question of how to combine the output of multiple models trained over training data? Stacking uses two levels of learners known as the base model (BM) and meta-model (MM). The BM consists of weak learners which are trained on the part of training data. Once BM is trained the prediction and label from the training dataset are fed into MM. The stacking method requires careful division of the training dataset. For this purpose, the training dataset is split into two further parts using K fold validation where out-of-fold predictions are fed to the MM. Stacking has been used for HAR by employing a combination of different machine learning algorithms, a few examples are [54, 186], and [55].

4 Deep learning approaches

Traditional machine learning approaches have shown immense progress in HAR by implementing diverse machine learning algorithms, as discussed in the above section. Deep learning algorithms enjoy success since they could automatically extract features using CNN. Besides Recurrent Neural Networks (RNN) can model sequences in very efficient ways. Sequences are the primary element of activity modeling and recognition. Different variants of RNN such as Long Short-Term Memory (LSTM) and transformers provide improved performance over traditional RNN algorithms. Furthermore, a deep learning-based paradigm called transfer learning allows pre-trained models to use in related HAR tasks. This reduces training time as well as improves performance with limited training data. HAR generates a lot of data since it continuously senses the environment. Autoencoders provide ways to reduce dimensions of data by learning efficient encoding representation of the data.

Due to the popularity of the internet of things (IoT) and edge computing distributed machine learning has gained popularity. HAR is also studied in a distributed setting, deep learning paradigm known as federated learning provides a way to work in distributed settings. Another popular area known as Reinforcement Learning (RL) works in an environment with limited training data. In RL an agent learns with evaluative feedback and employs a trial and error paradigm. RL models such as the actor-critic model, DQN, and monte Carlo-based models are employed for HAR. Deep learning has provided state-of-the-art performance in the domain of HAR in comparison to classical machine learning [6, 109, 138, 170].

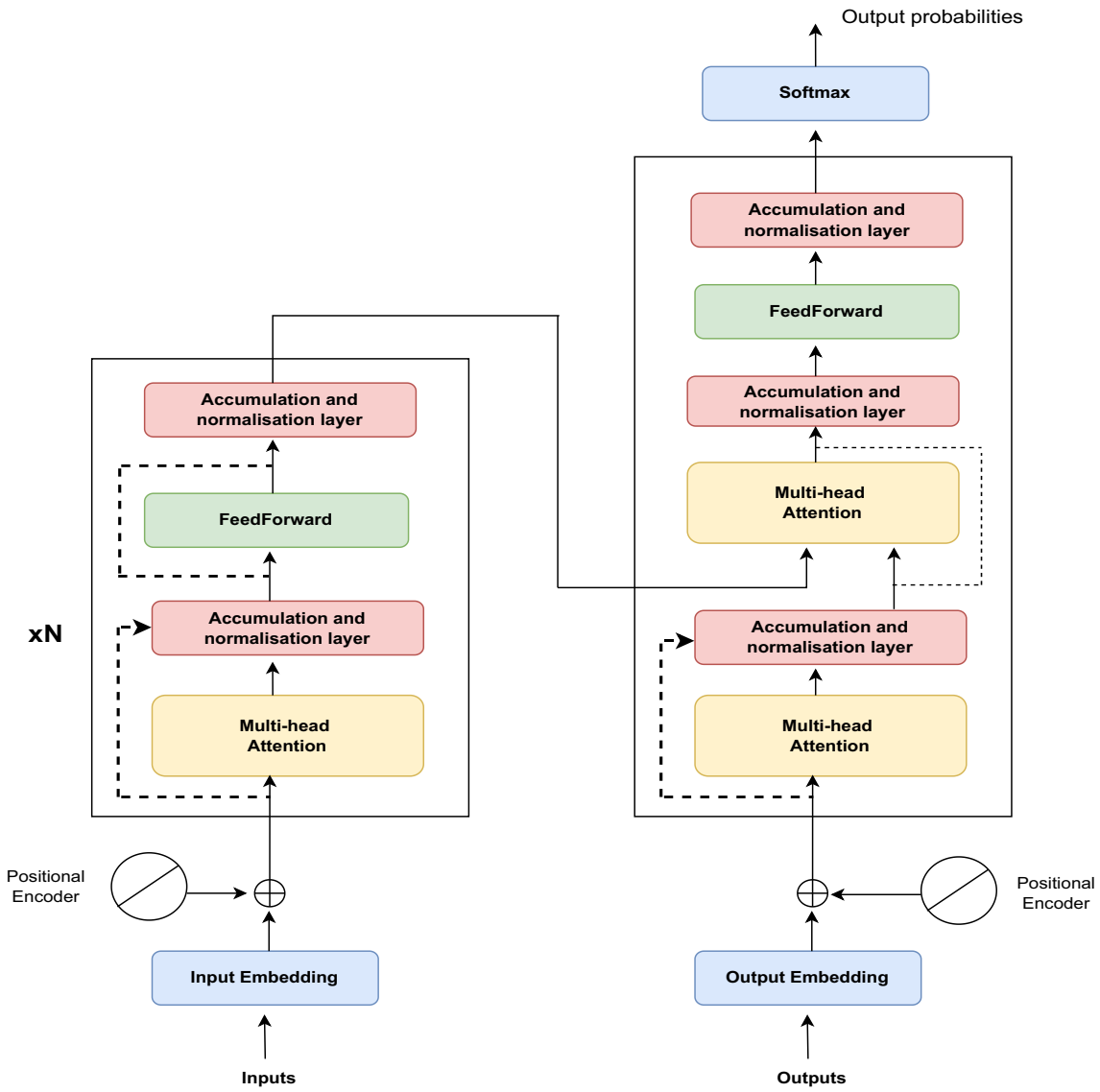


Fig. 14 Illustration of transformers

4.1 RNN and LSTM

Sequences are an integral part of several applications and standard neural networks cannot handle sequence data. This is because standard neural networks do not have memory and only make decisions based on the current input. RNN is a special type of neural network that can handle temporal sequences since it can maintain states [72]. RNN plays a vital role in HAR since each action depends on previous actions and sequence-based operations are vital for HAR pipelines [142]. Figure 12 shows cascaded RNN cells spanned over time intervals T_0 to T_N. Each consecutive cell maintains cell state H_{Ti} at i_{th} time interval. The input, output, and intermediate states are weighted by W_{XH}, W_{YH}, and W_{HH}. The next state and output are calculated using the weights as shown in Eqs. 10 and 11 respectively.

$$H_T = TanH[W_{HH}H_{T-1} + W_{XH}X_T] \tag{10}$$

$$Y_T = W_{YH}H_T \tag{11}$$

During the backpropagation phase, each RNN not only reduces loss function through its cells but also across time known as backpropagation through time (BPTT). BPTT suffers the problem of vanishing gradient, the problem gets more elevated for long-term dependencies. To solve the problems in RNNs LSTM is proposed. LSTM is a modified RNN with four stages having forgotten, store, update, and output mechanism in it as shown in Fig. 13. The F_T forget part decides which information should be removed at a certain point in time as shown in Eq. 12.

$$F_T = \sigma(W_F.[H_{T-1}, X_T] + B_F) \tag{12}$$

The second part called store has two parts, the first part shown in Eq. 13 has sigmoid while another part shown in Eq. 14 has *tanh*. The sigmoid part decides which value to let through while the *tanh* gives weightage to the value based on its importance. The Eq. 16 shows the next cell output, where O_T is given in Eq. 15.

$$\bar{S_T} = \sigma(W_{\bar{S}}.[H_{T-1}, X_T] + B_{\bar{S}}) \tag{13}$$

$$S_T = TanH[W_S.[H_{T-1}, X_T] + B_S] \tag{14}$$

$$O_T = \sigma(W_O.[H_{T-1}, X_T] + B_O) \tag{15}$$

$$H_T = O_T * TanH(S_T) \tag{16}$$

Another popular variant of LSTM is gated recurrent units (GRU). It can capture the dependencies of data in a much better way, besides they are computationally efficient as compared to LSTM [30]. LSTM and GRU share some similarities, the major difference is between how both control memory content sent to the output gate. LSTM has been recently applied widely for human activity recognition [142, 229, 28] and [208]. A research work employed structural RNN for group activity recognition in videos. They used spatiotemporal attention, as well as a semantic graph for group activity recognition, [155]. Another work used deep RNN (DRNN) for HAR, they showed that DRNN has much better performance than bidirectional and cascaded RNN architectures [129]. It is also proved in research that combining LSTM with ensembling learning can improve the results as compared to a single LSTM network [59].

LSTM also suffers problems due to the sequential nature of the operation, each next LSTM unit requires all previous LSTM units to be activated. This results in slow speed and requires efficient convolutional network layers to extract features before LSTM can provide reasonable performance. Recently transformers are purposed to solve the sequential nature of LSTM [188].

4.2 Transformers

Transformers are an extension of LSTM which could take data in parallel as compared to sequential data input in LSTM. However, feeding data in parallel is challenging as it requires efficient position encoding to keep track of the sequence of data. Moreover, embedding should be performed in a very efficient way. Transformers employ self-attention mechanisms to weigh a significant part of data more hence they do not need to process data in order like RNN or LSTM. The transformer uses an encoder and decoder to achieve this task as shown in Fig. 14. The encoder and decoder consist of multi-head attention blocks. The encoder consists of multiple stages of multi-head attention blocks each finding the relevant part of information. The decoder uses one multi-head attention block to encode output, while the second to learn encodings of the inputs. Over a period of training, the decoder adapts to input embeddings and learns to decoder sequences in the correct way. The Decoder has a feed-forward network on the end of the pipeline to perform given machine learning tasks such as classification.

The transformers have resulted in high accuracies in pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) [40]. The original work was initially proposed for natural language processing (NLP) problems however, recently variants of transformers for image processing have shown high accuracy [41]. Subsequently, transformers have been recently applied in human activity recognition resulting in much better and more accurate results as compared to traditional RNN [119, 56]. In HAR transformers are used for capturing spatiotemporal relationships between data as well as for data augmentation. Some research worked used attention models to extract feature-based spatiotemporal context [116, 130, 114]. Transformers are also widely used for data augmentation to improve the accuracy of the trained classifier [4]. Another work explored transformers for data augmentation using the self-attention mechanism to track long-term dependencies [223].

4.3 Deep belief network

A deep belief network (DBN) is a type of DNN, it consists of multiple hidden layers, and these layers are only visible to the next layer. To form learning more manageable, the property is restricted, i.e., there's no affiliation between hidden units. DBNs will be divided into two significant components. The primary one consists of multiple layers of Restricted Boltzmann Machine (RBMs) to pre-train the network, whereas the second could be a feed-forward back-propagation network that may refine the RBM stack results [71]. The authors in [7] have presented the DBN-based model; in their work, the model was trained by employing greedy layer-wise training of RBM. Hence, human activity recognition accuracy was improved in comparison to expensive handcrafted features. Then [16] have practiced RBM-based pipeline for activity recognition and have shown their approach outperforms other modeling alternatives. Deep learning-based algorithms mostly evolve in HAR using RGB video sequences based on the belief that every human action is composed of many small actions. A temporal structure is usually considered to enhance the classification of human actions/activities. Therefore, DBN approaches aim to develop a DL structure to the problem; it

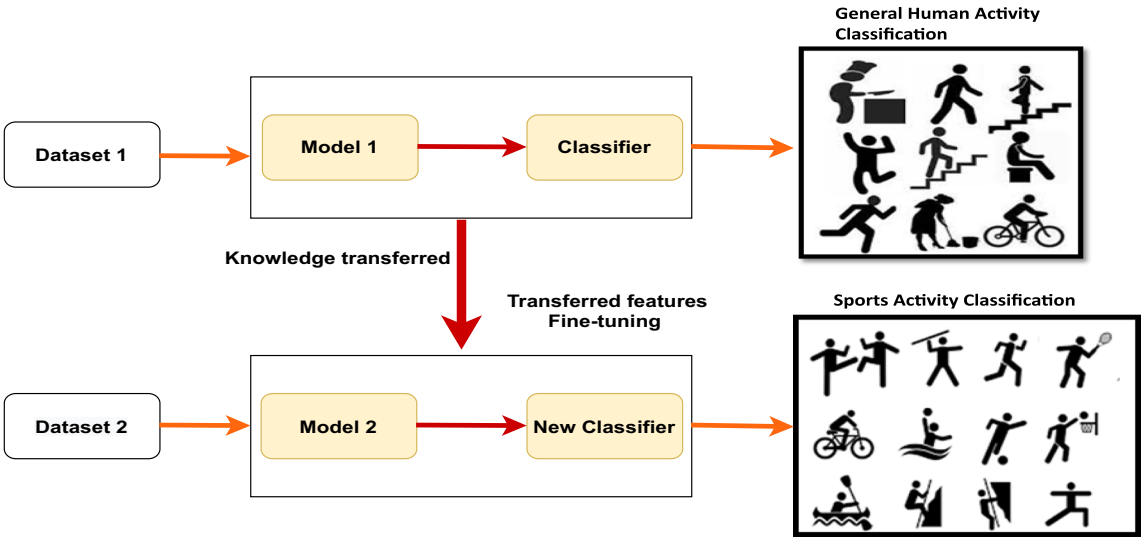


Fig. 15 Human activity classification by transfer learning

can be perceived as a DL structural architecture. To manipulate the DL network, the activation functions yield the hidden variables at every hidden layer [133]. DBN approach outperforms the methods constructed upon engineered features since it uses the skeleton coordinates extracted from depth images. In [211], it is observed that the DBN approach produces better recognition rates compared to those of other state-of-art methods. In [71] Hinton introduced the idea of deep belief networks, which were inspired by the backpropagation network. Although the multilayer perceptron and DBN are incredibly the same in terms of network structure, their coaching method is entirely different. In fact, the distinction in coaching technique is a vital issue that permits DBN to vanquish this shallow counterpart.

4.4 Autoencoders

Autoencoders (AE) consist of two units encoding units used to transform input data into features. And the decoding unit regenerates input based on learned features. AE is trained by minimizing the loss between actual data and regenerated input. AE is quite close to RBM however, they used deterministic units as compared to stochastic units. If the sparse constraint is introduced in the autoencoder, it could even improve HAR results. However, it is a very robust tool for feature extraction. The only drawback of AE is it depends too much on its layers and activation function, which may sometimes be hard to find, the most suitable one.

A research work proposed the stacked autoencoder-based model for optimal recognition accuracy along with reduced recognition time [5]. In [138], autoencoders performed very well with the Neural network for data compression using machine learning. The work further concluded that the autoencoder learns compressed distributed representation of input data for backpropagation. Another work used stacked encoders for four types of data including accelerometer, gyroscope, magnetometer, and barometer [5]. A similar work used data from four types of sensors built in a smartphone, including accelerometer, gyroscope, and magnetometer [201]. Besides mentioned work, there is rich literature on the use of autoencoders for dimensionality reduction as well as efficient feature encoding using AE. Especially in the case of HAR where data is sensed at very high rates and has high sparsity in its structures.

4.5 Convolutional neural network

Advancements in computer vision with deep learning have been established and enhanced through time. CNN is one of the most popular deep learning architectures and it has improved state-of-the-art dramatically in processing images, video, audio, and speech [103]. It is a neural network with an input and an output layer and many intermediate hidden layers. Thus, CNN is similar to regular ANN and is comprised of neurons that self-optimize through learning [143]. The primary difference between CNN to other neural networks is that instead of only using the typical activation function, convolution and pooling functions are also computed on hidden layers as indicated in Fig. 17. By performing a convolution operation on data, the convolutional layer detects distinct features from input. The very first convolutional layer detects low-level features, while the subsequent convolutional layers detect higher-level features. The activation functions used by the convolutional layers then contribute nonlinearities to the model

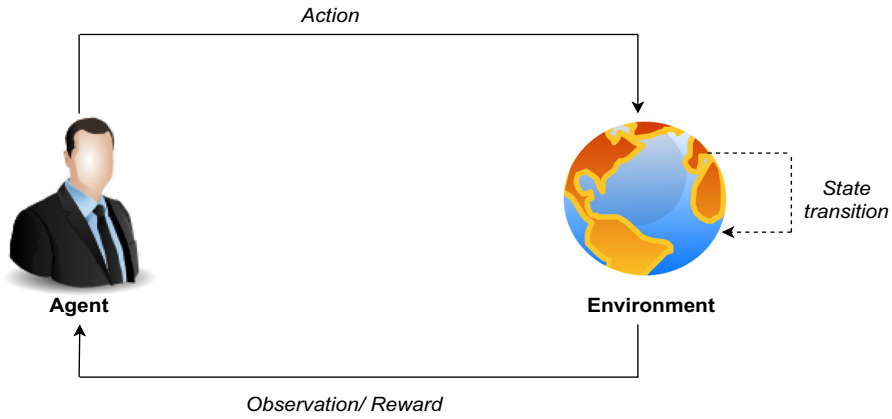


Fig. 16 Reinforcement learning

[132]. The convolutional layers then introduce nonlinearities to the model by using activation functions. In general, the 1D CNN in HAR is used for signal data [82, 104, 128] and 2D CNN takes the input data in the form of images [36, 191]. Alternatively, the 3D CNN fetch the input as 3D volume or a sequence of 2D frames (e.g. slices in a CT scan) [3, 80, 86, 92, 196, 233].

A study in [8] demonstrated a detailed overview of the evolution of DCNN architectures and how they met the object recognition and detection challenges. They used DCNN for object/activity detection and recognition from images. Generally, **R-CNNs:** is used to locate and classify the main object by localization. CNN architectures are capable of learning powerful features from weakly-labeled data that far surpass feature-based methods in performance and that these benefits are surprisingly robust to details of the connectivity of the architectures in time [184]. Authors in [158] have given energy and memory-efficient solution to recognize human activity by employing adaptive CNN architecture.

4.6 Hybrid DL approaches

An increasing number of studies also reveal that researchers have proposed and developed several various Deep learning hybrid approaches for HAR. In [115], authors have proposed a novel LSTM-CNN model combining the merits of LSTM and CNN for collaborative learning. Furthermore, their work also demonstrates that the proposed LSTM-CNN model outperforms standalone LSTM, CNN, and Deep Belief Network. In their research, the combination of RG+RP and LSTM-CNN provides a privacy-preserving collaborative learning framework that is both accurate and privacy-preserving. Similar approach is proposed in [208] using LSTM-CNN combination. In the proposed architecture for HAR, sensor-based HAR is focused where two-layer LSTM is followed by convolutional layers. In addition, a global average pooling layer (GAP) is applied to replace the fully connected layer after convolution for reducing model parameters. Many attempts have been made where initial layers are based on CNN and upper layers are based on diametrically different models [1, 173, 222]. For instance, the researchers in [197] a 1D CNN-LSTM network to learn local features and model the time dependence between features. In the first step, they used CNN for extracting features from the data collected by sensors. Subsequently, the long short-term memory (LTSM) network is developed over the learned features to capture long-term dependencies between two actions to further improve the HAR identification rate. The researchers have also improved the accuracy of HAR detection by proposing CNN-LSTM-ELM-based classifier [178].

5 Transfer learning

The transfer learning approach is used in machine learning (ML) to learn the model from one problem and use the same model for other related ML techniques [190]. Recently transfer learning is employed in deep learning, where a pre-trained model is reused as the starting point for a model that is under consideration for another task [183]. Thus, previously learned knowledge is utilized to model a new but relevant background. The learning of a new task relies on the previously known tasks, as shown in Fig. 15. For example, in the initial model, task-related in daily lives have been classified. This learned information is transferred to another task for sports activity recognition where the model learned from daily task classification is reused. Thus, the learning process becomes faster, more accurate, and requires less amount of data. In this way, Transfer learning saves huge computation and time resources required to develop neural network models. Transfer learning can be classified under three sub-settings, inductive transfer learning, transductive transfer learning, and unsupervised transfer learning, based on different

situations between the source and target domains and tasks [146].

Mostly HAR yields better performance through supervised machine-learning approaches. Although the cost of gathering and labeling data is high [63] due to the diverse, interleaved, and dynamic nature of human behavior. Therefore, transfer learning (TL) can be applied whenever there is a lack of sufficient labeled training data. In HAR, TL can use the existing knowledge to identify activities performed by different types of users, which might be using different sensor technology and in diverse environmental conditions. In some cases, when the source domain and target domain are not related to each other, instead of applying brute-force transfer, it is highly important to explore whether transfer learning is feasible or not to avoid the negative transfer. Therefore, there are two important parts, "what to transfer" and" how to transfer." The part of knowledge to be transferred across domains or tasks depends on "What to transfer." Once it is clear which knowledge can be transferred, then the learning algorithms need to be developed to transfer the knowledge, which corresponds to the "how to transfer" issue. Then the next point is what to transfer across these categories; the following approaches are adopted. Primarily, Feature-representation transfer: The advantage of using this approach is to reduce error rates by identifying good feature representations that can be utilized from the source to target domains. Depending upon the availability of labeled data, supervised or unsupervised methods may be applied for feature-representation-based transfers. Secondly, Instance transfer: Mostly, the source domain data is inadequate and not suitable to reuse directly. Therefore, instead of selecting the whole information, only a few instances are selected for transfer. Thirdly, Parameter transfer: In this approach, there is an assumption that the models for related tasks share some parameters or prior distribution of hyper-parameters. Lastly, Relational-knowledge transfer: Unlike the preceding three approaches, it prefers the data, which is not independent and identically distributed.

TL has been extensively used in video-based activity recognition, and it is one of the first sensor modalities where TL was initially applied [108]. The labeling of video sequences is quite exhausting and time-intensive job due to the detailed spatial locations and association of time durations [90]. A huge amount of research indicates the use of transfer learning with vision-based activity recognition [160]. Nevertheless, researchers are applying transfer learning techniques to both activity recognition using wearable accelerometers as well as activity recognition using smartphones [161, 174] and Ambient sensors.

6 Reinforcement learning (RL)

Unlike supervised learning algorithms, which has labeled training data set, or unsupervised algorithm which learn from the data, RL learns from continuous interaction with the environment. The problem solver in RL is called an *agent* while everything around the agent is known as the *environment*. The agent takes *actions*, against each action the environment transacts through its state and generates *reward* and *observation* to the agent as shown in Fig. 16. The reward is positive or negative reinforcement which tells the agent how good the previous action was. While the observation is sampled version of the internal state of the environment.

RL works in two ways, firstly when the environment is fully observable and secondly if the environment is unknown. Markov decision process (MDP) is used to represent a case when the environment is fully observable, observations, and the state of the environment will be the same in this case. The memoryless property of MDP means that the probability of the next state depends on the current state and action and not on the history of interactions in past. In Eq. 17 shows probability of next state $P(S_{t+1})$ given current state S_t and action A_t is actually chain of all states and actions before $t+1$. Equation 18 shows MDP where $P(S_{t+1}$ depends only on last state $P(S_t)$ and action A_t and not the entire chain of states and actions before time t.

$$P(S_{t+1} \mid S_t, A_t) = P(S_{t+1} \mid S_t, A_t, S_{t-1}, A_{t-1}, \ldots) \tag{17}$$

$$P(S_{t+1} \mid S_t, A_t) = P(S_{t+1} \mid S_t, A_t, \ldots) \tag{18}$$

Before proceeding further it is important to formally define all terminologies

- S : states of the environment, in case the environment is fully observable, the state of observations becomes the same.
- A : Defines a set of actions, against each observation, the agent takes one of the actions.
- r : Reward signal is provided by the environment against the action taken by the environment.
- γ : discount factor defines the worth of reward in the future.
- $p(s_{t+1}|s_t, a_t)$: defines the state transition model, how the next state will be transacted provided the environment is in state s and action a is performed.

The major job of the agent is to maximize the overall reward on average also known as a return. The return is the sum of all rewards obtained by the agent at the given time as shown in Eq. 19.

$$R = r_t + r_{t+1} + r_{t+2} + \ldots r_{t+H} \tag{19}$$

Where H is known as the horizon and defines the total

number of iterations or episodes. However, this sum might easily become infinite if the process continues forever. Therefore, a discount factor γ is added into the Eq. 19 to ensure the convergence as shown in Eq. 20 or in closed form in Eq. 21.

$$R = \gamma^0 \times r_t + \gamma^1 \times r_{t+1} + \gamma^2 \times r_{t+2} + \ldots \gamma^{H-1} \times r_{t+H} \tag{20}$$

$$R = \sum_{k=0}^{H} \gamma^k r_{t+k} \tag{21}$$

Equation 21 also reflects the fact the importance of reward becomes less significant as time passes. The procedure which is used by the agent to determine its next action is called policy π. This function maps the current state to the action which the agent should choose to reach the goal. The most well-known algorithms to find an optimal policy if the agent knows the environment is known as policy iteration (PI) and value iteration (VI). Before presenting VI and PI, it is important to define the value function (VF). The VF $V(S)$ defines how good it is for the agent to be in-state S. It is the average of total rewards if the agent starts from the state S and performs a certain set of actions chosen from policy π as shown in Eq. 22. Or in other words average return is obtained by the agent while being in some state S as shown in Eq. 23.

$$V_\pi(S) = E\left\{\sum_{k=0}^{H} \gamma^k r_{t+k} | S_t\right\} \tag{22}$$

$$V_\pi(S) = E\{G_t | S_t\} \tag{23}$$

Among all possible VF, there is one VF that has the maximum accumulative rewards and is represented by V^* as shown in Eq. 24. The corresponding optimal policy π^* having V^* is shown in 25.

$$V^*(S) = \max_\pi V^\pi(S) \quad \forall s \in S \tag{24}$$

$$\pi^* = arg\max_\pi V^\pi(S) \quad \forall s \in S \tag{25}$$

While VF only determines how good it is for an agent to be in-state S, the Q-function $Q(S, a)$ also tells the agent how good it is state S and take action a. The Q^*_π is an optimal q-function under policy π while being in-state S and taking action a. Since *V*(s)* is the maximum expected total reward when starting from state s, it will be the maximum of *Q*(s, a)* overall possible actions. Therefore, the relationship between *Q*(s, a)* and *V*(s)* is easily obtained, as shown in Eqs. 26 and 27.

$$V^*(S) = \max_a \quad Q^*(S, a) \quad \forall s \in S \tag{26}$$

$$\pi^* = arg\max_a \quad Q^*(S, a) \quad \forall s \in S \tag{27}$$

The Q function can also be expressed as in Eq. 23 for VF, in this case the value of action a under state s under policy π is given in Eq. 28.

$$Q_\pi(S, a) = E\{G_t | S_t, A_t\} = E\left\{\sum_{k=0}^{H} \gamma^k r_{t+k} | S_t, a_t\right\} \tag{28}$$

Computing summation multiple times for VF as in Eq. 22 and Q function as in Eq. 28 is not a simple and efficient solution. To solve this problem a dynamic programming (DP)-based solution is employed. Dp breaks the difficult problem into subproblems and solves them recursively. A well-known formulation known as the Bellman equation (BE) is used for DP. The BE breaks down value function into immediate reward and discounted future values as shown in Eq. 29.

$$R + \gamma V_\pi(S') \tag{29}$$

The Bellman equation for VF as shown in Eq. 30 is weighted with all possible actions given a certain state $\sum_a \pi(a, s)$ and probability or next state and rewards given current state and action $\sum_{s',r} p(s', r|s, a)$. Similarly, for q function the Bellman-based solution can be given as 31.

$$v_\pi(s) = \sum_a \pi(a, s) \quad \sum_{s',r} p(s', r|s, a) \quad [r + \gamma v_\pi(s')] \tag{30}$$

$$q_\pi(s, a) = \sum_{s',r} p(s', r|s, a) \quad [r + \gamma v_\pi(s')] \tag{31}$$

In order to find the optimal policy, two well-known techniques value iteration (VI) and policy iteration (PI) are presented in subsequent Sects. 6.1 and 6.1.1 respectively [180].

Generally, an RL agent's job is to make a policy that maximizes overall system rewards. While employing RL in HAR agents are trained in a way that they define the policy which enhances HAR accuracy. Activity learning with the mobile robot is challenging, the objective is to learn the activity with a high level of accuracy and least energy consumption. In [98] RL-based algorithm is used to control the motion of the robot which is observing the activities.

Human activity and behavior are considered better estimated and recognized with RL [168, 228]. For example, in [168] human arm movement has been recognized with RL. Commercial sensors have been deployed to sense human arm acceleration and agents of RL learn the pattern of motion. Then in [228] human behavior is observed and predicted with the help of deep-RL in a smart home-based environment. RL is better than its counterpart supervised and deterministic algorithms in the sense that agents can learn and predict the event by themselves, even if the

suitable action has not been provided to the agent. Because of this capability, it is very much appropriate for the applications which expect to have scenarios, never encountered before.

6.1 Value iteration

VI computes $V^*(S)$ recursively to improve the estimated value of V(S). It uses Eq. 31 until the $V(S)$ converges. The algorithm for VI is shown in algorithm 4.

Algorithm 4 Value Iteration algorithm

$V(S) \leftarrow$ random value
while $V(S)$ converges **do**
  for $\forall s_i$ **do**
    for $\forall a_i$ **do**
      $q_\pi(s,a) \leftarrow \sum_{s',r} p(s',r|s,a) \quad [r + \gamma v_\pi(s')]$
    $V^*(S) \leftarrow \max_a \quad Q^*(S,a)$

6.1.1 Policy iteration

One of the well-known algorithms to find the optimal policy is called policy iteration, in this algorithm, a random policy π is selected and is evaluated and improved iteratively. The major problem with VI algorithms is that it keeps on improving the VF until the VF converges. Since the major goal of the agent is to find the optimal policy, which in some cases will converge before VF. The PI redefines policy instead of improving VF at each step as shown in algorithm 5.

Algorithm 5 Policy Iteration algorithm

$\pi' \leftarrow$ random value
while $\pi = \pi'$ **do**
  for $\forall s_i$ **do**
    $v_\pi(s) \leftarrow \sum_a$
$\pi(a,s) \quad \sum_{s',r} p(s',r|s,a) \quad [r + \gamma v_\pi(s')]$

Other popular techniques for reinforcement learning if the environment is not known and can be observed through observations are actor-critic methods [95], Monte Carlo methods [204], Temporal difference (TD) learning-based RL [179].

RL has been used for a variety of tasks in HAR. For example, it is used to select appropriate features. A research work selected features for HAR based on the cost of feature selection and improving classifier performance [78]. Another work employed deep RL to drive policy from two activity recognition. The first one is the motion predictor using LSTM second vision predictor using CNN and LSTM [152]. A similar work used RL for feature selection by finding the right balance between power consumption and accuracy [212]. Recently robot-assisted life and HAR have gained attention and RL has been traditionally used in several areas of robotics. In this sense RL-based HAR using robots is one of the recent popular areas of research [107, 19].

7 Other related machine learning techniques

This section introduces Self-organizing maps (SOMs), Multiple classifiers systems (MCS), and multiple instance learning.

7.1 Self-organizing maps (SOMs)

Self-organized Maps (SOM) are unsupervised learning techniques that are also used in ANN. Unlike traditional ANNs, they are not trained using backpropagation neural networks; instead, they utilize competitive learning. SOM is used in [76] for identifying the basic posture prototypes of all the actions. The cumulative fuzzy distances from the SOM are calculated to achieve time-invariant action representations. After that, the Bayesian framework is applied to combine the recognition results produced for each camera. The solution to the camera viewing angle identification problem using combined neural networks. Rigorous experiments based on four datasets, KTH, Weizmann, UT-interaction, and TenthLab were carried out to assess the performance of the approach proposed in [110]. This resulted in the accuracy of 98.83%, 99.10%, 99.00%, and 97.00%, respectively for the abovementioned datasets.

7.2 Multiple classifier systems (MCS)

Multiple classifier systems (MCS) employ different prediction/classification algorithms to achieve a more accurate and reliable decision. There are three main multiple classifier systems, known as ensemble methods, a committee of classifiers, and a mixture of experts. Ensemble methods consist of multiple independent learning models to predict class labels based on the prediction made by multiple models as already discussed in Sect. 3.3. This technique is more popular for reducing total error, including decreased variance (bagging) and bias (boosting). The random forest algorithm comes in the category of an ensemble approach. Thus, a random forest algorithm creates multiple decision trees on data samples, as shown in Fig. 9. Subsequently, the prediction from each tree is counted, and the optimal solution is selected through voting.

Table 6 Best dataset for HAR ADL with highest accuracy level

Refs.	ML approach	Accuracy	Dataset	Num of ADL
[185]	HMM	92.50 %	Global silhouette features	6ADL
[135]	MLR	93.44 %	Self	10 ADL
[88]	SVM	94.07%	i3DPost and IXMAS	10 ADL
[102]	Additive Logistic Regression algorithm	95.7%	Dvs dataset	5 ADL
[182]	NB classifier	80% Subject dependent	Self	30ADL
		56% Subject independent		
[81]	Bagged Tree Model	95.7 %	Self	9 ADL
[21]	IBK classifier	99.9 % 6 Common ADL	MobiAct	6 ADL
		96.8 % all ADLs and 4 falls		12 ADL

The researchers have also opted for the approach of information fusion in multimodal biometrics by using pre-classification. Then they further classified it into sensor level and feature level extraction, which is helpful in video surveillance [52]. The well-known examples of classifier combinations based on resampling strategies are stacking, bagging, and boosting [169], as already discussed in detail in Sect. 3.3. The study conducted in [81] also attempted to train a bagged tree (Random Forest), and Boosted Tree (Gentle Adaboost), with a different number of trees for HAR. As mentioned earlier, the authors have also used SVM and KNN for comparative analysis. Results demonstrate that bagged trees with 300 trees achieved the lowest error rate of 4.3%. Apart from this, the multiclass classification for human activity classification based on micro-Doppler signatures was implemented using a decision-tree structure. In this research work, the classification accuracy based on the six features was achieved at around 90% [94].

7.3 Multiple instance learning

Multiple Instance Learning (MIL) has been used for human action recognition in video and image sequences. HOG and T-HOG (HOG-based texture descriptor) model is used for extracting space-time feature; the optical flow model is used for extracting motion features which are used to characterize human action. In action modeling and recognition, MIL is combined with AnyBoost and proposed the MIL boost for human action recognition [87]. They propose a novel multiple-instance Markov model to overcome the disadvantages of the traditional Markov model for human action recognition. This model's silent features are: First, it has a multiple-instance formulation, which makes this model select elementary actions with stable state variables. Second, this method gives a novel activity representation: a Markov chain bag, which encodes both local and long-range temporal information among elementary actions. Finally, this model explores the most discriminative Markov chain for action representation.

7.4 Spatial temporal pattern

As stated earlier, naturally, there is a temporal correlation present in human activities. With these temporal correlation properties, the next action can be predicted and recognized without intensive training. HMM, [31, 185], DTW, Fourier Temporal Pyramid, and Actionlet Ensemble Model have been used in the literature to detect a temporal pattern. HMM is well known for its capability to recognize temporal relations [185], although it requires extensive training. DTW is applied to find the distance between two temporal actions, then actions are recognized based on nearest-neighbor classification. Fourier Temporal Pyramid is very efficient for noise removal and discriminative for action recognition. But it is insensitive if there is any temporal misalignment. In contrast to the Action, let Ensemble Model is invariant to temporal misalignment and is also robust to noise. The researchers in [135] presented an unsupervised learning approach, i.e., a "bag of spatial-temporal words" model combined with a space-time interest points detector, for human action categorization and localization. The algorithm can also localize multiple actions in complex motion sequences containing multiple actions, and their results are promising. For similar actions (e.g., "running" and "walking"), the classification may benefit from a discriminative model. Additionally, few methods are based on the use of temporal characteristics in the recognition task. Relatively simple activities such as walking are typically used as test scenarios; the systems may use low-level or high-level data [123]. Low-level recognition is typically based on spatiotemporal data without much processing. The data are spatiotemporal templates [222] and motion templates [118]. The goal is usually to recognize whether a human is walking in the

scene or not [69]. More high-level methods are usually based on pose estimated data. These methods includes correlation, silhouette matching [93], HMMs [185] and neural networks [69, 169]. The objective is to recognize actions such as walking, carrying objects, removing and placing objects, pointing and waving [35], gestures for control [27], standing vs walking, walking vs jogging, walking vs running [61], and classifying various aerobic exercises [9, 182], or ballet dance steps [49].

8 Performance analysis of HAR systems

The current section will discuss the performance, accuracy, and challenges of HAR-based systems.

8.1 Performance and accuracy

This subsection provides an in-depth analysis of well-known algorithms for HAR with the particular application area and boundaries. Indeed, the algorithm selection depends on many factors, including the nature of the activity, such as speed of action and its complexity, and the amount of training data available. When there is insufficient training data, the training model can not attain a proper distribution trend and results in overfitting for decision trees, neural networks, and underfitting for SVM. Primarily, for detecting activity, probability-based algorithms work well to learn from actions and recognize the activity. But these probability-based methods are usually complicated and computationally inefficient. HMM is an example of such a probability-based algorithm that can estimate many parameters. Generally, because of its Markovian property, HMM calculated the conditionally independent features, but we cannot generalize it for all applications. Because of the normalization issue, the current observation sequence is mostly overlooked and ends in incorrect detection. Hence, whenever some application has a series of complex events, HMM is not a good choice. If these complex events can be decomposed into sub-events with simpler activities, HMM may work better. Moreover, if global normalization is applied, we can solve the label bias problem, in which the current observation has low entropy, and due to that, it is ignored. Another popular choice for HAR for different applications is SVM. It works well for data whose distribution is not known. Once its decision boundary is determined, it is a robust classifier and scales well for high-dimensional data.

For health-related applications, mostly deep learning classifiers are considered the favorite. There are multiple reasons for this choice; firstly, these classifiers are instantly capable of learning from raw data. So there is no requirement to extract handcrafted input features. Secondly, these models are capable of exploring and obtaining the advantage of temporal correlation between internal events. Hence, the model is well fitted; complex activity can be recognized efficiently due to deep layers; so simple to complex features is scalable. Further, the accuracy of different ADLs on the dataset obtained from various sources is shown in Table 6. This table summarizes the best research work based on model accuracy. It can be observed from table that bagged tree-based model [81] and logistic Regression-based [102] are giving same 95.7 % accuracy; however, [102] is based on fewer activities than [81]. Then [21] utilized IBK classifier (instance-based learner using 1 nearest neighbor). This research work outperforms all of its counterparts with 99.9 % accuracy for 6 common activities and 96.8 % for a bit complex HAR task, including 12 ADL and 4 falls.

8.2 Challenges and limitations of HAR systems

The following subsections will describe Non-vision-based and vision-based HAR and their associated challenges in detail. The vision-based HAR generally has better performance as compared to Non-vision-based HAR, though vision-based techniques are more challenging. Firstly, vision-based techniques have privacy issues, as not every person is willing to be observed constantly and recorded by cameras. Secondly, it is not practically manageable to record the intended area of interest for certain applications during the entire recording period. Finally, the vision-based techniques are generally computationally costly and require much more preprocessing before HAR can be done.

Non-vision-based sensors used for human activity recognition also undergo some limitations, such as the number of sensors employed, which affects the measurement's granularity. The second limitation involves the sensor's location, which influences the readings' precision and accuracy, for example, in smart home monitoring scenarios. The third factor concerns deployment obstacles, such as in human body implants. In high mobility scenarios such as sports activity recognition the sensors might move or get displaced, specially if they are deployed on the body. Other issues are the environment's influences affecting maintenance, such as temperature, humidity, power supply, etc. In some cases, the cost becomes a bottleneck since high precision may require sensors that exceed the total budget of bulk production of the product. Finally, the sensors used for HAR might hinder or obstruct the subject's daily chores or normal life activities, such as in HAR for elderly persons.

There are several challenges common to both vision and non vision-based HAR, which can dramatically degrade the system's performance. Ideally, the extracted features may include several variations, including human appearances,

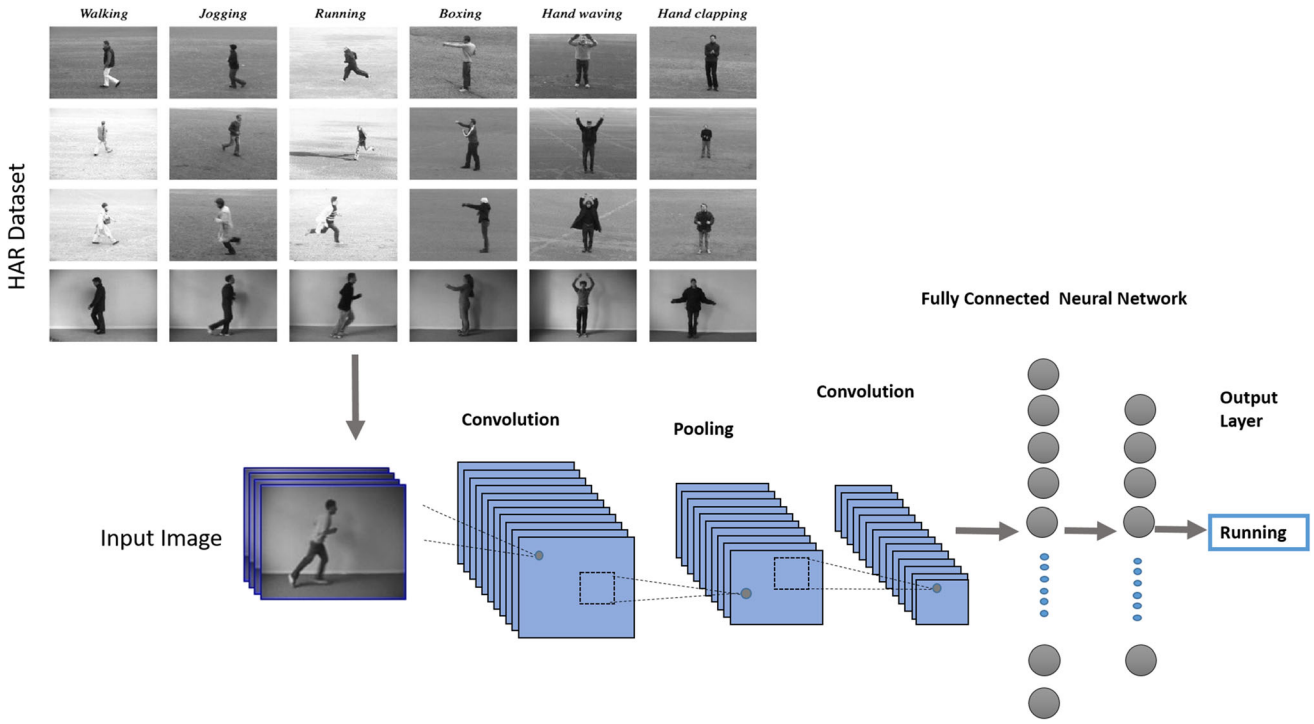


Fig. 17 Structure of deep convolutional neural network (DCN)

points of view, and diverse backgrounds. But in reality, each action could be performed in a different background, situation, and diverse rotations, illuminations and translations. Besides, the complexity of recognition may also depend on the speed of activity, the number of actions for an activity, the type of device used, and the sensing device's energy consumption. In applications like video surveillance and fault detection, offline processing adversely affects the characteristics of surveillance. For example, in sports event broadcast is a typical case of dynamic recording.

With the popularity of smart devices such as smart glasses and smartphones, people tend to record videos with embedded cameras from wearable devices. Most of these real-world videos have complex dynamic backgrounds. Therefore the main challenges include varying backgrounds, as in realistic cases, videos are broadcasted in complex backgrounds. Furthermore, these videos may have occlusions, brightness, and viewpoint variations, which introduce complications, thus requiring a high level of signal processing and machine learning algorithms to detect action in such dynamic conditions. Another significant challenge is due to the long-distance and low quality of videos recorded by surveillance cameras. In most cases, the cameras are installed at an elevated place; therefore, they cannot provide high-quality videos comparable to the offline dataset in which the targeted person is apparent and obvious.

8.3 Future directions and open issues

8.3.1 Simplification of complex models

Video-based human activity recognition is daunting and challenging in terms of model and training. One of the key simplifications can be achieved by using transfer learning and utilizing image models for video or transferring the knowledge learned from related video sequences.

8.3.2 Exploring temporal correlations among actions

Human activity recognition is the well-explored area for sensors and videos-based data. However, besides individual action recognition, it is critical to understand the correlation between different actions. Besides, various uncorrected activities might also have a time-based relationship between them. For example, climbing down the stairs and opening doors might be a different event; however, this is related to a person opening a door for his guest. One of the future directions involves finding a frame of action in uncorrelated temporal sequences. Consequently, this results in event recognition using uncorrelated activities.

8.3.3 Association between environment and human activity

Various actions occur in a specific space or Environment; although a lot of work has been done on HAR, very little attention has been paid to scene integration with human activities. One of the key future directions involves object correlation and integration with HAR. For example, in elderly HAR, the object around which a certain action occurred plays a vital role in the correct understanding of the action itself. Besides, the objects around the action might help understand the action's potential causes, which improves HAR.

8.3.4 Multiassociation of actions using big data

Most of the current literature focuses on human activity recognition from a particular scene. The upcoming 5G technology enables big data acquisition and processing. Big data can be applied to HAR by collecting data from multiple scenes having similar actions and finding spatial aggregates or filters to this data. This also helps to find interpolations to missing data or associations among data.

8.3.5 Real-time HAR and online learning

While offline approaches are useful in several application settings, some applications need real time processing for HAR. The online and real-time systems are constrained by power consumption and the short processing time. Nevertheless, some new techniques utilize inertial sensors to improve the accuracy of online approaches. Model development, machine learning, and better accuracy with constraint resources for real time systems is a challenging, an open and developing area of research.

9 Conclusion

HAR is an active area of research in computer science as it finds applications in numerous domains. The data for HAR is acquired using vision and non-vision-based sensors. Both types of sensors are relevant and suitable for certain application domains. This survey analyzes different application domains and certain sensors suitable for each case. We have provided pros and cons for vision as well as non-vision-based sensors. Several datasets are available in the literature and are suitable for different application areas; each provides a different set of actions, sensors, and data sampled at different rates. We reviewed available datasets for various application domains. We concluded that if a HAR application requires a short time of monitoring wearable sensors are used on the other hand with applications requiring long-term monitoring ambiance sensors are installed.

The survey provides various machine learning approaches to recognize human activities, including SVM, decision tree and KNN, bagged tree-based model, HMM, and GMM. After conducting a detailed analysis of traditional methods, we observed that SVM, decision tree, and KNN with the bagged tree work best for most application areas. This survey covers deep learning literature including CNN, transfer learning, reinforcement learning, RNN, and autoencoders. We identified that in presence of limited training dataset reinforcement learning can be used. CNN is one of the most suited techniques for feature extraction however, in the case of high dimensional feature space autoencoders can be used. In case HAR contains sequences RNN and its variants such as LSTM and GRU can be used. Another variant called transformers can be used in case the sequential nature of LSTM could affect the computational time of the system. We conclude that deep learning methods have much higher performance and accuracy as compared to traditional machine learning approaches.

It is advantageous for researchers to get a clearer picture of the current trends and research techniques in human activity recognition and know which devices, datasets, and algorithms are most suitable for the particular application area. Finally, we provided future directions, limitations, and openings in the area of HAR. In summary, we have seen HAR in terms of application areas and analyzed available datasets, algorithms, and sensors so that it could help researchers of HAR to choose from state-of-the-art.

Declarations

Conflict of interest All the authors declare no conflict of interest.

References

1. Abbaspour S, Fotouhi F, Sedaghatbaf A, Fotouhi H, Vahabi M, Linden M (2020) A comparative analysis of hybrid deep learning models for human activity recognition. Sensors 20(19):5707
2. Aggarwal JK, Xia L (2014) Human activity recognition from 3D data: a review. Pattern Recogn Lett 48:70–80
3. Alakwaa W, Nassef M, Badr A (2017) Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). Lung Cancer 8(8):409
4. Alawneh L, Alsarhan T, Al-Zinati M, Al-Ayyoub M, Jararweh Y, Hongtao L (2021) Enhancing human activity recognition using deep learning and time series augmented data. J Ambient Intell Humaniz Comput 12(12):10565–10580
5. Almaslukh B, AlMuhtadi J, Artoli A (2017) An effective deep autoencoder approach for online smartphone-based human activity recognition. Int J Comput Sci Netw Secur 17(4):160–165

6. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, Hasan M, Van Essen BC, Awwal AAS, Asari VK (2019) A state-of-the-art survey on deep learning theory and architectures. Electronics 8(3):292
7. Abu AM, Ahmed S, Dusit N, Linda D, Shaowei L, Hwee-Pink T (2016) Deep activity recognition models with triaxial accelerometers. In: Workshops at the thirtieth AAAI conference on artificial intelligence
8. Altenberger F, Lenz C (2018) A non-technical survey on deep convolutional neural network architectures. arXiv preprint arXiv:1803.02129
9. Amor BB, Jingyong S, Srivastava A (2015) Action recognition using rate-invariant analysis of skeletal shape trajectories. IEEE Trans Pattern Anal Mach Intell 38(1):1–13
10. Anitha G, Baghavathi Priya S (2019) Posture based health monitoring and unusual behavior recognition system for elderly using dynamic Bayesian network. Clust Comput 22(6):13583–13590
11. Ann OC, Theng LB (2014) Human activity recognition: a review. In: 2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014), pp. 389–393. IEEE
12. Attal F, Mohammed S, Dedabrishvili M, Chamroukhi F, Oukhellou L, Amirat Y (2015) Physical human activity recognition using wearable sensors. Sensors 15(12):31314–31338
13. Avci A, Bosch S, Marin-Perianu M, Marin-Perianu R, Havinga P (2010) Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In: 2010 23th International conference on architecture of computing systems pp. 1–10. VDE
14. Baloch Z, Shaikh FK, Unar MA (2019) Deep architectures for human activity recognition using sensors. 3C Tecnol 8:14–35
15. Banou S, Swaminathan M, Reus Muns G, Duong D, Kulsoom F, Savazzi P, Vizziello A, Chowdhury KR (2019) Beamforming galvanic coupling signals for IoMT implant-to-relay communication. IEEE Sens J 19(19):8487–8501
16. Bhattacharya S, Lane ND (2016) From smart to deep: Robust activity recognition on smartwatches using deep learning. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 1–6. IEEE
17. Bin Abdullah MFA, Negara AFP, Sayeed MS, Choi DJ, Muthu KS (2012) Classification algorithms in human activity recognition using smartphones. Int J Comput Inf Eng 6(77-84):106
18. Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (2015) Activitynet: a large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 961–970
19. Cai L, Boukhechba M, Kaur N, Wu C, Barnes LE, Gerber MS (2019) Adaptive passive mobile sensing using reinforcement learning. In: 2019 IEEE 20th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM), pp. 1–6. IEEE
20. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633–659
21. Chatzaki C, Pediaditis M, Vavoulas G, Tsiknakis M (2017) Human daily activity and fall recognition using a smartphone's acceleration sensor. In: Information and communication technologies for ageing well and e-Health, pp. 100–118. Springer International Publishing, Cham
22. Chaudhry HN, Javed Y, Kulsoom F, Mehmood Z, Khan ZI, Shoaib U, Janjua SH (2021) Sentiment analysis of before and after elections: Twitter data of us election 2020. Electronics 10(17):2082
23. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion maps-based local binary patterns. In: 2015 IEEE Winter Conference on Applications of Computer Vision, pp. 1092–1099. IEEE
24. Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps. J Real-Time Image Proc 12(1):155–163
25. Chen C, Zhu Z, Hammad A (2020) Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Autom Constr 110:103045
26. Chen IZC, Hengjinda P (2021) Early prediction of coronary artery disease (cad) by machine learning method-a comparative study. J Artif Intell 3(01):17–33
27. Chen L, Wei H, Ferryman J (2013) A survey of human motion analysis using depth imagery. Pattern Recogn Lett 34(15):1995–2006
28. Chen Y, Zhong K, Zhang J, Sun Q, Zhao X (2016) LSTM networks for mobile human activity recognition. In: 2016 International conference on artificial intelligence: technologies and applications. Atlantis Press
29. Cheng X, Huang B, Zong J (2021) Device-free human activity recognition based on GMM-HMM using channel state information. IEEE Access 9:76592–76601
30. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
31. Clark VNJJ (2002) Automated visual surveillance using hidden markov models. In: International conference on vision interface, pp. 88–93
32. Cui W, Li B, Zhang L, Chen Z (2021) Device-free single-user activity recognition using diversified deep ensemble learning. Appl Soft Comput 102:107066
33. Cumani S, Laface P, Kulsoom F (2015) Speaker recognition by means of acoustic and phonetically informed GMMS. In: Sixteenth annual conference of the international speech communication association
34. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886–893. IEEE
35. Darrell T, Maes P, Blumberg B, Pentland A (1996) A novel environment for situated vision and behavior. In: Exploratory Vision, pp. 319–331. Springer
36. Das S, Thonnat M, Sakhalkar K, Koperski M, Bremond F, Francesca G (2019) A new hybrid architecture for human activity recognition from RGB-D videos. In: International conference on multimedia modeling, pp. 493–505. Springer
37. Daverio P, Chaudhry HN, Margara A, Rossi M (2021) Temporal pattern recognition in graph data structures. In: 2021 IEEE International conference on big data (Big Data), pp. 2753–2763. IEEE
38. Davis K, Owusu E, Bastani V, Marcenaro L, Hu J, Regazzoni C, Feijs L (2016) Activity recognition based on inertial sensors for ambient assisted living. In: 2016 19th International conference on information fusion (FUSION), pp. 371–378
39. Devanne M, Wannous H, Berretti S, Pala P, Daoudi M, Del Bimbo A (2014) 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold. IEEE Trans Cybern 45(7):1340–1352
40. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
41. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S,

et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
42. Du Y, Chen F, Xu W, Li Y (2006) Recognizing interaction activities using dynamic Bayesian network. In: 18th International conference on pattern recognition (ICPR'06), vol. 1, pages 618–621. IEEE
43. Elangovan V (2021) Indoor group activity recognition using multi-layered HMMS. arXiv preprint arXiv:2101.10857
44. Ellis C, Masood SZ, Tappen MF, LaViola JJ, Sukthankar R (2013) Exploring the trade-off between accuracy and observational latency in action recognition. Int J Comput Vis 101(3):420–436
45. Epstein D, Chen B, Vondrick C (2019) Oops! predicting unintentional action in video. arXiv preprint arXiv:1911.11206
46. Evangelidis GD, Singh G, Horaud R (2014) Continuous gesture recognition from articulated poses. In: European conference on computer vision, pages 595–607. Springer
47. Fahad LG, Rajarajan M (2015) Integration of discriminative and generative models for activity recognition in smart homes. Appl Soft Comput 37:992–1001
48. Fanello SR, Gori I, Metta G, Odone F (2013) Keep it simple and sparse: real-time action recognition. J Mach Learn Res 14(44):2617–2640
49. Faridee AZM, Ramamurthy SR, Hossain HMS, Roy N (2018) Happyfeet: Recognizing and assessing dance on the floor. In: Proceedings of the 19th international workshop on mobile computing systems and applications, pp. 49–54
50. Ferreira PJS, Cardoso JMP, Mendes-Moreira J (2020) Knn prototyping schemes for embedded human activity recognition with online learning. Computers 9(4):96
51. Feuz KD, Cook DJ (2014) Heterogeneous transfer learning for activity recognition using heuristic search techniques. In: International journal of pervasive computing and communications
52. Fierrez J, Morales A, Vera-Rodriguez R, Camacho D (2018) Multiple classifiers in biometrics. part 1: Fundamentals and review. Inf Fus 44:57–64
53. Gaglio S, Re GL, Morana M (2014) Human activity recognition process using 3-D posture data. IEEE Trans Human-Mach Syst 45(5):586–597
54. Gao X, Haiyong Luo Q, Wang FZ, Ye L, Zhang Y (2019) A human activity recognition algorithm based on stacking denoising autoencoder and lightGBM. Sensors 19(4):947
55. Garcia-Ceja E, Galván-Tejada CE, Brena R (2018) Multi-view stacking for activity recognition with sound and accelerometer data. Inf Fus 40:45–56
56. Gavrilyuk K, Sanford R, Javan M, Snoek CGM (2020) Actor-transformers for group activity recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 839–848
57. Gong D, Medioni G, Zhao X (2013) Structured time series analysis for human action segmentation and recognition. IEEE Trans Pattern Anal Mach Intell 36(7):1414–1427
58. Gordon J, Hernández-Lobato JM (2020) Combining deep generative and discriminative models for Bayesian semi-supervised learning. Pattern Recogn 100:107156
59. Guan Yu, Plötz T (2017) Ensembles of deep LSTM learners for activity recognition using wearables. Proc ACM Interact Mob Wear Ubiq Technol 1(2):1–28
60. Gusain K, Gupta A, Popli B (2018) Transition-aware human activity recognition using extreme gradient boosted decision trees. In: Advanced computing and communication technologies, pp. 41–49. Springer
61. Ha JM, Yun S, Choi S (2015) Multi-modal convolutional neural networks for activity recognition. In: 2015 IEEE International conference on systems, man, and cybernetics, pp. 3017–3022. IEEE
62. Hannink J, Kautz T, Pasluosta CF, Gaßmann K-G, Klucken J, Eskofier BM (2016) Sensor-based gait parameter extraction with deep convolutional neural networks. IEEE J Biomed Health Inform 21(1):85–93
63. Hantke S, Abstreiter A, Cummins N, Schuller B (2018) Trustability-based dynamic active learning for crowdsourced labelling of emotional audio data. IEEE Access 6:42142–42155
64. Hartmann Y, Liu H, Schultz T (2021) Feature space reduction for human activity recognition based on multi-channel biosignals. In: Biosignals, pp. 215–222
65. Hayashi T, Nishida M, Kitaoka N, Takeda K (2015) Daily activity recognition based on dnn using environmental sound and acceleration signals. In: 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 2306–2310. IEEE
66. He ZY, Jin LW (2008) Activity recognition from acceleration data using AR model representation and SVM. In: 2008 International conference on machine learning and cybernetics, vol. 4, pp. 2245–2250
67. He Z, Jin L (2009) Activity recognition from acceleration data based on discrete consine transform and SVM. In: 2009 IEEE international conference on systems, man and cybernetics, pp. 5041–5044
68. Heckerman D (2008) A tutorial on learning with Bayesian networks. In: Innovations in Bayesian networks, pp. 33–82. Springer
69. Heisele B, Woehler C (1998) Motion-based recognition of pedestrians. In: Proceedings of the fourteenth international conference on pattern recognition (Cat. No. 98EX170), volume 2, pp. 1325–1330. IEEE
70. Helmi AM, Al-Qaness MAA, Dahou A, Damaševičius R, Krilavičius T, Elaziz MA (2021) A novel hybrid gradient-based optimizer and grey wolf optimizer feature selection method for human activity recognition using smartphone sensors. Entropy 23(8):1065
71. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
72. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
73. Hsu Y-L, Yang S-C, Chang H-C, Lai H-C (2018) Human daily and sport activity recognition using a wearable inertial sensor network. IEEE Access 6:31715–31728
74. Hu DH, Yang Q(2011) Transfer learning for activity recognition via sensor mapping. In: Twenty-second international joint conference on artificial intelligence
75. Iloga S, Bordat A, Le Kernec J, Romain O (2021) Human activity recognition based on acceleration data from smartphones using HMMS. IEEE Access 9:139336–139351
76. Iosifidis A, Tefas A, Pitas I (2012) View-invariant action recognition based on artificial neural networks. IEEE Trans Neural Netw Learn Syst 23(3):412–424
77. Iqbal JLM, Lavanya J, Arun S (2015) Abnormal human activity recognition using scale invariant feature transform. Int J Curr Eng Technol 5(6):3748–3751
78. Janisch J, Pevnỳ T, Lisỳ V (2019) Classification with costly features using deep reinforcement learning. Proc AAAI Conf Artif Intell 33:3959–3966
79. Jhuang H, Gall J, Zuffi S, Schmid C, black MJ (2013) Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 3192–3199
80. Ji S, Wei X, Yang M, Kai Yu (2012) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
81. Ji W, Liu H, Fisher J (2016) Human activity recognition via cellphone sensor data. In: Stanford University, 2016

82. Jiang W, Yin Z (2015) Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of the 23rd ACM international conference on multimedia, pp. 1307–1310
83. Kabir MH, Hoque MR, Thapa K, Yang S-H (2016) Two-layer hidden Markov model for human activity recognition in home environments. Int J Distrib Sens Netw 12(1):4560365
84. Kalischewski K, Wagner D, Velten J, Kummert A (2017) Activity recognition for indoor movement and estimation of travelled path. In: 2017 10th international workshop on multidimensional (nD) systems (nDS)
85. Kalsum T, Mehmood Z, Kulsoom F, Chaudhry HN, Khan AR, Rashid M, Saba T (2021) Localization and classification of human facial emotions using local intensity order pattern and shape-based texture features. J Intell Fuzzy Syst 40:9311–9331
86. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 36:61–78
87. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1725–1732
88. Käse N, Babaee M, Rigoll G (2017) Multi-view human activity recognition using motion frequency. In: 2017 IEEE international conference on image processing (ICIP), pp. 3963–3967. IEEE
89. Ke S-R, Le Uyen H, Thuc Y-JL, Hwang J-N, Yoo J-H, Choi K-H (2013) A review on video-based human activity recognition. Computers 2(2):88–131
90. Khan HAAF, Roy N (2017) Transact: transfer learning enabled activity recognition. In: 2017 IEEE International conference on pervasive computing and communications workshops (PerCom Workshops), pp. 545–550. IEEE
91. Khan SD, Basalamah S (2021) Scale and density invariant head detection deep model for crowd counting in pedestrian crowds. Vis Comput 37(8):2127–2137
92. Kim J, Li G, Yun I, Jung C, Kim J (2021) Weakly-supervised temporal attention 3D network for human action recognition. In: Pattern Recognition p. 108068
93. Kim K, Jalal A, Mahmood M (2019) Vision-based human activity recognition system using depth silhouettes: a smart home system for monitoring the residents. J Electr Eng Technol 14(6):2567–2573
94. Kim Y, Ling H (2009) Human activity classification based on micro-doppler signatures using a support vector machine. IEEE Trans Geosci Remote Sens 47(5):1328–1337
95. Konda V, Tsitsiklis J (1999) Actor-critic algorithms. Adv Neural Inf Process Syst, 12
96. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: 2011 International conference on computer vision, pp. 2556–2563. IEEE
97. Kulkarni S, Jadhav S, Adhikari D (2020) A survey on human group activity recognition by analysing person action from video sequences using machine learning techniques. In: Optimization in machine learning and applications, pp. 141–153. Springer
98. Kumrai T, Korpela J, Maekawa T, Yu Y, Kanai R (2020) Human activity recognition with deep reinforcement learning using the camera of a mobile robot. In: 2020 IEEE international conference on pervasive computing and communications (PerCom), pp. 1–10. IEEE
99. Kwon MC, Choi S (2018) Recognition of daily human activity using an artificial neural network and smartwatch. Wireless Commun Mob Comput 2018
100. Lafferty J, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: 2001 Proceedings of the 18th international conference on machine learning
101. Lara OD, Labrador MA (2012) A survey on human activity recognition using wearable sensors. IEEE Commun Surveys Tutor 15(3):1192–1209
102. Lara OD, Pérez AJ, Labrador MA, Posada JD (2012) Centinela: a human activity recognition system based on acceleration and vital sign data. Pervasive Mob Comput 8(5):717–729
103. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
104. Lee SM, Yoon SM, Cho H (2017) Human activity recognition from accelerometer data using convolutional neural network. In: 2017 IEEE international conference on big data and smart computing (bigcomp)
105. Li R, Liu Z, Tan J (2018) Exploring 3D human action recognition: from offline to online. Sensors 18(2):633
106. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. In: 2010 IEEE Computer society conference on computer vision and pattern recognition-workshops, pp. 9–14. IEEE
107. Xing L, Junpei Z, Kamruzzaman MM (2021) Complicated robot activity recognition by quality-aware deep reinforcement learning. Futur Gener Comput Syst 117:480–485
108. Lin XM, Li SZ (2009) Transfer adaboost learning for action recognition. In: 2009 IEEE international symposium on IT in edicine and education, vol. 1, pp. 659–664. IEEE
109. Litjens G, Ciompi F, Wolterink JM, de Vos BD, Leiner T, Teuwen J, Išgum I (2019) State-of-the-art deep learning in cardiovascular image analysis. JACC Cardiovasc Imag 12(8):1549–1565
110. Liu C, Ying J, Yang H, Hu X, Liu J (2020) Improved human action recognition approach based on two-stream convolutional neural network model. In: The visual computer, pp. 1–15
111. Liu H, Hartmann Y, Schultz T (2021) Motion units: generalized sequence modeling of human activities for sensor-based activity recognition. In: 2021 29th European signal processing conference (EUSIPCO), pp. 1506–1510
112. Liu Z, Li S, Hao J, Hu J, Pan M (2021) An efficient and fast model reduced kernel knn for human activity recognition. J Adv Transport, 2021
113. Luo J, Wang W, Qi H (2013) Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: Proceedings of the IEEE international conference on computer vision, pp. 1809–1816
114. Luptáková IDi, Kubovčík M, Pospíchal J (2022) Wearable sensor based human activity recognition with transformer. Preprint, 2022
115. Lyu L, He X, Law YW, Palaniswami M (2017) Privacy-preserving collaborative deep learning with application to human activity recognition. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp. 1219–1228
116. Mahmud S, Tonmoy M, Bhaumik KK, Rahman AKM, Amin MA, Shoyaib M, Khan MA, Ali AA (2020) Human activity recognition from wearable sensor data using self-attention. arXiv preprint arXiv:2003.09018
117. Manosha CKG, Rodrigo R (2012) Faster human activity recognition with SVM. In: International conference on advances in ICT for emerging regions (ICTer2012), pp. 197–203
118. Martinez J, Black MJ, Romero J (2017) On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2891–2900

119. Mazzia V, Angarano S, Salvetti F, Angelini F, Chiaberge M (2022) Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recogn 124:108487
120. Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J (2019) Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE international conference on computer vision, pp. 2630–2640
121. Minarno AE, Kusuma WA, Wibowo H (2020) Performance comparisson activity recognition using logistic regression and support vector machine. In: 2020 3rd International conference on intelligent autonomous systems (ICoIAS), pp. 19–24
122. Mliki H, Bouhlel F, Hammami M (2020) Human activity recognition from UAV-captured video sequences. Pattern Recogn 100:107140
123. Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Underst 81(3):231–268
124. Mojarad R, Attal F, Chibani A, Amirat Y (2020) Automatic classification error detection and correction for robust human activity recognition. IEEE Robot Autom Lett 5(2):2208–2215
125. Monfort M, Andonian A, Zhou B, Ramakrishnan K, Bargal SA, Yan T, Brown L, Fan Q, Gutfreund D, Vondrick C et al (2019) Moments in time dataset: one million videos for event understanding. IEEE Trans Pattern Anal Mach Intell 42(2):502–508
126. Morris BT, Trivedi MM (2011) Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach. IEEE Trans Pattern Anal Mach Intell 33(11):2287–2301
127. Moya Rueda F, Grzeszick R, Fink GA, Feldhorst S, Michael Ten Hompel (2018) Convolutional neural networks for human activity recognition using body-worn sensors. Informatics 5(2):26
128. Münzner S, Schmidt P, Reiss A, Hanselmann M, Stiefelhagen R, Dürichen R (2017) CNN-based sensor fusion techniques for multimodal human activity recognition. In: Proceedings of the 2017 ACM international symposium on wearable computers, pp. 158–165
129. Murad A, Pyun J-Y (2017) Deep recurrent neural networks for human activity recognition. Sensors 17(11):2556
130. Murahari VS, Plötz T (2018) On attention models for human activity recognition. In: Proceedings of the 2018 ACM international symposium on wearable computers, pp. 100–103
131. Muralikrishna SN, Muniyal B, Acharya UD, Holla R (2020) Enhanced human action recognition using fusion of skeletal joint dynamics and structural features. J Robot, 2020
132. Namatēvs I (2017) Deep convolutional neural networks: structure, feature extraction and training. Inf Technol Manag Sci (Sciendo) 20(1):40–47
133. Narejo S, Pasero E, Kulsoom F (2016) EEG based eye state classification using deep belief network and stacked autoencoder. Int J Electr Comput Eng 6(6):3131–3141
134. Naveenkumar M, Domnic S (2020) Deep ensemble network using distance maps and body part features for skeleton based action recognition. Pattern Recogn 100:107125
135. Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vision 79(3):299–318
136. Nurwulan NR, Selamaj G (2021) Human daily activities recognition using decision tree. J Phys: Conf Series 1833:012039
137. Nurwulan NR, Selamaj G (2021) A comparative evaluation of acceleration and jerk in human activity recognition using machine learning techniques. In: Proceedings of the 1st international conference on electronics, biomedical engineering, and health informatics, pp. 55–61. Springer
138. Nweke HF, Teh YW, Al-Garadi MA, Alo UR (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Syst Appl 105:233–261
139. Ogbuabor G, La R (2018) Human activity recognition for healthcare using smartphones. In: Proceedings of the 2018 10th international conference on machine learning and computing, pp. 41–46
140. Ohn-Bar E, Trivedi M (2013) Joint angles similarities and hog2 for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 465–470
141. Onofri L, Soda P, Pechenizkiy M, Iannello G (2016) A survey on using domain and contextual knowledge for human activity recognition in video streams. Expert Syst Appl 63:97–111
142. Ordóñez FJ, Roggen D (2016) Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1):115
143. O'Shea K, Nash R (2015) An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458
144. Osmani V, Balasubramaniam S, Botvich D (2008) Human activity recognition in pervasive health-care: supporting efficient remote collaboration. J Netw Comput Appl 31(4):628–655
145. Palaniappan A, Bhargavi R, Vaidehi V (2012) Abnormal human activity recognition using SVM based approach. In: 2012 International conference on recent trends in information technology, pp. 97–102
146. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
147. Paul P, George T (2015) An effective approach for human activity recognition on smartphone. In: 2015 IEEE International conference on engineering and technology (ICETECH), pp. 1–3
148. Paul P, George T (2015) An effective approach for human activity recognition on smartphone. In: 2015 IEEE International conference on engineering and technology (ICETECH), pp. 1–3. IEEE
149. Piyathilaka L, Kodagoda S (2013) Gaussian mixture based HMM for human daily activity recognition using 3d skeleton features. In: 2013 IEEE 8th conference on industrial electronics and applications (ICIEA), pp. 567–572
150. Piyathilaka L, Kodagoda S (2015) Human activity recognition for domestic robots. In: Field and service robotics, pp. 395–408. Springer
151. Plötz T, Hammerla NY, Olivier PL (2011) Feature learning for activity recognition in ubiquitous computing. In: Twenty-second international joint conference on artificial intelligence
152. Possas R, Caceres SP, Ramos F (2018) Egocentric activity recognition on a budget. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5967–5976
153. Pourbabaee B, Roshtkhari MJ, Khorasani K (2018) Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Trans Syst Man Cybern Syst 48(12):2095–2104
154. Purwar RK, Verma S, Jain A et al (2021) Crowd abnormality detection in video sequences using supervised convolutional neural network. Multimed Tools Appl 81:5259–5277
155. Qi M, Wang Y, Qin J, Li A, Luo J, Van Gool L (2019) stagnet: an attentive semantic RNN for group activity and individual action recognition. IEEE Trans Circuits Syst Video Technol 30(2):549–565
156. Raghavan VV, Gudivada VN, Govindaraju V, Rao CR (2016) Cognitive computing: theory and applications. Elsevier

157. Ramamurthy SR, Roy N (2018) Recent trends in machine learning for human activity recognition-a survey. Wiley Interdiscipl Rev Data Min Knowl Discov 8(4):e1254
158. Rashid N, Demirel BU, Faruque MA (2022) Ahar: Adaptive CNN for energy-efficient human activity recognition in low-power edge devices. In: IEEE Internet Things J, pp. 1–1
159. Ravi D, Wong C, Lo B, Yang GZ (2016) Deep learning for human activity recognition: a resource efficient implementation on low-power devices. In: 2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN), pp. 71–76. IEEE
160. Rokni SA, Ghasemzadeh H (2018) Autonomous training of activity recognition algorithms in mobile sensors: a transfer learning approach in context-invariant views. IEEE Trans Mob Comput 17(8):1764–1777
161. Rokni SA, Nourollahi M, Ghasemzadeh H (2018) Personalized human activity recognition using convolutional neural networks. In: Thirty-second AAAI conference on artificial intelligence
162. San-Segundo R, Montero JM, Moreno-Pimentel J, Pardo JM (2016) HMM adaptation for improving a human activity recognition system. Algorithms 9(3):60
163. Sanabria R, Caglayan O, Palaskar S, Elliott D, Barrault L, Specia L, Metze F (2018) How2: a large-scale dataset for multimodal language understanding. arXiv preprint arXiv:1811.00347
164. Sani S, Wiratunga N, Massie S, Cooper K (2017) knn sampling for personalised human activity recognition. In: International conference on case-based reasoning, pp. 330–344. Springer
165. Schuldt C Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: ICPR 2004 Proceedings of the 17th International Conference on Pattern Recognition, vol. 3, pp. 32–36. IEEE
166. Sebbak F, Chibani A, Amirat Y, Mokhtari A, Benhammadi F (2013) An evidential fusion approach for activity recognition in ambient intelligence environments. Robot Auton Syst 61(11):1235–1245
167. Sekiguchi R, Abe K, Yokoyama T, Kumano M, Kawakatsu M (2020) Ensemble learning for human activity recognition. In: Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers, pp. 335–339
168. Seok W, Park C (2018) Recognition of human motion with deep reinforcement learning. IEIE Trans Smart Process Comput 7(3):245–250
169. Shafiq M, Xiangzhan Yu, Bashir AK, Chaudhry HN, Wang D (2018) A machine learning approach for feature selection traffic classification using security analysis. J Supercomput 74(10):4867–4892
170. Shi S, Wang Q, Xu P, Chu X (2016) Benchmarking state-of-the-art deep learning software tools. In: 2016 7th International conference on cloud computing and big data (CCBD), pp. 99–104. IEEE
171. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJM (2015) A survey of online activity recognition using mobile phones. Sensors 15(1):2059–2085
172. Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A (2016) Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European conference on computer vision, pp. 510–526. Springer
173. Siraj MS, Shahid O, Ahad MAR (2020) Cooking activity recognition with varying sampling rates using deep convolutional GRU framework. In: Human activity recognition challenge, pp. 115–126. Springer
174. Soleimani E, Nazerfard E (2019) Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. arXiv preprint arXiv:1903.12489
175. Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402
176. Srivastava P, Wong WC (2012) Hierarchical human activity recognition using GMM. In: AMBIENT 2012: the second international conference on ambient computing, applications, services and technologies, pp. 32–37
177. Subetha T, Chitrakala S (2016) A survey on human activity recognition from videos. In: 2016 International conference on information communication and embedded systems (ICICES), pp. 1–7. IEEE
178. Sun J, Fu Y, Li S, He J, Xu C, Tan L (2018) Sequential human activity recognition based on deep convolutional network and extreme learning machine using wearable sensors. J Sensors, 2018
179. Sutton RS (1988) Learning to predict by the methods of temporal differences. Mach Learn 3(1):9–44
180. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT press
181. Tang Y, Ding D, Rao Y, Zheng Y, Zhang D, Zhao L, Lu J, Zhou J (2019) Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1207–1216
182. Tapia EM, Intille SS, Haskell W, Larson K, Wright J, King A, Friedman R (2007) Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: 2007 11th IEEE international symposium on wearable computers, pp. 37–40. IEEE
183. Torrey L, Shavlik J (2010) Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242–264. IGI global
184. Tripathi M (2021) Analysis of convolutional neural network based image classification techniques. J Innov Image Process (JIIP) 3(02):100–117
185. Uddin MZ, Thang ND, Kim TS (2010) Human activity recognition via 3-D joint angle features and hidden Markov models. In: 2010 IEEE international conference on image processing, pp. 713–716. IEEE
186. Ullah M, Ullah H, Khan SD, Cheikh FA (2019) Stacked LSTM network for human activity recognition using smartphone data. In: 2019 8th European workshop on visual information processing (EUVIP), pp. 175–180. IEEE
187. Usman Sarwar M, Rehman Javed A, Kulsoom F, Khan S, Tariq U, Kashif Bashir A (2021) Parciv: Recognizing physical activities having complex interclass variations using semantic data of smartphone. Softw Pract Exp 51:532–549
188. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Pocess Syst, 30
189. Vemulapalli R, Arrate F Chellappa R (2014) Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 588–595
190. Ventura D, Warnick S (2007) A theoretical foundation for inductive transfer. In: Brigham Young University, College of Physical and Mathematical Sciences, 19
191. Verma KK, Singh BM, Mandoria HL, Chauhan P (2020) Two-stage human activity recognition using 2D-convnet. Int J Interact Multimedia Artif Intell, 6(2)
192. Vijayakumar T, Vinothkanna R, Duraipandian M (2021) Fusion based feature extraction analysis of ECG signal interpretation-a systematic approach. J Artif Intell 3(01):1–16

193. Vishwakarma S, Agrawal A (2013) A survey on activity recognition and behavior understanding in video surveillance. Vis Comput 29(10):983–1009
194. Vizziello A, Savazzi P, Kulsoom F, Magenes G, Gamba P (2019) A novel galvanic coupling testbed based on pc sound card for intra-body communication links. In: EAI international conference on body area networks, pp. 135–149. Springer
195. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2:28
196. Wang H, Baosheng Yu, Xia K, Li J, Zuo X (2021) Skeleton edge motion networks for human action recognition. Neurocomputing 423:1–12
197. Wang H, Zhao J, Li J, Tian L, Tu P, Cao T, An Y, Wang K, Li S (2020) Wearable sensor-based human activity recognition using hybrid deep learning techniques. Secur Commun Netw, 2020
198. Wang J, Liu Z, Ying W, Yuan J (2013) learning actionlet ensemble for 3D human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914–927
199. Wang J, Chen Y, Hao S, Peng X, Lisha H (2019) Deep learning for sensor-based activity recognition: a survey. Pattern Recogn Lett 119:3–11
200. Wang L, Gu T, Tao X, Lu J (2009) Sensor-based human activity recognition in a multi-user scenario. In: European conference on ambient intelligence, pp. 78–87. Springer
201. Wang L (2016) Recognition of human activities using continuous autoencoders with wearable sensors. Sensors 16(2):189
202. Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona PO (2015) Action recognition from depth maps using deep convolutional neural networks. IEEE Trans Human-Mach Syst 46(4):498–509
203. Wang Y, Cang S, Yu H (2019) A survey on wearable sensor modality centred human activity recognition in health care. Expert Syst Appl 137:167–190
204. Wang Y, Won KS, Hsu D, Lee WS (2012) Monte carlo bayesian reinforcement learning. arXiv preprint arXiv:1206.6449
205. Wu C, Zhang J, Savarese S, Saxena A (2015) Watch-n-patch: Unsupervised understanding of actions and relations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4362–4370
206. Wu LF, Wang Q, Jian M, Qiao Y, Zhao BX (2021) A comprehensive review of group activity recognition in videos. Int J Autom Comput 18:334–350
207. Wu S, Oreifej O, Shah M (2011) Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: 2011 International conference on computer vision, pp. 1419–1426. IEEE
208. Xia K, Huang J, Wang H (2020) LSTM-CNN architecture for human activity recognition. IEEE Access 8:56855–56866
209. Xia L, Aggarwal JK (2013) Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2834–2841
210. Xin X, Tang J, Zhang X, Liu X, Zhang H, Qiu Y (2013) Exploring techniques for vision based human activity recognition: methods, systems, and evaluation. Sensors 13(2):1635–1650
211. Yalçın H (2016) Human activity recognition using deep belief networks. In: 2016 24th Signal processing and communication application conference (SIU), pp. 1649–1652
212. Yamagata T, Santos-Rodríguez R, McConville R, Elsts A (2019) Online feature selection for activity recognition using reinforcement learning with multiple feedback. arXiv preprint arXiv:1908.06134
213. Yan Y, Ricci E, Liu G, Sebe N (2015) Egocentric daily activity recognition via multitask clustering. IEEE Trans Image Process 24(10):2984–2995
214. Yang C, Wang Z, Wang B, Deng S, Liu G, Kang Y, Men H (2017) CHAR-HMM: an improved continuous human activity recognition algorithm based on hidden markov model. In: International conference on mobile ad-hoc and sensor networks, pp. 271–282. Springer
215. Yang J, Nguyen MN, San PP, Li XL, Krishnaswamy S (2015) Deep convolutional neural networks on multichannel time series for human activity recognition. In: Twenty-fourth international joint conference on artificial intelligence
216. Yang X, Tian YL (2012) Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, pp. 14–19. IEEE
217. Yang X, Zhang C, Tian YL (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM international conference on multimedia, pp. 1057–1060
218. Yin J, Yang Q, Pan JJ (2008) Sensor-based abnormal human-activity detection. IEEE Trans Knowl Data Eng 20(8):1082–1090
219. Yu G, Liu Z, Yuan J (2014) Discriminative orderlet mining for real-time recognition of human-object interaction. In: Asian conference on computer vision, pp. 50–65. Springer
220. Zanfir M, Leordeanu M, Sminchisescu C (2013) The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2752–2759
221. Zeng M, Nguyen LT, Yu B, Mengshoel OJ, Zhu J, Wu P, Zhang J (2014) Convolutional neural networks for human activity recognition using mobile sensors. In: 6th International conference on mobile computing, applications and services, pp. 197–205. IEEE
222. Zhang B, Xu H, Xiong H, Sun X, Shi L, Fan S, Li J (2020) A spatiotemporal multi-feature extraction framework with space and channel based squeeze-and-excitation blocks for human activity recognition. J Ambient Intell Human Comput 12:7983–7995
223. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning, pp. 7354–7363. PMLR
224. Zhang L, Suganthan PN (2014) Oblique decision tree ensemble via multisurface proximal support vector machine. IEEE Trans Cybern 45(10):2165–2176
225. Zhang L, Varadarajan J, Nagaratnam Suganthan P, Ahuja N, Moulin P (2017) Robust visual tracking using oblique random forests. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5589–5598
226. Zhang L, Wu X, Luo D (2015) Human activity recognition with HMM-DNN model. In: 2015 IEEE 14th International conference on cognitive informatics cognitive computing (ICCI*CC), pp. 192–197
227. Zhang L, Wu X, Luo D (2015) Recognizing human activities from raw accelerometer data using deep neural networks. In: 2015 IEEE 14th International conference on machine learning and applications (ICMLA), pp. 865–870. IEEE
228. Zhang WW, Li W (2019) A deep reinforcement learning based human behavior prediction approach in smart home environments. In: 2019 International conference on robots and intelligent system (ICRIS), pp. 59–62. IEEE
229. Zhao Y, Yang R, Chevalier G, Xu X, Zhang Z (2018) Deep residual bidir-lstm for human activity recognition using wearable sensors. In: Mathematical problems in engineering, 2018
230. Zheng Y, Liu Q, Chen E, Ge Y, Zhao JL (2014) Time series classification using multi-channels deep convolutional neural networks. In: International conference on web-age information management, pp. 298–310. Springer

231. Zhou L, Xu C, Corso JJ (2018) Towards automatic learning of procedures from web instructional videos. In: Thirty-second AAAI conference on artificial intelligence
232. Zhou W, Zhang Z (2014) Human action recognition with multiple-instance Markov model. IEEE Trans Inf Forensics Secur 9(10):1581–1591
233. Zou L, Zheng J, Miao C, Mckeown MJ, Wang ZJ (2017) 3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI. IEEE Access 5:23626–23636

Authors and Affiliations

Farzana Kulsoom[2] · Sanam Narejo[3] · Zahid Mehmood[1] · Hassan Nazeer Chaudhry[4] · butt Aisha[3] · Ali Kashif Bashir[5]

✉ Zahid Mehmood
zahid.mehmood@uettaxila.edu.pk

Farzana Kulsoom
farzana.kulsoom@uettaxila.edu.pk

Sanam Narejo
Sanam.narejo@faculty.muet.edu.pk

Hassan Nazeer Chaudhry
hassannazeer.chaudhry@polimi.it

butt Aisha
aishabutt004@yahoo.com

Ali Kashif Bashir
dr.alikashif.b@ieee.org

[1] Department of Computer Engineering, University of Engineering and Technology, Taxila, Pakistan

[2] Department of Telecommunication Engineering, University of Engineering and Technology, Taxila, Pakistan

[3] Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan

[4] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

[5] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK