Pattern Recognition 108 (2020) 107561

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Sensor-based and vision-based human activity recognition: A comprehensive survey

L. Minh Dang[a], Kyungbok Min[a], Hanxiang Wang[a], Md. Jalil Piran[a], Cheol Hee Lee[b], Hyeonjoon Moon[a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul 143-747(05006), South Korea
[b] Deep inspection, A-1314, Beopwon-ro 11 gil 25, Songpa-gu, Seoul 143-747(05006), South Korea

ARTICLE INFO

Article history:
Received 4 August 2019
Revised 19 May 2020
Accepted 21 July 2020
Available online 23 July 2020

Keywords:
Human activity recognition
Action recognition
Sensors
Vision
Human-centric sensing
Deep learning
Context-awareness

ABSTRACT

Human activity recognition (HAR) technology that analyzes data acquired from various types of sensing devices, including vision sensors and embedded sensors, has motivated the development of various context-aware applications in emerging domains, e.g., the Internet of Things (IoT) and healthcare. Even though a considerable number of HAR surveys and review articles have been conducted previously, the major/overall HAR subject has been ignored, and these studies only focus on particular HAR topics. Therefore, a comprehensive review paper that covers major subjects in HAR is imperative. This survey analyzes the latest state-of-the-art research in HAR in recent years, introduces a classification of HAR methodologies, and shows advantages and weaknesses for methods in each category. Specifically, HAR methods are classified into two main groups, which are sensor-based HAR and vision-based HAR, based on the generated data type. After that, each group is divided into subgroups that perform different procedures, including the data collection, pre-processing methods, feature engineering, and the training process. Moreover, an extensive review regarding the utilization of deep learning in HAR is also conducted. Finally, this paper discusses various challenges in the current HAR topic and offers suggestions for future research.

1. Introduction

During the past decade, sensor technology has achieved exceptional developments in multiple perspectives, including computational power, size, accuracy, and manufacturing costs [1]. These advancements enable a wide range of sensors to be integrated into smartphones and other portable devices to make them smarter and more useful. In addition, the evolution of video surveillance or closed-circuit television (CCTV) technology [2,3] has led to better video quality, more straightforward setup, lower cost, and secure communication. Therefore, an increasing number of applications utilizing CCTV systems for security and monitoring goals have been proposed recently [4,5]. Although each type of sensor aims at specific services and applications, sensors generally collect raw data from their target ubiquitously [6], and general knowledge is acquired by analyzing the collected data.

Human activity recognition or HAR, allows machines to analyze and comprehend several human activities from input data sources, such as sensors, and multimedia contents [3]. The initial work on HAR dates back to the beginning of the 1990s when Foerster et al. showed an accuracy of over 95% for HAR under a trickly controlled data collection [7]. Since then, the fast development of smartphones, wearable devices, and CCTV systems has motivated researchers to improve HAR systems under practical situations. HAR is applied in surveillance systems [8,9], behavior analysis [10], gesture recognition [11–13], patient monitoring systems [14,15], ambient assisted living (AAL) [16,17], and a variety of healthcare systems [18,19] that involve direct interaction or indirect interaction between human and smart devices. For instance, patients with obesity, diabetes, or cardiovascular diseases have to strictly follow a healthy, well-balanced diet and a regular exercise schedule [20]. Hence, tracking daily activities is necessary to give real-time feedback to patients about their progress and provide up-to-date reports to clinicians. Similarly, patients who have declined in mental ability or mental disorders must be monitored continuously to identify unusual actions in time and thus prevent unwanted consequences [21]. Real-time feedback on soldiers' actions, positions, and vital status in tactical situations is fundamental to develop their skills and ensure safety. Moreover, the feedback is also a valuable parameter to assist the commander in giving orders in training as well as combat situations [22]. However, it is considered a challenging research problem because there is no standard

* Corresponding author.
E-mail address: hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.patcog.2020.107561
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

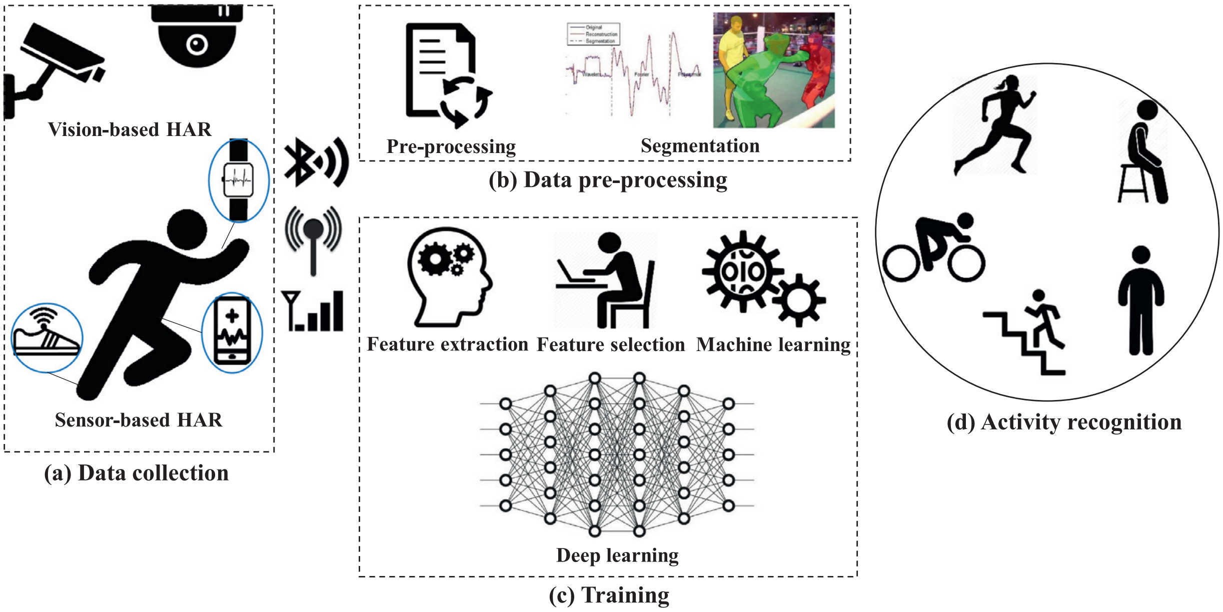


Fig. 1. Human activity recognition framework comprises of four main parts, (a) data collection for vision-based HAR (visual sensing technology) and sensor-based HAR (various types of sensors); (b) data pre-processing, which performs essential pre-processing steps for the collected data; (c) training phase, which utilizes machine learning (ML) or deep learning approaches to learn patterns from the collected data, and (d) activities recognition.

procedure to associate the collected data to a specific action, and it is also technically challenging because of the massive volume of data collected.

As classified previously in several studies, HAR can be divided into two primary approaches: vision-based HAR and sensor-based HAR based on the type of data being processed [23,24], as shown in Fig. 1. The former analyzes images or videos obtained from optical sensors [9,25], whereas the latter concentrates on investigating raw data extracted from wearable sensors and environmental sensors [11,18]. Optical sensors can be differentiated from other types of sensors based on the data type. While wearable sensors generate one-dimensional signal data, optical sensors create 2D images, 3D images, or videos. Wearable devices are representative examples of the sensor-based HAR because they are worn by users to automatically detect and track many activities, such as sitting, jogging, running, and sleeping [26]. However, a sensor does not work when a subject is either beyond its range [27] or performs unidentified actions [28]. On the other hand, CCTV systems have been applied in vision-based HAR systems for a long time [9]. The gestures and activities recognition systems based on analyzing the videos have been intensively studied [15,29]. Furthermore, security, surveillance [30,31], and interactive applications [32,33] are particularly benefited from this topic. During the last few years, a vast majority of research has focused on vision-based HAR, because the vision-based data are affordable and easy to collect compared to the sensor-based data. Thus, this study only covers a small and representative portion of the vision-based HAR studies.

For a long time, machine learning (ML) algorithms, such as random forest (RF) [34], Bayesian networks [35], Markov models [36,37], and support vector machine (SVM) [38,39] have been applied to solve the HAR problem. Under strictly controlled environments and limited input data, traditional ML algorithms have achieved remarkable performance. However, they require multiple pre-processing steps and proper hand-crafted features, which is inefficient and time-consuming [40]. In addition, the use of shallow features leads to poor performance on incremental learning or unsupervised learning [40,41]. In recent years, deep learning has received great interest from the community because deep learning-based research has achieved exceptional performance in various research topics, including object detection and recognition [42], image classification [43], and natural language processing (NLP) [44,45]. Compared to traditional ML algorithms, deep learning considerably reduces the effort of choosing the right features by automatically extracting abstract features through several hidden layers, and the deep learning structure has been proved to work well with unsupervised learning [46,47] and reinforcement learning [42]. Therefore, there is a growing number of deep learning-based HAR frameworks, which have been introduced recently. For example, Oyedotun et al. trained a stacked autoencoder-based convolutional neural network (CNN) on a public ASL static hand gesture dataset containing 24 hand gestures [12]. After the training process, the model achieved a recognition rate of 91.33%, which has proved the potential of deep learning in HAR. In another research, Pigou et al. examined deep learning for HAR in video and introduced a novel deep neural network architecture that incorporated bidirectional recurrence and temporal convolutions [13]. After that, the experimental results showed that the proposed model achieved state-of-the-art results on the publicly available HAR database. Even though many reviews on deep learning [48,49] and HAR [24,50] have been conducted, only a limited number of them cover both topics, and the fact that HAR is still in its development leads to the introduction of many new concepts.

1.1. Real-world applications

The last decade has witnessed significant growth in the number of HAR publications, and each study was dedicated to recognizing specific activity types or behaviors. Table 1 discusses trending HAR domains, including smart homes, healthcare, security and surveillance, autonomous driving, and human-robot interaction. For each area, two representative studies are selected, and the main research contents are discussed in detail.

1.2. Relevant surveys

Table 2 investigates different aspects of HAR and provides detailed contributions from eleven well-known research papers. Between 2012 and 2015, two comprehensive surveys on sensor-based

Table 1
Common research topics that frequently involve HAR. Two related work is described for each topic.

ID	Domain	Ref	Research content
1	Security and surveillance	[30]	• Automatically identifies suspicious or violent activities from the surveillance video. • An alarm is raised to alert the operator when potential suspicious activities happen.
		[31]	• Identifies anomalies in CCTV videos based on deep learning. • Collects and manually validates a huge dataset that contains real-world anomalies in CCTV videos.
2	Healthcare	[51]	• Uses multimodal wearable sensors to perform complex indoor HAR. • Plants various object sensors in the environment to analyze those activities more effectively.
		[17]	• Uses automated feature engineering to improve activity recognition in the AAL application. • Activities can be recognized using only smartphones or smartwatch with high accuracy.
3	Autonomous driving	[52]	• Recognizes the driver's activity automatically by tracking eye and head status. • Proposes a set of novel eye-based and head-based features for the driver's activity classification.
		[53]	• Highlights research issues related to seamless interaction between human agents and automated vehicles. • Discusses scientific tools and frameworks required for the future development of smart vehicles.
4	Human-robot interaction	[32]	• Describes a novel robotic system for multi-sensor signal processing. • Shows promising results in multimodal HAR.
		[54]	• Proposes a model and a new algorithm that can automatically process videos recorded by a robot, while it is interacting with people. • Analyzes and recognizes a set of human emotions during the human-robot interaction.
5	Smart home	[55]	• Proposes a smart agent architecture and recognition mechanism in the smart home environment. • Solves common issues in the smart home system, such as privacy, reusability, applicability and scalability.
		[56]	• Proposes a dynamic segmentation method for the data collected by sensors, which allows near real-time activities classification. • Introduces two new datasets that can reduce the undesired effects from sensor misconfiguration.
6	Entertainment	[57]	• Applies interaction force model to low-level features to detect group activity. • Achieves good performance regardless of the surrounding conditions.
		[33]	• Proposes three full-body 3D gesture recognition algorithms. • Introduces a simple gesture collection game prototype to evaluate the proposed algorithms.

Table 2
Summary of previous HAR reviews, which includes main contributions, publication year, and HAR group (S=Sensor-based HAR, V=Vision-based HAR).

ID	Ref	Year	V/S	Contribution
1	Wang et al. [64]	2019	S	• Reviews deep learning-based HAR models and real-world applications for different kinds of sensors. • Concentrates on applying ML and particularly deep learning for the HAR topic. • Shows public HAR datasets frequently used by researchers. • Discusses some limitations of deep learning-based HAR and proposes practical approaches to overcome these limitations.
2	Abdallah et al. [23]	2018	V	• Surveys two areas of HAR and data stream mining. • Reviews the adaptation capabilities of HAR in the streaming environment. • Categorizes the topic based on feature types. • Shows the advantages and drawbacks of algorithms in each group.
3	Raman et al. [63]	2018	S	• Summarizes common data mining techniques and ML algorithms for HAR. • Summarizes crucial issues and difficulties that exist in previous research. • Discusses recently proposed technologies and new approaches for HAR.
4	Wang et al. [62]	2018	V	• Describes recent research in RGB-D based action detection and classification. • Classifies HAR into four groups based on the sensor types, including visual-based, depth-based, skeleton-based, and the combination of visual and depth camera. • Investigates the strengths and weaknesses of previous systems. • Highlights the importance of spatial temporal-structural data in the video sequence.
5	Morales et al. [61]	2017	S	• Provides an overview of relevant signals, data capture method, and pre-processing process in sensor-based HAR. • Examines standard benchmarks to evaluate HAR. • Repetitive activities, postures, falls, and inactivity was also studied.
6	Cornacchia et al. [28]	2017	S	• A comprehensive review of HAR and classification using wearable sensors. • Reviews many types of sensors, such as accelerometer, gyroscope, pressure sensors, depth-based, and hybrid modality systems. • Categorizes previous works based on the ML algorithms, and whether the sensors data processing is performed on-board or remotely.
7	Herath et al. [24]	2017	V	• Conduct a comprehensive review of general processes for recognizing HAR. • Shows pioneering methods in handcrafted representations and focuses on deep learning-based approaches.
8	Onofri et al. [60]	2016	V	• Focuses on HAR methods in video streams. • Categorizes previous research by the way HAR is handled. • Analyzes the main contributions and shows future research.
9	Wang et al. [58]	2015	S	• Divides radio based-HAR into four categories: ZigBee, WiFi, radio-frequency identification (RFID), and others. • Compares with state-of-the-art research to show the strengths and limitations. • Provides future research directions for the HAR research topic.
10	Tsitsoulis et al. [59]	2013	V	• Reviews existing research and current approaches on vision-based HAR. • Proposes a first-level self-evaluation method that includes a set of important features. • Provides discussion for future research.
11	Lara et al. [27]	2012	S	• Reviews state-of-the-art HAR systems using wearable sensors. • Proposes a two-level taxonomy according to different learning approaches. • Discusses current issues, challenges, and possible solutions. • Evaluates twenty-eight HAR systems regarding classification accuracy, computational complexity, and adaptability.

HAR [27,58] and one review on vision-based HAR [59] were conducted. Radio communication techniques for sensor-based HAR were discussed in [58], whereas a survey on state-of-the-art wearable sensors-based HAR applications and the evaluation of twenty-eight HAR systems on different perspectives were introduced [27]. For the vision-based review [59], the authors evaluate existing research and recent vision-based HAR methods. In 2016, a survey on the trending video streaming techniques for HAR was conducted [60]. One year later, two reviews on sensor-based HAR [28,61] and a study on vision-based HAR [24] were published. The role of different sensor types, which includes an accelerometer, gyroscope, torque, and hybrid sensors, were analyzed in [28], whereas relevant signals, data capture methods, and a pre-processing process were investigated in [61]. In addition, pioneering methods in handcrafted representations and deep learning approaches for vision-based HAR was analyzed [24]. In 2018, recent research in RGB-depth-based (RGB-D) HAR was discussed [62]. Another analysis classified vision-based HAR based on different features in both data streams and activity recognition [23]. In the same year, a survey on standard ML algorithms and data mining techniques applied in HAR was introduced [63]. Most recently, recent deep learning-based HAR models and applications for different kinds of sensors were analyzed [64].

1.3. Contributions

As explained in the previous section, each study analyzes a particular aspect of HAR. Moreover, the introduction of various deep learning-based HAR frameworks has recently added many new hypotheses, procedures, and applications. As a result, a comprehensive survey on recent HAR research is crucial for stakeholders, physicians, and researchers, who are preparing to integrate HAR into existing systems or conducting new HAR research. This survey summarizes previous research, covers many aspects of HAR, such as dataset collection, feature engineering methods, and activity recognition models, and draws an overall evaluation of HAR research.

- A comprehensive review of sensor-based and vision-based HAR.
- Summarize and discuss publicly available datasets that have been used in vision-based HAR and sensor-based HAR.
- Categorize and analyze standard HAR data processing and feature engineering.
- Investigate HAR research that implemented ML algorithms and deep learning.
- Discuss existing challenges and show future directions for HAR research.

1.4. Review techniques

During the preparation of this paper, a critical obstacle was to search and filter the latest HAR literature. Firstly, research papers were downloaded using relevant keywords, such as *action recognition, activity recognition, action feature representation, interaction recognition, activity detection, gesture recognition*, and *action detection*. Secondly, relevant papers from the reference section of initially selected literature were added. After that, research that is related to the HAR datasets was also included. Finally, the overall layout for the manuscript was determined to cover all the selected literature. In addition, standard and representative methods are discussed in detail to help readers gain insights into those methods.

The remainder of this survey is divided into eight sections. In Section 2, background information of human activities, sensor-based HAR, vision-based HAR, and ML are provided. The data collection process for sensor-based HAR and vision-based HAR is explained thoroughly in Section 3. After that, the pre-processing and feature engineering processes are shown in Section 4. In Section 5, ML and deep learning algorithms that are applied to HAR frameworks are discussed. In addition, challenges and future research trends for the HAR topic are provided in Section 6. Finally, the conclusion, which includes the strengths and weaknesses of this survey, is given in Section 7.

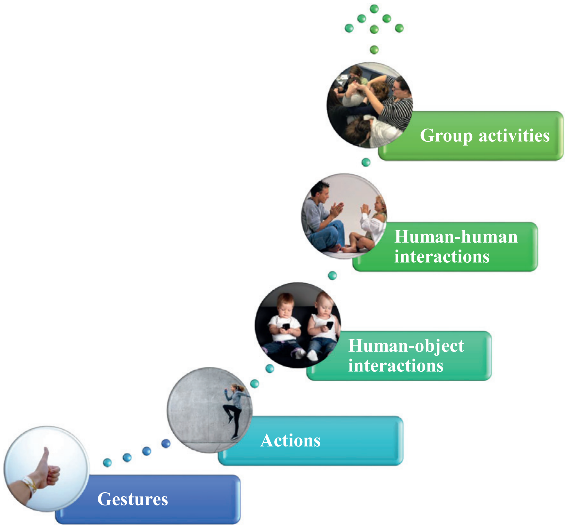


Fig. 2. Five degrees of human activities, including gestures, actions, human-object interactions, human-human interactions, and complex group activities.

2. Background

2.1. Level of human activities

As described in Fig. 2, human activity can be divided into five separate types/levels of activities varying from simple actions like hand gestures to advanced group activities based on the difficulty level and the activity length [65]. The final objective of HAR is to train machines to identify and recognize activities accurately.

The description of each type of activity is shown below, where the human-object interactions and human-human interactions are combined and described as interaction.

- A gesture is a simple hand movement or other parts of the human body to convey an idea or meaning. Facial expressions, hand-waving, and head shake are examples of gestures. A gesture is usually performed within a short period and is the simplest activity among the four groups.
- An action is a simple activity that is carried out by humans and involves several gestures. Examples of action are knocking, swimming, and running.
- An interaction is an activity conducted by two agents. Human is one of the agents, while the other can be an object or also human. Based on the agents' nature, the interaction can be classified into human-object and human-human interaction. Wrestling, hugging, and shaking hands are the human-human interaction examples, whereas an example of the human-object interaction is the interaction between a person and a mobile phone or a laptop.
- A group activity is the most complicated kind of activity that requires over two people and may include interaction with one or many objects. It involves a series of gestures, actions, and interactions. A group study, a football match, and a presentation are group activities examples.

2.2. Sensor-based HAR and vision-based HAR

Assume that a person is doing a predefined set of n activities A, which can be denoted as

$$A = A_0, \ldots, A_{n-1} \tag{1}$$

Multiple sensors are utilized to measure a list of attributes S of k time series within a time interval $I = [t_\alpha, t_\omega]$ to recognize these activities:

$$S = S_0, \ldots, S_{k-1} \tag{2}$$

The goal of HAR is to discover a temporal partition $< I_0, \ldots, I_{r-1} >$ of I, using the list of attributes S, and a collection of classes describing the activities that were conducted within each time partition I_j. The assumption indicates that the time partition I_j is continuous, non-overlapping, and $\bigcup_{j=0}^{r-1} I_j = I$.

Vision-based HAR relies on visual sensing technologies, such as CCTV and camera, to record human activities [2]. Although this approach does not require wearable or smartphone sensors, it relies heavily on image quality. The image resolution, lighting environments, illumination changes, among others, are elements that decide image quality. Since the collected data is a sequence of images or digitalized audio data, computer vision (CV) and audio signal processing technology are usually applied to perform data analysis and recognize the activities. Some primary steps include feature engineering, modelling, segmentation of activities, activity classification, and activity tracking. The latter demands various sensors to perform HAR [1]. The sensor is attached to a human body or the surroundings to collect user's activities continuously. The extracted raw data are mostly time series data and additional environment parameters, which are analyzed using statistical analysis, data fusion, and probabilistic approaches.

2.3. Machine leaning and deep learning-based HAR

For a long time, signal processing technology [66] has been used to analyze raw data collected from sensors. In contrast, CV technology [9] was implemented to preprocess and extract hand-crafted features from images or videos. These techniques could be efficiently applied in feature engineering to create sensor-specific, signal-specific, or domain-specific features. Next, the chosen features were trained using ML algorithms to provide the classification decision. However, the main weakness of the feature engineering technique is that the dataset is analyzed manually to select the suitable features set, and then feature engineering is implemented to extract the features and reduce features space [67]. This lengthy process is compulsory whenever new datasets or new sensors are used, which is complicated and unscalable.

For the past decades, deep learning has become a dominated research topic that has reached human-level performance in various research topics, including HAR. It performs well on big datasets and automatically extract abstract features from sensor signals or a sequence of images. Deep learning has outperformed traditional ML algorithms, which were trained on hand-crafted and domain-specific features [12,68]. As a result, it promotes the new solutions for existing HAR problems, the introduction of larger datasets, and real-time HAR systems. For example, the deep learning-based HAR framework using smartphone sensors proposed by Hassan et al. [69] outperformed all typical multiclass ML algorithms, including SVM and artificial neural network (ANN). In another study, a CNN model was used to extract local features and simple statistical features [70]. The results showed that the proposed framework reached state-of-the-art real-time performance with low computational cost. Finally, human fatigue expression recognition based on deep bimodal learning was introduced in [71]. The mentioned model overcame all previous algorithms with the recognition rate of over 96%.

Table 3 shows the main differences between machine learning and deep learning HAR research on three fundamental processes, including data pre-processing, feature engineering, and learning algorithm.

3. Data collection and benchmark datasets

Data collection, which refers to the acquisition of sensor signals or videos, is a crucial part and foundation of any HAR system. As a result, this section gives a thorough overview of important characteristics of sensor-based data collection and vision-based data collection. Moreover, a detailed description of benchmark datasets for both sensor-based HAR and vision-based HAR is also provided.

3.1. Sensor-based HAR

The sensor-based HAR approach has been applied to various real-world applications, especially smart home and healthcare. In addition, the rapid development of wireless sensor network (WSN) has led to a large amount of data being gathered from different sensors, such as wearable sensors, object sensors, and environmental sensors. Table 4 describes the advantages and disadvantages of three main categories of sensors, common sensor brands, and what they measure.

3.1.1. Wearable sensors

The evolution of the Internet of Things (IoT) and mobile computing in recent years [72] has created a perfect environment for the development of wearable sensors. Wearable sensors are the most prevalent sensor-based HAR. Three standard wearable sensors include accelerometer, magnetometer, and gyroscope, which can be conveniently worn by users [46,70] or integrated into portable devices, such as smartphones, smartwatches, smart bands, glasses, or helmets [73]. Human activities can then be detected by measuring the signal differences before and after an activity. For example, Hegde et al. introduced a wrist-worn sensor, which has an accelerometer and gyroscope. This device performance is examined by identifying a predefined set of activities of daily living (ADL). The obtained result was over 94%, which proved that the proposed device can classify ADL with high accuracy [74].

3.1.2. Object sensors

Object sensors refer to sensors attached to a particular object to identify activities related to that object [75,76]. While wearable sensors measure human activities directly, object sensors detect specific objects movement to infer human activity. For example, an accelerometer can be attached to a smart drinking cup to efficiently analyze the user's drinking habit and notify the user if the daily water intake is insufficient [77]. In another scenario, RFID is usually implemented in the IoT environment [26,78] and healthcare monitoring [79,80] to track and identify human or objects. Object sensors are utilized less often than wearable sensors due to high costs and setup challenges.

3.1.3. Environmental sensors

Different from wearable sensors and object sensors, environmental sensors are usually planted in the surroundings to sense accurate data on fundamental environmental parameters such as humidity, temperature, CO2, and particulate matter (PM2.5). Environmental sensors are used to monitor changes in the environmental parameters when physical activities occur [81]. Since environmental sensors are highly sensitive to the change in the surroundings, the adoption of suitable environmental sensors needs to

Table 3
Comparison of machine learning-based HAR and deep learning-based HAR characteristics.

Process	Machine learning	Deep learning
Data pre-processing	Demands thorough data pre-processing and normalization methods to improve the performance.	Data pre-processing and normalization methods are not required.
Feature engineering	• Relies on manually extracted features. • Depends on applications. • Fails to deal with complicated activities. • Requires feature selection, and dimensionality reduction approaches. • Fails to handle the inter-class variability and inter-class similarity.	• Learns abstract features from raw input data automatically. • Discovers spatial, temporal dependencies and scale invariant features from the input data automatically.
Learning process	• Works well on small training data. • Requires limited computation time and memory usage.	• Requires large dataset to prevent overfitting. • High computational complexity. • Requires specialized hardware to accelerate the training process.

Table 4
Measurement, common brands, advantages, and disadvantages of three primary sensor categories, which include wearable sensors (WS), object sensors (ObS), and environmental sensors (ES).

Category	Sensor	Measurement	Common brands	Advantages	Disadvantages
WS	Accelerometer	Acceleration force (force and gravity)	ActivPAL [20,74], ActiGraph [73,82]	Cheap, durable, high sensitivity, high impedance and high-frequency response.	Hysteresis error, sensitive to temperature, and efficiency decreased over time.
	Gyroscope	Angular velocity	InterSense [22,83], Sparkfun [71]	Fast, lightweight, measures/maintains rotational motion and higher resolution compared to other force or tilt sensors.	Expensive, dependence on the rotation of the earth, subjected to relative azimuth drift, and does not measure linear motion in any direction.
	Magnetometer sensor	Geomagnetic field	InterSense [84]	Cheap, easy to set up, consumes low power, and offers wide magnetic field range.	Sensitive, low precision and cannot be used with magneto torquers.
	Global positioning system (GPS)	Geo-location, velocity, and timing information	Garmin [67,85], Polar [86] and Raveon [87]	Free and directly measures global 3D positioning.	Drains battery and does not work indoors.
ObS	RFID	Radio-frequency	UHF [76,88]	Easy to install, secure, and can store up to 2 KB of data.	Costly, sensitive to external electromagnetic interference, and limited coverage range.
	WiFi	Wireless signal	Wireless router [75,89]	Easy to install, secure, and wireless signal range is better than RFID.	Requires more resources.
ES	Pressure	Pressure	Barometric [90,91]	High signal-to-noise ratio, real-time interface and less user intervention.	Localized sensing, more intrusive, and requires the mold.
	Barometer	Atmospheric pressure	Bosch [92,93]	Measures altitude coordinates and also supports the rapid acquisition of atmospheric pressure.	Low accuracy and easily to be influenced by unfavorable environment conditions.
	Temperature (Thermocouple)	Temperature	Infrared array sensor [86,94]	High-temperature range, direct contact, cheap, and provides fast response.	Corrosion and hard to calibrate.
	Sound	Air pressure	Mouser [95], Sparkfun [94]	Affordable and less user intervention.	Requires more memory and has limited coverage range.
	Radar	Detection of any object within radar's detection cones	Radar sensor [58,96]	Can penetrate mediums/insulators, provides accurate distance, velocity and amplitude information.	Expensive, low spatial resolution, and has limited range.

be carefully planned based on the activities. Roy et al. applied environmental sensors in multi-inhabitant smart environments. Experimental results show high activity classification accuracy as the environmental sensors produced remarkably richer data.

3.1.4. Hybrid sensors

Recently, researchers have increasingly utilized hybrid sensors, a combination of different types of sensors for HAR applications to improve activity recognition rate and model robustness [18,69]. Fig. 3 shows a hybrid sensors framework that can recognize 21 complicated indoor activities [97]. It includes three kinds of sensing contexts, including body sensing (wearable sensors), environmental sensing (environmental sensors), and location-sensing (object sensors). Experimental results showed that the framework reached state-of-the-art indoor activity recognition accuracy of 95%.

Table 5 summarizes ten publicly available datasets, which are usually used for training the sensor-based HAR models. Most of the datasets were collected using wearable sensors, such as accelerometer, gyroscope, and magnetometer. Table 5 also reports the number of participants in the data collection phase, the number of activities, the number of attributes, and the total sam-

Table 5
Detailed description of publicly available datasets for sensor-based HAR, which contains dataset name, number of subjects (#Sub), number of activities (#Act), number of attributes (#Attr), number of data instances (#Ins), sampling rate, year, devices (smartwatch (SW), wearable sensors (WS), and smartphone (SP)), and sensors (accelerometer (A), gyroscope (G), magnetometer (M), and object sensor (O)).

ID	Name	#Sub	#Act	#Attr	#Ins	Sampling rate	Year	Ref	Devices	Sensors
S1	UniMiB SHAR	30	17	NA	11,771	1–32 KHz	2017	[98]	SP	A
S2	Real world	15	8	7	NA	50 Hz	2016	[99]	SP & SW	A
S3	UCI Heterogeneity AR	9	5	16	43,930,257	100–200 Hz	2015	[100]	SP & SW	A, G
S4	HASC	5	6	4	NA	10–100 Hz	2015	[101]	SP	A, G, M, GPS
S5	UCI M-HEALTH	10	12	23	120	50 Hz	2014	[102]	WS	A, G, M
S6	UCI AR-HOP	14	7	9	75,128	NA	2013	[103]	RFID	O
S7	UCI HAR	30	6	561	10,299	50 Hz	2013	[104]	SP	A, G
S8	UCI OPPORTUNITY	4	6	242	2551	NA	2012	[105]	WS	Hybrid
S9	WISDM	29	6	46	5424	20 Hz	2012	[106]	WS	A

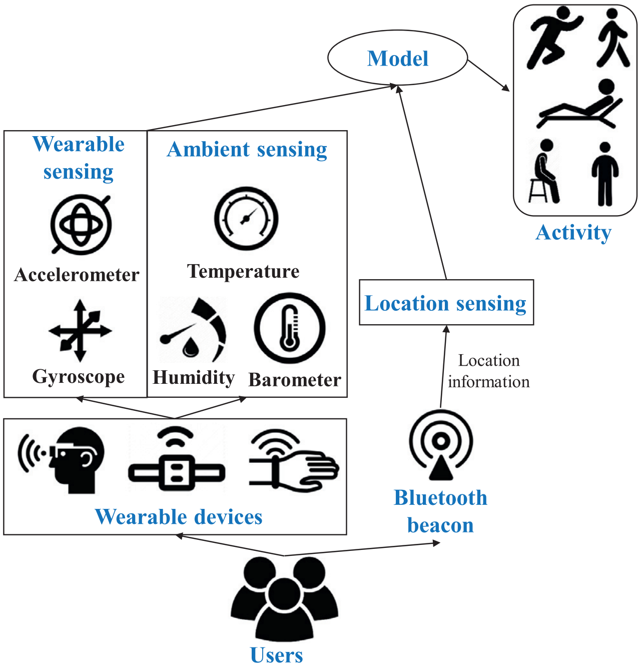


Fig. 3. A hybrid sensor-based indoor HAR framework (WS=Wearable sensors, ObS=Object sensors, ES=Environmental sensors). Object sensors are used to collect location information, whereas multiple wearable sensors and environmental sensors are applied to monitor the activities and sense the surroundings. After that, the data are fed into a model for classifying human activities.

ples. In addition, device, sensor types, and sampling rate are also described.

Among the mentioned datasets, WISDM and UCI datasets (HAR-AAL, HAR-US, and UCI HAR-US) are the standard ones, which are frequently used to evaluate sensor-based HAR models. Recently, they have also been utilized to measure the effectiveness of deep learning-based models. The classification accuracy for the WISDM dataset has exceeded 93%, while the classification accuracy was over 97% for UCI datasets [70].

3.2. Vision-based HAR

In the last few decades, vision-based HAR has become a trending topic because it is applied in various real-world applications. For example, it is essential to implement HAR for CCTV systems in public places to ensure surveillance and security. Vision-based HAR research can be divided based on data type, which includes RGB data [12,107,108] and RGB-D data [9,25]. Generally, vision-based HAR frameworks based on RGB data have achieved lower accuracy compared to the RGB-D data [12,109], because multi-modal data provide extra information and depth channels. However, configuration complexity (complicated settings for each use-case of HAR), computation complexity (big datasets), and high costs are why RGB data are still being used extensively in the current HAR frameworks.

3.2.1. RGB data

An RGB image contains red, green, and blue bands in the visible spectrum, which can be recorded using cameras equipped with a regular complementary metal-oxide-semiconductor (CMOS) sensor. For example, Zerrouki et al. implemented an efficient AdaBoost classifier for HAR using CCTV video [108]. The proposed framework was tested using two publicly available datasets, including the URFDD dataset and the Universidad de Malaga fall detection dataset. The experimental results proved that the model obtained high classification accuracy on RGB datasets.

RGB data are highly available, affordable, and produce rich texture data of the subjects. However, the sensor has a limited range, susceptible to calibration, and relies heavily on the environmental conditions, such as lighting, illumination, and cluttered background.

3.2.2. RGB-D data

Thanks to the development of depth sensors and range imaging techniques [110], scientists can perform HAR more accurately. As shown in Fig. 4, besides the original RGB data, RGB-D cameras also capture depth information, which can help the algorithms to recognize human activities more accurately.

In addition, skeleton data can also be extracted from the depth data to provide a compact rendering of the human body's skeleton, as shown in Fig. 5. Skeleton data have a low-dimensional space[8], which allows HAR models to perform faster. Thus, exploiting the 3D human joint from depth cameras is an attractive research direction because it can be applied in numerous applications.

Cippitelli et al. proposed an effective HAR algorithm based on skeleton data extracted from an RGB-D camera. The proposed model achieved state-of-the-art results in two benchmark datasets, the KARD and CAD-60 [112]. In another research, Jalal et al. introduced a multi-fused features-based online for HAR based on RGB-D sequences recorded by a Kinect device. The multi-fused features contained two main feature types, including depth silhouettes and human skeletons. Extensive investigations on three benchmark depth datasets proved that the introduced method obtained state-of-the-art results [9]. RGB-D data have many advantages compared to RGB data, such as robust against lighting conditions, illumination changes, color and texture change, working well even in the pitch-dark environment, and providing depth data. However, RGB-D data have low resolution, introducing noise to the images due to low sensitivity, and can be easily affected by some materials, such as light-absorbing and transparent materials.

Table 6 presents 20 publicly available datasets, which are regularly used in vision-based HAR. It includes 13 RGB datasets and

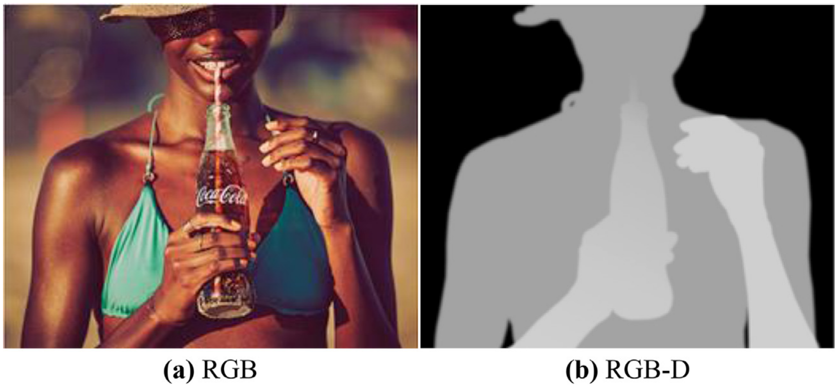


Fig. 4. Comparison of the two types of data, (a) RGB image and (b) RGB-D that contains the corresponding depth data for the RGB. In an RGB-D image, each pixel indicates a distance between the screen space and an object in the RGB image. Bright pixels, which are close to the camera, are having the highest value, and dark pixels, which are far from the camera, are having the lowest values.

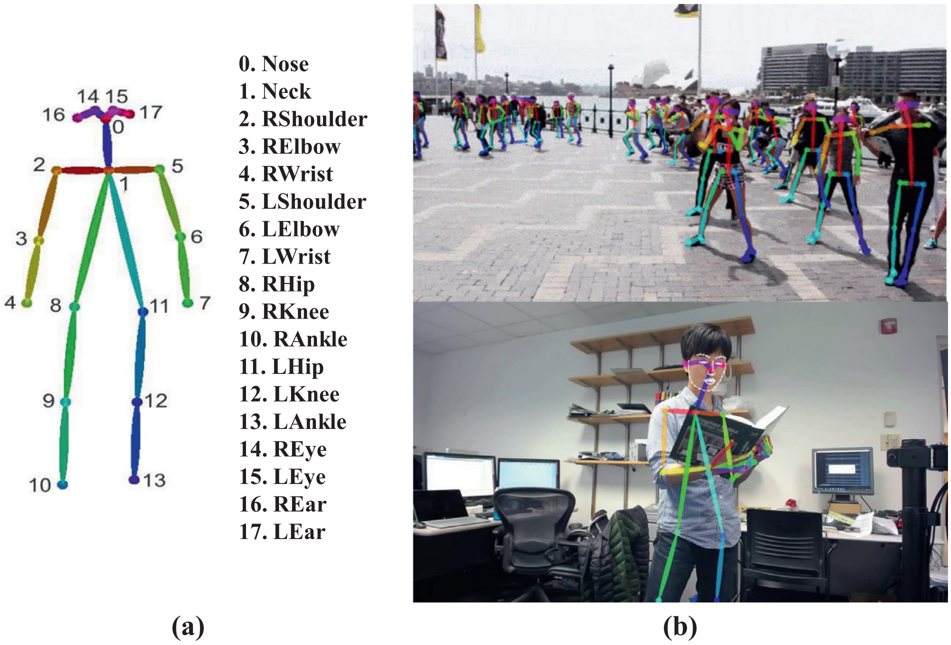


Fig. 5. Skeleton model represented by the OpenPose system [111]. (a) 18 key points that can be identified by the OpenPose framework; (b) Two examples of skeleton detected by the OpenPose.

Table 6
Detailed description of publicly available datasets for vision-based HAR, which includes dataset name, number of activities (#Act), number of videos (#Video), and year.

ID	Name	#Act	#Video	Year	Depth	Ref
V1	HACS	200	1,550,000	2019		[115]
V2	Moments in Time	339	1,000,000	2019		[116]
V3	AVA	80	430	2018		[5]
V4	Kinetics-700	700	650,000	2019	✓	[113]
V5	MultiTHUMOS	65	400	2018		[117]
V6	20BN-something	174	220,847	2017		[118]
V7	Charades-Ego	157	7860	2016		[119]
V8	DALY	10	8133	2016		[120]
V9	ActivityNet 200	200	19,994	2016		[121]
V10	NTU RGB+D	60	56,880	2016	✓	[114]
V11	UTD-MHAD	27	861	2015	✓	[122]
V12	Sports-1M	487	1,100,000	2014		[123]
V13	Berkeley MHAD	11	660	2014	✓	[124]
V14	CAD-120	4	120	2013	✓	[125]
V15	UCF101	101	13,320	2012		[126]
V16	SBU Kinect interaction	7	300	2012	✓	[127]
V17	HMDB51	51	7000	2011		[109]
V18	UT-Interaction	10	180	2010		[128]
V19	Hollywood2	12	3669	2009		[129]
V20	HDM05	70	1500	2007	✓	[130]

7 RGB-D datasets. Most of the previously introduced RGB datasets (HMDB [109] and UCF-101), and RGB-D datasets (Kinetics-700 [113] and NTU RGB+D [114]) contain a small number of activities (from 10 to 60 activities). However, bigger and more challenging datasets have been proposed recently. For example, the Kinetics-700 dataset was introduced in 2017 [113], which has about 700 activities with a total of 650,317 videos. In 2019, two new RGB datasets were proposed, which contained a huge number of videos (HACS with 1,550,000 videos [115] and Moments in time with 1,000,000 videos [116]).

3.3. Discussion

Data collection is a crucial part of every HAR system because the input data determine how the subsequent steps are conducted and can influence the overall performance of HAR frameworks [131]. In this section, the characteristics of standard sensors used in sensor-based HAR and vision-based HAR are presented. Wearable sensors, object sensors, environmental sensors, and hybrid sensors are common types of sensors, which are being used in sensor-based HAR. Among them, wearable sensors have been used the most because they have become cheaper, smaller, so they can be integrated into other devices. In contrast, hybrid sensors are increasingly used in complex activity recognition applications because the combination of multiple sensors improves the model's robustness and performance. On the other hand, vision-based HAR is performed on two data types: RGB and RGB-D data. Although depth information from RGB-D data has proved to improve the HAR performance, low-resolution data and the introduction of noise are the reasons that current HAR frameworks are still using RGB data.

For a long time, a host of carefully collected and validated benchmark datasets has been used to verify HAR models' effectivenesss. Self-collected datasets and publicly available datasets are two common types of datasets. In the self-collected dataset, a dataset [70,132] and an optimal algorithm that works well with that dataset are proposed. However, self-collected datasets require extensive effort during the data collection process because it is time-consuming to collect and validate the dataset manually. On the other hand, benchmark datasets are often used by the researchers to verify the proposed frameworks [98,99]. They are larger than self-collected datasets, already been validated by experts, and have been widely applied in many HAR frameworks. As a result, this section also discusses standard benchmark datasets (9 sensor-based HAR dataset, and 20 vision-based HAR dataset), which have been utilized to verify the effectiveness of HAR frameworks.

4. Pre-processing and feature engineering

This section shows the pre-processing and feature engineering processes that need to be conducted before feeding data into training algorithms. Data pre-processing is considered the most important processes affecting the overall performance of HAR frameworks. After a dataset is collected, it should be pre-processed to minimize noise introduced during the data collection process and the sensors themselves before feeding into an ML algorithm. For sensor-based HAR, the pre-processing process involves both denoising and segmentation, whereas segmentation is the most common pre-processing method implemented for vision-based HAR.

After the pre-processing process, feature engineering, which involves feature extraction and feature selection, are implemented [17]. While feature extraction converts the input into a set of features without losing valuable information, feature selection performs dimensionality reduction on the extracted features to reduce the high-dimensional data and prevent the overfitting problem. Fig. 6 shows standard pre-processing and feature engineering methods for sensor-based HAR and vision-based HAR.

4.1. Sensor-based HAR

After the data collection process, data pre-processing, which is the initial step in the sensor-based HAR pipeline, is implemented. After that, the feature engineering is conducted to choose the most suitable feature set to train ML models.

4.1.1. Pre-processing

Two primary data pre-processing methods for sensor-based HAR are denoising and segmentation.

- Denoising: Sensor data usually contain noise due to miscalibration, malfunction, placement errors, noisy ambient environments, and multiple activities. As a result, data pre-processing techniques have an important role in mitigating the generated noise. Standard denoising methods are low-pass filter, mean filter [133], linear filter [134], wavelet filter, and Kalman filter[135]. Ignatov et al. noticed that noise occurred during the data collection process. Therefore, they implemented the singular value decomposition technique to reduce noise [136]. In other research, the pre-processing process for sensor data was proposed [137]. For each input signal, the authors generated a new signal based on integrating white noise as random noise into the target signal. White noise reduces the noise caused by human's dynamic actions and maintains components that have a low frequency.
- Segmentation: Activity usually lasts for a comparatively long time compared to the sensors sampling rates. Therefore, a single sample extracted from a sensor at a specific time instance does not give sufficient data to identify an activity. As a result, collected signals must be segmented using the segmentation approach instead of relying solely on a sample basis. Data segmentation splits the data stream into various fragments, which then can be mapped into a specific activity [138]. It can be categorized into time-driven windows segmentation, event-driven windows segmentation, and action-driven windows segmentation. Time-driven windows segmentation separates the signal into numerous consecutive windows of fixed-size time intervals, while the event-driven windows method applies estimation methods to separate sensor signals into event-based windows. Finally, action-driven windows segmentation detects the windows where individual activity occurs. Although these methods work well with real-time applications and do not require any pre-processing techniques, they are especially sensitive to the window size. As a result, Hammerla et al. introduced a 1-second sliding window with a 50% overlap to create a HAR dataset that contained approximately 650,000 samples using wearable sensors [139]. The preliminary results showed that the proposed model outperformed previous studies on the collected dataset. On the other hand, an adaptive sliding window segmentation approach for physical HAR based on a tri-axial accelerometer was introduced to solve the weaknesses of fixed-size sliding window methods [140]. The window size is adaptable by analyzing information from the sensor signal.

4.1.2. Feature extraction

Feature extraction is then implemented to extract essential features from the pre-processed data based on distinctive characteristics, such as signal frequency and signal phase. Based on given signal properties, feature extraction methods can be divided into frequency-domain and time-domain approaches. Time-domain methods usually extract median, variance, mean, range, kurtosis, and skewness features [9,141], whereas spectral entropy,

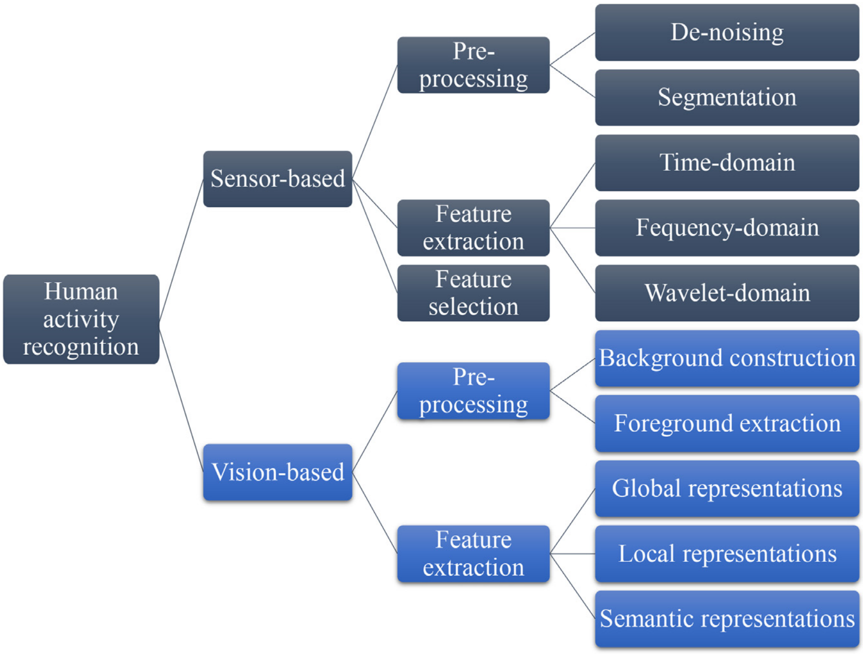


Fig. 6. Standard pre-processing and feature engineering methods for sensor-based HAR and vision-based HAR.

spectral power, peak power, and peak frequency are features belonging to the frequency domain approach [69,70]. The wavelet-domain features approaches have recently gained more attention because wavelets can be used for the feature extraction by calculating wavelet energy in the components (mean-absolute-value, root-mean-square) [142].

- Time-domain approach (TD): In this approach, time-domain features (TDFs) are extracted based on the amplitude variations of signal over time [143]. TD methods can perform real-time computation because no transformation is required on embedded systems. In addition, the amplitude and phase of a signal at any given instance can be analyzed quickly using extracted TDFs. However, signal frequency information is missing in the TD approach.
- Frequency-domain approach (FD): Frequency-domain features (FDFs) are extracted based on the frequency variations of a signal over time. While TD approach shows how a signal changes over time, the FD approach presents how much of a signal remains inside each frequency band over a range of frequencies [144]. Therefore, various transform operators are necessary to convert a signal into a variety of frequencies in the frequency domain approach. The FD methods need enormous computational power, so it is not suitable for low-power wearable devices.
- Wavelet transform (WT): Wavelet transform decomposes a signal into a set of basic functions referred to as wavelets. For example, discrete wavelet transform (DWT) [145], an extension of WT, converts a discrete-time signal into a discrete wavelet representation. Although the higher resolution is obtained when a signal is decomposed into wavelets rather than frequencies [142], WT requires extra computational power and demands a long time to select proper wavelet energy.

4.1.3. Feature selection

Feature selection chooses a subset of features that is important for classification algorithms to make decisions [146]. In addition, it reduces high-dimensional spaces and time complexity by removing irrelevant features. An alternative approach to feature selection is representation learning, where the models concentrate on analyzing the data to extract a good feature set [107].

Wrapper methods [147], filter methods [14], and embedded methods [148] are three main methods of the feature selection process. Filter methods exploit the inherent characteristics of variables/features to select a subset of features using the correlation coefficient to rank the original features. Moreover, the filter-based feature selection does not use any classifier to evaluate the extracted feature subset. Unlike filter methods, wrapper methods have been proved to achieve better performance because many classifiers are applied to judge the performance of the selected subsets [147]. On the other hand, embedded methods choose the best feature subset by calculating the optimal weights of a function that has achieved high results previously. Embedded methods are similar to wrapper approaches and can be used in multiclass and regression problems. Dawn et al. conducted a comprehensive survey of feature selection methods for HAR applications [149]. The authors categorized and analyzed each type of feature selection method and also gave advantages and disadvantages of each feature selection method, which can help readers gain an insight into the feature selection process for HAR.

4.2. Vision-based HAR

Vision-based HAR is a challenging topic because it involves a variety of activities and interactions between subjects [2]. In addition, complex backgrounds, occlusion, viewpoint variations, and lighting conditions make the vision-based HAR even more difficult. Therefore, segmentation and feature engineering processes play a crucial role in improving the performance of HAR applications.

4.2.1. Segmentation

Segmentation is an essential process that extracts the target subjects from a sequence of images or videos. It is classified into background construction and foreground extraction [150]. Background construction-based methods first construct the background information, and then objects of interest are identified by analyzing the difference between the most recent frame and the extracted background [151].

Table 7
Advantages and limitations of two video segmentation approaches (background construction and foreground extraction).

Type	Model	Advantage	Disadvantage	Ref
Background construction	Basic	• Easy to implement • Low computing power	Weak against complicated and multimodal backgrounds	[153,154]
	Statistical	• Works well with multimodal backgrounds • Adaptive parameters	• Poor performance on videos with unstable illumination or lighting condition • Complicated • Gaussian function can impact overall performance	[155–157]
	Fuzzy	Works well with dynamic backgrounds, shadow and complex illumination	• Unable to detect objects that have the same gray level as the background • Manually threshold value	[158–160]
	Neural network	Handles unstable lighting, dynamic backgrounds and bootstrapping issues	Overfitting problem	[151,161,162]
	Others	Fast and easy to implement	Poor performance on dynamic background	[163–166]
Foreground extraction	Optical flow	• Deals with occlusion and distortion • Works well with video recorded by moving camera	Complex and time-consuming	[167,168]
	Temporal information	• Performs on video captured by moving camera, • Easy to implement • Low computing power	Sensitive to noise	[152,169]
	Markov Random Fields	• Preserves boundaries for segmented objects • Works well with complicated backgrounds	High computing power	[150,170]

Table 7 presents the main advantages and limitations of different video segmentation methods. The background construction-based segmentation is highly efficient in tracking fast-moving objects recorded by fixed cameras, requires low computing power, and is simple to deploy. On the other hand, for foreground extraction-based segmentation, human activities are recorded by a pan-tilt-zoom camera or camera mounted on moving objects, such as moving robots, cars, and unmanned aerial vehicles (UAVs) [152]. It is challenging to perform segmentation in this scenario compared to the one recorded by fixed cameras as the background and foreground keep changing. Therefore, foreground extraction-based segmentation is implemented instead of the background construction-based technique. Temporal-information, spatial-information, or spatiotemporal information are analyzed to extract the object from the video, and then subsequent frames are used to obtain feature-based information.

4.2.2. *Feature extraction*

Handcrafted feature-based representation is a traditional feature extraction method that has been used for a long time, and it has produced remarkable results in many HAR applications [29,137]. This approach assumes that each dataset has a representative feature set, enabling a learning model to achieve the best performance. However, it is time-consuming and inefficient because the feature set is selected and validated manually by experts. Three conventional feature-based representation approaches are global feature extraction, local feature extraction, and depth-based feature extraction, which are shown in Table 8.

- Global feature representation approaches collect and encode global descriptors as a set of features directly from videos or a series of images. In addition, background subtraction is applied to localize and extract silhouettes region of interests (ROIs) and shapes of a target subject. Many studies encoded the ROIs using optical flow, corners, edges, or ridges as global descriptors [171,172], while other work [173–175] stacked the silhouettes ROIs alongside the time axis to create the 3D space-time dimensions. Furthermore, DFT [46,176], which extracts FD data from the ROI, is also considered a global features representation method. Global features representation methods were applied in initial HAR research and have slowly become obsolete because of the poor performance on videos with occlusions, noise, and changes in viewpoints.
- Local feature representation methods use local descriptors to represent input images and encode them as single features instead of extracting the shapes or silhouettes. They concentrate on particular local patches, which are exploited by dense sampling or interest point detectors. Histogram of oriented gradients (HOG) is among the basic methods to extract local descriptors by counting gradient orientation occurrences in localized parts of an image. Although the scale-invariant feature transform (SIFT) and HOG are both local feature detectors, SIFT converts an input image into a huge collection of local feature vectors and requires more computing power. As a result, SIFT is unsuitable for real-time applications. An improved approximation of SIFT, the speed-up robust feature (SURF), was proved to run faster than SIFT and preserved the detected points' quality [178]. The shape-based local feature descriptor was initially developed to find matching points between object shapes. It mainly creates a log-polar histogram of edges for each point around a shape. Edge structures near the reference point are sampled with greater detail than structures that are far away because the histogram bins get larger with a growing radius. Most local feature representation methods demonstrate their robustness against partial occlusions and noise.
- Semantic feature representation approaches imitate the human perception of an activity. Perception is an essential element that helps humans perceive and recognize any action based on the visual analysis of body postures obtained from videos. The human perception is also based on additional information, such as scenes or contexts, visual features of the activity, and objects that are usually related to activities [65]. Three main semantic feature representation methods are pose estimation, appearance-based approach, and the 3D approach. The pose estimation relies on deformable part models, such as body part detectors that consist of multi-stage processing [182,183]. On the other hand, appearance-based methods extract visual features that link high-level semantic features to low-level appearance features [107,184]. The appearance-based techniques are divided into two subgroups, local features or global features.

Table 8
Advantages and limitations of the three features representation methods (global features, local features and semantic-based features).

Type	Model	Advantage	Disadvantage	Ref
Global feature	2D silhouette	Obtained easily from the depth data	• Depends heavily on the background model • View-dependent	[171,172]
	Space-time volumes	Does not need background subtraction	Sensitive to noise and occlusion	[173–175]
	Discrete Fourier transform	• Simple implementation • Low computing power	No spatial/temporal localization	[46,176]
Local feature	SIFT	• Invariant to geometric and photometric transformation • Invariant to 3D projection and illumination changes	• High-dimensional features • High computing power	[4,177]
	SURF	Fast and robust	• Patented • Poor performance on high-dimensional data	[2,178]
	HOG	Invariant to geometric and photometric transformations	• Only used for human detection • Subject size can impact the performance	[29,30,179]
	Shape-based features	• Robustness to noise • Similar to human perceptions	Depends on silhouette segmentation	[180,181]
Semantic feature	Pose estimation	Robustness to inter-class variations	Hard to extract the pose accurately	[182,183]
	Appearance-based features	Includes contextual information	Sensitive to intra-class variations	[107,184]
	Depth map	Provides both geometry and visual information	• Occlusion problem • Introduces noise in both spatial and temporal cases	[8,180]

For the 3D semantic feature representation, the depth map data are collected together with the RGB data using the RGB-D camera. After that, human activity analysis can be implemented by operating directly on the depth map data or extract skeleton data and perform joints position detection [8,180].

4.3. Discussion

This section introduces and analyzes the pre-processing and feature engineering processes of a typical HAR framework. The output is an optimal set of features that improves both the model performance and reduces computational complexity. For sensor-based HAR, denoising is applied to reduce noise introduced during the data collection process. After that, the segmentation process is implemented for both vision-based HAR and sensor-based HAR. Based on the applications, the time-domain feature approach, frequency-domain feature approach, or wavelet-domain feature approach is applied to extract distinctive features from the pre-processed signal. On the other hand, the global representation approach, local representation approach, or semantic representation approach is implemented to extract features from vision-based HAR.

5. Learning algorithms

In recent years, the exponential growth of sensors has brought a more efficient and straightforward data collection processes for HAR applications, so the amount of data produced is truly mind-boggling [118]. In order to analyze, recognize patterns, and to acquire general knowledge about a dataset, useful features are extracted from the dataset to feed into ML algorithms.

Generally, ML algorithms are grouped into generative models, discriminative models, and template-based models, as shown in Table 9. For a set of input data X and activity classes Y, the generative approach models the joint probability distribution $Pr(X|Y)$, then derives the posterior distribution $Pr(Y|X)$ to select a correct activity class Y [185]. On the contrary, the discriminative approach uses a conditional probability distribution $Pr(Y|X)$, which indicates the natural distribution to classify a given input X to an activity class Y instantly. Standard generative algorithms are Markov models, Bayesian models, mixture models, and deep belief networks (DBN). In contrast, decision tree, evolutionary algorithms, K-nearest neighbors (KNN), SVM, fuzzy logic, regression, and neural networks are typical discriminative algorithms. In addition, template-based approaches, such as template matching and dynamic time warping, are also used for vision-based HAR.

5.1. Generative model

The generative model is the earliest semi-supervised learning approach, and it has been applied in statistics for a long time [185,186]. Generative approaches output the joint probability $Pr(X|Y)$ based on a set of inputs X and a set of labels Y. Then Bayes rules are used to obtain the posterior distribution $Pr(Y|X)$ to predict the most suitable class y for an input x. The generative model achieves good performance when the training dataset is small [202]. Thus, it is robust and less prone to the overfitting problem. However, it has fewer degrees of freedom compared to the discriminative model.

5.1.1. Bayesian networks

Bayesian networks are probabilistic graphical models that use Bayesian inference to compute the probability [186]. Bayesian networks express conditional dependencies through edges in a directed graph. After that, the probabilistic inferences of random variables in the graph can be efficiently derived through the conditional dependencies. Although Bayesian networks have all advantages of the generative model, attribute independence is its most significant weakness. For example, the Naïve Bayes (NB) model only works when all features under consideration are independent [203]. However, the assumption causes many problems when the features are extracted from acceleration signals and physiological signals because these signals are highly correlated.

A generative framework based on the Bayesian network was proposed to address the structural variabilities of complicated HAR [185]. The authors applied a new process to describe the distinctive

Table 9
Description of three main learning approaches, which includes generative methods, discriminative methods, and others.

Group	Type	Algorithm	Ref
Generative	Bayesian	Bayesian network	[186]
		Naïve bayes	[1,9]
	Markov model	Hidden Markov Model	[9,187]
	Mixture model	Gaussian mixture model	[188]
		K-means	[188]
	Restricted Boltzman Machines	Deep belief network	[189]
		Deep Boltzmann machine	[69]
	Deep autoencoder	Sparse autoencoder	[190,191]
		Denoising autoencoder	[46,192]
Discriminative	Decision tree	C4.5	[193]
		Random forest	[194]
	Random fields	Conditional random field	[195]
	Lazy learning	K-nearest neighbors (KNN)	[1,196]
	Transform domain	Support vector machine (SVM)	[1,4,30]
	Fuzzy logic	Fuzzy inference	[10,142]
	Regression	Logistic regression	[193]
		Adaptive regression	[197]
	Neural network	Multilayer perceptron	[198]
		Convolutional neural network	[68,199]
		Recurrent neural network	[200,201]
Others	Template-based	Template matching	[186]
		Dynamic time warping	[1,9]

configurations of the activity explicitly. In another work, a graph-based HAR framework was demonstrated [35] that combined the hierarchical action semantic dictionary and Bayesian graph model inference. Moreover, the authors applied a recursion-based approach to perform vision-based HAR.

5.1.2. Markov model

A Markov model is a stochastic model that contains states and events represented by transitions and is often applied to temporal and sequential data because it can adequately describe the dependencies of current data with previous data. Hidden Markov model (HMM) belong to statistical Markov models, which presumes that the states of the Markov process are unobservable, and each state emits a discrete random output [37]. A good fine-tune HMM model has been proven to obtain better compression than a simple Markov model, allowing more significant sequences to be detected [204]. Although HMM is a well-known solution to solve the speech recognition problem [204], it has recently been adopted to recognize human activities. For each video, features that represent all pixels in each mesh are extracted from each frame. Next, the extracted features (confusion matrix, the initial probability of hidden states, and the transition matrix) represent each activity are fed into the HMM. However, HMM needs to be trained on a set of sequences and require a bigger seed than simple Markov models. Moreover, there are many possible HMMs for a given set of sequences, and it is challenging to select the best one.

An activity recognition framework for spinal cord injury patients based on augmenting an RF classifier with an HMM static state estimator was introduced [37]. The results showed that the proposed system obtained 88.9%%, which demonstrated a significant improvement compared to when only the RF classifier was used. In addition, a human sensing system based on HMM was proposed to categorize six human actions using data collected from the smartphone accelerometers and gyroscopes [205]. The model showed the best recognition error rate of 2.5% on the proposed dataset.

5.1.3. Mixture model

A mixture model is a probabilistic model representing the appearance of sub-populations in the overall population without knowing the sub-population information [206]. A mixture model correlates with the mixture distribution that depicts the probabilities of occurrence of observations in the entire population.

The Gaussian mixture model (GMM) assumes that all data points are the results of a combination of a finite Gaussian distributions number with unknown parameters [207]. For HAR applications, a separate GMM can be trained for each activity, and the final classification model is based on the GMM that achieved the highest probability. Global minimum convergence is not guaranteed, and the expectation-maximization (EM) algorithm needs to be implemented are two main weaknesses of the GMM [206]. A temporal GMM was introduced by Muaaz et al. to match actions in an unsupervised temporal segmentation and recognition for human motion data [188]. The technique obtained an average accuracy of 72% on merged activities conducted by several participants.

K-means is a particular case of GMM that attempts to divide a given dataset into a fixed number (k) of clusters [208]. Initially, a k number of centroids are randomly selected, and data from the dataset is assigned to the cluster centroids according to the distance, such as Euclidean, until it converges. K-means performs poorly when overlapping clusters occur. A K-mean clustering framework was proposed to differentiate three basic human forearm movements based on a single wearable wrist-worn device [209]. Four healthy participants and four survivors of strokes were involved in a series of experiments, and the obtained results showed that the mentioned framework correctly identified three types of actions with an average accuracy of 88%.

5.1.4. Restricted Boltzmann machines

Restricted Boltzmann machines (RBMs) proposed by Geoff Hinton [210] are the original deep learning model applied to unsupervised learning. This model recognizes data patterns by reconstructing input data with just two layers (the visible and hidden layers). When an input is fed into an RBM model during the forward pass, it is encoded by the visible layer. After that, the visible layer is translated into a set of numbers. On the other hand, RBM takes the translated set of numbers and converts them to the visible layer to restore the inputs.

Since the introduction of RBMs, many deep learning methods for HAR have been introduced. For example, Deep belief network

(DBN) is a generative graphical model and a class of RBM [189]. DBN contains a stack of restricted Boltzmann machines in charge of detecting features, each RBM layer has connections with both the previous and the next layers, but hidden units in each layer have no relationship with each other. The stacked structure allows DBN to be applied in both supervised and unsupervised scenarios. A functional classifier can be created in the supervised applicationsby adding a final softmax layer to the existing DBN structure [189]. In contrast, DBN can be applied to segment unlabeled data in the unsupervised scenario. The HMM can be considered as a reduced DBN, which has fixed graph structures and a restrained number of random variables.

The deep belief network was applied to obtain different features from input data recorded by different sensors and then form a stochastic temporal HAR based on HMM [37]. Moreover, research by Triboan et al. [138] utilizes DBN for real-time feature exploration for HAR. However, the proposed framework was trained locally with initialized parameters, backpropagation, and the classification performed by SoftMax Regression due to the computational complexity of deep learning.

5.1.5. Autoencoder

Autoencoder (AE) has recently become a trending deep learning approach in unsupervised learning [211]. It consists of two symmetrical DBN that has multiple layers representing the encoding half of the net and second set of multiple layers that make up the decoding half. It is trained to learn useful components that can be used to reconstruct the inputs and discard any components that are not typical features. As a result, AE tends to acquire representation in the hidden layer and rejects noise from the input.

Two common extensions of the autoencoder model are sparse autoencoder [46], and denoising autoencoder [191]. Denoising autoencoder was proposed by intentionally adding noise to the inputs, and then the autoencoder is trained to recover the original, nonperturbed signal. The authors found that they could improve the robustness of their internal layers (i.e., latent-space representation) by purposely introducing noise to their signal. Allow the autoencoder's hidden layers to learn more robust filters and prevent it from learning a simple function, and reduce the risk of overfitting in the autoencoder. With a different approach, sparse autoencoder is introduced to deal with sparse and overcomplete features from raw input by customizing the model loss function to deal with sparsity and assign different active units close to zero. It is highly efficient in extracting low dimensional features from complicated and high dimensional data, including videos, images, sensors.

A new HAR approach based on wearable sensors was introduced by Wang [190]. The authors applied continuous autoencoder (CAE) as a new stochastic function, which significantly improved the model classification rate to 99.3%.

5.2. Discriminative model

The discriminative model is considered more straightforward than the generative model, and it has achieved high performance when the training data is sufficient [202]. The discriminative model is usually used in supervised ML by learning a decision boundary using the posterior probability $Pr(Y|X)$ directly from the training data samples. After a model is trained, it can classify unobserved input data X into a class Y. Linear regression and logistic regression are standard classifiers of the discriminative model.

5.2.1. Decision tree

A decision tree is a tree-like algorithm where non-leaf nodes denote attributes or features, branches represent a conjunction of features that lead to the decision, and leaves represent the class label [212]. A decision tree that uses a discrete set of values is referred to as a classification tree. On the other hand, a regression tree is a decision tree that is fed continuous variables.

RF is a supervised ML algorithm that contains many decision trees. The training process gradually improves system performance by applying randomization and bootstrap aggregating methods to generate several decision trees [34]. RF decides a class for a new sample based on a majority vote of all decision trees from the forest. RF can be implemented quickly for both regression and classification problems. However, it requires huge computing power and a large dataset to achieve good performance.

A novel framework to identify important poses for each action window was proposed in [194]. The authors extracted geometrical and temporal features from each action window. After that, they extended RF classifier by applying a differential evolution metaheuristic algorithm to split the tree node. The proposed algorithm was trained in a considerably short time, and it achieved similar accuracy as previous state-of-the-art frameworks.

5.2.2. Conditional random fields

Conditional random fields (CRFs) is a discriminative model that is usually used for labeling and clustering structured data, such as lattices, trees, and sequences [213]. It is based on the conditional distribution concept with an associated graphical structure. The conditional distribution characteristic allows the implicit representation of input variables, enables the extraction of rich and global features from the input data, and prevents the label bias problem.

Liu et al. proposed a coupled hidden CRFs model for HAR by combining RGB and depth information [195]. It improves the original hidden-state CRFs model from one-chain structure to multiple-chain structure (multimodel data). The authors also introduced inference methods and a training model to explore potential associations between depth information and RGB and the temporal context of the model for a particular modality. Experimental results on three benchmark datasets showed that the proposed model outperformed state-of-the-art models by analyzing additional characteristics of both RGB and depth modalities.

5.2.3. Transform domain

The SVM algorithm is a well-known discriminative model and has been extensively applied in HAR. It discovers an optimal hyperplane that separates the training data into two distinct classes and gives a maximum margin between the hyperplane and any sample from the training set [38]. Traditionally, SVM is a linear classifier that is utilized to perform the binary classification problem. However, SVM can also perform a non-linear classification problem by using a kernel trick [214]. Non-linear kernel functions are applied to transform the data in the original dimensional space into a higher-dimensional space. The kernel trick allows the SVM classifier to adapt well to relatively high dimensional data. However, it is challenging to choose an appropriate kernel function and optimal hyper-parameters. Moreover, non-linear SVM requires more training time when the number of training samples in the dataset increase.

Cippitelli et al. used an RGB-D camera to extract the skeleton data [112]. After that, feature vectors are obtained from the skeleton data to train a multi-class SVM model. The experiments conducted on five publicly available datasets showed that the model performed well and overcame state-of-the-art results for both KARD and CAD-60 datasets.

5.2.4. Lazy learning

K-nearest neighbors (KNN) algorithm is a lazy learning approach that can directly classify a new test sample based on existing samples [215]. Euclidean distance [215] is often used to compute the similarity distance between samples. The label for a new sample is decided using a majority vote of its k nearest neighbors based on the Euclidean distance. For example, if k is equal to 1, then the label is set similar to the label of that nearest neighbor. One noticeable limitation is that the computational complexity increases when a new class is assigned, and a new sample is added to the existing dataset.

An optimized KNN classifier, which automatically searches for an optimal k parameter to minimize the misclassification errors of the KNN classifier, was proposed [196]. The authors examined the proposed model on ten standard activity recognition datasets and proved that the proposed model obtained lower error rates than GA and artificial bee colony algorithms.

5.2.5. *Deep learning*

Deep learning is a subset of ML that imitates the network of neurons in the human brain, enabling the computer to deal with complicated problems and reach human-level performance. Through the backpropagation process, it automatically extracts abstract features from input data [216]. Deep learning algorithms work well with large datasets, and they can even work with interconnected and unstructured datasets. The more data deep learning algorithms are trained on, the better the performance they achieve [216].

In the last few years, many kinds of deep learning models have been introduced, and each kind was dedicated to a particular problem with a set of representative datasets. Multilayer perceptron (MLP) [217], CNN [218], and recurrent neural network (RNN) [201] have been confirmed to perform well in various applications. Table 10 discusses the main contributions, datasets, experimental results, field, and year published of various deep learning-based HAR research.

A collection of input and output pairs are usually fed to MLP, so the model can learn to find the dependencies between the input and output pairs. MLP only uses the fully connected layers, where each perceptron connects with every other perceptron. The learning process includes the backpropagation that adjusts the weights, biases, and parameters of the model by finding the minimal error (root mean squared error (RMSE)). Because MLP only contains fully connected layers, the total number of parameters can increase exponentially when the number of layers increases. Another problem of MLP is that it uses flattened vectors as inputs and ignores spatial information [217].

CNN is currently the most used deep learning approach to deal with many practical problems, such as image classifications [218], object detections, and NLP topics [189]. It stacks a list of convolutional layers, pooling layers, fully connected layers, and a function to give the final output. The CNN architecture was constructed to completely utilize the 2D data structure, such as images, speech, or signal. CNN has inbuilt translation invariance because there are robust connections and tied weights between hidden units.

RNN [201] is a subset of deep learning based on the principle of extracting the output of a layer and feeding this back as the input of another layer to predict the output of the current layer. In each time step, RNN considers both the current input and the previously received inputs, because it memorizes previous inputs using the hidden state features. Although RNN is specifically designed to handle sequential data, it suffers from the vanishing/exploding gradient problem. As a result, RNN fails to deal with long sequences if *tanh* is applied as the activation function, whereas the model is unstable if *relu* is used. In addition, RNN layers cannot be stacked into a very deep model because the saturated activation functions make the gradient decay over layers.

5.3. *Other models*

5.3.1. *Template-based approach*

Template matching techniques verify parts of a source image resemble a given template image [223]. There are two primary inputs (a) a source image and (b) a template image. Template matching techniques depend solely on the image, so they are susceptible to background changes, background clutter, scale changes, and illumination. In HAR applications, a template matching technique is applied to recognize similar activities in a source image and template images based on a pixel-by-pixel basis. However, there are some obvious limitations, such as affine variant and highly inefficient computational power due to multiple templates comparison.

A template matching-based sports activities classification framework using data recorded by accelerometer sensors was proposed in [223]. The proposed method showed robust results through various experiments, even with data generated by a new subject. A multi-view HAR model based on temporal template matching was introduced by Kushwaha and Srivastava [224], where activity templates were constructed using spatial pose information. The experimental results on three primary datasets proved that the demonstrated framework was robust and achieved high recognition rates.

Dynamic time warping (DTW) is a robust template matching algorithm based on time series data. It has been extensively applied in speech recognition topics because it can minimize the distortion and shifting effect in time series data [225]. DTW algorithm is usually applied to warp and align segments in time, and it can be adapted and applied in HAR because the activities can be considered a series of keyframes. The DTW method is remarkably effective in analyzing the similarity of time series data using a warping path that can detect identical shapes at different time steps. However, DTW belongs to the template-based approach, so the algorithm complexity increases sharply when the framework processes more activity classes with significant intraclass and interclass variance.

The DTW algorithm was applied to time-phased data and the signal magnitude of an on-body creeping wave to perform HAR [225]. Several experiments showed that a 10-second window for the DTW algorithm brought a good trade-off between model performance and computational efficiency. In another research, the DTW technique was applied to process different shapes of foot movements, which was captured using wearable sensors [21]. The obtained results showed the effectiveness of the proposed method in detecting early signs of Alzheimer's disease.

6. Challenges and future work

Table 11 presents current challenges of both sensor-based HAR and vision-HAR, and discuss possible solutions for each challenge based on recent HAR research.

By discussing the challenges described in Table 11, many HAR future research topics that are interesting to investigate are discussed.

6.1. *Transfer learning*

In recent years, deep learning has taken over HAR research, similar to the current trend in the CV community. However, it is challenging to train a new deep learning-based model from scratch. As a result, the implementation of the HAR model based on previous pre-trained models is a good approach because these models have already obtained objects' spatial relationships. It is interesting to explore some trending topics in transfer learning, such as inflation or domain adaptation [226].

Table 10
Detailed description of various HAR studies grouped by model types, including study field, the proposed model, publication year, devices (A=accelerometer, G=gyroscope, M=magnetometer, I=Inertial, and K=Kinect camera), datasets, experimental results, and main contributions.

Type	Field	S/V	Model	Year	Devices	Dataset	Results	Main contribution
CNN	Indoor HAR	V	5-CNNs [199]	2018	RGB-D	V10, V14 and V16	V10-95.11% and V16-96.67%	• Combines individual CNN classifiers of RGB, depth and skeletal data to classify activities. • Introduces a new processing method for skeleton data.
	Mobile HAR	S	Customized CNN [70]	2018	A	S7 and S10	S7-97.62% and S10-93.32%	• Applies a shallow CNN architecture to extract features automatically. • Combines statistical features and global features of the sensor's time-series data. • Examines the influence of time-series duration on real-time HAR. • Achieves state-of-the-art performance on both WISDM and UCI HAR datasets.
	Intelligent vehicles	V	Customized CNN [219]	2019	K	Self	Detection rate of 91%	• Proposes a customized deep learning model to identify driver's behaviors. • Introduces an unsupervised Gaussian mixture segmentation model to extract the driver's body region from the background.
RNN	HAR	V	RNN tree [200]	2017	NA	Self and V9	V9-0.832%	• Introduces an adaptive framework for fine-grained activity recognition. • Combines multiple RNN models in a tree structure to improve the overall performance. • Uses transfer learning to train the proposed model so that it is more adaptable when a new class is added. • Collects a large-scale dataset for sensor-based HAR.
	Gesture recognition	S	LSTM [220]	S	A, G and M	S8	S9-80%	• Designs an RNN-based HAR to classify six different hand activities. • Collected data from inertial sensors can be fed directly to the proposed model without pre-processing.
	HAR	S	Residual Bidir-LSTM [221]	2018	A, G and M	S7 and S8	S7-93.6% and S8-90%	• Increases learning speed by customizing the LSTM model and window size parameter. • Studies the importance of window size for HAR.

(*continued on next page*)

Table 10 (*continued*)

Type	Field	S/V	Model	Year	Devices	Dataset	Results	Main contribution
AE	HAR	S	Continuous AE [190]	2016	A, G and M	Swiss-roll	98.4%	• Introduces a continuous autoencoder with fast stochastic gradient descent to reduce the training time. • Introduces time and frequency domain feature extraction (TFFE) methods to extract features from sensors effectively.
	Fall detection	S	Customized AE [192]	2017	A, G and M	DLR and COV	Good tradeoff between TPR and FPR	• Trains different auto-encoder models using data collected by various wearable devices to predict fall. • Proposes a threshold tightening method to identify unseen falls accurately.
	HAR	V	AE [211]	2018	NA	V20	Good results on data corrupted by noise	• Proposes a coupled stacked denoising tensor autoencoder (DTAE) for HAR. • Proposed model handles temporal and spatial corruption effectively. • For temporal corruption, each DTAE in stacked DTAE deals with different corruption ratio. • For spatial corruption, each DTAE processes the same temporal corruption ratio but different spatial noise.
DBN	HAR	S	DBN [68]	2018	A, G and M	S5	97.5%	• Proposes a DBN-based HAR framework. • Proposed model achieves higher performance compared to the previous methods.
	HAR	S	DBN [69]	2018	A, G and M	S5	95.85%	• Introduces DBN-based HAR model using smartphone inertial sensors. • Proposed approach outperforms traditional recognition approaches, such as SVM and ANN.
Hybrid	Indoor HAR	S	CNN+LSTM [222]	2016	I and 3-axis A	S8	0.91%	• Proposes a deep learning model that combines convolutional layers and LSTM recurrent layers. • The model is trained instantly on the sensor data with minimum pre-processing processes. • Outperforms previous research on the OPPORTUNITY dataset challenge.
	Video streaming	V	CNN and AE [191]	2019	A, G and M	Self	97.8%	• Designs an efficient and optimized HAR model to process video streaming data. • Applies a fast dynamic frame skipping technique to improve the model speed. • Reaches the state-of-the-art results on real-time CCTV systems.
	Surveillance	V	CNN and LSTM [168]	2018	NA	V15	94.4%	• Applies a CNN-based optical flow model to extract temporal features from CCTVs. • Trains a pre-trained MobileNet model on the extracted saliency features for HAR.

Table 11
Issues and possible solutions for several topics in HAR.

Name	S	V	Issue	Solution	Ref
Unsupervised HAR	✓	✓	• Relies heavily on labeled data • Acquires sufficient training data is tedious and costly	• Crowd-sourcing • Deep transfer learning	[131,227]
Standard benchmarks	✓	✓	• No universally recognized benchmark • Cannot evaluate HAR models effectively	• A uniform protocol (performance metrics and dataset) that allows the quantitative comparison between different approaches	[228]
Activity prediction	✓	✓	• Early prediction is particularly necessary for CCTV systems • Subtle details in human movements need to be captured to predict a future movement • Predicts unfinished activity with limited observations	• Chooses accurate and distinctive features • Proposes a new method for HAR	[229]
Intra-class variation and inter-class similarity	✓	✓	• The same activity can differ between subjects • Different activities may contain similar shapes	• Needs to figure out unique and distinctive features. • Develops a deep learning model that adapts to these challenges • Synthetic data generation	[230]
Multi-subject interactions		✓	• The activities commonly involve the interaction between many people and objects • Detects and tracks several subjects at the same time, such as group activities recognition is challenging	• Spatio-temporal relations among subjects • Develops an appropriate learning model that focuses on differentiating higher-level activities	[57]
Composite activities	✓	✓	• Human activities are usually overlapping and concurrent • The recognition of composite activities brings additional uncertainty	• Recognizes fine-grained activities using hybrid devices	[231]
Non-invasive HAR	✓		• Humans have to adhere to sensor-specific conditions • Uncomfortable	• Non-invasive approach needs to be developed • Development of sensor technology	[232,233]
Real-world videos		✓	• Dynamic backgrounds • Occlusions, illumination variance, and viewpoint changes occur frequently • CCTV systems usually record low-quality videos and occlusions can appear in the recorded videos. • It is even more challenging when the activities are at a long distance.	• Use of multi-sensor systems • Combination of RGB video and depth sensors	[9,113]
Energy and resource constrains	✓	✓	• Both sensor-based and vision-based HAR require real-time sensing, which is energy-consuming • They also require significant computing resources	• Adopts a lower sampling frequency • Consider adaptive segmentation methods	[234,235]

6.2. Interpretable video model

Interpretable image models have been studied extensively in recent years. However, there is limited research on interpretable video models. As explained in [4,183], in a series of frames extracted from a video, there are only some keyframes crucial for the identification of activities. In addition, activities are different in their temporal characteristics. It is possible to identify some activities using frames extracted at the beginning of the video. The interpretability of complicated activities based on the keyframes is a good research topic to answers questions, such as how these frames are organized in the temporal domain, how they contribute to the classification task, and can these frames be selected to train the model faster without affecting the HAR performance. This type of understanding can help researchers develop more efficient HAR frameworks.

6.3. Multimodal data

Humans perceive multimodal data such as image, audio, and text every day, and the multimodal data allow human to understand other types of data. For example, reading enables the reconstruction of the corresponding part of the human's visual sense. Therefore, it is beneficial if the multimodal data are used to interpret complicated activities because multimodal data contain rich semantic knowledge [176,222].

Multimodal data also enables the discovery of long-term temporal relationships between objects from the multimodal data because it can be challenging to extract from multimodal data directly [236]. Long-term temporal relationships can show the sequential order of activities that occur during a prolonged sequence comparable to how the human brain works. When a human remembers something, one sequence evokes the next sequence from a prolonged sequence, like a long-term video. In addition, the communications between entities are also important to understand long-term relationships. For example, pre-defined object interactions happen in a particular activity under specific scene settings. As a result, HAR should analyze both activities and the multimodal data, such as the interpretation of objects, scenes, and temporal relationships of activities. The analysis of multimodal data also supports the long-duration activity prediction.

6.4. Physical aspect of activities

There has been growing interest in researching the physical aspects of activities, such as fine-grained activities. For example, the 20BN-something-something dataset [118] is introduced to stimulate human-object interactions research. The aforementioned mentioned dataset contains label templates or textual descriptions, such as ”Putting something next to something” to define human-object interaction or object-object interaction. The dataset enables the development of systems that comprehend the physical aspects of activities, including human-object interactions and their spatial correlations. Although much information is inferred from the CCTV videos, some physical aspects, such as the movement style, force, and acceleration, are hard to be inferred. Therefore, it is crucial to propose new HAR datasets that include such information.

6.5. Learning actions without labels

For an increasing number of huge activity recognition datasets, such as something-something [118] and sports-1M [123], the manual labeling method is inefficient and costly. Although automatic labeling using search engines and video subtitles [109] is achievable in specific areas, it still needs to be validated manually. Crowdsourcing [131] is a better choice. However, it is challenging because of the label diversity problem, which can produce incorrect activity labels. As a result, researchers need to introduce a more effective and robust HAR method that automatically processes unlabeled data [116].

7. Conclusion

A comprehensive survey of state-of-the-art methods, along with their pros and cons for vision-based HAR and sensor-based HAR has been provided in this paper. These methods have become particularly influential in recent decades thanks to their potential integration in emerging activity recognition applications. The complete descriptions, analyses, and highlights of their features help researchers gain general knowledge in the field of activity recognition.

We covered several perspectives of existing work, including handcrafted feature designs, models, deep architectures, datasets, and evaluation protocols. We emphasized up-to-date development in both sensor-based HAR and vision-based HAR. Different datasets were investigated by reflecting the main requirements of activity recognition applications: real-time operation with limited onboard computational resources and constrained observational conditions (e.g., limited camera resolution). The characteristics, strengths, and weaknesses of traditional machine learning and deep learning models used in HAR were also analyzed. In addition, the review also addressed challenges in the HAR topic and possible solutions for these challenges.

Aside from activity recognition widespread applications in pattern recognition and image processing, there exist various problems for future research, such as activity tracking, system design, and speed. This survey is expected to encourage further research in the activity recognition area.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation).

References

[1] Y. Liu, L. Nie, L. Liu, D.S. Rosenblum, From action to activity: sensor-based activity recognition, Neurocomputing 181 (2016) 108–115.

[2] A. Bux, P. Angelov, Z. Habib, Vision based human activity recognition: a review, in: Advances in Computational Intelligence Systems, Springer, 2017, pp. 341–371.

[3] L.M. Dang, S.I. Hassan, S. Im, H. Moon, Face image manipulation detection based on a convolutional neural network, Expert Syst. Appl. 129 (2019) 156–168.

[4] B. Jagadeesh, C.M. Patil, Video based human activity detection, recognition and classification of actions using SVM, Trans. Mach. Learn. Artif.Intell. 6 (6) (2019) 22.

[5] C. Gu, C. Sun, D.A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al., AVA: A video dataset of spatio-temporally localized atomic visual actions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6047–6056.

[6] J.A. Muñoz-Cristóbal, M.J. Rodríguez-Triana, V. Gallego-Lema, H.F. Arribas-Cubero, J.I. Asensio-Pérez, A. Martínez-Monés, Monitoring for awareness and reflection in ubiquitous learning environments, Int. J. Hum.–Comput.Interact. 34 (2) (2018) 146–165.

[7] F. Foerster, M. Smeja, Joint amplitude and frequency analysis of tremor activity., Electromyogr. Clin. Neurophysiol. 39 (1) (1999) 11–19.

[8] X. Ji, J. Cheng, W. Feng, D. Tao, Skeleton embedded motion body partition for human action recognition using depth sequences, Signal Process. 143 (2018) 56–68.

[9] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, D. Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, Pattern Recognit. 61 (2017) 295–308.

[10] G. Batchuluun, J.H. Kim, H.G. Hong, J.K. Kang, K.R. Park, Fuzzy system based human behavior recognition by combining behavior prediction and recognition, Expert Syst. Appl. 81 (2017) 108–133.

[11] C. Xu, L.N. Govindarajan, L. Cheng, Hand action detection from ego-centric depth sequences with error-correcting hough transform, Pattern Recognit. 72 (2017) 494–503.

[12] O.K. Oyedotun, A. Khashman, Deep learning in vision-based static hand gesture recognition, Neural Comput. Appl. 28 (12) (2017) 3941–3951.

[13] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, J. Dambre, Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video, Int. J. Comput. Vis. 126 (2–4) (2018) 430–439.

[14] N.A. Capela, E.D. Lemaire, N. Baddour, Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients, PLoS ONE 10 (4) (2015) e0124414.

[15] A. Prati, C. Shan, K.I.-K. Wang, Sensors, vision and networks: from video surveillance to activity recognition and health monitoring, J. Ambient Intell. Smart Environ. 11 (1) (2019) 5–22.

[16] S. Sankar, P. Srinivasan, R. Saravanakumar, Internet of things based ambient assisted living for elderly people health monitoring, Res. J. Pharm. Technol. 11 (9) (2018) 3900–3904.

[17] E. Zdravevski, P. Lameski, V. Trajkovik, A. Kulakov, I. Chorbev, R. Goleva, N. Pombo, N. Garcia, Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering, IEEE Access 5 (2017) 5262–5280.

[18] J. Qi, P. Yang, M. Hanneghan, S. Tang, B. Zhou, A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors, IEEE Internet Things J. 6 (2) (2018) 1384–1393.

[19] C. Aviles-Cruz, E. Rodriguez-Martinez, J. Villegas-Cortez, A. Ferreyra-Ramirez, Granger-causality: an efficient single user movement recognition using a smartphone accelerometer sensor, Pattern Recognit. Lett. 125 (2019) 576–583.

[20] G. Plasqui, Smart approaches for assessing free-living energy expenditure following identification of types of physical activity, Obes. Rev. 18 (2017) 50–55.

[21] R. Varatharajan, G. Manogaran, M. Priyan, R. Sundarasekar, Wearable sensor devices for early detection of alzheimer disease using dynamic time warping algorithm, Cluster Comput. 21 (1) (2018) 681–690.

[22] E.E. Cust, A.J. Sweeting, K. Ball, S. Robertson, Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance, J. Sports Sci. 37 (5) (2019) 568–600.

[23] Z.S. Abdallah, M.M. Gaber, B. Srinivasan, S. Krishnaswamy, Activity recognition with evolving data streams: areview, ACM Comput. Surv. (CSUR) 51 (4) (2018) 71.

[24] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: a survey, Image Vis. Comput. 60 (2017) 4–21.

[25] X. Yang, Y. Tian, Super normal vector for human activity recognition with depth cameras, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2017) 1028–1039.

[26] B. Alsinglawi, Q.V. Nguyen, U. Gunawardana, A. Maeder, S.J. Simoff, Rfid systems in healthcare settings and activity of daily living in smart homes: a review, E-Health Telecommun. Syst. Netw. 6 (2017) 1–17.

[27] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutor. 15 (3) (2012) 1192–1209.

[28] M. Cornacchia, K. Ozcan, Y. Zheng, S. Velipasalar, A survey on activity detection and classification using wearable sensors, IEEE Sens. J. 17 (2) (2017) 386–403.

[29] K.S. Kumar, R. Bhavani, Human activity recognition in egocentric video using hog, gist and color features, Multimed. Tools Appl. (2018) 1–17.

[30] P.K. Roy, H. Om, Suspicious and violent activity detection of humans using hog features and SVM classifier in surveillance videos, in: Advances in Soft Computing and Machine Learning in Image Processing, Springer, 2018, pp. 277–294.

[31] A. Thyagarajmurthy, M. Ninad, B. Rakesh, S. Niranjan, B. Manvi, Anomaly detection in surveillance video using pose estimation, in: Emerging Research in Electronics, Computer Science and Technology, Springer, 2019, pp. 753–766.

[32] L. Martínez-Villaseñor, H. Ponce, A concise review on sensor signal acquisition and transformation applied to human activity recognition and human–robot interaction, International Journal of Distributed Sensor Networks 15 (6) (2019). 1550147719853987

[33] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, S.J. Maybank, Asymmetric 3d convolutional neural networks for action recognition, Pattern Recognit. 85 (2019) 1–12.

[34] C. Hu, Y. Chen, L. Hu, X. Peng, A novel random forests based class incremental learning method for activity recognition, Pattern Recognit. 78 (2018) 277–290.

[35] Q. Xiao, R. Song, Action recognition based on hierarchical dynamic bayesian network, Multimed. Tools Appl. 77 (6) (2018) 6955–6968.

[36] C.A. Ronao, S.-B. Cho, Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models, International Journal of Distributed Sensor Networks 13 (1) (2017). 1550147716683687

[37] P. Sok, T. Xiao, Y. Azeze, A. Jayaraman, M.V. Albert, Activity recognition for incomplete spinal cord injury subjects using hidden Markov models, IEEE Sens. J. 18 (15) (2018) 6369–6374.

[38] B.M. Abidine, L. Fergani, B. Fergani, M. Oussalah, The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition, Pattern Anal. Appl. 21 (1) (2018) 119–138.

[39] Z. Chen, Y. Wang, Infrared–ultrasonic sensor fusion for support vector machine–based fall detection, J. Intell. Mater. Syst. Struct. 29 (9) (2018) 2027–2039.

[40] I. Portugal, P. Alencar, D. Cowan, The use of machine learning algorithms in recommender systems: a systematic review, Expert Syst. Appl. 97 (2018) 205–227.

[41] T.N. Nguyen, H. Nguyen-Xuan, J. Lee, A novel data-driven nonlinear solver for solid mechanics using time series forecasting, Finite Elem. Anal. Des. 171 (2020) 103377.

[42] E.P. Ijjina, K.M. Chalavadi, Human action recognition in RGB-D videos using motion sequence information and deep learning, Pattern Recognit. 72 (2017) 504–516.

[43] T.-H. Tan, M. Gochoo, S.-C. Huang, Y.-H. Liu, S.-H. Liu, Y.-F. Huang, Multi--resident activity recognition in a smart home using RGB activity image and DCNN, IEEE Sens. J. 18 (23) (2018) 9718–9727.

[44] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput Intell Mag 13 (3) (2018) 55–75.

[45] A. Angeleas, N. Bourbakis, A two formal languages based model for representing human activities, in: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2016, pp. 779–783.

[46] M.S. Seyfioğlu, A.M. Özbayoğlu, S.Z. Gürbüz, Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities, IEEE Trans. Aerosp. Electron. Syst. 54 (4) (2018) 1709–1723.

[47] T.N. Nguyen, S. Lee, H. Nguyen-Xuan, J. Lee, A novel analysis-prediction approach for geometrically nonlinear problems using group method of data handling, Comput. Methods Appl. Mech. Eng. 354 (2019) 506–526.

[48] Q. Zhang, L.T. Yang, Z. Chen, P. Li, A survey on deep learning for big data, Inform. Fusion 42 (2018) 146–157.

[49] M. Mohammadi, A. Al-Fuqaha, S. Sorour, M. Guizani, Deep learning for IoT big data and streaming analytics: a survey, IEEE Commun. Surv. Tutor. 20 (4) (2018) 2923–2960.

[50] C. Chen, R. Jafari, N. Kehtarnavaz, A survey of depth and inertial sensor fusion for human action recognition, Multimed. Tools Appl. 76 (3) (2017) 4405–4425.

[51] M. Ehatisham-Ul-Haq, A. Javed, M.A. Azam, H.M. Malik, A. Irtaza, I.H. Lee, M.T. Mahmood, Robust human activity recognition using multimodal feature-level fusion, IEEE Access 7 (2019) 60736–60751.

[52] T. Billah, S.M. Rahman, M.O. Ahmad, M. Swamy, Recognizing distractions for assistive driving by tracking body parts, IEEE Trans. Circuits Syst. Video Technol. 29 (4) (2018) 1048–1062.

[53] E. Ohn-Bar, M.M. Trivedi, Looking at humans in the age of self-driving and highly automated vehicles, IEEE Trans. Intell. Veh. 1 (1) (2016) 90–104.

[54] R. Mojarad, F. Attal, A. Chibani, S.R. Fiorini, Y. Amirat, Hybrid approach for human activity recognition by ubiquitous robots, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 5660–5665.

[55] J. Rafferty, C.D. Nugent, J. Liu, L. Chen, From activity recognition to intention recognition for assisted living within smart homes, IEEE Trans. Hum. Mach. Syst. 47 (3) (2017) 368–379.

[56] J. Wan, M.J. O'grady, G.M. O'hare, Dynamic sensor event segmentation for real-time activity recognition in a smart home context, Pers. Ubiquitous Comput. 19 (2) (2015) 287–301.

[57] C. Wateosot, N. Suvonvorn, Group activity recognition with an interaction force based on low-level features, IEEJ Trans. Electr. Electron. Eng. 14 (7) (2019) 1061–1073.

[58] S. Wang, G. Zhou, A review on radio based activity recognition, Digit. Commun. Netw. 1 (1) (2015) 20–29.

[59] A. Tsitsoulis, N. Bourbakis, A first stage comparative survey on vision-based human activity recognition, J. AI Tools 24 (6) (2013).

[60] L. Onofri, P. Soda, M. Pechenizkiy, G. Iannello, A survey on using domain and contextual knowledge for human activity recognition in video streams, Expert Syst. Appl. 63 (2016) 97–111.

[61] J. Morales, D. Akopian, Physical activity recognition by smartphones, a survey, Biocybern. Biomed. Eng. 37 (3) (2017) 388–400.

[62] P. Wang, W. Li, P. Ogunbona, J. Wan, S. Escalera, RGB-D-based human motion recognition with deep learning: a survey, Comput. Vis. Image Underst. 171 (2018) 118–139.

[63] S. Ramasamy Ramamurthy, N. Roy, Recent trends in machine learning for human activity recognition–a survey, Wiley Interdiscip. Rev: Data Min. Knowl. Discov. 8 (4) (2018) e1254.

[64] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: a survey, Pattern Recognit. Lett. 119 (2019) 3–11.

[65] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: a literature review, Pattern Recognit. 48 (8) (2015) 2329–2345.

[66] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, D. Anguita, Transition-aware human activity recognition using smartphones, Neurocomputing 171 (2016) 754–767.

[67] J. Ye, G. Qi, N. Zhuang, H. Hu, K.A. Hua, Learning compact features for human activity recognition via probabilistic first-take-all, IEEE Trans. Pattern Anal. Mach. Intell. 42 (1) (2018) 126–139.

[68] M.M. Hassan, S. Huda, M.Z. Uddin, A. Almogren, M. Alrubaian, Human activity recognition from body sensor data using deep learning, J. Med. Syst. 42 (6) (2018) 99.

[69] M.M. Hassan, M.Z. Uddin, A. Mohamed, A. Almogren, A robust human activity recognition system using smartphone sensors and deep learning, Future Gener. Comput. Syst. 81 (2018) 307–313.

[70] A. Ignatov, Real-time human activity recognition from accelerometer data using convolutional neural networks, Appl. Soft Comput. 62 (2018) 915–922.

[71] L. Zhao, Z. Wang, X. Wang, Y. Qi, Q. Liu, G. Zhang, Human fatigue expression recognition through image-based dynamic multi-information and bimodal deep learning, J. Electron. Imaging 25 (5) (2016) 053024.

[72] L.M. Dang, M. Piran, D. Han, K. Min, H. Moon, et al., A survey on internet of things and cloud computing for healthcare, Electronics 8 (7) (2019) 768.

[73] R. Yao, G. Lin, Q. Shi, D.C. Ranasinghe, Efficient dense labelling of human activity sequences from wearables using fully convolutional networks, Pattern Recognit. 78 (2018) 252–266.

[74] N. Hegde, M. Bries, T. Swibas, E. Melanson, E. Sazonov, Automatic recognition of activities of daily living utilizing insole-based and wrist-worn wearable sensors, IEEE J. Biomed. Health Inform. 22 (4) (2018) 979–988.

[75] W. Wang, A.X. Liu, M. Shahzad, K. Ling, S. Lu, Device-free human activity recognition using commercial WiFi devices, IEEE J. Sel. Areas Commun. 35 (5) (2017) 1118–1131.

[76] W. Ruan, Q.Z. Sheng, L. Yao, X. Li, N.J. Falkner, L. Yang, Device-free human localization and tracking with UHF passive RFID tags: a data-driven approach, J. Netw. Comput. Appl. 104 (2018) 78–96.

[77] L. Roland, L. Lidauer, G. Sattlecker, F. Kickinger, W. Auer, V. Sturm, D. Efrosinin, M. Drillich, M. Iwersen, Monitoring drinking behavior in bucket-fed dairy calves using an ear-attached tri-axial accelerometer: a pilot study, Comput. Electron. Agric. 145 (2018) 298–301.

[78] X. Fan, F. Wang, F. Wang, W. Gong, J. Liu, When RFID meets deep learning: exploring cognitive intelligence for activity identification, IEEE Wirel. Commun. (2019) 2.

[79] J. Qi, P. Yang, A. Waraich, Z. Deng, Y. Zhao, Y. Yang, Examining sensor-based physical activity recognition and monitoring for healthcare using internet of things: a systematic review, J. Biomed. Inform. 87 (2018) (2018) 138–153.

[80] J. Hao, A. Bouzouane, S. Gaboury, Recognizing multi-resident activities in non-intrusive sensor-based smart homes by formal concept analysis, Neurocomputing 318 (2018) 75–89.

[81] N. Roy, A. Misra, D. Cook, Ambient and smartphone sensor assisted ADL recognition in multi-inhabitant smart environments, J. Ambient Intell. Humaniz. Comput. 7 (1) (2016) 1–19.

[82] Y. Athavale, S. Krishnan, A device-independent efficient actigraphy signal-encoding system for applications in monitoring daily human activities and health, Sensors 18 (9) (2018) 2966.

[83] M. Westhoven, C. Plegge, T. Henrich, T. Alexander, Posture based recognition of the visual focus of attention for adaptive mobile information systems, in: International Conference on Augmented Cognition, Springer, 2016, pp. 416–427.

[84] C. Xu, J. He, X. Zhang, C. Yao, P.-H. Tseng, Geometrical kinematic modeling on human motion using method of multi-sensor fusion, Inf. Fusion 41 (2018) 243–254.

[85] P. Davidson, H. Virekunnas, D. Sharma, R. Piché, N. Cronin, Continuous analysis of running mechanics by means of an integrated INS/GPS device, Sensors 19 (6) (2019) 1480.

[86] X. Luo, Q. Guan, H. Tan, L. Gao, Z. Wang, X. Luo, Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors, Sensors 17 (8) (2017) 1738.

[87] A.M. Wempe, R.F. Keefe, S.M. Newman, T.B. Paveglio, Intent to adopt location sharing for logging safety applications, Safety 5 (1) (2019) 7.

[88] A. Jayatilaka, D.C. Ranasinghe, Real-time fluid intake gesture recognition based on batteryless UHF RFID technology, Pervasive Mob. Comput. 34 (2017) 146–156.

[89] Y. Gu, F. Ren, J. Li, PAWS: Passive human activity recognition based on WiFi ambient signals, IEEE Internet Things J. 3 (5) (2016) 796–805.

[90] D. Rodríguez-Martín, A. Samà, C. Pérez-López, A. Català, J. Cabestany, Posture transition analysis with barometers: contribution to accelerometer-based algorithms, Neural Comput. Appl. (2018) 1–15.

[91] F. Masse, R. Gonzenbach, A. Paraschiv-Ionescu, A.R. Luft, K. Aminian, Wearable barometric pressure sensor to improve postural transition recognition of mobility-impaired stroke patients, IEEE Trans. Neural Syst. Rehabil. Eng. 24 (11) (2016) 1210–1217.

[92] M. Gochoo, T.-H. Tan, S.-H. Liu, F.-R. Jean, F.S. Alnajjar, S.-C. Huang, Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN, IEEE J. Biomed. Health Inform. 23 (2) (2018) 693–702.

[93] S. Chernbumroong, S. Cang, H. Yu, Genetic algorithm-based classifiers fusion for multisensor activity recognition of elderly people, IEEE J. Biomed. Health Inform. 19 (1) (2015) 282–289.

[94] C. Galván-Tejada, F. López-Monteagudo, O. Alonso-González, J. Galván-Tejada, J. Celaya-Padilla, H. Gamboa-Rosales, R. Magallanes-Quintanar, L. Zanella–Calzada, A generalized model for indoor location estimation using environmental sound from human activity recognition, ISPRS Int. J. Geoinf. 7 (3) (2018) 81.

[95] J. Kang, J. Kim, K. Kim, M. Sohn, Complex activity recognition using polyphonic sound event detection, in: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Springer, 2018, pp. 675–684.

[96] W. Li, B. Tan, Y. Xu, R.J. Piechocki, Log-likelihood clustering-enabled passive RF sensing for residential activity recognition, IEEE Sens. J. 18 (13) (2018) 5413–5421.

[97] P. Bharti, D. De, S. Chellappan, S.K. Das, HuMAn: complex activity recognition with multi-modal multi-positional body sensing, IEEE Trans. Mob. Comput. 18 (4) (2019) 857–870.

[98] D. Micucci, M. Mobilio, P. Napoletano, UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones, Appl. Sci. 7 (10) (2017) 1101.

[99] T. Sztyler, H. Stuckenschmidt, On-body localization of wearable devices: an investigation of position-aware activity recognition, in: 2016 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, 2016, pp. 1–9.

[100] A. Stisen, H. Blunck, S. Bhattacharya, T.S. Prentow, M.B. Kjærgaard, A. Dey, T. Sonne, M.M. Jensen, Smart devices are different: assessing and mitigatingmobile sensing heterogeneities for activity recognition, in: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, ACM, 2015, pp. 127–140.

[101] T. Hayashi, M. Nishida, N. Kitaoka, K. Takeda, Daily activity recognition based on DNN using environmental sound and acceleration signals, in: 2015 23rd European Signal Processing Conference (EUSIPCO), IEEE, 2015, pp. 2306–2310.

[102] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J.A. Holgado-Terriza, S. Lee, H. Pomares, I. Rojas, Design, implementation and validation of a novel open framework for agile development of mobile health applications, Biomed. Eng. Online 14 (2) (2015) S6.

[103] R.L.S. Torres, D.C. Ranasinghe, Q. Shi, A.P. Sample, Sensor enabled wearable RFID technology for mitigating the risk of falls near beds, in: 2013 IEEE International Conference on RFID (RFID), IEEE, 2013, pp. 191–198.

[104] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, A public domain dataset for human activity recognition using smartphones., in: 2013 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (Esann), 2013, pp. 24–26.

[105] R. Chavarriaga, H. Sagha, A. Calatroni, S.T. Digumarti, G. Tröster, J.d.R. Millán, D. Roggen, The opportunity challenge: a benchmark database for on–body sensor-based activity recognition, Pattern Recognit. Lett. 34 (15) (2013) 2033–2042.

[106] J.R. Kwapisz, G.M. Weiss, S.A. Moore, Activity recognition using cell phone accelerometers, ACM SigKDD Explor. Newsl. 12 (2) (2011) 74–82.

[107] D. Xu, Y. Yan, E. Ricci, N. Sebe, Detecting anomalous events in videos by learning deep representations of appearance and motion, Comput. Vis. Image Underst. 156 (2017) 117–127.

[108] N. Zerrouki, F. Harrou, Y. Sun, A. Houacine, Vision-based human action classification using adaptive boosting algorithm, IEEE Sens. J. 18 (12) (2018) 5115–5121.

[109] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2556–2563.

[110] J.C. Nunez, R. Cabido, J.J. Pantrigo, A.S. Montemayor, J.F. Velez, Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition, Pattern Recognit. 76 (2018) 80–94.

[111] S. Ghazal, U.S. Khan, M.M. Saleem, N. Rashid, J. Iqbal, Human activity recognition using 2d skeleton data and supervised machine learning, IET Image Proc. 13 (13) (2019) 2572–2578.

[112] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, A human activity recognition system using skeleton data from RGBD sensors, Comput. Intell. Neurosci. 2016 (2016) 21.

[113] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset, arXiv:1907.06987(2019).

[114] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+D: A large scale dataset for 3d human activity analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[115] H. Zhao, A. Torralba, L. Torresani, Z. Yan, HACS: Human action clips and segments dataset for recognition and temporal localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8668–8678.

[116] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S.A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, et al., Moments in time dataset: one million videos for event understanding, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1–8, doi:10.1109/TPAMI.2019.2901464.

[117] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, L. Fei-Fei, Every moment counts: dense detailed labeling of actions in complex videos, Int. J. Comput. Vis. 126 (2–4) (2018) 375–389.

[118] R. Goyal, S.E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al., The something something video database for learning and evaluating visual common sense, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 5843–5851.

[119] G.A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari, Actor and observer: Joint modeling of first and third-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7396–7404.

[120] P. Weinzaepfel, X. Martin, C. Schmid, Human action localization with sparse spatial supervision, arXiv:1605.05197(2016).

[121] F.C. Heilbron, V. Escorcia, B. Ghanem, J.C. Niebles, ActivityNet: A large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[122] C. Chen, R. Jafari, N. Kehtarnavaz, UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 168–172.

[123] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[124] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley MHAD: A comprehensive multimodal human action database, in: 2013 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2013, pp. 53–60.

[125] H.S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from RGB-D videos, Int. J. Rob. Res. 32 (8) (2013) 951–970.

[126] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, arXiv:1212.0402(2012).

[127] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2012.

[128] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities., in: 2009 IEEE International Conference on Computer Vision (ICCV), vol. 1, Citeseer, 2009, p. 2.

[129] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: IEEE Conference on Computer Vision & Pattern Recognition, 2009.

[130] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, A. Weber, Documentation mocap database HDM05(2007).

[131] G.A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta, Hollywood in homes: crowdsourcing data collection for activity understanding, in: European Conference on Computer Vision, Springer, 2016, pp. 510–526.

[132] S. Bhattacharya, N.D. Lane, 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), IEEE, 2016, pp. 1–6.

[133] H. Castro, V. Correia, E. Sowade, K. Mitra, J. Rocha, R. Baumann, S. Lanceros-Méndez, All-inkjet-printed low-pass filters with adjustable cutoff frequency consisting of resistors, inductors and transistors for sensor applications, Org. Electron. 38 (2016) 205–212.

[134] S. Wang, X. Zhang, X. Liu, J. Zhang, S. Ma, W. Gao, Utility-driven adaptive preprocessing for screen content video compression, IEEE Trans. Multimed. 19 (3) (2016) 660–667.

[135] C. Ma, A. Wang, G. Chen, C. Xu, Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network, Vis. Comput. 34 (6–8) (2018) 1053–1063.

[136] A.D. Ignatov, V.V. Strijov, Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer, Multimed. Tools Appl. 75 (12) (2016) 7257–7270.

[137] Z. Wang, D. Wu, J. Chen, A. Ghoneim, M.A. Hossain, A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection, IEEE Sens. J. 16 (9) (2016) 3198–3207.

[138] D. Triboan, L. Chen, F. Chen, Z. Wang, A semantics-based approach to sensor data segmentation in real-time activity recognition, Future Gener. Comput. Syst. 93 (2019) 224–236.

[139] N.Y. Hammerla, S. Halloran, T. Plötz, Deep, convolutional, and recurrent models for human activity recognition using wearables, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, in: IJCAI'16, AAAI Press, 2016, pp. 1533–1540.

[140] M.H.M. Noor, Z. Salcic, I. Kevin, K. Wang, Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer, Pervasive Mob. Comput. 38 (2017) 41–59.

[141] Z. Qin, Y. Zhang, S. Meng, Z. Qin, K.-K.R. Choo, Imaging and fusing time series for wearable sensor-based human activity recognition, Inf. Fusion 53 (2020) 80–87.

[142] H. He, Y. Tan, W. Zhang, A wavelet tensor fuzzy clustering scheme for multi-sensor human activity recognition, Eng. Appl. Artif. Intell. 70 (2018) 109–122.
[143] S.U. Jan, Y.-D. Lee, J. Shin, I. Koo, Sensor fault classification based on support vector machine and statistical time-domain features, IEEE Access 5 (2017) 8682–8690.
[144] F. Gu, K. Khoshelham, S. Valaee, J. Shang, R. Zhang, Locomotion activity recognition using stacked denoising autoencoders, IEEE Internet Things J. 5 (3) (2018) 2085–2093.
[145] J. Wang, X. Zhang, Q. Gao, X. Ma, X. Feng, H. Wang, Device-free simultaneous wireless localization and activity recognition with wavelet feature, IEEE Trans. Veh. Technol. 66 (2) (2017) 1659–1669.
[146] J. Guo, Y. Mu, M. Xiong, Y. Liu, J. Gu, Activity feature solving based on TF-IDF for activity recognition in smart homes, Complexity 2019 (2019).
[147] A. Bhavan, S. Aggarwal, Stacked generalization with wrapper-based feature selection for human activity recognition, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, pp. 1064–1068.
[148] F. Viegas, L. Rocha, M. Gonçalves, F. Mourão, G. Sá, T. Salles, G. Andrade, I. Sandin, A genetic programming approach for feature selection in highly dimensional skewed data, Neurocomputing 273 (2018) 554–569.
[149] D.D. Dawn, S.H. Shaikh, A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector, Vis. Comput. 32 (3) (2016) 289–306.
[150] B. Sun, Y. Li, C. Guosheng, J. Zhang, B. Chang, C. Min, Moving target segmentation using Markov random field-based evaluation metric in infrared videos, Opt. Eng. 57 (1) (2018) 013106.
[151] M. Babaee, D.T. Dinh, G. Rigoll, A deep convolutional neural network for video sequence background subtraction, Pattern Recognit. 76 (2018) 635–649.
[152] H. Mliki, F. Bouhlel, M. Hammami, Human activity recognition from UAV-captured video sequences, Pattern Recognit. 100 (2019) 107140.
[153] W. Kim, Y. Kim, Background subtraction using illumination-invariant structural complexity, IEEE Signal Process. Lett. 23 (5) (2016) 634–638.
[154] K. Goyal, J. Singhai, Texture-based self-adaptive moving object detection technique for complex scenes, Comput. Electr. Eng. 70 (2018) 275–283.
[155] B.N. Subudhi, S. Ghosh, S.C. Shiu, A. Ghosh, Statistical feature bag based background subtraction for local change detection, Inf. Sci. 366 (2016) 31–47.
[156] Y. Shen, W. Hu, M. Yang, J. Liu, B. Wei, S. Lucey, C.T. Chou, Real-time and robust compressive background subtraction for embedded camera networks, IEEE Trans. Mob. Comput. 15 (2) (2016) 406–418.
[157] S. Jiang, X. Lu, WeSamBE: A weight-sample-based method for background subtraction, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2105–2115.
[158] D.K. Panda, S. Meher, Detection of moving objects using fuzzy color difference histogram based background subtraction, IEEE Signal Process. Lett. 23 (1) (2016) 45–49.
[159] Z. Zeng, J. Jia, D. Yu, Y. Chen, Z. Zhu, Pixel modeling using histograms based on fuzzy partitions for dynamic background subtraction, IEEE Trans. Fuzzy Syst. 25 (3) (2017) 584–593.
[160] A. Darwich, P.-A. Hébert, A. Bigand, Y. Mohanna, Background subtraction based on a new fuzzy mixture of gaussians for moving object detection, J. Imaging 4 (7) (2018) 92.
[161] D. Sakkos, H. Liu, J. Han, L. Shao, End-to-end video background subtraction with 3d convolutional neural networks, Multimed. Tools Appl. 77 (17) (2018) 23023–23041.
[162] T. Minematsu, A. Shimada, H. Uchiyama, R.-i. Taniguchi, Analytics of deep neural network-based background subtraction, J. Imaging 4 (6) (2018) 78.
[163] H. Yong, D. Meng, W. Zuo, L. Zhang, Robust online matrix factorization for dynamic background subtraction, IEEE Trans. Pattern Anal. Mach. Intell. 40 (7) (2018) 1726–1740.
[164] M. Chen, X. Wei, Q. Yang, Q. Li, G. Wang, M.-H. Yang, Spatiotemporal GMM for background subtraction with superpixel hierarchy, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1518–1525.
[165] X. Liu, J. Yao, X. Hong, X. Huang, Z. Zhou, C. Qi, G. Zhao, Background subtraction using spatio-temporal group sparsity recovery, IEEE Trans. Circuits Syst. Video Technol. 28 (8) (2018) 1737–1751.
[166] D. Berjón, C. Cuevas, F. Morán, N. García, Real-time nonparametric background subtraction with tracking-based foreground update, Pattern Recognit. 74 (2018) 156–170.
[167] A. Ladjailia, I. Bouchrika, H.F. Merouani, N. Harrati, Z. Mahfouf, Human activity recognition via optical flow: decomposing activities into basic actions, Neural Comput. Appl. (2019) 1–14.
[168] A. Ullah, K. Muhammad, J. Del Ser, S.W. Baik, V. Albuquerque, Activity recognition using temporal optical flow convolutional features and multi-layer LSTM, IEEE Trans. Ind. Electron. 66 (12) (2018) 9692–9702.
[169] R. Singh, J.K. Dhillon, A.K.S. Kushwaha, R. Srivastava, Depth based enlarged temporal dimension of 3d deep convolutional network for activity recognition, Multimed. Tools Appl. 78 (21) (2019) 30599–30614.
[170] L. Liu, S. Wang, Y. Peng, Z. Huang, M. Liu, B. Hu, Mining intricate temporal rules for recognizing complex activities of daily living under uncertainty, Pattern Recognit. 60 (2016) 1015–1028.
[171] C. Dhiman, D.K. Vishwakarma, A review of state-of-the-art techniques for abnormal human activity recognition, Eng. Appl. Artif. Intell. 77 (2019) 21–45.
[172] K. Kim, A. Jalal, M. Mahmood, Vision-based human activity recognition system using depth silhouettes: a smart home system for monitoring the residents, J. Electr. Eng. Technol. 14 (6) (2019) 2567–2573.
[173] F. Baumann, A. Ehlers, B. Rosenhahn, J. Liao, Recognizing human actions using novel space-time volume binary patterns, Neurocomputing 173 (2016) 54–63.
[174] O. Kihl, D. Picard, P.-H. Gosselin, Local polynomial space–time descriptors for action classification, Mach. Vis. Appl. 27 (3) (2016) 351–361.
[175] Y. Fu, T. Zhang, W. Wang, Sparse coding-based space-time video representation for action recognition, Multimed. Tools Appl. 76 (10) (2017) 12645–12658.
[176] A. Shahroudy, T.-T. Ng, Q. Yang, G. Wang, Multimodal multipart learning for action recognition in depth videos, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2016) 2123–2129.
[177] G.V. Kale, Human activity recognition on real time and offline dataset, Int. J. Intell. Syst. Appl. Eng. 7 (1) (2019) 60–65.
[178] M.A.R. Ahad, J. Tan, H. Kim, S. Ishikawa, Activity representation by SURF-based templates, Comput. Methods Biomech. Biomed. Eng. Imaging Visual. 6 (5) (2018) 573–583.
[179] C.I. Patel, S. Garg, T. Zaveri, A. Banerjee, R. Patel, Human action recognition using fusion of features for unconstrained video sequences, Comput. Electr. Eng. 70 (2018) 284–301.
[180] F. Patrona, A. Chatzitofis, D. Zarpalas, P. Daras, Motion analysis: action detection, recognition and evaluation based on motion capture data, Pattern Recognit. 76 (2018) 612–622.
[181] J. Tang, H. Cheng, Y. Zhao, H. Guo, Structured dynamic time warping for continuous hand trajectory gesture recognition, Pattern Recognit. 80 (2018) 21–31.
[182] M. Li, Z. Zhou, X. Liu, Multi-person pose estimation using bounding box constraint and LSTM, IEEE Trans. Multimed. 21 (10) (2019) 2653–2663.
[183] K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognit. 71 (2017) 402–413.
[184] Y. Zhang, H. Lu, L. Zhang, X. Ruan, Combining motion and appearance cues for anomaly detection, Pattern Recognit. 51 (2016) 443–452.
[185] L. Liu, S. Wang, G. Su, Z.-G. Huang, M. Liu, Towards complex activity recognition using a bayesian network-based probabilistic generative framework, Pattern Recognit. 68 (2017) 295–309.
[186] L. Liu, S. Wang, B. Hu, Q. Qiong, J. Wen, D.S. Rosenblum, Learning structures of interval-based bayesian networks in probabilistic generative model for human complex activity recognition, Pattern Recognit. 81 (2018) 545–561.
[187] R. Akhavian, A.H. Behzadan, Smartphone-based construction workers' activity recognition and classification, Autom. Constr. 71 (2016) 198–209.
[188] M. Muaaz, R. Mayrhofer, Accelerometer based gait recognition using adapted gaussian mixture models, in: Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi Media, ACM, 2016, pp. 288–291.
[189] M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, R. Guan, Text classification based on deep belief network and softmax regression, Neural Comput. Appl. 29 (1) (2018) 61–70.
[190] L. Wang, Recognition of human activities using continuous autoencoders with wearable sensors, Sensors 16 (2) (2016) 189.
[191] M. Gnouma, A. Ladjailia, R. Ejbali, M. Zaied, Stacked sparse autoencoder and history of binary motion image for human activity recognition, Multimed. Tools Appl. 78 (2) (2019) 2157–2179.
[192] S.S. Khan, B. Taati, Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders, Expert Syst. Appl. 87 (2017) 280–290.
[193] J.D. Farah, N. Baddour, E.D. Lemaire, Design, development, and evaluation of a local sensor-based gait phase recognition system using a logistic model decision tree for orthosis-control, J. Neuroeng. Rehabil. 16 (1) (2019) 22.
[194] U.M. Nunes, D.R. Faria, P. Peixoto, A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier, Pattern Recognit. Lett. 99 (2017) 21–31.
[195] A.-A. Liu, W.-Z. Nie, Y.-T. Su, L. Ma, T. Hao, Z.-X. Yang, Coupled hidden conditional random fields for RGB-D human action recognition, Signal Process. 112 (2015) 74–82.
[196] A. Tharwat, H. Mahdi, M. Elhoseny, A.E. Hassanien, Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm, Expert Syst. Appl. 107 (2018) 32–44.
[197] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J.-H. Lai, J. Zhang, Early action prediction by soft regression, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2018) 2568–2583.
[198] M.A. Khan, T. Akram, M. Sharif, M.Y. Javed, N. Muhammad, M. Yasmin, An implementation of optimized framework for action classification using multilayers neural network on selected fused features, Pattern Anal. Appl. 22 (4) (2019) 1377–1397.
[199] P. Khaire, P. Kumar, J. Imran, Combining CNN streams of RGB-D and skeletal data for human activity recognition, Pattern Recognit. Lett. 115 (2018) 107–116.
[200] W. Li, L. Wen, M.-C. Chang, S. Nam Lim, S. Lyu, Adaptive RNN tree for large-scale human action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1444–1452.
[201] M. Inoue, S. Inoue, T. Nishida, Deep recurrent neural network for mobile human activity recognition with high throughput, Artif. Life Rob. 23 (2) (2018) 173–185.
[202] J. Dou, Q. Qin, Z. Tu, Robust visual tracking based on generative and discriminative model collaboration, Multimed. Tools Appl. 76 (14) (2017) 15839–15866.
[203] H.J. Escalante, E.F. Morales, L.E. Sucar, A naive bayes baseline for early gesture recognition, Pattern Recognit. Lett. 73 (2016) 91–99.

[204] M.K. Mustafa, T. Allen, K. Appiah, A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition, Neural Comput. Appl. 31 (2) (2019) 891–899.
[205] R. San-Segundo, J.D. Echeverry-Correa, C. Salamea, J.M. Pardo, Human activity monitoring based on hidden Markov models using a smartphone, IEEE Instrum. Meas. Mag. 19 (6) (2016) 27–31.
[206] G.J. McLachlan, S.X. Lee, S.I. Rathnayake, Finite mixture models, Annu. Rev. Stat. Appl. 6 (2019) 355–378.
[207] N. Ding, H. Ma, H. Gao, Y. Ma, G. Tan, Real-time anomaly detection based on long short-term memory and gaussian mixture model, Comput. Electr. Eng. 79 (2019) 106458.
[208] P. Arora, S. Varshney, et al., Analysis of k-means and k-medoids algorithm for big data, Procedia Comput. Sci. 78 (2016) 507–512.
[209] D. Biswas, A. Cranny, N. Gupta, K. Maharatna, J. Achner, J. Klemke, M. Jöbges, S. Ortmann, Recognizing upper limb movements with wrist worn inertial sensors using k-means clustering classification, Hum. Mov. Sci. 40 (2015) 59–76.
[210] G.E. Hinton, A practical guide to training restricted Boltzmann machines, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 599–619.
[211] C. Jia, M. Shao, S. Li, H. Zhao, Y. Fu, Stacked denoising tensor auto-encoder for action recognition with spatiotemporal corruptions, IEEE Trans. Image Process. 27 (4) (2018) 1878–1887.
[212] B. Kamiński, M. Jakubczyk, P. Szufel, A framework for sensitivity analysis of decision trees, Cent. Eur. J. Oper. Res. 26 (1) (2018) 135–159.
[213] J.Y. Chang, Nonparametric feature matching based conditional random fields for gesture recognition from multi-modal video, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1612–1625.
[214] A.M. Murugavel, S. Ramakrishnan, Hierarchical multi-class SVM with ELM kernel for epileptic eeg signal classification, Med. Biol. Eng. Comput. 54 (1) (2016) 149–161.
[215] Z. Deng, X. Zhu, D. Cheng, M. Zong, S. Zhang, Efficient kNN classification algorithm for big data, Neurocomputing 195 (2016) 143–148.
[216] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: a brief review, Comput. Intell. Neurosci. 2018 (2018).
[217] A.A. Heidari, H. Faris, I. Aljarah, S. Mirjalili, An efficient hybrid multilayer perceptron neural network with grasshopper optimization, Soft Comput. 23 (17) (2019) 7941–7958.
[218] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[219] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, F.-Y. Wang, Driver activity recognition for intelligent vehicles: a deep learning approach, IEEE Trans. Veh. Technol. 68 (6) (2019) 5379–5390.
[220] P. Rivera, E. Valarezo, T.-S. Kim, Recognition of human hand activities based on a single wrist IMU using recurrent neural networks, Int. J. Pharma Med. Biol. Sci. 6 (4) (2017) 114–118.
[221] Y. Zhao, R. Yang, G. Chevalier, X. Xu, Z. Zhang, Deep residual Bidir-LSTM for human activity recognition using wearable sensors, Math. Prob. Eng. 2018 (2018).
[222] F. Ordóñez, D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.
[223] J. Margarito, R. Helaoui, A.M. Bianchi, F. Sartor, A.G. Bonomi, User-independent recognition of sports activities from a single Wrist-Worn accelerometer: a template-matching-based approach, IEEE Trans. Biomed. Eng. 63 (4) (2016) 788–796.
[224] A.K.S. Kushwaha, R. Srivastava, Multiview human activity recognition system based on spatiotemporal template for video surveillance system, J. Electron. Imaging 24 (5) (2015) 051004.
[225] Y. Li, D. Xue, E. Forrister, G. Lee, B. Garner, Y. Kim, Human activity classification based on dynamic time warping of an on-body creeping wave signal, IEEE Trans. Antennas Propag. 64 (11) (2016) 4901–4905.
[226] S.M. Salaken, A. Khosravi, T. Nguyen, S. Nahavandi, Seeded transfer learning for regression problems with deep learning, Expert Syst. Appl. 115 (2019) 565–577.
[227] L. Guo, L. Wang, J. Liu, W. Zhou, B. Lu, HuAc: Human activity recognition using crowdsourced WiFi signals and skeleton data, Wirel. Commun. Mob. Comput. 2018 (2018).
[228] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, ActivityNet: A large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 961–970.
[229] Z. Pei, X. Qi, Y. Zhang, M. Ma, Y.-H. Yang, Human trajectory prediction in crowded scene using social-affinity long short-term memory, Pattern Recognit. 93 (2019) 273–282.
[230] K. Akila, S. Chitrakala, Highly refined human action recognition model to handle intraclass variability & interclass similarity, Multimed. Tools Appl. (2019) 1–19.
[231] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, B. Schiele, Recognizing fine-grained and composite activities using hand-centric features and script data, Int. J. Comput. Vis. 119 (3) (2016) 346–373.
[232] H. Zhang, J. Smeddinck, R. Malaka, Y. Shu, C. Chen, B. He, Z. Fu, M. Lawo, Wireless non-invasive motion tracking of functional behavior, Pervasive Mob. Comput. 54 (2019) 29–44.
[233] M.Á. Antón, J. Ordieres-Meré, U. Saralegui, S. Sun, Non-invasive ambient intelligence in real life: dealing with noisy patterns to help older people, Sensors 19 (14) (2019) 3113.
[234] L. Zheng, D. Wu, X. Ruan, S. Weng, A. Peng, B. Tang, H. Lu, H. Shi, H. Zheng, A novel energy-efficient approach for human activity recognition, Sensors 17 (9) (2017) 2064.
[235] C. Xu, D. Chai, J. He, X. Zhang, S. Duan, InnoHAR: A deep neural network for complex human activity recognition, IEEE Access 7 (2019) 9893–9902.
[236] T.N. Nguyen, C.H. Thai, A.-T. Luu, H. Nguyen-Xuan, J. Lee, NURBS-based post-buckling analysis of functionally graded carbon nanotube-reinforced composite shells, Comput. Methods Appl. Mech. Eng. 347 (2019) 983–1003.

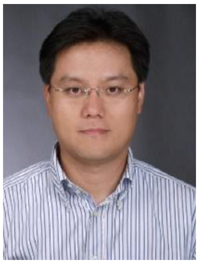
L. Minh Dang received the B.S. degree majoring in Information Systems in 2016 from the University of Information Technology, VNU HCMC, Vietnam. He is currently pursuing the Ph.D. degree in Computer Science from Sejong University, Seoul, South Korea. He joined the Computer Vision Pattern Recognition Laboratory (CVPR Lab) at the beginning of 2017. His current research interests include computer vision, natural language processing, video coding, and deep learning.

Kyungbok Min received the B.S. degree in Electronics and Computer Engineering from Korea University. He is currently pursuing Ph.D. in Computer Science from Sejong University, Seoul, South Korea. He joined Computer Vision Pattern Recognition Laboratory (CVPR Lab) from 2017. His current research interests include computer vision, natural language processing and artificial intelligence.

Hanxiang Wang received the B.S. degree in Software engineering in 2018 from the LinYi University. He is currently pursuing Ph.D. degree in Computer Science from Sejong University, Seoul, South Korea. He joined Computer Vision Pattern Recognition Laboratory (CVPR Lab) at the beginning of 2018. His current research interests include computer vision, deep learning, image processing, video coding and artificial intelligence

Md. Jalil Piran (S'10-M'16) received the Ph.D. degree in electronics and radio engineering from Kyung Hee University, South Korea, in 2016. Subsequently, he continued his work as a Postdoctoral Research Fellow in the field of resource management and quality of experience in 5G-Cellular Networks, and the Internet of Things with the Networking Laboratory, Kyung Hee University. He is currently a Professor with the Department of Computer Science and Engineering, Sejong University, Seoul, South Korea. He has published substantial number of technical papers in well-known international journals and conferences in research fields of: resource allocation and management in; 5G mobile and wireless communication, HetNet, the Internet of Things (IoT), multimedia communication, streaming, adaptation, and QoE, cognitive radio networks, wireless sensor networks, machine learning, fuzzy logic, and neural networks.

Cheol Hee Lee received the B.S. degree in Civil Engineering from Chungbuk National University. He is currently pursuing Ph.D. in Structural Engineering from Chungbuk National University, South Korea. He has founded AI start-up specialized in inspection for public facilities, Deep Inspection in 2015. He has completed the expert course regarding machine learning, deep learning, and neural network through deeplearning.ai. His current research interests include computer vision, artificial intelligence, XAI, and meta-learning.



Hyoenjoon Moon received the B.S. degree in Electronics and Computer Engineering from Korea University in 1990. He received the M.S. and the Ph.D. degrees from Electrical and Computer Engineering at State University of New York at Buffalo in 1992 and 1999, respectively. From January 1996 to October 1999, he was a senior research in Electro-Optics/Infrared Image Processing Branch at U.S. Army Research Laboratory (ARL) in Adelphi, MD. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) program. From November 1999 to February 2003, he was a principal research scientist at Viisage Technology in Littleton, MA. His main interest is on research and development is on real-time facial recognition system for access control, surveillance, and big database applications. He has extensive background on still image and real-time video based computer vision and pattern recognition. Since March 2004, he has joined the Department of Computer Science and Engineering at Sejong University, where he is currently a professor and chairman. His current research interests include image processing, biometrics, artificial intelligence and machine learning.