Archives of Computational Methods in Engineering (2024) 31:179–219
https://doi.org/10.1007/s11831-023-09986-x

REVIEW ARTICLE

Human Activity Recognition (HAR) Using Deep Learning: Review, Methodologies, Progress and Future Research Directions

Pranjal Kumar[1] · **Siddhartha Chauhan**[1] · **Lalit Kumar Awasthi**[1]

Received: 29 December 2022 / Accepted: 2 July 2023 / Published online: 12 August 2023

Abstract
Human activity recognition is essential in many domains, including the medical and smart home sectors. Using deep learning, we conduct a comprehensive survey of current state and future directions in human activity recognition (HAR). Key contributions of deep learning to the advancement of HAR, including sensor and video modalities, are the focus of this review. A wide range of databases and performance metrics used in the implementation of HAR methodologies are described in depth. This paper explores the wide range of HAR's potential uses, from healthcare, emotion calculation and assisted living to security and education. The paper provides an in-depth analysis of the most significant works that employ deep learning techniques for a variety of HAR downstream tasks across both the video and sensor domains including the most recent advances. Finally, it addresses problems and limitations in the current state of HAR research and proposes future research avenues for advancing the field.

1 Introduction

Identifying and comprehending human actions, also known as Human Activity Recognition (HAR), is essential for a wide range of practical uses. It is possible to integrate it into automated navigation systems [1] in order to recognise human behaviours for the purpose of ensuring safe operation, as well as surveillance systems [2] in order to recognise potentially hazardous activities involving humans. A great number of other applications, such as human-robot interaction [3], video retrieval [4] and entertainment [5], are dependent on it. Health monitoring, home automation, fitness, traffic scheduling and control, augmented reality, precise advertising, and security are just a few examples of the many services that rely on an understanding of human activity [6]. For instance, a person's activity log can be used to determine his caloric intake for the day, leading to advice on how to improve his diet and fitness levels; similarly, monitoring elderly people's fall activity can prompt immediate help in the event of a fall, preventing potentially catastrophic injuries.

Wearables, environmental sensors, and computer vision systems all feed data into HAR systems, which then use a machine learning or deep learning model to recognize the activities [7]. Installing environmental sensors in a home is a costly endeavor [8]. Vision systems, which rely on cameras for recognition, are often seen as invasive [9]. Wearable devices are another option, and they're attracting researchers' attention because of how widely they're used. Since sensors like the accelerometer, gyroscope, and compass are already built into and integrated into wearable devices like fitness trackers and smartwatches, these devices are primarily used for recognition. Figure 1 demonstrates a general framework for HAR. Smartphones are used instead of wearable devices for activity recognition because they are convenient, can be used anywhere, are inexpensive, have the same kinds of embedded sensors that wearable devices do, and are often used in real-time applications [10]. The visibility of data can also be used to roughly categorise modalities into visual and non-visual categories. RGB, depth, skeleton, point cloud, infrared sequence, and event stream are just some of the visual modalities that can be used to accurately depict human actions. When it comes to HAR, visual modalities tend to perform better than others. In particular, HAR has found widespread use in monitoring and surveillance systems [11], where RGB video data predominates. A person's joint trajectories can be represented by their skeleton data. If the action being performed has nothing to do with

✉ Pranjal Kumar
pranjal@nith.ac.in

[1] National Institute of Technology Hamirpur, Hamirpur, Himachal Pradesh 177005, India

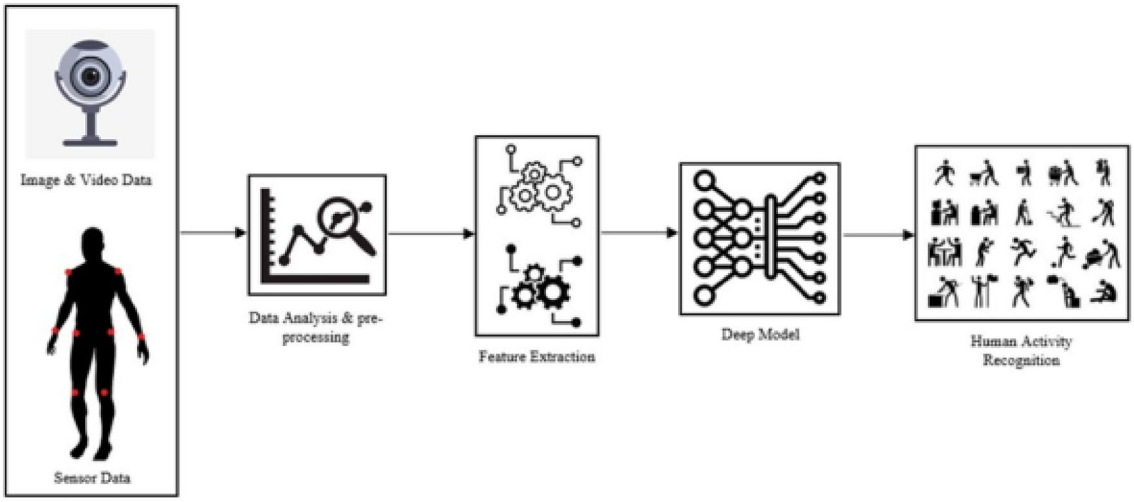


Fig. 1 General Human activity recognition (HAR) framework

objects or the scene context, HAR can do it quickly and easily. Point clouds and depth data capture the 3D structure and distance information used for HAR in robot navigation and self-driving applications. Additionally, infrared data can be used for HAR even in low-light settings, and the event stream is well-suited for HAR as it maintains the foreground motion of the human subjects while removing distracting background elements. Human behaviour is not visually "intuitive" to represent in non-visual modalities like sound, acceleration, WiFi, radar etc. However, in situations where subject confidentiality must be maintained, these modalities can also be used for HAR. While acceleration data can be used to implement fine-grained HAR, audio data can be used to pinpoint events in time sequences. Since radar is a non-visual modality, its data can also be used for HAR in through-wall applications.

Traditional machine learning techniques can detect human action. The problem with the standard machine learning approaches to HAR is that they necessitate manually designing and selecting features to use. To accomplish this requires time-consuming human involvement and specialized knowledge, and even then, the resulting feature set may not function as optimally as possible. In recent years, deep learning approaches have been proposed [12, 13] to alleviate the need for human intervention in the feature engineering process. The application of deep learning techniques to HAR has the potential to improve the field in a variety of ways. For one, it eliminates the need for the time-consuming and often-complex process of designing features by hand. Second, it has proven to be more precise in HAR than traditional methods [14–16]. Finally, it can learn from unlabeled data, which is particularly helpful for HAR because it is impractical to collect a large amount of labeled activity data. Fourth, it has the robust capability of learning useful features from raw data, and it can process activity-related data from a wide range of people, device models, and device poses. Furthermore, the machine learning-based solutions rely entirely on pre-processed data from raw signals, which contains valuable and remarkable features that can enhance the performance of classification algorithms. Deep learning models can be used to quickly address or circumvent these difficulties [17]. Recent improvements and promising results on various benchmark datasets used by machine learning-based solutions have been achieved by deep learning models. In the data pre-processing and feature extraction stage, it can help reduce the workload. In addition, it can strengthen the deep learning model's generalization abilities and make it less prone to breakdowns.

1.1 Contributions

In this work, we contribute significantly to the literature by looking at a wider perspective on the overall development HAR research from both sensor and video modalities over the last decade. We don't just focus on algorithmic information, contrary to current surveys. As explained in the last section, most of the studies/works examined only particular machine-learning aspects of HAR. More recently, the introduction of a variety of deep learning frameworks and methodologies for HPE modeling has also added various new hypotheses, procedures, and applications. Therefore, a thorough HAR survey is important and crucial to collaborators/contributors, physicians, and researchers who attempt to formulate and integrate these methods with existing systems or carry out ameliorated HAR research. In this survey, we recapitulate both past and current research and cover a broad range of aspects of HAR, including datasets, methods, and

human activity recognition models. The following key points highlight our contributions:

- A detailed discussion on the variety of sensor-based and video-based databases and performance metrics incorporated is presented for a better understanding of the frontier ideas in HAR.
- A comparative review of all the major works that use deep learning models for various downstream tasks in each domain for both sensor-based and video-based HAR is conducted.
- An overview of a variety of applications of HPE across domains like surveillance and security, emotional calculation, healthcare and rehabilitation, education, etc., is presented along with the most recent advances in the field of HAR.
- Several unresolved problems in this area have been examined and the future direction for deep learning-based HAR is discussed.

1.2 Organisation of Paper

The paper is organised in the following manner: In Sect. 2, we compare and contrast the findings of several recent large-scale surveys of HAR. Section 3 explains in depth the many data sources and Sect. 4 discusses the pre-processing techniques used in HAR. Following this, Sects. 5 and 6 discuss the history of research in various fields and a wide variety of classifications and deep learning-based methodologies for sensor-based and video-based HAR, respectively. In Fig. 2 we can see how various deep-learning strategies for HAR have been categorized. In Sect. 7, we discuss about various metrics of performance measurement used in HAR. The eighth section focuses on the many practical uses of this technology, including healthcare, emotion calculation, assisted living, security and education. In the final section of the paper, we discuss some of the most contentious issues and their potential future evolution as a means of wrapping up the discussion.

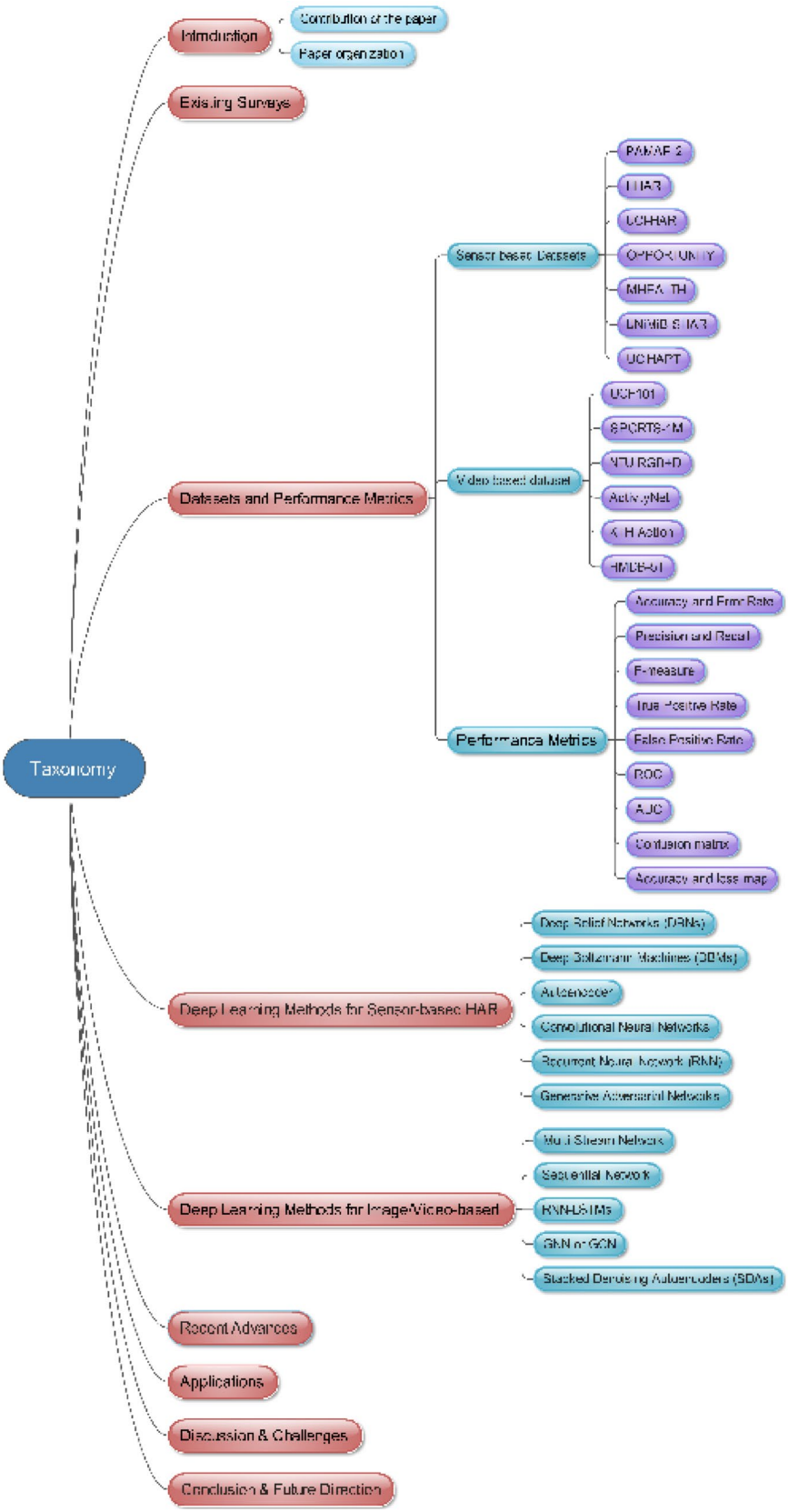
2 Existing Surveys

Multiple uses, including smart healthcare services and smart home systems, can benefit from HAR. Wearable sensors, smartphones, RF sensors (Wi-Fi, RFID), LED light sensors, cameras, etc., are just some of the sensors that have been used for human activity recognition. As wireless sensor networks have evolved quickly, a wealth of information has been gathered to aid in the identification of human activities using various sensors. Traditional shallow learning algorithms like support vector machine and random forest necessitate the manual extraction of some representative features from large and noisy sensory data. Manual feature engineering, on the other hand, is time-consuming and prone to missing implicit features because it relies on specialised domain expertise. In recent years, deep learning has seen tremendous success in many difficult research domains, including image recognition and natural language processing. The ability to automatically learn representative features from massive data sets is the primary benefit of deep learning [18, 19]. HAR may be an appropriate application for this technology. As a result, it is crucial to record the successes and think critically about them in order to achieve even more. Vision-based HAR and sensor-based HAR are the two main types of HAR currently available. Preprocessing data, object segmentation, feature extraction, and classifier implementation are the integral parts of the vision-based processing phase. Many researchers over the past few decades have proposed numerous video-based HAR technologies that can achieve the rapid recognition of human behaviour by using video and motion sensors in response to the enormous market demand and economic value of such technologies. However, when privacy is a major concern, the shadow of the object, the colour of the background, and the intensity of the light can all negatively impact the accuracy of vision-based HAR. This privacy concern, however, can be avoided when smartphones and wearable sensors are used for HAR in smart homes.

In the last few years, various survey studies have been published related to human activity recognition. The first HAR system was proposed by authors in [20], which uses five wearable dual-axis accelerometers and machine learning classifiers to recognise 20 ADLs with an impressive 84% classification accuracy. Combining accelerometers with gyros has been shown to boost recognition performance [21, 22]. Data from smartphone inertial sensors were used for classification alongside expert hybrid models to create a HAR system in [23] that could be used to identify five transport activities. The authors of [24] proposed a system for offline HAR that makes use of a smartphone equipped with a three-axis accelerometer. The smartphone was hidden in a pocket during the experiment. An activity recognition system was developed in [25] by attaching a smartphone to the user's waist and using the device's inertial sensors.

Single-data modality and multi-data modality approaches, such as fusion-based and co-learning-based frameworks, are presented by the authors of [11, 26]. In [27–30], the authors analyse several prominent studies that employ various sensing technologies to carry out HAR tasks by means of machine learning (ML) methods. Improved recognition accuracy in HAR has been achieved through the application of deep learning techniques in recent years. The accuracy of these deep learning models is vastly superior to that of more conventional recognition strategies. In their survey of the relevant literature, the authors of [31] find that

Fig. 2 Taxonomy of this review



Convolutional Neural Networks (CNNs), Long Short-Term Memories (LSTMs), and Support Vector Machines (SVMs) are the most effective methods. In [32, 33], the benefits and advantages of multi-user activity recognition are laid out, along with the sensing methods, recognition approaches, and practical applications that make use of them, as well as the challenges and techniques involved in data fusion. [34] summarises the deep learning techniques used in smartphone and wearable sensor-based recognition systems. In [35], authors concentrated primarily on techniques for recognising human actions and interacting with inanimate objects. For the purpose of action classification inferred from time series of 3D skeletons, Presti et al. [36] provided a survey of human action recognition based on 3D skeletons, summarising the main technologies, including hardware and software. Further, Kang and Wildes [37] presented the results of another survey. It provided a concise summary of algorithms for recognising and detecting actions, with an emphasis on feature encoding and classification.

3 Datasets

As interest in human action recognition algorithms has grown, many datasets have been recorded and made available to the research community. Improvements in action recognition have largely been shown on industry-standard benchmark datasets. With these data sets, we can test and compare various approaches to a problem. We provide a brief overview of the most relevant publicly available datasets in this area.

3.1 Sensor-Based Dataset

To impartially compare the efficacy of deep learning and machine learning-based solutions for HAR [39], researchers have compiled a wide range of benchmark datasets. The subject's head, shin, forearm, chest, upper arm, thigh, waist, and legs were all used to collect motion signals for the dataset's embedded sensors. Smartphones are tucked into pants or a shirt, and smartwatches are wrapped around the dominant hand. Sensors in these devices include things like accelerometers, gyroscopes, magnetometers, temperature sensors, and ambient light detectors, among others. Time series data from the MotionSense dataset [38], for example, includes 12 features (*attitude.roll, attitude.pitch, attitude.yaw, gravity.x, gravity.y, gravity.z, rotationRate.x, rotationRate.y, rotationRate.z, userAcceleration.x, userAcceleration.y, userAcceleration.z*), as shown in Fig. 3. Subjects' ages, heights, weights, and other biometric characteristics are described in different ways across datasets' collected signals. As part of the sensory data collection process, the subjects are given both simple and complex tasks to complete. Traveling by foot, jumping, lying down, running, jogging, ascending and descending stairs, and pedaling a bicycle are all easy. Complex tasks include preparing meals, laundering clothes, and cleaning the kitchen. Table 1 gives an overview of several representative benchmark datasets based on sensors for HAR.

3.1.1 PAMAP2

In the PAMAP2 Physical Activity Monitoring dataset, nine subjects wore three inertial measurement units and a heart

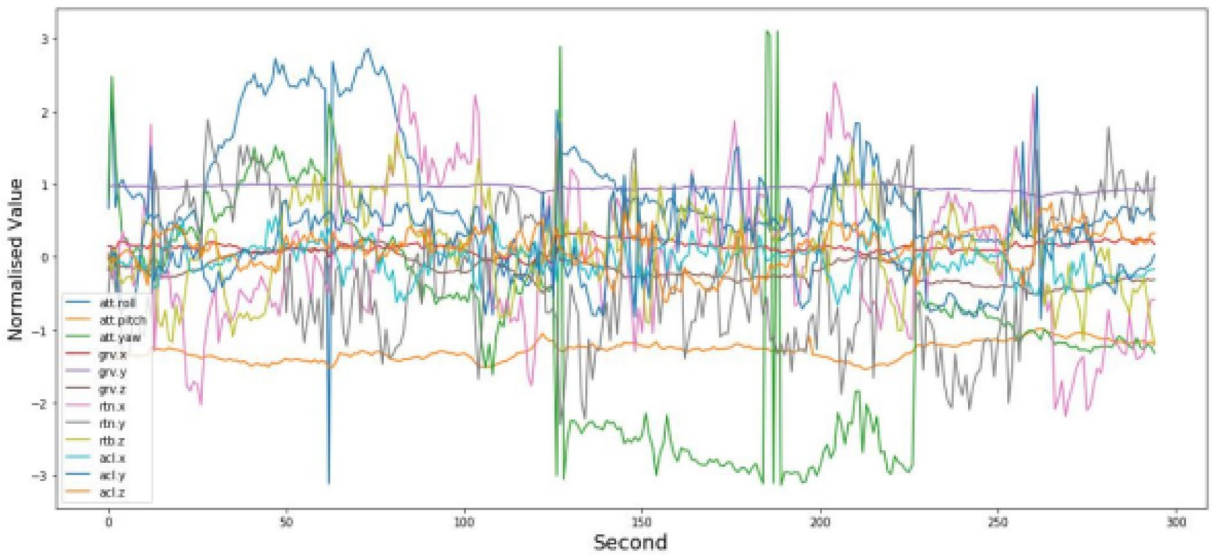


Fig. 3 Sample from MotionSense dataset [38]

Table 1 An overview of some representative benchmark sensor-based datasets for HAR (*IMU* inertial measurement unit, *HR* Heart rate, *ECG* Electrocardiogram, *Acc.* Accelerometer, *SP* Smartphone, *SW* smart watch)

Dataset	Year	Sampling frequency	Environment	Devices	Subjects	Activities	Key features
Daphnet [44, 45]	2010	64 Hz	Lab	3 Acc	3	10	A data set created to compare different automatic gait freeze recognition methods using data collected from accelerometers attached to the user's hips and legs
PAMAP2 [46–48]	2012	100Hz	Lab	3 IMU, 1 HR Monitor	18	9	Training and testing for data processing, segmentation, feature extraction, and classification, and estimating the intensity of activities
OPPORTUNITY [49, 50]	2012	50Hz	Home	9D IMU	4	6	Dataset from Wearable, Object, and Ambient Sensors to provide a standard against which human activity recognition algorithms can be evaluated
UCI-HAR [51–53]	2013	20Hz	Out of Lab	9D IMU	51	18	Time-series sensor information from a smartphone and smartwatch accelerometer and gyroscope
MHEALTH [54–56]	2014	50Hz	Out of Lab	9D IMU, ECG	10	12	Acceleration, Rotational velocity, and magnetic field orientation are all measured by sensors strapped to the subject's chest, right wrist, and left ankle
HHAR [57–59]	2015	100-200 Hz	Out of Lab	SP, SW	9	6	Acceleration, Developed to study how sensor heterogeneities affect human activity recognition algorithms
UniMiB-SHAR [60–62]	2017	50 Hz	Controlled	SP	30	17	Samples of acceleration collected using an Android smartphone specifically designed for human activity recognition and fall detection
Skoda checkpoint [63–65]	2017	98 Hz	Controlled	20 3D Acc	1	10	To address human movement complexity, noise from sensing devices, and individual differences in factory maintainence

rate monitor while engaging in eighteen distinct physical activities (such as walking, cycling, playing soccer, etc.). Dataset can be used for training and testing algorithms for data processing, segmentation, feature extraction, and classification; activity recognition; and intensity estimation. Three Colibri inertial measurement units operate on a wireless signal (IMU). The rate of sampling is 100 Hz. One IMU is worn on the dominant hand's wrist. One

inertial measurement unit is worn on the chest. One IMU is implanted in the ankle of the dominant foot The sampling rate of the HR monitor is 9Hz [40].

3.1.2 HHAR

Collected from Smartphones and Smartwatches, Heterogeneity HAR is designed to benchmark human activity recognition algorithms (classification, automatic data segmentation, sensor fusion, feature extraction, etc.) in real-world contexts; specifically, the dataset is gathered with a variety of different device models and use-scenarios, in order to reflect sensing heterogeneities to be expected in real deployments. Data from two smartphone motion sensors are included in the dataset. Smartwatches and smartphones were worn as readers carried out scripted activities in a random order. The accelerometer and gyroscope, both built into the device, are used as sensors, with their readings sampled as frequently as possible. There are eight mobile devices total, including four smartwatches (two LG watches and two Samsung Galaxy Gears) and four smartphones (two Samsung Galaxy S3 minis, two Samsung Galaxy S3, two LG Nexus 4s, and two Samsung Galaxy S+s). Nine people's recordings have been made [41].

3.1.3 UCI-HAR

The human Activity Recognition database collected data from 30 people while they went about their daily lives with a smartphone attached to their waists and equipped with inertial sensors. Thirty volunteers between the ages of 19 and 48 participated in the experiments. The participants moved through six different positions while wearing a Samsung Galaxy S II smartphone on their waist: walking, walking upstairs, walking downstairs, sitting, standing, and lying down. We recorded 3-axis linear acceleration and 3-axis angular velocity at a rate of 50 Hz using its built-in accelerometer and gyroscope. In order to manually label the data, the experiments have been filmed. This obtained dataset was then randomly split in half, with 70% of the volunteers used to produce the training data and 30% used to produce the test data. The accelerometer and gyroscope signals were pre-processed with noise filters and then sampled in fixed-width sliding windows of 2.56 seconds and 50% overlap (128 readings/window). Using a Butterworth low-pass filter, we were able to disentangle the gravitational and body-motion components of the acceleration signal recorded by the sensor. Since it is assumed that the gravitational force consists entirely of low-frequency components, a filter with a cut-off frequency of 0.3 Hz was employed. To create a vector of features, time and frequency domain variables were calculated for each window [42].

3.1.4 OPPORTUNITY

OPPORTUNITY dataset is created to evaluate and compare human activity recognition algorithms for HAR from Wearable, Object, and Ambient Sensors (classification, automatic data segmentation, sensor fusion, feature extraction, etc). The data set is made up of motion sensor readings gathered while users went about their daily routines. There are a total of 14 sensors–7 IMUs, 12 3D acceleration sensors, and 4 3D localization sensors-that can be worn on the body. There are a total of 12 objects with 3D acceleration and 2D rate of turn measured by the object sensors. There are 13 switches and 8 3D acceleration sensors that detect the surrounding environment. Six trials were recorded for each of four users. Five of these are ADLI runs, in which daily tasks are completed in an unforced and organic manner. The sixth iteration is a "drill" iteration, wherein users carry out a predetermined series of actions. The user's actions throughout the scenario are annotated at various tiers, and this is reflected in the classes that are used to describe them. There are 13 "low-level actions" that link 13 actions to 23 objects, 17 "mid-level gesture" classes, and 5 "high-level activity" classes that represent different "modes of locomotion." Activity recognition environments and scenarios are built to produce a large number of realistic activity primitives. Each participant worked in a space designed to replicate a studio apartment, complete with a deckchair, kitchen, doors leading to the outdoors, coffee machine, table, and chair. There are a total of 6 separate runs for each subject. Five of these tasks, known as ADLs (activities of daily living), were carried out in accordance with the conditions described below. The other option is a drill run, which is meant to produce many separate instances of the activity being tested. As the ADL run progresses, different events take place. Numerous action primitives happen in every context (like making a sandwich) (e.g. reaching for bread, moving to the bread cutter, operating the bread cutter) [43].

3.1.5 MHEALTH

The purpose of the MHEALTH (Mobile Health) dataset is to serve as a benchmark for methods of human behaviour analysis using multimodal body sensing. Ten volunteers representing a wide range of backgrounds participated in the MHEALTH (Mobile HEALTH) study, which recorded their body movements and vital signs as they engaged in a variety of physical activities. The subject wears sensors on their chest, wrist, and ankle, which record acceleration, rotational velocity, and magnetic field orientation. The sensor can also take 2-lead ECG readings when placed on the chest, which can be used for basic heart monitoring, checking for various arrhythmias, or studying the ECG's response to physical exertion. This dataset was

compiled from recordings of body motion and vital signs made by 10 volunteers with varying backgrounds as they engaged in 12 different types of physical activity. The data was collected using Shimmer2 [BUR10] wrist-worn sensors. The subject had elastic straps attached to their chest, right wrist, and left ankle where the sensors were placed. By employing several sensors, we are able to more accurately capture the body's dynamics by measuring the acceleration, rate of turn, and magnetic field orientation that are experienced by various parts of the body. The chest-mounted sensor also provides two-lead electrocardiogram readings, but these are not used in training the recognition model. As an example, this data can be used for routine heart monitoring, the diagnosis of arrhythmias, or the study of how physical activity affects the electrocardiogram (ECG). The sampling rate of 50 Hz is used for all sensing modalities because it is adequate for recording human activity. A video camera captured each meeting. Given the variety of body parts involved in each action (e.g., the frontal elevation of arms versus knees bending), the intensity of the actions (e.g., cycling versus sitting and relaxing), and their execution speed or dynamicity, this dataset is found to generalize to common activities of the daily living (e.g., running vs. standing still). Activities were collected in a non-laboratory setting with no requirements for how they should be performed beyond the subject's best effort [66].

3.1.6 UniMiB-SHAR

Android smartphones were utilised in the data collection process for UniMiB SHAR, a dataset designed for HAR and fall detection. Thirty people, ranging in age from 18 to 60, contributed 11,771 samples of human activities and falls. The samples are organised into 17 fine-grained classes that are then grouped into two coarse-grained classes, one of which includes examples of 9 different ADLs and the other of which includes examples of 8 different types of falls. The dataset was saved with all the information necessary to select samples based on various criteria, such as the type of ADL performed, the age, the gender, and so on. Finally, four distinct classifiers and two distinct feature vectors have been benchmarked on the dataset. Four different classification tasks (fall vs. no fall, 9 activities, 8 falls, 17 activities, and falls) were tested and analysed. We ran both a fivefold cross-validation (where all subjects' samples were used in both the training and test datasets) and a leave-one-subject-out cross-validation on each classification task (i.e., the test data include the samples of a subject only, and the training data, the samples of all the other subjects) [68]

3.1.7 UCIHAPT

Thirty subjects were recorded while they performed everyday tasks and posture changes while wearing a smartphone attached to a belt with inertial sensors to create an activity recognition data set. Thirty volunteers between the ages of 19 and 48 took part in the experiments. Six fundamental movements were performed, including three static postures (standing, sitting, and lying) and three dynamic activities (walking, walking downstairs, and walking upstairs). As part of the study, we also tracked the subjects' postural changes as they moved between the various static positions. The transitions include standing to sitting, sitting to standing, lying down to sitting, lying down to lying down, standing to lying down, and lying down to standing up. During the course of the experiment, each participant wore a smartphone (a Samsung Galaxy S II) at their waist. Using the device's built-in accelerometer and gyroscope, we recorded linear acceleration in all three directions and angular velocity in all three directions at a steady 50 Hz. Video recordings of the experiments were taken so that the data could be manually annotated. A random split was performed on the obtained dataset, with 70% of the volunteers used to produce the training data and 30% used to produce the test data. Noise filters were applied to the accelerometer and gyroscope signals before they were sampled in fixed-width sliding windows of 2.56 seconds with 50% overlap (128 readings per window). Using a Butterworth low-pass filter, we were able to disentangle the gravitational and body-motion components of the acceleration signal recorded by the sensor. As it is generally accepted that the gravitational force consists entirely of low-frequency components, a filter with a cutoff frequency of 0.3 Hz was employed. By summing up time- and frequency-domain variables for each window, a vector of 561 features was derived [23].

3.2 Video-Based Dataset

The goal of these datasets is to provide difficult videos of people acting in natural settings with varying backgrounds and lighting. But these deeds are not "real." Then, many scientists have created new realistic benchmark datasets by extracting realistic situations from movies or sports videos on social networks like YouTube. The general approach in these datasets is to collect videos from "in-the-wild" sources with many clips and action classes. Due to their massive size, it is easy to see that many datasets are created with deep learning algorithms in mind. Table 2 gives an overview of several representative benchmark datasets based on image/video for HAR.

3.2.1 UCF101

The UCF101 dataset was built using YouTube's realistic action videos and 101 distinct action categories to train a computer to recognize specific types of motion. The 50-category UCF50 data set has been expanded here. The UCF101 data set is the most difficult to date because it contains 13320 videos from 101 action categories and the widest range of challenges in terms of camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. UCF101 seeks to inspire more research into action recognition by learning and exploring new realistic action categories, as most existing data sets are not naturalistic and are staged by actors. The videos in each of the 101 action categories are further divided into 25 subcategories, with 4-7 videos per subcategory. There may be commonalities between the videos in a set, such as a shared setting or point of view [69].

3.2.2 SPORTS-1M

More than a million clips from YouTube's Sports channel make up the Sports-1M dataset. The authors provided a YouTube URL where users can access the dataset's video clips. Since the dataset was created, roughly 7% of the videos have been deleted by their creators on YouTube. In spite of this, the dataset still contains over a million videos, split across 487 distinct sports-related categories with anywhere from one thousand to three thousand clips in each. By analyzing the text metadata of the videos and using the YouTube Topics API, the videos are automatically categorized into 487 different types of sports (e.g. tags, and descriptions). Only about 5 percent of the videos have annotations for more than one category [67] as shown in Fig. 4.

3.2.3 NTU RGB+D

Large-scale RGB-D HAR dataset developed at NTU. There are 56,880 data points representing 60 different classes of action, gathered from 40 different people. The actions can be generally divided into three categories: 40 daily actions (e.g., drinking, eating, reading), nine health-related actions (e.g., sneezing, staggering, falling down), and 11 mutual actions (e.g., punching, kicking, hugging) (e.g., punching, kicking, hugging). There are 17 distinct scene conditions that these events that occur in across 17 videos (i.e., S001-S017). Three cameras were used to record the events, one each at a 45-degree, 0-degree, and +45-degree horizontal imaging viewpoint. Action characterization is supported by a wide variety of data types, from depth maps and 3D skeleton joint positions to RGB frames and infrared sequences. The performance evaluation is performed by a cross-subject test that split the 40 subjects into training and test groups, and by a cross-view test that employed one camera (+45-degree) for testing, and the other two cameras for training [70].

3.2.4 ActivityNet

The ActivityNet dataset includes 849 hours of videos culled from YouTube, in addition to 200 distinct categories of activities. ActivityNet is the largest benchmark for temporal activity detection to date in terms of both the number of activity categories and the number of videos, which makes the task particularly challenging. ActivityNet was developed by Microsoft Research and consists of a large collection of videos. The dataset, version 1.3, includes a total of 19994 unedited videos and is separated into three subsets: training, validation, and testing in the proportions of 2:1:1. Each activity category has, on average, 137 videos that have not

Fig. 4 Sample from SPORTS-1M dataset [67]

Table 2 An overview of some representative benchmark video-based datasets for HAR (*S* Skeleton, *D* Depth, *IR* Infrared, *Au* Audio, *Ac* Acceleration, *Gyr* Gyroscope)

Dataset	Year	Total samples	Modality	View	Devices	Subjects	Activities	Key features
KTH [72, 73]	2004	2391	RGB	Single	1 Camera	25	6	Variations in experimentation include outdoor, outdoor with scale, outdoor with different clothes, and indoor setting
UCF101 [74, 75]	2012	13320	RGB	Single	n/a	n/a	101	Variety in terms of actions, range of variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, and lighting conditions
HMDB-51 [76, 77]	2013	31838	RGB,S	Single	n/a	n/a	51	Each clip is tagged with a category label, an action label, and a meta-label that describes the clip's property
DHA [78, 79]	2012	357	RGB,D	Multiview	n/a	23	21	An efficient local spatiotemporal descriptor for recognising actions in 3D video
NTU RGB+D [80, 81]	2016	56880	RGB, S, D, IR	Multiview	3	40	60	3D skeletal data includes the 3D coordinates of 25 body joints at each frame, while RGB videos have a resolution of 1920×1080
Something-Something-v1 [82, 83]	2017	108499	RGB	n/a	n/a	n/a	174	Massive, meticulously labelled video clips of people using commonplace items to perform commonplace tasks
Kinetics-400 [84, 85]	2017	306245	RGB	Single	n/a	n/a	400	Include interactions between people and things, like playing an instrument, and between people, like shaking hands
MMAct [86, 87]	2019	36,764	RGB, S, Ac, Gyr	Egocentric	n/a	20	37	Includes four different scenarios that cover a wide variety of everyday uses, from desk-based tasks to check-ins
EPIC-KITCHENS-100 [88, 89]	2020	89979	RGB, Au, Ac	Egocentric	n/a	45	n/a	large-scale dataset recorded in first-person (egocentric) view; multimodal, naturalistic recordings in their habitats

been edited. On average, there are 1.41 activities that have temporal boundaries attached to them across all of the videos. Annotations of test videos' ground truth are not made available to the public [71].

3.2.5 KTH Action

In 2004, the KTH Royal Institute of Technology was the first institution to make an effort to develop a non-trivial dataset that was made available to the public for the purpose of action recognition. The KTH dataset is one of the most common datasets, and it includes six different actions: walking, jogging, running, boxing, and hand-clapping with both hands. In order to capture the nuances of each performance, each action is carried out by a different one of 25 different people, and the environment is systematically changed for each actor in each action. Variations on the set include the following: outdoors (s1), outdoors but with a

scale change (s2), outdoors but with different clothes (s3), and indoors (s4). The ability of each algorithm to recognize actions independent of the background, the appearance of the actors, and the scale of the actors are put to the test by these variations [90].

3.2.6 HMDB-51

A new frontier in computer vision research, video recognition, and search are becoming increasingly important as nearly one billion videos are viewed daily online. While large, static image datasets with thousands of categories have received a lot of attention, human action datasets have lagged far behind. In this article, we present HMDB compiled from a wide range of media, primarily motion pictures but also including some data from publicly available sources like the Prelinger archive, YouTube, and Google Videos. There are a total of 6849 clips in the dataset, and they've been broken down into 51 different categories of action. There are five distinct kinds of action categories: Expressions like smiling, laughing, chewing, and talking are examples of general facial actions. Smoking, eating, and drinking is examples of masticatory facial actions. Cartwheel, clap hands, climb, climb stairs, dive, land on your back, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, and wave is all examples of general body movements. Body motions involving the use of an object: brushing hair, drawing a sword, dribbling a ball, playing golf, hitting a ball, kicking a ball, picking up an object, pouring, pushing, riding a bike, riding a horse, shooting a bow, firing a gun, swinging a baseball bat, swinging a sword, and throwing. Human interaction body motions include fencing, hugging, kicking, kissing, punching, and shaking hands [91].

4 Pre-processing Methodologies

Certain pre-processing techniques must be used before feeding data to a deep model to achieve satisfactory performance. Here are some common pre-processing techniques used:

4.1 Data Segmentation

Typically, the duration of activity exceeds the sampling rates of the sensors. That's why you need more than just a single sample from a sensor at a single point in time to accurately identify an event. As a result, the segmentation method needs to be used to analyse the collected signals rather than relying solely on a sample basis. Segmenting data allows for individual data points to be associated with a given task [92]. Segmenting windows by time, events, or actions are the three main types. By contrast, the event-driven windows method uses estimation techniques to partition sensor signals into event-based windows, while time-driven windows segmentation splits the signal into many consecutive windows of fixed-size time intervals. Finally, individual activity windows are identified through action-driven windows segmentation. These techniques are sensitive to the window size, despite the fact that they are useful for real-time applications and don't necessitate any pre-processing steps. Alternatively, to address the shortcomings of fixed-size sliding window methods, an adaptive sliding window segmentation approach for physical HAR using a triaxial accelerometer was introduced [93]. By analysing data from the sensor signal, the window size can be adjusted. Segmentation is a necessary step in video-based human activity recognition (HAR). It involves dividing the video into segments, each of which represents a single action. Segmentation can be performed manually or automatically. Manual segmentation is typically done by a human observer who watches the video and identifies the start and end points of each action. This can be a time-consuming and labor-intensive process, but it can be very accurate. Automatic segmentation methods use computer algorithms to identify the start and end points of actions. These methods can be faster and more efficient than manual segmentation, but they may not be as accurate. The best method for segmentation depends on the specific application. For example, if accuracy is critical, manual segmentation may be the best option. However, if speed and efficiency are more important, automatic segmentation may be a better choice. In the case of video-based HAR or biosignal collection supplemented by the video camera(s) recording the whole process, the acquired dataset will be segmented by dedicated persons relying on the video. This is because manual segmentation is typically more accurate than automatic segmentation for this type of data [94, 95].

4.2 Data Scaling

Unless the raw attributes have meaning in the original domain, raw data are usually not sufficient for machine learning methods [96]. Because deep models typically perform best on inputs with low values, we often need to rescale the raw data to a certain range to make it usable by the models (e.g., between 0 and 1). It is computationally expensive and could cause overflow on digital computers if a model is trained with excessively large input values [97]. Normalization and standardisation are two common methods of scaling. The deep learning algorithms excel at processing time-series signals for the purposes of feature extraction and classification because of the advantages of local dependency and scaling invariance [34]. As a result, there has been a recent uptick in interest in using deep learning models

like CNN, LSTM, and hybrid models for human activity recognition.

4.3 Data Denoising

It is common for sensor data to contain artefacts like errors in calibration and operation, problems with placement, background noise, and concurrent uses. As a result, the generated noise can be reduced with the help of data pre-processing techniques. Low-pass filter, mean filter, linear filter, wavelet filter, and Kalman filter[98] are common denoising techniques. In their analysis, Ignatov et al. found that background noise was present during data collection. So, they used a method called singular value decomposition to cut down on the background commotion [99]. Pre-processing techniques for sensor data have been proposed in other studies [100]. The authors generated a new signal by incorporating white noise as random noise into the desired signal for each input signal. White noise dampens the clamour of humans' kinetic actions while keeping low-frequency elements intact.

4.4 Data Label Encoding

Categorical labels are typically used to describe activities like walking and shopping; however, deep models require all input data to be numeric, so this eliminates them as potential sources of information. If we assign an integer value to each label, we can easily accomplish this. Since the model may try to learn an ordering relationship in categories, integer encoding may not perform well. Common practise suggests encoding the label with a single "hot" character instead [101]. One hot encoding relies on an identity matrix whose size is proportional to the number of activity types. Activites are represented in the table by rows, each of which contains exactly one element with the value 1.

4.5 Feature Selection

Selecting relevant features for classification algorithms to use is known as feature selection [102]. Furthermore, it simplifies high-dimensional spaces and saves time by discarding superfluous details. In representation learning, models focus on analysing data to extract a good feature set as an alternative to traditional feature selection [103]. In order to select a subset of features, filtering techniques use the correlation coefficient to rank the original features, taking advantage of the variables' and features' inherent characteristics. The extracted feature subset is not evaluated by a classifier in filter-based feature selection. As many classifiers are used to evaluate the selected subsets in wrapper methods, it has been shown to achieve better performance than filter methods [104]. Conversely, embedded methods pick the best feature subset by determining the optimal weights of a function that has shown to produce excellent results in the past. While wrapper approaches are limited to univariate problems, embedded methods can be applied to multiclass and regression issues.

4.6 Data Transformation

Before using the input data to train a deep model, it is often helpful to perform certain transformations on the data. The input data's correlations can be lowered with the help of transformations. As a generalisation of standardisation, "whitening" (also known as "sphering") is a linear transformation that returns a vector with the unit diagonal white covariance instead of the original vector's covariance. To help deep learning models learn features more quickly and accurately, PCA whitening is a common preprocessing technique [105, 106]. In order to reduce the input data's correlations, ZCA whitening is another common preprocessing method. There is a connection between PCA whitening and ZCA whitening [107], and the ZCA whitening matrix can be obtained by multiplying the PCA whitening matrix by an orthogonal matrix. Since a lot of HAR sensor data (like accelerometer readings) is typically a time series, spectrogram analysis could be useful for capturing variations in the input data. The Fourier transform [108] or the wavelet transform [109] can be used to create spectrograms, which are time-varying representations of the frequency spectrum of the input signal.

5 Deep Learning (DL) Techniques for Sensor-Based HAR

Over the past few years, DL methods have consistently outperformed traditional ML methods on a wide variety of HAR tasks. Increases in both the quantity and quality of available data, the speed with which computing hardware can process that data, and improvements in the underlying algorithms are all major contributors to deep learning's success. The proliferation of freely available datasets online has facilitated the rapid development of complex models by researchers and developers. The advent of graphics processing units (GPUs) and field-programmable gate arrays (FPGAs) has greatly reduced the length of time required to train elaborate and large models [110, 111]. Finally, technological progress in optimization and training methods has speed up the learning procedure. Here we will go over some of the deep learning-based sensor-based HAR initiatives that have been made.

5.1 Deep Belief Networks (DBNs)

Among the earliest and most promising deep models for HAR, DBNs stand out. The authors of [112] present a DBN-based method for activity detection in voice signals. DBNs with four hidden layers were used by the authors of [113] to detect routines in a smart home. The authors of [114] introduced the DBN method for facial expression recognition, which consists of three separate layers. Some researchers used EEG data in conjunction with a DBN to create a system for identifying feelings (see [115]). DBNs, unlike other directed generative models, are able to infer the states of hidden units with just a single forward pass [116]. The obtained weights can be used to initialise any number of feature-detection layers in a classification network. DBNs have existed for some time, but they are rarely employed due to the challenges inherent in both directed and undirected models, such as the inability to perform inference to marginalise out the hidden units and the inability to determine the partition function of the top two layers [117].

5.2 Deep Boltzmann Machines (DBMs)

The factorial nature of the conditional distribution over a single DBM layer is made possible by the fact that a DBM can be represented as a bipartite graph [119]. DBMs are easier to implement and implement, but they provide more accurate posterior approximations [117]. The log probability of the training data has a set of variational bounds [120] that cannot be explicitly optimized in DBNs. DBMs are distinct in that all the hidden units in a single layer are conditionally independent given the other layers, making it possible to optimize the variational bounds. Some DBM-based works have been completed for HAR so far. When it comes to recognizing gestures, transportation modes, and indoor/outdoor activities, Bhattacharya and Lane [121] used a three-layer model composed of RBMs. For automatic activity recognition, Plötz et al. [122] presented a DBM-based method for learning features from data. To perform a variety of audio sensing tasks (such as ambient scene analysis, emotion recognition, stress detection, and speaker identification), Lane et al. [123] proposed a deep model made up of three layers of RBMs. As for mobile HAR, Radu et al. [124] presented a DBM learning method that incorporates multiple modalities.

The fact that a DBM can be represented as a bipartite graph [119] allows for the factorial nature of the conditional distribution over a single DBM layer. DBMs offer more precise posterior approximations [117] while being simpler to implement and use. Because of these variational bounds [120], DBNs are unable to perform an explicit optimization of the log probability of the training data. DBMs are unique because it is possible to optimise the variational bounds because all the hidden units in a single layer are conditionally independent given the other layers. Existing work for HAR makes use of DBM. In [121], a three-layer RBM model was used for recognising gestures, transportation modes, and indoor/outdoor activities. In [122], the authors presented a DBM-based method for learning features from data that could be used for automatic activity recognition. It was proposed in [123] that a deep model consisting of three layers of RBMs be used to perform a wide range of audio sensing tasks, including ambient scene analysis, emotion recognition, stress detection, and speaker identification. In [124], the authors presented a DBM learning approach for mobile HAR that makes use of several different sensory modalities.

5.3 Autoencoder

The mean squared error along with KL divergence are the two loss functions that are utilised the most frequently when training autoencoders. To classify accelerometer and gyroscope sensor data. For the purpose of exploring useful feature representations, the authors of [125] introduce an autoencoder architecture that makes use of both a sparse autoencoder and a denoising autoencoder. Using a freely available HAR dataset [42] hosted in the UCI repository, authors evaluated the efficacy of the principal component analysis (PCA), the Fast Fourier Transform (FFT). The results show that the stacked autoencoder provides the greatest improvement (7%) and the highest accuracy (92.16%). (compared to conventional methods using hand-crafted features). Another common use for autoencoders is the cleaning and de-noising raw sensor data [118, 126, 127], as noise is a pertaining issue with the wearable signals as they induce obstruction in the ability to learn patterns from them, as shown in Fig. 5 . In [126], Mohammed and Tashev looked into the feasibility of using sensors sewn into everyday

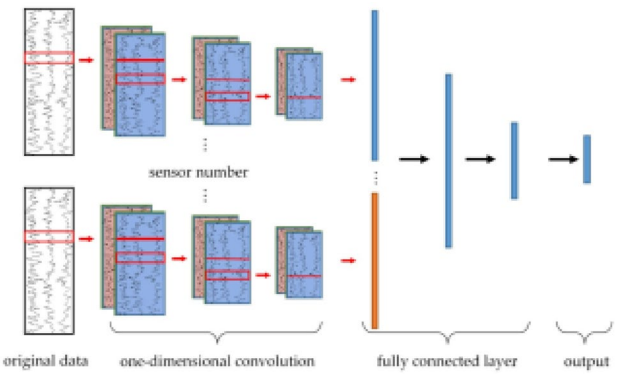


Fig. 5 Recognizing human actions through a convolutional neural network (CNN). One sensor dataset per convolutional network. Next, the fully-connected layer receives the combined convolutional network outputs as input [118]

garments for HAR. However, they found that the mean signal-to-noise ratio (SNR) of sensors worn on loose clothing is low due to the presence of numerous motion artefacts. Using the UCI dataset, Gao et al. [42, 118] investigate the potential of stacking autoencoders for de-noising raw sensor data in order to enhance HAR. As soon as the denoised signals are ready, we use LightGBM (LBG) to categorize the activities. Authors in [38] present a architecture i.e., Guardian-Estimator-Neutralizer (GEN) that identify tasks while protecting the identities of participants based on their gender. GEN's goal is to filter out any potentially sensitive information from the raw data and produce a new set of features. Data is transformed into an inference-specific representation by the Guardian, which is built using a deep denoising autoencoder. By making educated guesses about what parts of the transformed data are sensitive and what parts are not, the Estimator guides the Guardian. We try to identify an activity without revealing a person's gender so as to protect their anonymity. As an optimizer, the Neutralizer aids the Guardian in arriving at a transformation function that is nearly optimal. To gauge how well the proposed framework performs, it is tested on both the existing publically available MobiAct [128] and a new dataset called MotionSense.

5.4 Convolutional Neural Networks

Convolution layers, pooling layers, detector layers (like ReLU layers), and fully connected layers are the four standard types of layers in a basic CNN. A complex CNN can be constructed by stacking these layers. Due to CNNs' impressive results across many applications, especially in image classification, many different types of CNNs have been proposed [129]. CNN, one of the earliest and most successful deep learning models, has also seen extensive use in sensor-based HAR. Using a CNN, Ronao and Cho [130] were able to distinguish between six distinct locomotion activities and show that their method was superior to MLP, Naive Bayes, and SVM. Authors in [131] used a distributed CNN and analysed the effect of sensor location (such as the legs, body and arms) on activity recognition to identify some intermediate-level activities (e.g., opening a drawer). Improving performance, the authors of [132] combined handcrafted time and frequency domain features with features generated from a CNN, called HAR-Net, to classify six locomotion activities from smartphone accelerometer and gyroscope signals. The authors of [133] have proven that a shallow three-layer CNN network can successfully recognise activities occurring locally on a device running on a limited amount of system resources. Layers of the network are convolutional, fully connected, and softmax. Similarly, authors in [134] and in [135] employed a modest layer count (four layers). A crucial decision in training CNNs is the selection of the loss function to be used. Cross-entropy is typically used for classification tasks and mean-squared error for regression. In [136] authors propose the first shallow CNN to consider cross-channel communication, in contrast to traditional CNN models which process input data by extracting and learning channel-wise features independently. Different channels within the same layer work together to isolate specific features from sensor data. A convolutional neural network (CNN) was developed by authors in [137] to identify common actions and gestures. The authors of [138] presented a convolutional neural network (CNN) that uses partial weight sharing and full weight sharing for HAR, trained on multimodal data (such as accelerometers and gyroscope sensors). Using information gathered from mobile devices' sensors, Zeng et al. [134] developed a convolutional neural network (CNN) for HAR. Using information gleaned from an accelerometer, a magnetometer, a gyroscope, and a barometer, authors of [16] proposed using convolutional neural networks (CNNs) to identify locomotion activities.

5.5 Recurrent Neural Network (RNN)

Several RNN-based models, such as Continuous Time RNN (CTRNN) [141], Independently RNN (IndRNN) [142], and Personalized RNN (PerRNN) [143], have been proposed by researchers to enhance the effectiveness of RNN models for human activity recognition. In contrast to earlier models that only took into account a single dimension of time-series input, as shown in Fig. 6, the CNN layer of the CNN + RNN model developed by the authors in [139] receives stacked multisensor data from each channel for fusion. To solve the domain adaptation issue brought on by session-to-session, sensor-to-sensor, and subject-to-subject variations, Ketykó et al. [144] employ a recurrent neural network. Residual networks have the advantage of being much easier to train than convolutional networks because gradients can pass through the addition operator more directly. Gradients are not hindered by residual connections, and the layer outputs may be improved with their help. The accuracy of a model trained with raw accelerometer and gyroscope data is improved from 80% to 92% by combining LSTM with batch normalisation, as proposed in [145], while the harmonic loss function is proposed in [146]. Activity recognition with data from multiple wearable sensors is proposed in [147], where a convolutional neural network (CNN) and long short-term memory model (LSTM) are suggested. RNNs have been widely used for HAR because activity recognition can be viewed as a sequential problem. Although RNNs can be used as generative models [148], they are more commonly thought of as a type of discriminative model. The discriminative RNNs are employed in the field of HAR. It is trained using supervised methods, which aim to reduce a cost function associated with network output and its associated label. Using a deep recurrent neural network (DRNN) made up of LSTMs,

Murad and Pyun [149] were able to recognize actions from a variety of publicly available datasets. They proved that the one-way DRNN performed better than both the two-way and the cascaded versions. The authors of [150] also used an LSTM-based DRNN for human activity recognition based on acceleration signals. The authors of [151] used a HAR method to show that LSTM networks working together produce better results than those working alone. RNNs are typically fed raw time series data from IMUs and EMGs [152, 144]. Besides the raw time series data [153], RNNs usually take in both raw time series data and unique features as inputs. Similar performance for gesture recognition was achieved when training an RNN on raw data or with simple custom features, as demonstrated by the authors of [154].

5.6 Generative Adversarial Networks

As collecting labelled data in HAR is difficult and expensive, GANs and their variants have great potential for widespread use in HAR but have only been used in a small number of works so far. Using GANs could drastically lessen the time spent on gathering labelled data [155]. While GAN has seen a lot of success in a variety of settings, the initial implementation has a few issues, including gradient vanishing, lack of diversity, and unstable training. For this reason, numerous improvements upon the first GAN [156] have been proposed. In tests, GAN has proven its ability to produce synthetic sensor data that is both realistic and well-balanced. Using GANs with a tailored network, Wang et al. [157] generated synthetic data from the publicly available HAR dataset. In addition, the researchers improved performance by balancing out the initial imbalanced training set through methods including oversampling and the incorporation of synthetic sensor data into the training process. They created genuine data of various pursuits in [158, 140] as shown in Fig. 7. To combat the dramatic performance drop when pre-trained models are tested against unseen data from new users, GAN has been widely applied in transfer learning in HAR due to its ability to generate new data. Because it would be impractical to collect data for each new user, [159] is an effort that used a GAN to perform cross-subject transfer learning for HAR. Recent studies on sensor-based HAR that employed deep learning methods are summarised in Table 3.

6 Deep Learning Methods for Image/Video-Based HAR

Human action is a set of coordinated movements that occurs over space and time. The literature provides a wide variety of definitions of action [161–163]. Here, "an action" refers to a single movement or a series of movements carried out by one or more people. Individual actions are seen as snapshots of human dynamics, each of which begins and ends at a specific point in time. Given an image sequence containing one or more actions, human action recognition attempts to assign an action name to each frame or sequence of frames. Recognition of human actions is typically a multi-step process, with the first two steps focusing on human detection and segmentation. The goal of those tiers is to learn how to detect and isolate the ROIs in the video that corresponds to still or moving human figures. The next level involves extracting the visual information of actions and representing it using features. Then, the action recognition system uses these features to make sense of everything that's happening. Thus, it is possible to view action recognition as a classification problem based on the features used. Human action recognition systems have evolved over the years, with earlier attempts relying on frame-by-frame analysis methods like shape matching techniques [164], and more recent studies focusing on Spatio-temporal analysis of human motions.

Human action recognition is just one area where multiple DL architectures have been proposed and proven to achieve state-of-the-art results. Here, we outline the most pivotal DL architectures for recognising human actions.

6.1 Multi-stream Network

Two-stream convolutional neural networks are a relatively new but increasingly popular method, with the first stream dedicated to the spatial features of a video and the second to its temporal aspects. The spatial stream does action recognition in the form of sparse optical flow, while the temporal stream does action recognition from still images [165]. In the end, late fusion is used to combine the two streams; this approach to action recognition has been shown to be superior to handcrafted approaches. The procedure was developed by Carreira et al. [166] and implemented with Inception-V1. Before reaching Inception-final V1's average pooling layer, the spatial and temporal streams travelled through the network's 3D convolutional layer. The goal of the ConvNet stream of motion is to differentiate between actions that involve similar pose changes but differ in velocity or orientation. Consider the differences between those two common forms of transportation: walking and running, and pushing and pulling. These kinds of motions are managed by the movement ConvNet. The source of the current is the geometric centre of human-inhabited areas. Finally, the hinge loss classifier is applied to all three streams to reliably categorise people's actions. There was also a parallel effort in [167]. The Spatiotemporal Distilled Dense-Connectivity Network (STDDCN) model was proposed for HAR from video data by Hao et al. [160], as shown in Fig. 8. In part, this network was influenced by [168], which employs a similar strategy of knowledge distillation and dense connectivity.

Table 3 Synopsis of some of the recent studies on sensor-based HAR using deep learning techniques (*Ac.* Accelerometer, *Gyr* Gyrocscope, *Mag.* Magnetometer)

Name	Year	Methodology	Input	Evaluation	Limitation	Findings
Gochoo at el. [48]	2021	SGD	Ac., Gyr., Mag. data	Accuracy: 83.18%, 94.16%, 92.5%	Complex activities missing	Information is extracted using a feature-based, hierarchical approach. The time, wavelet, and time-frequency domains are all examples of these features
Gil-Martín at el. [47]	2021	CNN	Inertial signals	Accuracy: 94.27%, 84.46%	Analysis of sequential information is missing	Classifying the state of the system at each time window requires the model to learn features from the signal spectra and additional fully connected layers
Ronald at el. [45]	2021	Inception-ResNet	Group of signals	Loss: 0.1761, 0.4322 0.479, 0.2271	Subject dependent method	Model designed to achieve a higher predictive accuracy for devices with a restricted amount of computational resources
Mahmud at el. [49]	2020	self-attention Neural network	Sensor readings	Macro F1-Score: 0.96, 0.42, 0.67	Window-size classification on recently published models is missing	The model avoids recurrent architectures in favour of various attention mechanisms to produce a higher-dimensional feature representation for classification
Xia at el. [50]	2020	LSTM-CNN	Segmented raw data	F1 Score: 95.78%, 95.85%, 92.63%	Unable to capture temporal information from sequential data	After convolution, the fully connected layer was removed and replaced with a global average pooling layer (GAP) to bring down the number of model parameters
Khatun at el. [52]	2022	CNN-LSTM	Ac., Gyr., linear acceleration data	Accuracy: 99.93%, 98.76%, 93.11%	Real-time classification	The self-attention algorithm equips the model with the resources it needs to improve the system's predictive abilities
Challa at el. [51]	2021	CNN-BiLSTM	time-series data	Accuracy: 96.05%, 96.37%, and 94.29%	Accuracy for complex activities is low	The proposed model performs automatic feature extraction using raw sensor data with only a minimal amount of prior data preprocessing
Qin at el. [58]	2020	ResNet	Ac., Gyr. data	Macro F1-Score: 96.74%, 98.5%	Evaluation done on relatively smaller dataset	By fusing two networks and training the heterogeneous data with pixel-wise correspondence, a fusion residual network is adopted
Sa-nguannarm at el. [59]	2021	CNN, LSTM	Ac. data	Precision: 0.974, 0.975	Model is unable to recognise complex activities	The model fixes serious problems like data leakage and insufficient storage space

Table 3 (continued)

Name	Year	Methodology	Input	Evaluation	Limitation	Findings
Atkinson at el. [60]	2020	CNN	Ac. data	Recall: 0.927, 0.538	Results achieved on generalised synthetic dataset	The main goal of this model is to detect and eliminate label noise by adapting existing methods for single instances to time series
Atkinson at el. [62]	2020	CNN	Ac. data	Accuracy: 0.97, 0.81, 0.78, 0.93	Backward locking phenomenon	The model employs a CNN with a local loss function to facilitate HAR in the field of pervasive and wearable computing

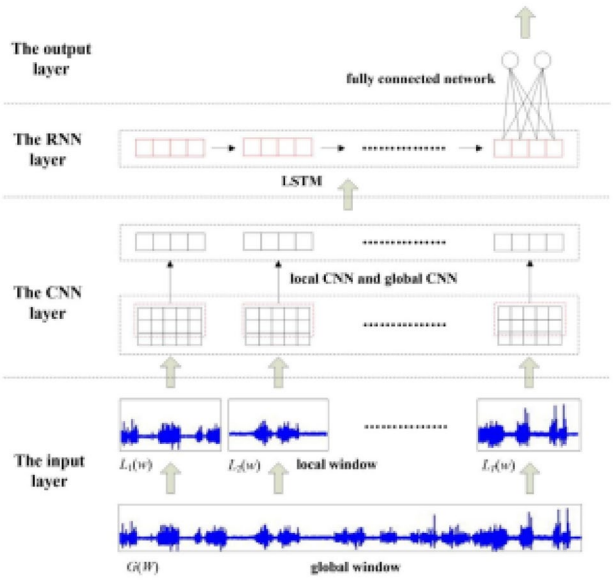


Fig. 6 The structure of the HConvRNN network. [139]

The goal of this model is to investigate the interplay between different features and streams of visual information, such as appearance and motion. Spatio-temporal feature relationships at feature representation layers are strengthened through the dense network in a more explicit manner. Both streams can talk to the final layers thanks to knowledge distillation within and between them. To mitigate the high computational cost of accurately computing optical flow, the authors of [169] attempted to simulate the knowledge of the flow stream during training in order to avoid using optical flow during testing. This was done so that optical flow wouldn't be used in any of the tests. One network is trained using optical flow data, while the other network is trained using motion vectors extracted from compressed videos with no additional computation required in [169]. For this purpose, the teacher model's generated soft labels were used to supplement the training of the student network and thus facilitate knowledge transfer. Unlike [169], the trainable flow layer proposed by Piergiovanni and Ryoo [170] can detect motion without computing optical flows. STDDCN is able to acquire high-level ordered spatiotemporal features thanks to its novel architecture. From RGB, Depth, and skeleton joint positions, [171] propose a fusion method with two 3D Convolutional Neural Networks (3DCNN) and a Long Short Term Memory (LSTM) network. To distinguish between activities performed with one or more limbs, the authors in [172] proposed a 2D convolutional neural network (CNN)-based algorithm.

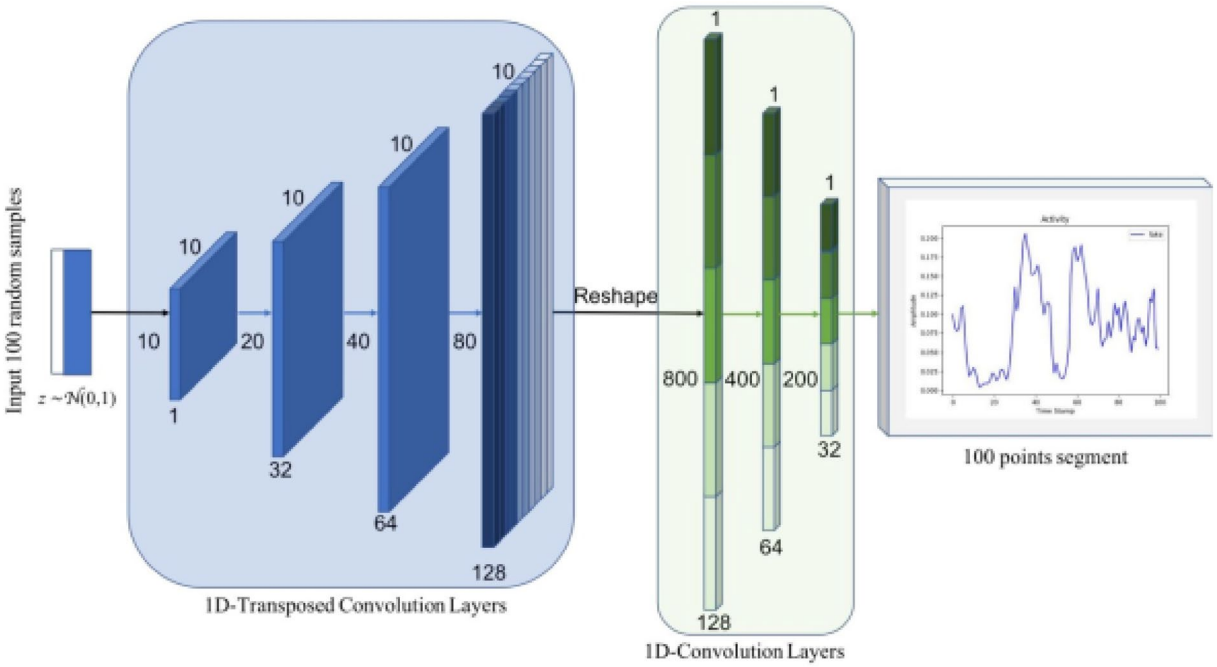


Fig. 7 Structure and organization of ActivityGAN's generator module [140]

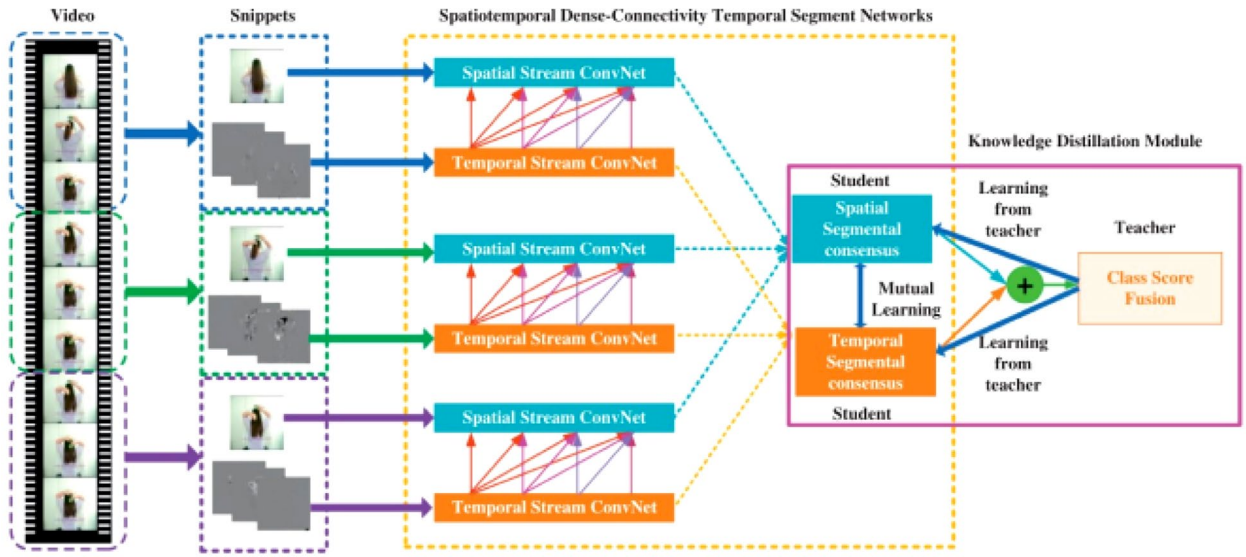


Fig. 8 STDDCN's fundamental pipeline, used for identifying activity in videos [160]

6.2 Sequential Network

Sequential networks based on convolutional neural networks (CNNs) have a unified data pipeline (either a single stream or a stacked one). Comparable to traditional convolutional networks, this 3D ConvNet architecture takes a more organic approach to video modelling by incorporating Spatiotemporal filters. Because of the unique properties of this network architecture, hierarchical representations of Spatio-temporal data can be built from the ground up[165].

New architecture for a two-stream convolutional neural network using long-short-term spatiotemporal features is presented by Varol et al. [173]. (LSF CNN). The goal of this network is to speed up and improve upon the process of recognising human action from video data. Two smaller networks were combined to form this larger one. An initial

LT-Net, or long-term spatiotemporal features extraction network, takes the RGB frames as inputs and processes them over time. To outperform a model that uses three independent CNN streams [174, 175], Zheng et al. [176] introduce a cross-modal architecture for human activity recognition. The first step in this model is to extract the information from the various modalities and map it into a shared subspace. The features are then combined after they have been aligned, creating representations that are correlated, consistent, and complimentary. The learned features are used as input to the classifier in the final layer, which is responsible for the actual action recognition. To train a regular 3D convolutional neural network (CNN), the MARS method [177] suggests using two different learning strategies. It functions on one RGB frame that is a direct analogue of the video stream. As a result, it reduces the cost of computing optical flow during testing. In order to evaluate the performance of a standard 3D convolution network, Yang et al. [178] propose an Asymmetric 3D CNN model that employs asymmetric single-direction 3D convolution architecture [179]. The capability of feature learning is improved in this model by the Asymmetric 3D convolutions network. Incorporating multi-scale 3D convolution branches, this model is a collection of local 3D convolutional networks or MicroNets. An asymmetric 3D-CNN deep network is built with these MicroNets to efficiently carry out the action recognition task. Authors of Principal Component Analysis Network (PCANet) propose a method for choosing a subset of frames from each action [180]. Concurrently, a feature vector is computed for each frame based on the PCANet's training data. The Whitening Principal Component Analysis (WPCA) algorithm is then applied to the combined feature vectors to reduce their dimensionality [181]. To enable HAR on videos with poor spatial resolution, the authors of [182] proposed two video super-resolution methods to produce high resolution videos. These 4K videos were used as input into a spatial and temporal model to determine action category. By learning different types of information (e.g., spatial and temporal) from the input videos through separate networks and then performing fusion to get the final result, two-stream 2D CNN architectures allow traditional 2D CNNs to efficiently manage the video data and achieve high HAR accuracy [11]. When it comes to effectively modelling the temporal information at the video level, temporal sequence modelling networks like LSTM can make up for the inefficiencies of these architectures.

6.3 RNN-LSTMs

RNN-LSTM's main proposition is in their modeling of the long-term contextual information of temporal sequences. This benefit makes RNN LSTM one of the best sequence learners for time-series data, including visual information of human action. Because RNNs' hidden layers contain recurrent connections, they can be deployed for temporal data analysis. Due to the vanishing gradient problem, however, the vanilla RNN has difficulty modelling the temporal dependency over longer time periods. In order to model the long-term temporal dynamics of video sequences, the majority of modern methods employ gated RNN architectures like LSTM [183–185]. The LSTM network's performance on the human action recognition task has been shown to be highly robust by Grushin et al. [186] using the hand-crafted feature HOF [187]. Evidence supports CNNs' ability to learn features from unlabeled data. For this reason, the works of Singh et al. [188], Wu et al. [189], Baccouche et al. [190], Ng et al. [191], Li et al. [192], Wang et al. [193], and Chen et al. The primary goal of these works is to extract motion features from input video using industry-standard CNN models like AlexNet [194], VGGNet [175], or GoogLeNet [195]. Next, an RNN-LSTM network is fed the results from the CNN so that sequences can be labeled with previously learned features. Even though RNN-LSTMs have been proposed in multiple studies as a comprehensive learning framework for skeleton-based action recognition, the aforementioned work only employs them for sequence classification. Improvements in HAR performance for LSTM-based frameworks can also be attributed to the incorporation of attention mechanisms, such as spatial attention [196, 197], temporal attention [198, 199], and combined spatial and temporal attention [200, 201]. Research conducted by Du et al. [202], Li et al. [203], and Liu et al. [204]. Using depth-sensor-provided 3D human skeleton sequences, RNNLSTMs are able to directly learn motion features and classify them into categories. The efficiency of these strategies is illustrated by experiments on state-of-the-art datasets. Action recognition using multi-source data was also investigated by Mahasseni et al. [205], who employed a parallel architecture. Unsupervised training is used to teach an RNN-LSTM how to interpret 3D sequences of human skeletons. Simultaneously, a CNN-equipped RNN LSTM is trained on 2D video. The system's performance is enhanced by comparing the results.

6.4 GNN or GCN

As a result of their expressive power, graph structures have recently sparked a renewed interest in employing learning models for analysis of graphs [11, 206]. Skeleton data cannot be adequately modelled by using RNNs to process a vector sequence or CNNs to process 2D/3D maps of the body's joints, as these representations do not capture the complex spatio-temporal configurations and correlations of the joints. This provides support for the idea that topological graphs are a more apt representation for the skeletal data. Many GNN and GCN-based HAR methods [207, 208] have been

proposed because the skeleton data can be represented as a graph with edges and nodes.

In more recent times, research into GCN-based HAR has started to pick up some steam [209–212]. Using GCNs as a basis for a skeleton-based HAR system. Spatial-temporal GCNs, also known as ST-GCNs, were first presented to the public by the authors in [208]. These GCNs have the ability to automatically learn both spatial and temporal patterns from skeleton data. We were able to generate action representations with robust generalisation capabilities for HAR by first estimating pose from the input videos and then processing the data using spatio-temporal graphs. This allowed to generate action representations for HAR. Because implicit joint correlations have been overlooked in previous works [208], the authors of [213] proposed an Actional-Structural GCN (AS-GCN), combining action links along with structural links into a generalised skeleton graph. The reason for this is that an AS-GCN combines actional links and structural links into one. High-order dependencies were represented by structural links, and latent dependencies on actions were captured by actional links. Peng et al. [214] used a neural architecture search scheme to decide on their GCN's architecture so that they could more effectively investigate the implicit joint correlations. In particular, they used a Chebyshev polynomial approximation to broaden the search space, enabling the implicit capture of joint correlations based on multiple dynamic graph sub-structures and higher-order connections. Further, integrated context information was used to model long-range dependencies, as shown in [215]. With a cross-domain spatial residual layer and a dense connection block based on ST-GCN for learning global information, the authors of [216] were able to successfully capture the spatio-temporal information. Skeleton and node trajectories from a skeleton sequence are fed to a spatial graph router and a temporal graph router, respectively. Using a skeleton-joint generative adversarial network (ST-GCN), the authors of [217] were able to classify newly generated skeleton-joint connectivity graphs. With the intention of developing a reliable feature extractor. The authors of [218] combined a multi-scale aggregation scheme that eliminated entanglements with a spatial-temporal graph convolutional operator called G3D.

In [220], authors introduced joint semantics at a high level for HAR. The mechanisms of attention were used in [219, 221] to extract global dependencies and information with discriminatory power as shown in Fig. 9. To further reduce the computational costs of GCNs, the authors of [222] developed a Shift GCN that swaps out regular graph convolutions for shift graph operations and lightweight point-wise convolutions. In this vein, the authors of [223] proposed a multi-stream GCN model that merges different types of inputs like joint positions, motion velocities, and bone features early on, and then uses distinct convolutional layers and a compound scaling strategy to significantly reduce redundant trainable parameters while increasing the model's capacity. By contrast, the symbiotic GCNs proposed by the authors in [224] can do both action recognition and motion prediction at the same time. For simultaneous performance of action recognition and motion prediction, the proposed Sym GNN utilises a multi-branch multi-scale GCN. This allows for a mutually beneficial relationship between the two pursuits.

6.5 Stacked Denoising Autoencoders (SDAs)

For deep learning, SDA is a must-have tool. Vincent et al. [225] first introduced this idea; it is an extension of a classical autoencoder [226]. The weights of an SDA are tuned with a back-propagation algorithm, and the architecture is built by stacking multiple autoencoders together [227]. Each autoencoder undergoes a greedy "unsupervised pre-training" procedure in which it is trained incrementally. A supervised learning algorithm for recognition tasks will take the SDAs's output as its input representation after it has been learned. In 2007, Huang et al. [228] presented the first successful application based on the encoder-decoder model for object recognition tasks. A few years after the publication of Huang et al.'s model [228], Baccouche et al. [229] proposed an autoencoder-based solution for learning sparse spatio-temporal features. When compared to methods employing hand-crafted features, experimental results on the KTH [90] and GEMEP-FERA datasets [230] were superior. Furthermore, autoencoder-based methods have been proposed by [231–233]. For instance, authors in [233] used Kinect [234] skeleton data to build a 3-layer SDA architecture for human action recognition. Similar research using an SDA model to learn skeleton feature for human body pose classification was conducted by Budiman et al. [231]. Xie et al. [240] used an SDA architecture with three hidden layers to learn contour features from a single depth frame in order to recognise human action. In [232], Hasan et al. presented an autoencoder-based framework for continuously learning human activity models from streaming videos. First, a sparse autoencoder will take a streaming video with some annotated activities and extract space-time interest points (STIP) [235] from the motion. However, SDAs have one major drawback when dealing with massive datasets: they take an extremely long time to train. Two 3D convolutional neural networks (CNNs) were used as the generator and joint discriminator in an adversarial framework presented by Mehta et al. [236]. The thermal data and optical flow were fed into the generator network, and the joint discriminator then attempted to tell the reconstructed data apart from the real. Recent research has focused on using deep learning to extract HAR from the CSI signal. Discriminative features for a deep sparse auto-encoder can be learned from CSI streams, as proposed by authors in [237]. With the CSI

signal converted to radio images, the authors of [238] fed them into a deep sparse auto-encoder to learn discriminative features for HAR. A novel variant of SDAs, dubbed "mSDA," was proposed by Chen et al. [239] to get around this restriction. On the same dataset, mSDA was shown to achieve parity with SDA's performance while requiring 450 times less time to train. Utilizing the mSDA, Gu et al. [240] trained a mSDA network for multi-view action recognition. In order to generate features for each camera view, a mSDA is first trained using all of the available camera views. The collected features from each camera view are then combined into a single integrated representation that can be fed into a classifier. The state-of-the-art recognition performance was demonstrated by testing the model on three benchmark multi-view action datasets. Table 4 provides a summary of some of the most recent research on image/video-based HAR that made use of deep learning techniques.

7 Performance Metrics

Accuracy, precision, recall, F1-score, confusion matrix, and accuracy/loss are the algorithm evaluation indicators that were utilized in this experiment. The following are some definitions that pertain to both of the aforementioned classification issues. Some common performance metrics used to assess the efficacy of HAR models is listed below.

7.1 Accuracy and Error Rate

The percentage of correct predictions relative to the total number of data examples is a common metric for evaluating a classification system's efficacy. This is how it is defined more specifically:

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{1}$$

Related to accuracy is the error rate, which measures how many incorrect predictions were made relative to the total number of data examples. following is a description of it:

$$\text{error } = \frac{fp + fn}{tp + fp + tn + fn} = 1 - \text{ accuracy.} \tag{2}$$

It is important to keep in mind that accuracy and error rate are not appropriate measures to use in situations where the data are very imbalanced (Fig. 10).

7.2 Precision and Recall

Precision and recall are another popular pairing of classification metrics. To calculate accuracy, we divide the number of confirmed positive cases (human presence) by the sum of all confirmed positive cases.

$$precision = \frac{tp}{tp + fp} \tag{3}$$

The recall rate is calculated by dividing the number of true positive predictions by the sum of all true positive predictions.

$$recall = \frac{tp}{tp + tn} \tag{4}$$

Precision tells us how well a model does in terms of false positives, while recall tells us how well it does in terms of false negatives [242].

7.3 F-Measure

It may be challenging to assess the impact of each model parameter using both precision and recall. One way to address this issue is with F-measure, which takes the harmonic mean of the two metrics. Specifically, it is described as follows:

$$F - measure = 2 \cdot \frac{\text{precision } \cdot \text{ recall}}{\text{precision } + \text{ recall}} \tag{5}$$

7.4 True Positive Rate

When calculating a test's sensitivity, one uses the true positive rate (*tpr*), which is the percentage of positive cases that were correctly identified relative to the total number of positive cases.

$$tpr = \frac{tp}{tp + fn} \tag{6}$$

7.5 False Positive Rate

The false positive rate, also called the fall-out rate, is the percentage of false negatives relative to the total number of true negatives.

$$fpr = \frac{fp}{fp + tn} \tag{7}$$

7.6 ROC

Visualizing the performance of a classifier can be accomplished with the help of a ROC graph, that shows how the true positive rate is related to the false positive rate as

Table 4 An overview of recent research into Deep Learning-based image/video-based HAR (*S* skeleton, *D* depth, *IR* infrared, *Au* audio, *Ac* acceleration, *Gyr* gyroscope)

Name	Year	Methodology	Input	Evaluation	Limitation	Findings
Jaouedi at el. [73]	2020	GRNN	video	Accuracy: 96.3%, 89.01%, 89.3%	Training time for video classification	Sequential data and video classification are two areas where GNN's increased computational powers are being adopted
Zhang at el. [85]	2021	Transformer	Image	Top-1 Accuracy: 70.9%, 46.1%, 87.2%, 87.9%, 79.8%	Limited ablation study	Video transformers learn better video representations when co-trained with images (as single-frame videos) and when trained on a variety of video datasets with varying label spaces
Wu at el. [75]	2021	GAN	video	cMAP: 77.5%	Privacy leakage risk is core to the framework	To learn an anonymization transform for input videos, a new adversarial training framework is developed
Varshney at el. [74]	2021	Multi-stream CNN	Cropped frames	Accuracy: 84.3%, 85.1%	Poor performance for complex activities	spatial and temporal data factored into the model. Two different fusion schemes, the average fusion and the convolution fusion of the spatial and temporal stream are used
Cheng at el. [80]	2022	CNN	Cropped frames	Accuracy: 93.6%, 94.2%	Memory and computation cost	An innovative method for human action recognition in RGB-D by combining spatial-temporal data with cross-modality interactive learning
Dong at el. [81]	2020	GCNN	Body joints	Accuracy: 90.5%, 86.4%	Issues with object-related individual actions	High-order spatial and temporal features derived from skeleton data, including velocity features, acceleration features, and relative distance between 3D joints
Li at el. [82]	2021	GCNN + 3D CNN	Video clips	Top-1 Accuracy: 52.5%, 65%, 75.1%	Limited ablation study	Model represent and encode the unique patterns of each action in the videos, a new design of sub-graphs has been developed
Wang at el. [83]	2021	2D CNN	Cropped frames	Top-5 Accuracy: 84.1%, 91.6%, 94.4%	Fail to capture long-range temporal data for complex activity	Model includes an effective temporal module (TDM) by explicitly leveraging a temporal difference operator, and systematically evaluates its impact on both short- and long-term motion modelling
Chen at el. [84]	2021	Transformer	Motion vector, Audio	Top-1 Accuracy: 66.7%, 83.5%, 98.9%	Computationally expensive	Self-attention is factorised across spatial and temporal dimensions, and the model can handle a large number of spatiotemporal tokens extracted from multiple modalities

Table 4 (continued)

Name	Year	Methodology	Input	Evaluation	Limitation	Findings
Islam at el. [87]	2021	Encode-decoder + Graphical attention	Multimodal data	F-1 Score: 75.24%, 97.56%	Poor performance in real-world setting	Disentangling and extracting important modality-specific features that allow for feature interactions using a multi-modal mixture-of-experts model
Duhme at el. [86]	2021	GCN	Skeleton data	Accuracy: 94.5%, F-1 Score: 89.6%	Few activities misclassified	For multi-modal action recognition, the model integrates different types of sensor data into a graph that is then trained with a GCN model

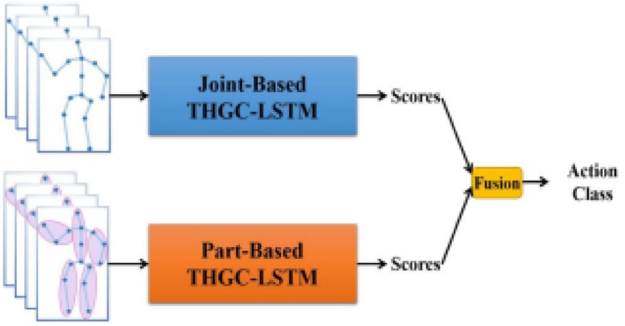


Fig. 9 Representation of the hybrid structure as a sum of parts and joints of human body [219]

shown in fig. 11. It offers more nuanced insights than simple numerical measures like accuracy or error rate.

7.7 AUC

Another performance indicator is the area under the ROC curve (AUC). Area under the curve (AUC) is a single scalar value that depicts classification performance in contrast to the ROC curve's two-dimensional representation [244]. The value of AUC can be anywhere between 0 and 1, and the area covered by guessing at random is half of that. The AUC value should be increased whenever possible for improved classification performance.

7.8 Confusion Matrix

Each column of the error matrix represents a classifier's prediction for the sample it was given. Indicating whether or not multiple categories are muddled can be done with ease thanks to the second matrix, where each row expresses the real category to which the version belongs. As shown in fig. 12, authors of [245] regularise the confusion matrix and convert the predicted value and the real value in the matrix into corresponding proportions so that the data sizes of the two datasets can be compared easily.

7.9 Accuracy and Loss Map

Reaction to fluctuating loss and accuracy while training a neural network model. Values for precision and error will be generated at the end of each epoch. The training of the network model can be visually reflected by plotting the accuracy diagram and the loss diagram. The trend can be used to judge the quality of the model's training, spot temporal outliers (like overfitting or underfitting), and fine-tune the model over time.

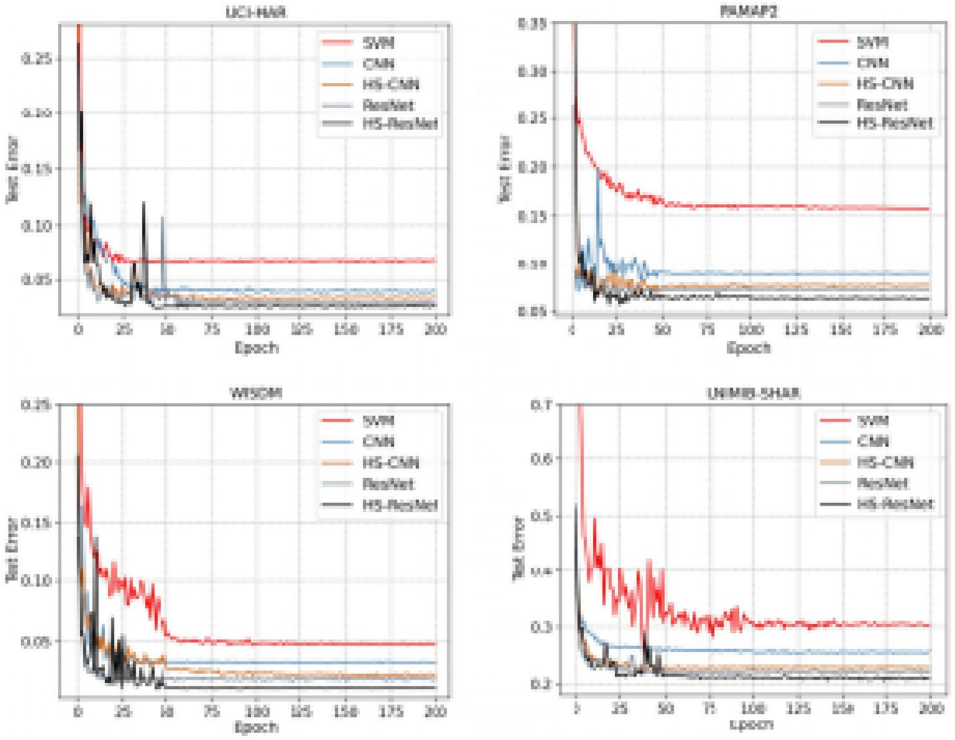


Fig. 10 Comparision of the relative error of various models on HAR data sets [241]

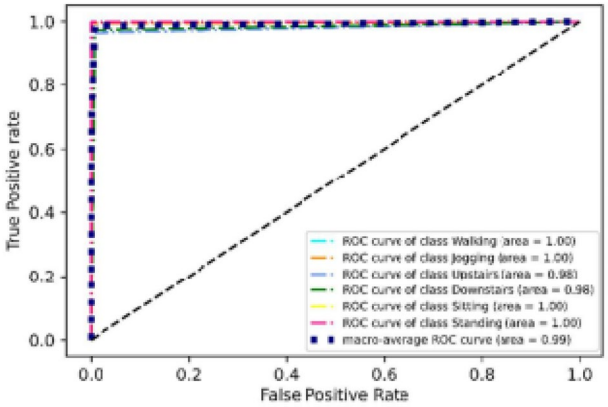


Fig. 11 Ensem-HAR [243] model's ROC curve on the WISDM dataset

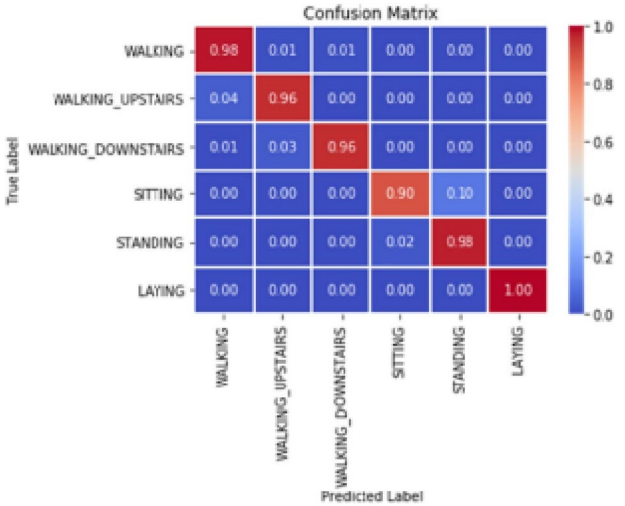


Fig. 12 Evaluation of the proposed model's [245] confusion matrix using data from the UCI-HAR dataset

8 Applications

This section discusses the significance of HAR in several different applications, including video analysis and retrieval, visual surveillance, HCI, education, medicine, and abnormal activity recognition.

8.1 Surveillance and Security

When an observer is not physically present at the recording location, they can still keep an eye on things with the help of a video surveillance system. Video can be analysed in real time to perform surveillance tasks, or it can be stored and analysed at a later time. Abnormal activity detection and gamer behaviour analysis are two other applications of

video surveillance technology. There are many recent developments in the field of user activity recognition, with surveillance being one of the most prominent examples. Recent studies [246] have concentrated on the use of cameras to record images or videos and the application of various algorithms to identify patterns of activity for the purposes of surveillance. In their paper [246], Deng et al. presented a hierarchical graphical model for identifying individuals and GAR in a surveillance scene that relies on deep neural networks. Problems with public safety, such as large-scale emergency management in the event of an evacuation, can be mitigated through the use of crowd monitoring. The effect of local interactions on the efficacy of evacuation was the subject of research presented by Braun et al. [247] . The authors of [248] described methods that use pedestrian behaviour to infer and visualise crowd conditions from GPS location traces. During city-wide mass gatherings, the method was used to detect developing, potentially critical crowd situations at an early stage. Because video surveillance is an important application for a variety of reasons related to security, it is essential to categorise activities as either typical or abnormal [249]. A technique for manually keeping an eye out for anomalous behaviour in crowded places like grocery stores, city squares, and college campuses was proposed by Mohan et al. [250]. PCA and CNN eliminate the need for laborious manual processes like false alarms and pinpoint the exact location of a video anomaly. PCA and SVM classifier are used to identify anomalous events in individual frames. Most surveillance-based security systems employ activity learning, monitoring, and recognition to address suspicious behaviour and identify potential dangers. For the purposes of surveillance and security monitoring, vision-based activity recognition employs the use of cameras. It has become popular due to its capacity for visually analysing patterns and trends [251]. Jiang et al. [252] proposed a method for real-time pedestrian detection that first extracts static sparse features using a fast feature pyramid, and then uses sparse optical flow to obtain sparse dynamical features between frames. Adaboost utilises a combination of these two kinds of features to make accurate classifications. The best experimental results were obtained on the TUD dataset. Automatic tracking and detection of criminal or brutal activities in videos was proposed by Basha et al. [253] using a CNN-DBNN. Features extracted from frames by CNN are sent to the discriminative Deep Belief Network (DDBN).

8.2 Healthcare and Rehabilitation

The capacity for diagnosis and data collection in the medical and rehabilitation fields has been significantly enhanced by HAR. Wearables have become an indispensable tool for doctors in assessing and monitoring patients' health because of their ability to record vital signs, store data, and transmit that information to hospitals and other medical facilities. Specifically, many publications have detailed different approaches to monitoring and assessing the signs and symptoms of Parkinson's disease (PD) [254, 255]. Many people's lives are cut tragically short by pulmonary diseases like COPD, asthma, and the coronavirus simian immunodeficiency virus (COVID-19). Coughing is a common symptom of pulmonary diseases, and recent works have used wearables to detect this symptom [256–258]. Because of their increased susceptibility to illness, the elderly have long been a focus of healthcare reform efforts. The detection of falls and other abnormal behaviours in the elderly requires constant monitoring with automatic surveillance systems. In [259], a method is mentioned for modelling the actions of those with dementia (such as Alzheimer's and Parkinson's). Vanilla RNNs, Long Short-Term Memory, and Gated Recurrent Units are all types of RNNs used for anomaly detection in the elderly with dementia (GRU). Methods for assessing depressive symptoms using wrist-worn sensors [260] and for monitoring infants for stroke using wearable accelerometers [261] have also been introduced in other works. Electromyography (EMG) sensors have been widely used to detect muscle activities and hand motions. This has resulted in improved prosthesis control for individuals who have missing or have damaged limbs [262, 263]. Equipment, such as wearable devices, can be placed on the body of the person to be monitored in real time to recognise a specific feature, such as falls, gait, and breathing disorders. However, the person being tracked may find these gadgets intrusive or forget to wear them. Taylor et al. [264] showed that a non-invasive method can detect human motion in a near-real-time scenario. To further evaluate the RF algorithm's performance while in either a standing or seated position, Taylor et al. [264] generated a dataset of radio wave signals with software-defined radios (SDRs).

8.3 Emotional Calculation

The seven basic emotions-happiness, anger, sadness, thought, grief, fear, and surprise-are all manifestations of emotion, which is a more nuanced and long-lasting physiological evaluation and experience of human attitude. Emotion computing entails primarily the following activities and processes: determining an individual's emotional state and its relationship to their physiology and behaviour; using a wide range of sensors to collect data on the behavioural characteristics and physiological changes associated with an individual's current emotional state (including voice signals, facial expressions, body postures, and other forms of body language; pulse; skin electricity; brain electricity; and olfactory signals); and analysing the mechanism by which emotions are triggered and processed[33]. Users' physiological signals related to emotional changes can be captured in real

time by emotionally interactive intelligent systems via smart wearable devices, and when the system is monitoring users' large emotional fluctuations, it can regulate users' emotions in real time to avoid health hazards or make health care suggestions. The use of computational emotion in distance education has the potential to enhance the effectiveness of computer-assisted human learning by piqueing students' interest and facilitating more effective learning. In online shopping, the system can record the customer's interest in products and automatically analyse their preferences based on their eye movement, focus, and other parameters while they browse design solutions. A new context-aware multimodal sentiment analysis framework was proposed by Dashtipour et al. [265] to integrate sentiment across modalities using fusion techniques at both the decision-level (late) and the feature-level (early). Recent research has made use of a variety of physiological parameters, such as EEG, ECG, EMG, photoelectron plethysmography (PPG), body temperature, facial features, and more [266, 267]. To estimate scores of good and bad effects, Hssayeni et al. [268] created a multi-modal physiological data fusion framework by using deep CNN to collect motion and physiological signals collected from wearable devices (such as respiration (RESP), electrocardiogram (ECG), electromyogram (EMG), and electrodermal activity (EDA). In doing so, they looked into two different types of data fusion met (gradient augmentation trees and convolutional neural networks). Data fusion was expanded to include electrooculography by Khezri et al. [269]. Wearable sensors record heart and respiratory activity, and Mohino Herranz et al. [270] analysed this data to determine three distinct emotions: apathy, sadness, and disgust.

8.4 Education

The capacity to recognise human actions depicted in videos is of tremendous value in the contexts of both learning and instruction. Through the analysis of student activities captured on video, it may be possible to recognise human behaviour and implement automated attendance tracking in educational institutions. Taking attendance manually can be a time-consuming process, during which the instructor may not be able to monitor what is going on in the classroom due to time constraints[271]. Because we now have the technology to do so, we are able to use a system that can monitor attendance automatically and in real time inside of the classroom. Authors in [272] suggests developing an automatic attendance monitoring system by making use of the Viola-Jones algorithm. Students and their movements in and around the classroom, such as entering and exiting, are logged in [273] which also keeps track of the classroom's layout. Because it can identify faces and track motion, the system can also recognise and identify actions. This capability comes from its ability to recognise motion and facial expressions. The Haar cascade classifier is utilised in order to recognise a person's face, and in order to train the system, a combination of the eigenfaces and fisherfaces algorithms is utilised. The motion analysis process necessitates the utilisation of three auxiliary modules, namely body detection, tracking, and motion recognition. In order to take attendance, it is necessary to make assumptions regarding the capacity of the classroom as well as the lighting.

8.5 Assisted Living

The importance of assisted living systems, which allow patients or the elderly some measure of autonomy, is growing. Smart environments and smart homes are the focus of these kinds of applications [32]. The need to worry about people's personal information makes cameras inconvenient. This is why surveillance cameras have been gradually being replaced by ambient, wearable, and RF-based sensors. The smart home is outfitted with a wide variety of Internet of Things (IoT) devices that coordinate their efforts to improve residents' ease of use, comfort, entertainment, privacy, and safety [274]. In order to create useful smart home services, activity recognition of residents as they go about their daily routines is essential. It's crucial for lowering healthcare expenditures [275], making at-home care and comfort possible [276], and cutting down on energy use [277]. In order to tell apart the various uses of a shared space, the authors of [278] implemented Indoor Mobility (IM) and FuzzyEn. More so, to ensure maximum relevance and minimum redundancy, the authors of [279] used a back-propagation neural network technique to pick the relevant features in MAR in the smart home. Both the execution time and the accuracy of the activity recognition were improved with the proposed method. For automatic health monitoring, Wilson et al. [280] presented a system that makes use of the relationship between location and activity. Wang et al. [281] proposed a method for MAR of Activities of Daily Living (ADLs) based on sensor reading, which can be used to keep an eye on senior citizens as they go about their day in a smart home. In a similar vein, Gu et al. [282] proposed an innovative activity model grounded in emerging patterns that could recognise users in both single- and multi-user settings, with the latter model being able to capture user interaction.

9 Recent Advances

The most recent developments in the field of HAR, as well as the most significant contributions made to it, are discussed in this section.

9.1 Parameter Efficient Transfer Learning (PETL)

Using pre-training on massive datasets, NLP researchers have developed large-scale models like BERT [285], GPT-3 [286], and PaLM [287]. Both [288]) and [289] provide a high-level summary of the various existing PETL techniques, such as Prefix-tuning, Adapter, LoRA, and Prompt-tuning. Using PETL methods, large-scale pre-trained language models have been successfully adapted to downstream NLP tasks like translation, question and answering and reading and comprehension. As a result of the success of these PETL techniques in natural language processing (NLP), researchers are beginning to look in the opposite direction and apply them to vision tasks; for instance, VPT employs prompt-tuning, and AdaptFormer makes use of an adapter as show in Fig. 13. Besides prompt-tuning, [290] also proposed a prompt matcher for semantic segmentation. The multimodal model tuning strategy necessitates training a new pre-trained model and may not smoothly apply to pre-trained vision models, but fine-tuning vision-language pre-training models has been proposed [291]. It shows promising performance via text prompts (e.g., text category label) [283].

9.2 Temporal Understanding of Neural Networks

Working memory has traditionally been implemented in the field of neural engineering through the use of recurrent connections, leading to the development of what are now called Recurrent Neural Networks (RNN). Following the success of RNNs, more complex memory cells were proposed, such as LSTM [292] and LMU [293], which further enhance the memory capacity of ANNs and its training via backpropagation. Recently, attention networks and their Transformer architectures [294] have proven successful at solving temporal processing tasks like sequence transduction [295], time series forecasting [296], and video action recognition [297]. Instead of presenting events in chronological order as they occur, these networks collect data over a period of time (or the entire sequence), which is then processed in a batch mode. By accumulating stimuli over time and then feeding them to the network as a single input, these methods can be viewed as an implementation of working memory outside of the neural network. For most temporal tasks today, the most precise systems are those based on Transformer ANNs [298].

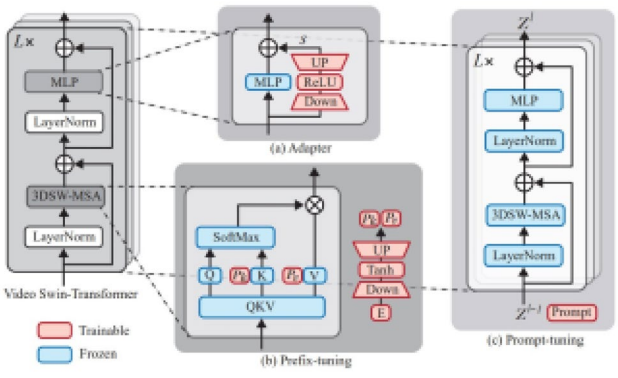


Fig. 13 The visual PETL methods in one coherent overview. The trainable parameters they introduce to the various nodes of the backbone model are implemented in different ways [283]

9.3 Audio-Visual Representation Learning

The use of audio-visual correspondence (AVC) to facilitate autonomous learning has also been investigated in the context of the auditory modality [299, 300]. Simply put, AVC is the task of determining whether or not a given video clip and its accompanying short audio clip belong to the same sequence. It has been demonstrated that similar tasks, such as temporal synchronisation [301] between audio and video, audio classification [302, 303], spatial alignment prediction between audio and 360 degree videos [304], and optimal combination of self-supervised tasks [305], are useful for learning efficient multi-modal video representations [284] as shown in Fig. 14. As a form of cross-modal instance discrimination, contrastive learning has been investigated in other works [306–308].

9.4 Multi-dataset Co-Training

Image detection [309, 310] and segmentation [311] are just two examples of the types of tasks where multi-dataset co-training has been investigated in the past. Multiple proposals [312, 313] have been made to train on merged versions of video datasets. Results tend to improve with increasing dataset size. The use of multiple datasets at once is likely to mitigate the negative impact of dataset bias, and the use of multiple datasets to increase data size and improve final performance [314]. To combat the potential for bias in the training data, OmniSource [315] includes web images as part of the training dataset. For self-supported pretraining and fine-tuning on downstream datasets, VATT [316] makes use of supplemental multi-modal data. Even in the final tuning phase, CoVeR [85] combines image and video training, and the results show a significant improvement in performance compared to training on individual datasets. The scope of PolyViT [317] is expanded to include training with image, video, and audio datasets of varying sampling sizes. In this paper, we propose a straightforward method (no multi-stage training, no complex dataset-wise sampling and hyper-parameter tuning) for training multi-action datasets, without the need for images or any other supplementary data [318].

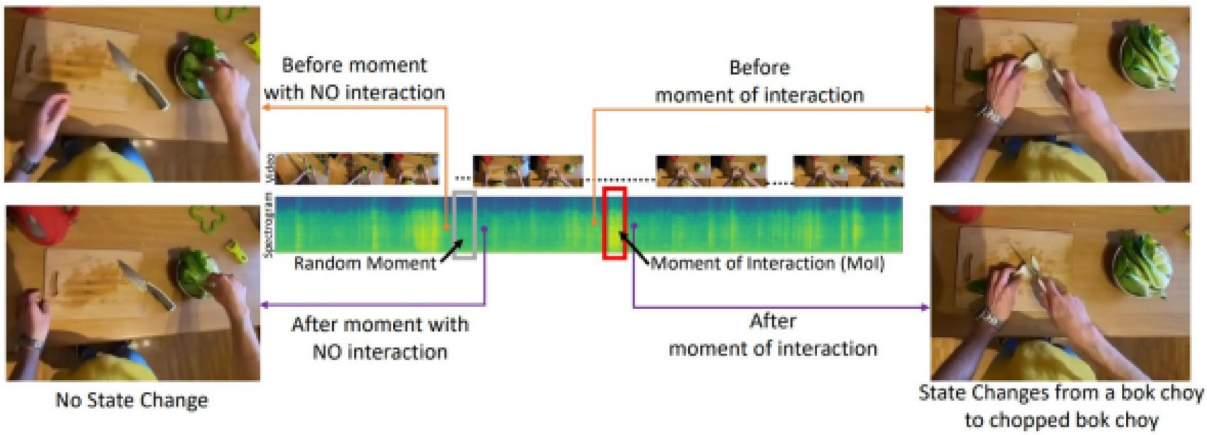


Fig. 14 There are no transitions between the initial and final states due to the lack of interactions. Representations can be learned, that are sensitive to the shift in visual state caused by interactions by sampling from moments of interaction (MoI), as indicated by the red box [284]

9.5 Self-supervised 3D Action Recognition

In order to learn 3D action representations without any external supervision, many previous works have proposed various methods. Autoencoder-based models are proposed in [319], and in LongT GAN an additional adversarial training strategy is proposed. This method of learning latent representation through sequential reconstruction is based on the generative paradigm. To predict and categorise skeleton sequences, P &C [320] also trains an encoder-decoder network. The authors also propose strategies to weaken the decoder, placing more demands on the encoder, so that more robust and distinguishable features can be learned. MS2L [321] integrates multiple pretext tasks in order to learn a better representation, in contrast to the previously mentioned methods which only adopt a single reconstruction task. Newer attempts [322, 323] have introduced contrastive learning based on momentum encoders, leading to improved performance. The first of these to conduct cross-modal knowledge mining is CrosSCLR [324]. Discovers false positives and rebalances training samples based on the context differences between skeleton modalities. However, since accurate initial representation is crucial for the successful positive mining in CrosSCLR, it is necessary to train in two phases [325].

10 Discussion and Challenges

In the last two decades, human action recognition has risen to prominence as a major area of study in computer vision. In particular, the advent of DL models and developments in parallel computing techniques, such as GPU computing, have ushered in a plethora of new possibilities in this area. There have been numerous DL-based methods developed and used for a wide range of human action recognition applications. Over the past few years, human action recognition has jumped from recognising actions in a controlled environment using small size benchmark datasets to recognising actions in realistic videos using very large-scale benchmarks. There would have been less progress without the use of DL methods. The field of HAR is expanding rapidly, but there are still some obstacles that, if solved, would make the field even better and encourage more novel HAR techniques to be implemented. These difficulties and prospects in HAR are discussed here.

10.1 Collecting Labeled Data

Lack of large-scale labelled datasets is a major obstacle to training robust Human-Activity Recognition models (HAR). As labelling massive amounts of data is a time-consuming and costly endeavour, unsupervised and semi-supervised learning techniques have emerged to learn useful features from data without the aid of labels [326]. It has been shown that generative deep models (such as AEs and GANs) can benefit from unsupervised data, but they are not directly applicable for HAR [110]. There is also promising potential in the development of semi-supervised deep models and active deep models [327], which are able to function with a reduced amount of labelled data. Building new deep models that can be taught with limited labelled data is an urgent task. Due to this difficulty, most HAR data collection efforts [68] are conducted on a relatively small scale, in a controlled or semi-controlled setting, leading to models that are not transferable to the real world. Combining generative and discriminative models into a single framework, called a hybrid model [328, 329] , shows great promise. While there have been some studies, they are all very early in their stages of development.

10.2 Robustness

The robustness and reliability of models is gaining more and more attention as a central issue in the community [33]. Multi-sensory systems, which combine the strengths of different kinds of sensors, are increasingly popular as a means of increasing robustness [330]. DeepFusionHAR is a proposed architecture by authors in [331] that combines manually crafted features with deep learning extracted features from multiple sensors to identify commonplace and athletic activities. Using the sensors already present in smartphones and smartwatches, authors in [332] proposed a multi-sensory approach to classifying 20 complex actions and 5 basic actions. Utilizing an accelerometer, gyroscope, magnetometer, microphone, and GPS, Pires et al. [333] demonstrated a mobile application on a multi-sensor mobile platform for activity classification in daily life. In some applications [334], multi-sensory networks are combined with attention modules to train on the most representative and discriminative sensor modality for distinguishing human activities.

10.3 Multi-modality Learning

To improve HAR, many have proposed using multi-modal learning techniques, such as multi-modal fusion and cross-modal transfer learning. Due to their complementary nature, multi-modality data fusion improves HAR performance, and co-learning can be used to address the issue of insufficient data for some modalities. Few-shot learning methods [335, 336] are one such method. Despite the fact that HAR has only been tried with a small number of shots [337, 338]. Given the importance of resolving data scarcity issues in many real-world scenarios, more sophisticated few-shot action analysis has yet to be fully explored.

10.4 Hybrid HAR

Despite the flexibility afforded by hybrid approaches, which can combine features and pre-processing steps, the high computational complexity of the target system may hinder both real-time and lengthy video processing. Long videos and real-time applications that require constant video streaming may experience issues due to these constraints. The computational expense of training the model is a difficulty of hybrid HAR [271].

10.5 Privacy Preservation

Users are starting to worry about their privacy [27]. In general, people are less willing to agree to data collection from a sensor if that sensor has a higher inference potential. Several works, such as the anonymizing autoencoder [339] and the GEN architecture [38], propose methods for human activity classification that are less invasive to people's privacy. It is possible to train replacement auto encoders to replace values that indicate non-sensitive inferences with features that correspond to sensitive inferences, as in the case of time-series data. Features that can be used to identify a specific person are obscured in these works, while those that are shared by different activities or movements are kept intact [111]. For learning problems with privacy concerns, federated learning is a promising new method [340, 341]. As a result, a global model can be learned collectively without users having to share their personal information. To boost the efficiency of the federated learning system, Xiao et al. [342] implemented a federated averaging technique in conjunction with a perceptive extraction network.

11 Conclusion and Future Direction

Because of its significance, HAR has been the subject of extensive study over the past few decades, and researchers have employed a wide range of data modalities, each with their own unique characteristics, to accomplish this goal [11]. Identifying human actions in wearables has opened up a wealth of possibilities for tracking and enhancing our daily lives. The use of AI and ML has been crucial to the development of wearables that support HAR. With the advent of DL, activity recognition performance has reached new heights in wearables-based HAR [111]. When it comes to identifying and categorising human actions and making predictions about human behaviour, deep learning (DL) based approaches and other techniques have proven to be the best option at the present time [343]. The accuracy of the HAR model was enhanced by using CNNs at the frame level instead of the conventional hand-crafted manual feature-based extraction methods. In the future, 3D-CNNs enhanced CNN's accuracy by applying and processing a batch of frames simultaneously. In order to effectively incorporate the temporal component of the videos, many cutting-edge HAR models have begun using RNNs and LSTMs. The Two Stream Fusion technique outperformed C3D without the need for the additional parameters required by C3D [165].

Despite the great development in the field of HAR along with deep learning, there still remains few open problems for better real-world applications, including the deployment of DNNs, domain adaptation, complex activity recognition etc,. Methodical network and sufficient training data for generalizability are the most important and prominent requirements involving deep-learning approaches. Some of the most interesting and potential directions for future research are as follows.

11.1 Event-Based Datasets

Commercial adoption of event-based sensors is still in its early stages, which prevents the collection of massive amounts of event-based data. Many datasets used in machine learning today are manufactured in a lab from simulated data or data captured within frames [298]. Recording existing frame-based datasets with a neuromorphic camera yielded three neuromorphic classification datasets: N-MNIST [344], N-Caltech101 [345], and DVS-CIFAR10 [346]. These techniques generated motion via saccadic eye movements, which provided the neuromorphic camera with the requisite brightness changes, thereby simulating biological vision. Using a pan-tilt platform, an Asynchronous Time-based Image Sensor [347] was relocated in the first two data sets. In the third data set, a fixed DVS camera [348] was used to capture a moving target image. Still, using data from a real-world acquisition is the ideal choice when creating event-based systems. At present, most native neuromorphic datasets are still relatively straightforward in comparison to conventional frame-based ones.

11.2 Cross-Modal Mutual Distillation

There is abundant supplementary data between skeleton modalities for use in 3D action recognition. For unsupervised 3D action representation learning, however, the question of how best to model and use this data persists as a significant obstacle. In recent years, thanks to developments in human pose estimation algorithms [349, 350], skeleton-based 3D human action recognition has gained popularity due to its portability and invulnerability to environmental factors. Several well-known pretexts have been extensively studied in the literature [321, 351, 320] to learn robust and discriminative representation, including motion prediction, jigsaw puzzle recognition, and masked reconstruction. While integrating multimodal information [352, 353] is crucial to enhancing the performance of 3D action recognition, this aspect of cross-modal interactive learning is largely overlooked in early contrastive learning-based attempts [321, 354].

11.3 Video-Based Self-supervised Learning

It's not easy to learn representations based on video. Selecting an appropriate SSL loss is the first obstacle. Multiple methods [355] have tried to learn representations that are independent of object transformations and viewpoints. However, representations that are sensitive to these deformations are necessary for many tasks further down the pipeline. Audio-based representation learning is another option that has been explored using multi-modal data [302, 356]. While these methods can produce invariant representations, their overarching goal is to harmonise audio and visual features in a single location. The second difficulty is managing the fact that state-of-the-art video-based SSL methods like Kinetics [166] rely on the human curation of video datasets. These methods are made to work with carefully chosen clips that show a single action or interaction between two objects. As opposed to the large egocentric datasets of daily activities, which typically contain unfiltered real-world data [284].

11.4 Model Generalizability

High generalizability is achieved when a model continues to perform well when presented with new data. Overfitting occurs when a model performs well on the training data but poorly on new data. These days, researchers are working hard to make HAR models more applicable in a wide variety of contexts [357, 58]. It is one of the main goals of generalizability studies in HAR to develop models that can be applied to a more extensive sample, but doing so typically necessitates extensive data and complicated models. DL-based HAR generally outperforms and generalises better than other types of methods when dealing with situations where high model complexity and data are not bottlenecks. New training methods, such as invariant risk minimization [358] and federated learning methods [359], can adapt and learn predictors across multiple environments, which is an unexplored avenue of generalizability. The incorporation of these domains into DL based HAR could not only increase the generalizability of HAR models, but do so in a manner that is not dependent on any particular model [111].

11.5 CNNs and Vision Transformers

In all computer vision tasks involving images and videos, CNNs serve as the standard backbones. To boost accuracy and efficiency, many novel convolutional neural architectures have been proposed (e.g., VGG [175], ResNet [360] and DenseNet [361]). Despite CNNs' continued dominance, Vision Transformers have proven to be a significant advancement in computer vision. To classify images, Vision Transformer (ViT [362]) uses the Transformer architecture directly, with promising results. Over the past few years, ViT and its variants [363–365] have produced remarkable results in the processing of still images and moving videos.

Acknowledgements NOT APPLICABLE

Funding NOT APPLICABLE

Data Availability NOT APPLICABLE

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Consent to Participate Not Applicable.

Consent for Publication Yes.

Ethical Approval Not Applicable.

References

1. Lu M, Hu Y, Lu X (2020) Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Appl Intell 50(4):1100–1111
2. Lin W, Sun M-T, Poovandran R, Zhang Z (2008) Human activity recognition for video surveillance. In: 2008 IEEE international symposium on circuits and systems (ISCAS), pp 2737–2740. IEEE
3. Rodomagoulakis I, Kardaris N, Pitsikalis V, Mavroudi E, Katsamanis A, Tsiami A, Maragos P (2016) Multimodal human action recognition in assistive human-robot interaction. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2702–2706. IEEE
4. Hu W, Xie D, Fu Z, Zeng W, Maybank S (2007) Semantic-based surveillance video retrieval. IEEE Trans Image Proces 16(4):1168–1181
5. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from single depth images. In: CVPR 2011, pp 1297–1304. IEEE
6. Lara OD, Labrador MA (2012) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor 15(3):1192–1209
7. Wang J, Chen Y, Hao S, Peng X, Hu L (2019) Deep learning for sensor-based activity recognition: a survey. Pattern Recognit Lett 119:3–11
8. Xu T, Zhou Y, Zhu J (2018) New advances and challenges of fall detection systems: a survey. Appl Sci 8(3):418
9. Sathyanarayana S, Satzoda RK, Sathyanarayana S, Thambipillai S (2018) Vision-based patient monitoring: a comprehensive review of algorithms and technologies. J Ambient Intell Hum Comput 9(2):225–251
10. Masoud M, Jaradat Y, Manasrah A, Jannoud I (2019) Sensors of smart devices in the internet of everything (IoE) era: big opportunities and massive doubts. J Sensors 15
11. Sun Z, Ke Q, Rahmani H, Bennamoun M, Wang G, Liu J (2022) Human action recognition from various data modalities: A review. IEEE Trans Pattern Analysis and machine intelligence. 22 IEEE
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
13. Sze V, Chen Y-H, Yang T-J, Emer JS (2017) Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE 105(12):2295–2329
14. Gu F, Khoshelham K, Valaee S (2017) Locomotion activity recognition: A deep learning approach. In: 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), pp 1–5. IEEE
15. Pei L, Xia S, Chu L, Xiao F, Wu Q, Yu W, Qiu R (2021) Mars: Mixed virtual and real wearable sensors for human activity recognition with multidomain deep learning model. IEEE Internet of Things Journal 8(11):9383–9396
16. Zhou B, Yang J, Li Q (2019) Smartphone-based activity recognition for indoor localization using a convolutional neural network. Sensors 19(3):621
17. Nweke HF, Teh YW, Al-Garadi MA, Alo UR (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl 105:233–261
18. Gao Z, Wang D, Wan S, Zhang H, Wang Y (2019) Cognitive-inspired class-statistic matching with triple-constrain for camera free 3d object retrieval. Future Gener Comput Syst 94:641–653
19. Yin Y, Chen L, Xu Y, Wan J, Zhang H, Mai Z (2020) Qos prediction for service recommendation with deep feature learning in edge computing environment. Mob Netw Appl 25:391–401
20. Bao L, Intille SS (2004) Activity recognition from user-annotated acceleration data. In: Pervasive computing: second international conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21–23, 2004. Proceedings 2, pp 1–17. Springer, Berlin
21. Wu W, Dasgupta S, Ramirez EE, Peterson C, Norman GJ et al (2012) Classification accuracies of physical activities using smartphone motion sensors. J Med Internet Res 14(5):e2208
22. Zhao Y, Li H, Wan S, Sekuboyina A, Hu X, Tetteh G, Piraud M, Menze B (2019) Knowledge-aided convolutional neural network for small organ segmentation. IEEE J Biomed Health Inf 23(4):1363–1373
23. Reyes-Ortiz J-L, Oneto L, Samà A, Parra X, Anguita D (2016) Transition-aware human activity recognition using smartphones. Neurocomputing 171:754–767
24. Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. ACM SigKDD Explor Newsl 12(2):74–82
25. Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL (2012) Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Ambient assisted living and home care: 4th international workshop, IWAAL 2012, Vitoria-Gasteiz, Spain, December 3–5, 2012. Proceedings 4, pp 216–223. Springer, Berlin
26. Wang Y, Cang S, Yu H (2019) A survey on wearable sensor modality centred human activity recognition in health care. Expert Syst Appl 137:167–190
27. Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y (2021) Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities. ACM Comput Surv 54(4):1–40
28. Demrozi F, Pravadelli G, Bihorac A, Rashidi P (2020) Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. IEEE Access 8:210816–210836
29. Fu B, Damer N, Kirchbuchner F, Kuijper A (2020) Sensing technology for human activity recognition: a comprehensive survey. IEEE Access 8:83791–83820
30. Sousa Lima W, Souto E, El-Khatib K, Jalali R, Gama J (2019) Human activity recognition using inertial sensors in a smartphone: an overview. Sensors 19(14):3213
31. Arshad MH, Bilal M, Gani A (2022) Human activity recognition: review, taxonomy and open challenges. Sensors 22(17):6463
32. Li Q, Gravina R, Li Y, Alsamhi SH, Sun F, Fortino G (2020) Multi-user activity recognition: challenges and opportunities. Inf Fusion 63:121–135
33. Qiu S, Zhao H, Jiang N, Wang Z, Liu L, An Y, Zhao H, Miao X, Liu R, Fortino G (2022) Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Inf Fusion 80:241–265
34. Ramanujam E, Perumal T, Padmavathi S (2021) Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review. IEEE Sens J 21(12):13029–13040
35. Subetha T, Chitrakala S (2016) A survey on human activity recognition from videos. In: 2016 international conference on information communication and embedded systems (ICICES), pp 1–7. IEEE

36. Presti LL, La Cascia M (2016) 3d skeleton-based human action classification: a survey. Pattern Recognit 53:130–147
37. Kang SM, Wildes RP (2016) Review of action recognition and detection methods. arXiv:1610.06906
38. Malekzadeh M, Clegg RG, Cavallaro A, Haddadi H (2018) Protecting sensory data against sensitive inferences. In: Proceedings of the 1st workshop on privacy by design in distributed systems, pp 1–6
39. Asuncion A, Newman D (2007) Uci machine learning repository
40. Reiss A, Stricker D (2012) Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th international symposium on wearable computers, pp 108–109. IEEE
41. Stisen A, Blunck H, Bhattacharya S, Prentow TS, Kjærgaard MB, Dey A, Sonne T, Jensen MM (2015) Smart devices are different: assessing and mitigatingmobile sensing heterogeneities for activity recognition. In: Proceedings of the 13th ACM conference on embedded networked sensor systems, pp 127–140
42. Anguita D, Ghio A, Oneto L, Parra Perez X, Reyes Ortiz JL (2013) A public domain dataset for human activity recognition using smartphones. In: Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning, pp 437–442
43. Chavarriaga R, Sagha H, Calatroni A, Digumarti ST, Tröster G, Millán JdR, Roggen D (2013) The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recognit Lett 34(15):2033–2042
44. Kleanthous N, Hussain AJ, Khan W, Liatsis P (2020) A new machine learning based approach to predict freezing of gait. Pattern Recognit Lett 140:119–126
45. Ronald M, Poulose A, Han DS (2021) isplinception: an inception-resnet deep learning architecture for human activity recognition. IEEE Access 9:68985–69001
46. Davidashvilly S, Hssayeni M, Chi C, Jimenez-Shahed J, Ghoraani B (2022) Activity recognition in parkinson's patients from motion data using a cnn model trained by healthy subjects. In: 2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp 3199–3202. IEEE
47. Gil-Martín M, San-Segundo R, Fernández-Martínez F, Ferreiros-López J (2021) Time analysis in human activity recognition. Neural Process Lett 53(6):4507–4525
48. Gochoo M, Tahir SBUD, Jalal A, Kim K (2021) Monitoring real-time personal locomotion behaviors over smart indoor-outdoor environments via body-worn sensors. IEEE Access 9:70556–70570
49. Mahmud S, Tonmoy M, Bhaumik KK, Rahman AM, Amin MA, Shoyaib M, Khan MAH, Ali AA (2020) Human activity recognition from wearable sensor data using self-attention. arXiv:2003.09018
50. Xia K, Huang J, Wang H (2020) LSTM-CNN architecture for human activity recognition. IEEE Access 8:56855–56866
51. Challa SK, Kumar A, Semwal VB (2021) A multibranch cnn-bilstm model for human activity recognition using wearable sensor data. Vis Comput 1–15
52. Khatun MA, Yousuf MA, Ahmed S, Uddin MZ, Alyami SA, Al-Ashhab S, Akhdar HF, Khan A, Azad A, Moni MA (2022) Deep cnn-lstm with self-attention model for human activity recognition using wearable sensor. IEEE J Transl Eng Health Med 10:1–16
53. Pang YH, Ping LY, Ling GF, Yin OS, How KW (2021) Stacked deep analytic model for human activity recognition on a uci har database. F1000Research 10
54. Albahri A, Zaidan A, Albahri O, Zaidan B, Alamoodi A, Shareef AH, Alwan JK, Hamid RA, Aljbory M, Jasim AN et al (2021) Development of iot-based mhealth framework for various cases of heart disease patients. Health Technol 11(5):1013–1033
55. Allgaier J, Schlee W, Langguth B, Probst T, Pryss R (2021) Predicting the gender of individuals with tinnitus based on daily life data of the trackyourtinnitus mhealth platform. Sci Rep 11(1):1–14
56. Tsang KC, Pinnock H, Wilson AM, Shah SA (2020) Application of machine learning to support self-management of asthma with mhealth. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC), pp 5673–5677. IEEE
57. Bock M, Hölzemann A, Moeller M, Van Laerhoven K (2021) Improving deep learning for har with shallow lstms. In: 2021 International symposium on wearable computers, pp 7–12
58. Qin Z, Zhang Y, Meng S, Qin Z, Choo K-KR (2020) Imaging and fusing time series for wearable sensor-based human activity recognition. Inf Fusion 53:80–87
59. Sa-nguannarm P, Elbasani E, Kim B, Kim E-H, Kim J-D (2021) Experimentation of human activity recognition by using accelerometer data based on lstm. In: Advanced multimedia and ubiquitous engineering, pp 83–89. Springer, Berlin
60. Atkinson G, Metsis V (2020) Identifying label noise in time-series datasets. In: Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers, pp 238–243
61. Liu H, Hartmann Y, Schultz T (2021) CSL-share: a multimodal wearable sensor-based human activity dataset
62. Teng Q, Wang K, Zhang L, He J (2020) The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. IEEE Sens J 20(13):7265–7274
63. Hoelzemann A, Van Laerhoven K (2020) Digging deeper: towards a better understanding of transfer learning for human activity recognition. In: Proceedings of the 2020 international symposium on wearable computers, pp 50–54
64. Zhang W, Zhu T, Yang C, Xiao J, Ning H (2020c) Sensors-based human activity recognition with convolutional neural network and attention mechanism. In: 2020 IEEE 11th international conference on software engineering and service science (ICSESS), pp 158–162. IEEE
65. Zhou Yu et al. (2022) A hybrid attention-based deep neural network for simultaneous multi-sensor pruning and human activity recognition. IEEE Int Things J 9(24):25363–25372
66. Banos O, Garcia R, Holgado-Terriza JA, Damas M, Pomares H, Rojas I, Saez A, Villalonga C (2014) mhealthdroid: a novel framework for agile development of mobile health applications. In: International workshop on ambient assisted living, pp 91–98. Springer, Berlin
67. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732
68. Micucci D, Mobilio M, Napoletano P (2017) Unimib shar: a dataset for human activity recognition using acceleration data from smartphones. Appl Sci 7(10):1101
69. Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
70. Shahroudy A, Liu J, Ng T-T, Wang G (2016) NTU RGB+D: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019
71. Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (2015) Activitynet: a large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 961–970
72. Afza F, Khan MA, Sharif M, Kadry S, Manogaran G, Saba T, Ashraf I, Damaševičius R (2021) A framework of human action recognition using length control features fusion and weighted

entropy-variances based feature selection. Image Vis Comput 106:104090
73. Jaouedi N, Boujnah N, Bouhlel MS (2020) A new hybrid deep learning model for human action recognition. Journal of King Saud University Comput Inf Sci 32(4):447–453
74. Varshney N, Bakariya B (2021) Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple cnn streams. Multimedia Tools Appl pp 1–13
75. Wu Zhenyu et al (2020) Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. IEEE Trans Pattern Analysis Mach Int 44(4): 2126–2139
76. Umar IM. Ibrahim KM, Gital AY, Zambuk FU, Lawal MA, Yakubu ZI (2022) Hybrid model for human activity recognition using an inflated I3-D two stream convolutional-LSTM network with optical flow mechanism. In: 2022 IEEE Delhi section conference (DELCON), pp 1–7. IEEE
77. Wang T, Ng WW, Li J, Wu Q, Zhang S, Nugent C, Shewell C (2021) A deep clustering via automatic feature embedded learning for human activity recognition. IEEE Trans Circuit Syst Video Technol 32(1):210–223
78. Bulbul MF, Ali H (2021) Gradient local auto-correlation features for depth human action recognition. SN Appl Sci 3(5):1–13
79. Khaled H, Abu-Elnasr O, Elmougy S, Tolba A (2021) Intelligent system for human activity recognition in iot environment. Complex Intell Syst pp 1–12
80. Cheng Q, Liu Z, Ren Z, Cheng J, Liu J (2022) Spatial-temporal information aggregation and cross-modality interactive learning for RGB-D-based human action recognition. IEEE Access 10:104190–104201
81. Dong J, Gao Y, Lee HJ, Zhou H, Yao Y, Fang Z, Huang B (2020) Action recognition based on the fusion of graph convolutional networks with high order features. Appl Sci 10(4):1482
82. Li D, Qiu Z, Pan Y, Yao T, Li H, Mei T (2021) Representing videos as discriminative sub-graphs for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3310–3319
83. Wang L, Tong Z, Ji B, Wu G (2021) TDN: temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1895–1904
84. Chen J, Ho CM (2022) Mm-vit: Multi-modal video transformer for compressed video action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1910–1921
85. Zhang B, Yu J, Fifty C, Han W, Dai AM, Pang R, Sha F (2021) Co-training transformer with videos and images improves action recognition. arXiv:2112.07175
86. Duhme M, Memmesheimer R, Paulus D (2021) Fusion-gcn: Multimodal action recognition using graph convolutional networks. In: DAGM German conference on pattern recognition, pp 265–281. Springer, Berlin
87. Islam MM, Iqbal T (2021) Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition. IEEE Robot Autom Lett 6(2):1729–1736
88. Damen D, Doughty H, Farinella GM, Furnari A, Kazakos E, Ma J, Moltisanti D, Munro J, Perrett T, Price W et al (2022) Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. Int J Comput Vis 130(1):33–55
89. Huang Z, Qing Z, Wang X, Feng Y, Zhang S, Jiang J, Xia Z, Tang M, Sang N, Ang Jr MH (2021) Towards training stronger video vision transformers for epic-kitchens-100 action recognition. arXiv:2106.05058
90. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition, 2004 (ICPR 2004), vol 3, pp 32–36. IEEE
91. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011). HMDB: a large video database for human motion recognition. In: 2011 international conference on computer vision, pp 2556–2563. IEEE
92. Triboan D, Chen L, Chen F, Wang Z (2019) A semantics-based approach to sensor data segmentation in real-time activity recognition. Future Gener Comput Syst 93:224–236
93. Noor MHM, Salcic Z, Kevin I, Wang K (2017) Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer. Pervasive Mob Comput 38:41–59
94. Liu H, Hartmann Y, Schultz T (2022) A practical wearable sensor-based human activity recognition research pipeline. In: HEALTHINF, pp 847–856
95. Sarapata G, Morinan G, Dushin Y, Kainz B, Ong J, O'Keeffe J (2022) Video-based activity recognition for automated motor assessment of Parkinson's disease. IEEE J Biomed Health Inform
96. García S, Luengo J, Herrera F (2015) Data preprocessing in data mining. Springer, Berlin
97. Heaton J (2018) Ian Goodfellow, Yoshua Bengio, and Aaron Courville: deep learning. Genetic Program Evolvable Mach 19(1–2):305–307
98. Castro H, Correia V, Sowade E, Mitra K, Rocha J, Baumann R, Lanceros-Méndez S (2016) All-inkjet-printed low-pass filters with adjustable cutoff frequency consisting of resistors, inductors and transistors for sensor applications. Org Electron 38:205–212
99. Ignatov AD, Strijov VV (2016) Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. Multimed Tools Appl 75:7257–7270
100. Wang Z, Wu D, Chen J, Ghoneim A, Hossain MA (2016) A triaxial accelerometer-based human activity recognition via eemd-based features and game-theory-based feature selection. IEEE Sens J 16(9):3198–3207
101. Müller AC, Guido S (2016) Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc
102. Guo J, Mu Y, Xiong M, Liu Y, Gu J (2019) Activity feature solving based on tf-idf for activity recognition in smart homes. Complexity 2019:1–10
103. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vision Image Understand 156:117–127
104. Bhavan A, Aggarwal S (2018) Stacked generalization with wrapper-based feature selection for human activity recognition. In: 2018 IEEE symposium series on computational intelligence (SSCI), pp 1064–1068. IEEE
105. Li Z, Fan Y, Liu W (2015) The effect of whitening transformation on pooling operations in convolutional autoencoders. EURASIP J Adv Signal Process 2015(1):1–11
106. Nam W, Dollár P, Han JH (2014) Local decorrelation for improved pedestrian detection. Adv Neural Inf Process Syst 27
107. Kessy A, Lewin A, Strimmer K (2018) Optimal whitening and decorrelation. Am Stat 72(4):309–314
108. Bracewell RN, Bracewell RN (1986) The Fourier transform and its applications, vol 31999. McGraw-Hill, New York
109. Sejdic E, Djurovic I, Stankovic L (2008) Quantitative performance analysis of scalogram as instantaneous frequency estimator. IEEE Trans Signal Process 56(8):3837–3845
110. Gu F, Chung M-H, Chignell M, Valaee S, Zhou B, Liu X (2021) A survey on deep learning for human activity recognition. ACM Comput Surv 54(8):1–34
111. Zhang S, Li Y, Zhang S, Shahabi F, Xia S, Deng Y, Alshurafa N (2022) Deep learning in human activity recognition with wearable sensors: a review on advances. Sensors 22(4):1476

112. Zhang X-L, Wu J (2012) Deep belief networks based voice activity detection. IEEE Trans Audio Speech Lang Process 21(4):697–710
113. Fang H, Hu C (2014) Recognizing human activity in smart home using deep learning algorithm. In: Proceedings of the 33rd Chinese control conference, pp 4716–4720. IEEE
114. Uddin MZ, Hassan MM, Almogren A, Alamri A, Alrubaian M, Fortino G (2017) Facial expression recognition utilizing local direction-based robust features and deep belief network. IEEE Access 5:4525–4536
115. Zheng W-L, Zhu J-Y, Peng Y, Lu B-L (2014) Eeg-based emotion classification using deep belief networks. In: 2014 IEEE international conference on multimedia and expo (ICME), pp 1–6. IEEE
116. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
117. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
118. Gao X, Luo H, Wang Q, Zhao F, Ye L, Zhang Y (2019) A human activity recognition algorithm based on stacking denoising autoencoder and lightgbm. Sensors 19(4):947
119. Tang Y, Salakhutdinov R, Hinton G (2012) Robust boltzmann machines for recognition and denoising. In: 2012 IEEE conference on computer vision and pattern recognition, pp 2264–2271. IEEE
120. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
121. Bhattacharya S, Lane ND (2016) From smart to deep: Robust activity recognition on smartwatches using deep learning. In: 2016 IEEE international conference on pervasive computing and communication workshops (PerCom Workshops), pp 1–6. IEEE
122. Plötz T, Hammerla NY, Olivier PL (2011) Feature learning for activity recognition in ubiquitous computing. In: Twenty-second international joint conference on artificial intelligence
123. Lane ND, Georgiev P, Qendro L (2015) Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, pp 283–294
124. Radu V, Lane ND, Bhattacharya S, Mascolo C, Marina MK, Kawsar F (2016) Towards multimodal deep learning for activity recognition on mobile devices. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct, pp 185–188
125. Li Y, Shi D, Ding B, Liu D (2014) Unsupervised feature learning for human activity recognition using smartphone sensors. In: Mining intelligence and knowledge exploration, pp 99–107. Springer, Berlin
126. Mohammed S, Tashev I (2017) Unsupervised deep representation learning to remove motion artifacts in free-mode body sensor networks. In: 2017 IEEE 14th international conference on wearable and implantable body sensor networks (BSN), pp 183–188. IEEE
127. Valarezo AE, Rivera LP, Park H, Park N, Kim T-S (2020) Human activities recognition with a single writs imu via a variational autoencoder and android deep recurrent neural nets. Comput Sci Inf Syst 17(2):581–597
128. Vavoulas G, Chatzaki C, Malliotakis T, Pediaditis M, Tsiknakis M (2016) The mobiact dataset: recognition of activities of daily living using smartphones. In: International conference on information and communication technologies for ageing well and e-health, vol 2, pp 143–151. SciTePress
129. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J et al (2018) Recent advances in convolutional neural networks. Pattern Recognit 77:354–377
130. Ronao CA, Cho S-B (2016) Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst Appl 59:235–244
131. Hughes D, Correll N (2018) Distributed convolutional neural networks for human activity recognition in wearable robotics. In: Distributed autonomous robotic systems, pp 619–631. Springer, Berlin
132. Dong M, Han J, He Y, Jing X (2018) Har-net: Fusing deep representation and hand-crafted features for human activity recognition. In: International conference on signal and information processing, networking and computers, pp 32–40. Springer, Berlin
133. Ravi D, Wong C, Lo B, Yang G-Z (2016) Deep learning for human activity recognition: a resource efficient implementation on low-power devices. In: 2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN), pp 71–76. IEEE
134. Zeng M, Nguyen LT, Yu B, Mengshoel OJ, Zhu J, Wu P, Zhang J (2014) Convolutional neural networks for human activity recognition using mobile sensors. In: 6th international conference on mobile computing, applications and services, pp 197–205. IEEE
135. Lee S-M, Yoon, SM, Cho H (2017) Human activity recognition from accelerometer data using convolutional neural network. In: 2017 IEEE international conference on big data and smart computing (bigcomp), pp 131–134. IEEE
136. Huang W, Zhang L, Gao W, Min F, He J (2021) Shallow convolutional neural networks for human activity recognition using wearable sensors. IEEE Trans Instrum Measur 70:1–11
137. Yang J, Nguyen MN, San PP, Li XL, Krishnaswamy S (2015) Deep convolutional neural networks on multichannel time series for human activity recognition. In: Twenty-fourth international joint conference on artificial intelligence
138. Ha S, Choi S (2016) Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: 2016 international joint conference on neural networks (IJCNN), pp 381–388. IEEE
139. Lv M, Xu W, Chen T (2019) A hybrid deep convolutional and recurrent neural network for complex activity recognition using multimodal sensors. Neurocomputing 362:33–40
140. Li X, Luo J, Younes R (2020) Activitygan: generative adversarial networks for data augmentation in sensor-based human activity recognition. In: Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM Iinternational symposium on wearable computers, pp 249–254
141. Bailador G, Roggen D, Tröster G, Triviño G (2007) Real time gesture recognition using continuous time recurrent neural networks. In: 2nd international ICST conference on body area networks
142. Zheng L, Li S, Zhu C, Gao Y (2019) Application of indrnn for human activity recognition: the sussex-huawei locomotion-transportation challenge. In: Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers, pp 869–872
143. Wang X, Liao W, Guo Y, Yu L, Wang Q, Pan M, Li P (2019c) PERRNN: personalized recurrent neural networks for acceleration-based human activity recognition. In: ICC 2019-2019 IEEE international conference on communications (ICC), pp 1–6. IEEE
144. Ketykó I, Kovács F, Varga KZ (2019). Domain adaptation for semg-based gesture recognition with recurrent neural networks. In: 2019 international joint conference on neural networks (IJCNN), pp 1–7. IEEE

145. Zebin T, Sperrin M, Peek N, Casson AJ (2018) Human activity recognition from inertial sensor time-series using batch normalized deep LSTM recurrent networks. In: 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 1–4. IEEE
146. Hu Y, Zhang X-Q, Xu L, He FX, Tian Z, She W, Liu W (2020) Harmonic loss function for sensor-based human activity recognition based on lstm recurrent neural networks. IEEE Access 8:135617–135627
147. Ordóñez FJ, Roggen D (2016) Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16(1):115
148. Sutskever I, Martens J, Hinton GE (2011) Generating text with recurrent neural networks. In: ICML
149. Murad A, Pyun J-Y (2017) Deep recurrent neural networks for human activity recognition. Sensors 17(11):2556
150. Inoue M, Inoue S, Nishida T (2018) Deep recurrent neural network for mobile human activity recognition with high throughput. Artif Life Robot 23(2):173–185
151. Guan Y, Plötz T (2017) Ensembles of deep lstm learners for activity recognition using wearables. Proc ACM Interact Mob Wear Ubiquitous Technol 1(2):1–28
152. Gupta R, Dhindsa IS, Agarwal R (2020) Continuous angular position estimation of human ankle during unconstrained locomotion. Biomed Signal Process Control 60:101968
153. Okai J, Paraschiakos S, Beekman M, Knobbe A, de Sá CR (2019) Building robust models for human activity recognition from raw accelerometers data using gated recurrent units and long short term memory neural networks. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 2486–2491. IEEE
154. Czuszyński K, Rumiński J, Kwaśniewska A (2018) Gesture recognition with the linear optical sensor and recurrent neural networks. IEEE Sens J 18(13):5429–5438
155. Shi J, Zuo D, Zhang Z (2021) A gan-based data augmentation method for human activity recognition via the caching ability. Internet Technol Lett 4(5):e257
156. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144
157. Wang J, Chen Y, Gu Y, Xiao Y, Pan H (2018) Sensorygans: an effective generative adversarial framework for sensor-based human activity recognition. In: 2018 international joint conference on neural networks (IJCNN), pp 1–8. IEEE
158. Chan MH, Noor MHM (2021) A unified generative model using generative adversarial network for activity recognition. J Ambient Intell Hum Comput 12(7):8119–8128
159. Soleimani E, Nazerfard E (2021) Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. Neurocomputing 426:26–34
160. Hao W, Zhang Z (2019) Spatiotemporal distilled dense-connectivity network for video action recognition. Pattern Recognit 92:13–24
161. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Understand 104(2–3):90–126
162. Poppe R (2010) A survey on vision-based human action recognition. Image Vision Comput 28(6):976–990
163. Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: a survey. IEEE Trans Circuit Syst Video Technol 18(11):1473–1488
164. Carlsson S, Sullivan J (2001) Action recognition by shape matching to key frames. In: Workshop on models versus exemplars in computer vision, volume 1. Citeseer
165. Sharma V, Gupta M, Pandey AK, Mishra D, Kumar A (2022) A review of deep learning-based human activity recognition on benchmark video datasets. Appl Artif Intell 36(1):2093705
166. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
167. Naieera P, Anu P, Sadiq M (2018) An intelligent action predictor from video using deep learning. In: 2018 international conference on emerging trends and innovations in engineering and technological research (ICETIETR), pp 1–4. IEEE
168. Zhang S, Wei Z, Nie J, Huang L, Wang S, Li Z (2017) A review on human activity recognition using vision-based method. J Healthc Eng
169. Zhang B, Wang L, Wang Z, Qiao Y, Wang H (2016). Real-time action recognition with enhanced motion vector CNNs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2718–2726
170. Piergiovanni A, Ryoo MS (2019) Representation flow for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9945–9953
171. Verma KK, Singh BM, Dixit A (2019) A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. Int J Inf Technol pp 1–14
172. Verma KK, Singh BM, Mandoria HL, Chauhan P (2020) Two-stage human activity recognition using 2D-convnet
173. Varol G, Laptev I, Schmid C (2017) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Machine Intell 40(6):1510–1517
174. Huang C-D, Wang C-Y, Wang J-C (2015) Human action recognition system for elderly and children care using three stream convnet. In: 2015 international conference on orange technologies (ICOT), pp 5–9. IEEE
175. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
176. Zheng H, Zhang X-M (2020) A cross-modal learning approach for recognizing human actions. IEEE Syst J 15(2):2322–2330
177. Crasto N, Weinzaepfel P, Alahari K, Schmid C (2019) Mars: Motion-augmented rgb stream for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7882–7891
178. Yang H, Yuan C, Li B, Du Y, Xing J, Hu W, Maybank SJ (2019) Asymmetric 3D convolutional neural networks for action recognition. Pattern Recognit 85:1–12
179. Ji S, Xu W, Yang M, Yu K (2012) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
180. Abdelbaky A, Aly S (2020) Human action recognition based on simple deep convolution network pcanet. In: 2020 international conference on innovative trends in communication and computer engineering (ITCE), pp. 257–262. IEEE
181. Thameri M, Kammoun A, Abed-Meraim K, Belouchrani A (2011) Fast principal component analysis and data whitening algorithms. In: International workshop on systems, signal processing and their applications, WOSSPA, pp 139–142. IEEE
182. Zhang H, Liu D, Xiong Z (2019) Two-stream action recognition-oriented video super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8799–8808
183. Meng Y, Lin C-C, Panda R, Sattigeri P, Karlinsky L, Oliva A, Saenko K, Feris R (2020). AR-NET: adaptive frame resolution for efficient action recognition. In: European conference on computer vision, pp 86–104. Springer, Berlin
184. Perrett T, Damen D (2019) DDLSTM: dual-domain lstm for cross-dataset action recognition. In: Proceedings of the IEEE/

CVF conference on computer vision and pattern recognition, pp 7852–7861
185. Sun L, Jia K, Chen K, Yeung D-Y, Shi BE, Savarese S (2017). Lattice long short-term memory for human action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2147–2156
186. Grushin A, Monner DD, Reggia JA, Mishra A (2013) Robust human action recognition via long short-term memory. In: The 2013 international joint conference on neural networks (IJCNN), pp 1–8. IEEE
187. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–8. IEEE
188. Singh B, Marks TK, Jones M, Tuzel O, Shao M (2016) A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1961–1970
189. Wu J, Wang G, Yang W, Ji X (2016) Action recognition with joint attention on multi-level deep features. arXiv:1607.02556
190. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. In: International workshop on human behavior understanding, pp 29–39. Springer, Berlin
191. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
192. Li Q, Qiu Z, Yao T, Mei T, Rui Y, Luo J (2016) Action recognition by learning deep multi-granular spatio-temporal video representation. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval, pp 159–166
193. Wang Y, Wang S, Tang J, O'Hare N, Chang Y, Li B (2016) Hierarchical attention network for action recognition in videos. arXiv:1607.06416
194. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
195. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
196. Ge H, Yan Z, Yu W, Sun L (2019) An attention mechanism based convolutional lstm network for video action recognition. Multimed Tools Appl 78(14):20533–20556
197. Sudhakaran S, Escalera S, Lanz O (2019) LSTA: long short-term attention for egocentric action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9954–9963
198. Meng L, Zhao B, Chang B, Huang G, Sun W, Tung F, Sigal L (2019) Interpretable spatio-temporal attention for video action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
199. Wu Z, Xiong C, Ma C-Y, Socher R, Davis LS (2019) Adaframe: adaptive frame selection for fast video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1278–1287
200. Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CG (2018) Videolstm convolves, attends and flows for action recognition. Comput Vision Image Understand 166:41–50
201. Liu Z, Li Z, Wang R, Zong M, Ji W (2020) Spatiotemporal saliency-based multi-stream networks with attention-aware lstm for action recognition. Neural Comput Appl 32(18):14593–14602
202. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118
203. Li Y, Lan C, Xing J, Zeng W, Yuan C, Liu J (2016) Online human action detection using joint classification-regression recurrent neural networks. In: European conference on computer vision, pp 203–220. Springer, Berlin
204. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision, pp 816–833. Springer, Berlin
205. Mahasseni B, Todorovic S (2016) Regularizing long short term memory with 3d human-skeleton sequences for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3054–3062
206. Zhao L, Song Y, Zhang C, Liu Y, Wang P, Lin T, Deng M, Li H (2019) T-GCN: a temporal graph convolutional network for traffic prediction. IEEE Trans Intell Transp Syst 21(9):3848–3858
207. Si C, Jing Y, Wang W, Wang L, Tan T (2018) Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the European conference on computer vision (ECCV), pp 103–118
208. Yan S, Xiong Y, Lin D (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence
209. Cheng K, Zhang Y, Cao C, Shi L, Cheng J, Lu H (2020) Decoupling gcn with dropgraph module for skeleton-based action recognition. In: European conference on computer vision, pp 536–553. Springer, Berlin
210. Korban M, Li X (2020) DDGCN: a dynamic directed graph convolutional network for action recognition. In: European conference on computer vision, pp 761–776. Springer, Berlin
211. Yu P, Zhao Y, Li C, Yuan J, Chen C (2020) Structure-aware human-action generation. In: European conference on computer vision, pp 18–34. Springer, Berlin
212. Zhang X, Xu C, Tian X, Tao D (2019) Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Trans Neural Netw Learn Syst 31(8):3047–3060
213. Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q (2019) Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3595–3603
214. Peng W, Hong X, Chen H, Zhao G (2020) Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 2669–2676
215. Zhang X, Xu C, Tao D (2020d) Context aware graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14333–14342
216. Wu C, Wu X-J, Kittler J (2019) Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
217. Li B, Li X, Zhang Z, Wu F (2019) Spatio-temporal graph routing for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8561–8568
218. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W (2020) Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 143–152
219. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1227–1236

220. Zhang P, Lan C, Zeng W, Xing J, Xue J, Zheng N (2020) Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1112–1121
221. Wen Y-H, Gao L, Fu H, Zhang F-L, Xia S (2019) Graph cnns with motif and variable temporal block for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence 33:8989–8996
222. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H (2020) Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 183–192
223. Song Y.-F, Zhang Z, Shan C, Wang L (2022) Constructing stronger and faster baselines for skeleton-based action recognition. IEEE Trans Pattern Anal Machine Intell
224. Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q (2021) Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Trans Pattern Anal Mach Intell 44(6):3316–3333
225. Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning, pp 1096–1103
226. Holyoak KJ (1987) Parallel distributed processing: explorations in the microstructure of cognition. Science 236:992–997
227. Cilimkovic M (2015) Neural networks and back propagation algorithm. Institute of Technology Blanchardstown, Blanchardstown Road North Dublin 15(1)
228. Ranzato M, Huang FJ, Boureau Y-L, LeCun Y (2007) Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: 2007 IEEE conference on computer vision and pattern recognition, pp 1–8. IEEE
229. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2012) Spatio-temporal convolutional sparse auto-encoder for sequence classification. In: BMVC, pp 1–12. Citeseer
230. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer K (2011) The first facial expression recognition and analysis challenge. In: 2011 IEEE international conference on automatic face & gesture recognition (FG), pp 921–926. IEEE
231. Budiman A, Fanany MI, Basaruddin C (2014) Stacked denoising autoencoder for feature representation learning in pose-based action recognition. In: 2014 IEEE 3rd global conference on consumer electronics (GCCE), pp 684–688. IEEE
232. Hasan M, Roy-Chowdhury AK (2014) Continuous learning of human activity models using deep nets. In: European conference on computer vision, pp 705–720. Springer, Berlin
233. Wu DX, Pan W, Xie LD, Huang CX (2014) An adaptive stacked denoising auto-encoder architecture for human action recognition. In: Applied mechanics and materials, vol 631, pp 403–409. Trans Tech Publ
234. Zhang Z (2012) Microsoft kinect sensor and its effect. IEEE Multimed 19(2):4–10
235. Laptev I (2005) On space-time interest points. Int J Computer Vis 64(2):107–123
236. Mehta V, Dhall A, Pal S, Khan SS (2021) Motion and region aware adversarial learning for fall detection with thermal imaging. In: 2020 25th international conference on pattern recognition (ICPR), pp 6321–6328. IEEE
237. Wang J, Zhang X, Gao Q, Yue H, Wang H (2016) Device-free wireless localization and activity recognition: a deep learning approach. IEEE Trans Veh Technol 66(7):6258–6267
238. Gao Q, Wang J, Ma X, Feng X, Wang H (2017) Csi-based device-free wireless localization and activity recognition using radio image features. IEEE Trans Veh Technol 66(11):10346–10356
239. Chen M, Xu Z, Weinberger KQ, Sha F (2012) Marginalized stacked denoising autoencoders. In: Proceedings of the learning workshop, Utah, UT, USA, vol 36
240. Gu F, Flórez-Revuelta F, Monekosso D, Remagnino P (2015) Marginalised stacked denoising autoencoders for robust representation of real-time multi-view action recognition. Sensors 15(7):17209–17231
241. Tang Y, Zhang L, Min F, He J (2022) Multiscale deep feature learning for human activity recognition using wearable sensors. IEEE Trans Ind Electron 70(2):2106–2116
242. Sunasra M (2017) Performance metrics for classification problems in machine learning. Medium recuperado de. https://medium.com/thalusai/performance-metrics-for-classification-problems-in-machine-learningpart-i-b085d432082b
243. Bhattacharya D, Sharma D, Kim W, Ijaz MF, Singh PK (2022) Ensem-har: An ensemble deep learning model for smartphone sensor-based human activity recognition for measurement of elderly health monitoring. Biosensors 12(6):393
244. Fawcett T (2006) An introduction to roc analysis. Pattern Recognit Lett 27(8):861–874
245. Challa SK, Kumar A, Semwal VB (2022) A multibranch cnn-bilstm model for human activity recognition using wearable sensor data. Vis Comput 38(12):4095–4109
246. Deng Z, Zhai M, Chen L, Liu Y, Muralidharan S, Roshtkhari MJ, Mori G (2015) Deep structured models for group activity recognition. arXiv:1506.04191
247. Braun A, Musse SR, de Oliveira LPL, Bodmann BE (2003) Modeling individual behaviors in crowd simulation. In: Proceedings 11th IEEE international workshop on program comprehension, pp 143–148. IEEE
248. Wirz M, Franke T, Roggen D, Mitleton-Kelly E, Lukowicz P, Tröster G (2012) Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings. In: 2012 IEEE 21st international workshop on enabling technologies: infrastructure for collaborative enterprises, pp 367–372. IEEE
249. Sunil A, Sheth MH, Shreyas E, et al (2021) Usual and unusual human activity recognition in video using deep learning and artificial intelligence for security applications. In: 2021 fourth international conference on electrical, computer and communication technologies (ICECCT), pp 1–6. IEEE
250. Mohan A, Choksi M, Zaveri MA (2019) Anomaly and activity recognition using machine learning approach for video based surveillance. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), pp 1–6. IEEE
251. Igwe OM, Wang Y, Giakos GC, Fu J (2020) Human activity recognition in smart environments employing margin setting algorithm. J Ambient Intell Hum Comput pp 1–13
252. Jiang Y, Wang J, Liang Y, Xia J (2019) Combining static and dynamic features for real-time moving pedestrian detection. Multimed Tools Appl 78(3):3781–3795
253. Parthasarathy P, Vivekanandan S, et al (2019) Detection of suspicious human activity based on cnn-dbnn algorithm for video surveillance applications. In: 2019 innovations in power and advanced computing technologies (i-PACT), vol 1, pp 1–7. IEEE
254. Hammerla NY, Fisher J, Andras P, Rochester L, Walker R, Plötz T (2015) Pd disease state assessment in naturalistic environments using deep learning. In: Twenty-Ninth AAAI conference on artificial intelligence
255. Um TT, Pfister FM, Pichler D, Endo S, Lang M, Hirche S, Fietzek U, Kulić D (2017) Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 216–220

256. Nemati E, Zhang S, Ahmed T, Rahman MM, Kuang J, Gao A (2021) Coughbuddy: multi-modal cough event detection using earbuds platform. In: 2021 IEEE 17th international conference on wearable and implantable body sensor networks (BSN), pp 1–4. IEEE
257. Xu X, Nemati E, Vatanparvar K, Nathan V, Ahmed T, Rahman MM, McCaffrey D, Kuang J, Gao JA (2021) Listen2cough: leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. Proc ACM Interact Mob Wear Ubiquitous Technol 5(1):1–22
258. Zhang S, Nemati E, Ahmed T, Rahman MM, Kuang J, Gao A (2021) A novel multi-centroid template matching algorithm and its application to cough detection. In: 2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC), pp 7598–7604. IEEE
259. Arifoglu D, Bouchachia A (2017) Activity recognition and abnormal behaviour detection with recurrent neural networks. Procedia Comput Sci 110:86–93
260. Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C, Sontag D, Picard R (2017) Objective assessment of depressive symptoms with machine learning and wearable sensors data. In: 2017 seventh international conference on affective computing and intelligent interaction (ACII), pp 325–332. IEEE
261. Gao Y, Long Y, Guan Y, Basu A, Baggaley J, Ploetz T (2019) Towards reliable, automated general movement assessment for perinatal stroke screening in infants using wearable accelerometers. Proc ACM Interact Mob Wearable Ubiquitous Technol 3(1):1–22
262. Parajuli N, Sreenivasan N, Bifulco P, Cesarelli M, Savino S, Niola V, Esposito D, Hamilton TJ, Naik GR, Gunawardana U et al (2019) Real-time emg based pattern recognition control for hand prostheses: a review on existing methods, challenges and future implementation. Sensors 19(20):4596
263. Samuel OW, Asogbon MG, Geng Y, Al-Timemy AH, Pirbhulal S, Ji N, Chen S, Fang P, Li G (2019) Intelligent emg pattern recognition control method for upper-limb multifunctional prostheses: advances, current challenges, and future prospects. IEEE Access 7:10150–10165
264. Taylor W, Shah SA, Dashtipour K, Zahid A, Abbasi QH, Imran MA (2020) An intelligent non-invasive real-time human activity recognition system for next-generation healthcare. Sensors 20(9):2653
265. Dashtipour K, Gogate M, Cambria E, Hussain A (2021) A novel context-aware multimodal framework for persian sentiment analysis. Neurocomputing 457:377–388
266. Wu W, Zhang H, Pirbhulal S, Mukhopadhyay SC, Zhang Y-T (2015) Assessment of biofeedback training for emotion management through wearable textile physiological monitoring system. IEEE Sens J 15(12):7087–7095
267. Yin Z, Zhao M, Wang Y, Yang J, Zhang J (2017) Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Comput Methods Progr Biomed 140:93–110
268. Hssayeni MD, Ghoraani B (2021) Multi-modal physiological data fusion for affect estimation using deep learning. IEEE Access 9:21642–21652
269. Khezri M, Firoozabadi M, Sharafat AR (2015) Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals. Compur Methods Progr Biomed 122(2):149–164
270. Mohino-Herranz I, Gil-Pita R, García-Gómez J, Rosa-Zurera M, Seoane F (2020) A wrapper feature selection algorithm: an emotional assessment using physiological recordings from wearable sensors. Sensors 20(1):309
271. Pareek P, Thakkar A (2021) A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif Intell Rev 54(3):2259–2322
272. Chintalapati S, Raghunadh M (2013) Automated attendance management system based on face recognition algorithms. In: 2013 IEEE international conference on computational intelligence and computing research, pp 1–5. IEEE
273. Lim JH, Teh EY, Geh MH, Lim CH (2017) Automated classroom monitoring with connected visioning system. In: 2017 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), pp 386–393. IEEE
274. Aldrich FK (2003) Smart homes: past, present and future. In: Inside the smart home, pp. 17–39. Springer, Berlin
275. Rodriguez E, Chan K (2016) Smart interactions for home healthcare: a semantic shift. Int J Arts Technol 9(4):299–319
276. Moreno LV, Ruiz MLM, Hernández JM, Duboy MÁV, Lindén M (2017) The role of smart homes in intelligent homecare and healthcare environments. In: Ambient assisted living and enhanced living environments, pp 345–394. Elsevier
277. Alsamhi SH, Ma O, Ansari M, Meng Q et al (2019) Greening internet of things for greener and smarter cities: a survey and future prospects. Telecommun Syst 72(4):609–632
278. Howedi A, Lotfi A, Pourabdollah A (2019). Distinguishing activities of daily living in a multi-occupancy environment. In: Proceedings of the 12th ACM international conference on pervasive technologies related to assistive environments, pp 568–574
279. Oukrich N, Maach A, Sabri E, Mabrouk E, Bouchard K (2016) Activity recognition using back-propagation algorithm and minimum redundancy feature selection method. In: 2016 4th IEEE international colloquium on information science and technology (CiSt), pp 818–823. IEEE
280. Wilson DH, Atkeson C (2005) Simultaneous tracking and activity recognition (star) using many anonymous, binary sensors. In: International conference on pervasive computing, pp 62–79. Springer, Berlin
281. Wang L, Gu T, Tao X, Lu J (2009) Sensor-based human activity recognition in a multi-user scenario. In: European conference on ambient intelligence, pp 78–87. Springer, Berlin
282. Gu T, Wu Z, Wang L, Tao X, Lu J (2009). Mining emerging patterns for recognizing activities of multiple users in pervasive computing. In: 2009 6th annual international mobile and ubiquitous systems: networking and services, MobiQuitous, pp 1–10. IEEE
283. Yu BX, Chang J, Liu L, Tian Q, Chen CW (2022) Towards a unified view on visual parameter-efficient transfer learning. arXiv:2210.00788
284. Mittal H, Morgado P, Jain U, Gupta A (2022) Learning state-aware visual representations from audible interactions. arXiv:2209.13583
285. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
286. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Information Process Syst 33:1877–1901
287. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, et al (2022) Palm: Scaling language modeling with pathways. arXiv:2204.02311
288. He J, Zhou C, Ma X, Berg-Kirkpatrick T, Neubig G (2021) Towards a unified view of parameter-efficient transfer learning. arXiv:2110.04366
289. Bruce X, Liu Y, Zhang X, Zhong S-h, Chan KC (2022) Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. IEEE Trans Pattern Anal Mach Intell

290. Liu L, Yu BX, Chang J, Tian Q, Chen C-W (2022) Prompt-matched semantic segmentation. arXiv:2208.10159
291. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PMLR
292. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
293. Voelker A, Kajić I, Eliasmith C (2019) Legendre memory units: continuous-time representation in recurrent neural networks. Adv Neural Inf Process Syst 32
294. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł. Polosukhin I (2017) Attention is all you need. Adv Neural Information Process Syst 30
295. Takase S, Kiyono S (2021) Lessons on parameter sharing across layers in transformers. arXiv:2104.06022
296. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 11106–11115
297. Wei C, Fan H, Xie S, Wu C-Y, Yuille A, Feichtenhofer C (2022) Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14668–14678
298. Vicente-Sola A, Manna DL, Kirkland P, Di Caterina G, Bihl T (2022) Evaluating the temporal understanding of neural networks on event-based action recognition with dvs-gesture-chain. arXiv:2209.14915
299. Arandjelovic R, Zisserman A (2017) Look, listen and learn. In: Proceedings of the IEEE international conference on computer vision, pp 609–617
300. Arandjelovic R, Zisserman A (2018) Objects that sound. In: Proceedings of the European conference on computer vision (ECCV), pp 435–451
301. Korbar B, Tran D, Torresani L (2018) Cooperative learning of audio and video models from self-supervised synchronization. Adv Neural Inf Process Syst 31
302. Alwassel H, Mahajan D, Korbar B, Torresani L, Ghanem B, Tran D (2020) Self-supervised learning by cross-modal audio-video clustering. Adv Neural Inf Process Syst 33:9758–9770
303. Asano Y, Patrick M, Rupprecht C, Vedaldi A (2020) Labelling unlabelled videos from scratch with multi-modal self-supervision. Adv Neural Inf Process Syst 33:4660–4671
304. Morgado P, Li Y, Nvasconcelos N (2020) Learning representations from audio-visual spatial alignment. Adv Neural Inf Process Syst 33:4733–4744
305. Piergiovanni A, Angelova A, Ryoo MS (2020) Evolving losses for unsupervised video representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 133–142
306. Morgado P, Misra I, Vasconcelos N (2021) Robust audio-visual instance discrimination. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12934–12945
307. Morgado P, Vasconcelos N, Misra I (2021) Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12475–12486
308. Patrick M, Asano YM, Kuznetsova P, Fong R, Henriques JF, Zweig G, Vedaldi A (2020) Multi-modal self-supervision from generalized data transformations. arXiv:2003.04298
309. Wang X, Cai Z, Gao D, Vasconcelos N (2019) Towards universal object detection by domain attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7289–7298
310. Zhou X, Koltun V, Krähenbühl P (2022) Simple multi-dataset detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7571–7580
311. Lambert J, Liu Z, Sener O, Hays J, Koltun V (2020) MSEG: a composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2879–2888
312. Munro J, Damen D (2020). Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 122–132
313. Song X, Zhao S, Yang J, Yue H, Xu P, Hu R, Chai H (2021) Spatio-temporal contrastive domain adaptation for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9787–9795
314. Ghadiyaram D, Tran D, Mahajan D (2019) Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12046–12055
315. Duan H, Zhao Y, Xiong Y, Liu W, Lin D (2020) Omni-sourced webly-supervised learning for video recognition. In: European conference on computer vision, pp 670–688. Springer, Berlin
316. Akbari H, Yuan L, Qian R, Chuang W-H, Chang S-F, Cui Y, Gong B (2021) Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Adv Neural Inf Process Syst 34:24206–24221
317. Likhosherstov V, Arnab A, Choromanski K, Lucic M, Tay Y, Weller A, Dehghani M (2021) Polyvit: co-training vision transformers on images, videos and audio. arXiv:2111.12993
318. Liang J, Zhang E, Zhang J, Shen C (2022) Multi-dataset training of transformers for robust action recognition. arXiv:2209.12362
319. Zheng N, Wen J, Liu R, Long L, Dai J, Gong Z (2018) Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
320. Su K, Liu X, Shlizerman E (2020). Predict & cluster: unsupervised skeleton based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9631–9640
321. Lin L, Song S, Yang W, Liu J (2020) Ms2l: multi-task self-supervised learning for skeleton based action recognition. In: Proceedings of the 28th ACM international conference on multimedia, pp 2490–2498
322. Guo T, Liu H, Chen Z, Liu M, Wang T, Ding R (2022) Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. Proc AAAI Conf Artif Intell 36:762–770
323. Thoker FM, Doughty H, Snoek CG (2021) Skeleton-contrastive 3d action representation learning. In: Proceedings of the 29th ACM international conference on multimedia, pp 1655–1663
324. Li L, Wang M, Ni B, Wang H, Yang J, Zhang W (2021c) 3D human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4741–4750
325. Mao Y, Zhou W, Lu Z, Deng J, Li H (2022). CMD: self-supervised 3d action representation learning with cross-modal mutual distillation. arXiv:2208.12448
326. Jain Y, Tang CI, Min C, Kawsar F, Mathur A (2022) Collossl: Collaborative self-supervised learning for human activity recognition. Proc ACM Interact Mob Wear Ubiquitous Technol 6(1):1–28

327. Tran T, Do T-T, Reid I, Carneiro G (2019) Bayesian generative active deep learning. In: International conference on machine learning, pp 6295–6304. PMLR
328. Seyfioğlu MS, Özbayoğlu AM, Gürbüz SZ (2018) Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. IEEE Trans Aerosp Electron Syst 54(4):1709–1723
329. Zou H, Zhou Y, Yang J, Jiang H, Xie L, Spanos CJ (2018) Deepsense: device-free human activity recognition via autoencoder long-term recurrent convolutional network. In: 2018 IEEE international conference on communications (ICC), pp 1–6. IEEE
330. Abedin A, Ehsanpour M, Shi Q, Rezatofighi H, Ranasinghe DC (2021) Attend and discriminate: beyond the state-of-the-art for human activity recognition using wearable sensors. Proc ACM Interact Mob Wear Ubiquitous Technol 5(1):1–22
331. Huynh-The T, Hua C-H, Tu NA, Kim D-S (2020) Physical activity recognition with statistical-deep fusion model using multiple sensory data for smart health. IEEE Internet Things J 8(3):1533–1543
332. Hanif MA, Akram T, Shahzad A, Khan MA, Tariq U, Choi J-I, Nam Y, Zulfiqar Z (2022) Smart devices based multisensory approach for complex human activity recognition
333. Pires IM, Pombo N, Garcia NM, Flórez-Revuelta F (2018) Multi-sensor mobile platform for the recognition of activities of daily living and their environments based on artificial neural networks. In: IJCAI, pp 5850–5852
334. Zhang X, Yao L, Huang C, Wang S, Tan M, Long G, Wang C (2018) Multi-modality sensor data classification with selective attention. arXiv:1804.05493
335. Wang Y, Yao Q, Kwok JT, Ni LM (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv 53(3):1–34
336. Xie T-T, Tzelepis C, Fu F, Patras I (2021) Few-shot action localization without knowing boundaries. In: Proceedings of the 2021 international conference on multimedia retrieval, pp 339–348
337. Liu J, Shahroudy A, Perez M, Wang G, Duan L-Y, Kot AC (2019) Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell 42(10):2684–2701
338. Zhang H, Zhang L, Qi X, Li H, Torr PH, Koniusz P (2020) Few-shot action recognition with permutation-invariant attention. In: European conference on computer vision, pp 525–542. Springer, Berlin
339. Dai R, Lu C, Avidan M, Kannampallil T (2021) Respwatch: Robust measurement of respiratory rate on smartwatches with photoplethysmography. In: Proceedings of the international conference oninternet-of-things design and implementation
340. Li C, Niu D, Jiang B, Zuo X, Yang J (2021) Meta-har: federated representation learning for human activity recognition. In: Proceedings of the web conference, pp 912–922
341. Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G (2021) A survey on security and privacy of federated learning. Future Gener Comput Syst 115:619–640
342. Xiao Z, Xu X, Xing H, Song F, Wang X, Zhao B (2021) A federated learning system with enhanced feature extraction for human activity recognition. Knowl-Based Syst 229:107338
343. Pham HH, Khoudour L, Crouzil A, Zegers P, Velastin SA (2022) Video-based human action recognition using deep learning: a review. arXiv:2208.03775
344. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp 448–456. PMLR
345. Orchard G, Jayawant A, Cohen GK, Thakor N (2015) Converting static image datasets to spiking neuromorphic datasets using saccades. Front Neurosci 9:437
346. Li H, Liu H, Ji X, Li G, Shi L (2017) Cifar10-dvs: an event-stream dataset for object classification. Front Neurosci 11:309
347. Posch C, Matolin D, Wohlgenannt R (2010) A QVGA 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. IEEE J Solid-State Circuit 46(1):259–275
348. Lichtsteiner P, Posch C, Delbruck T (2008) A 128 128 120 db 15s latency asynchronous temporal contrast vision sensor. IEEE J Solid-State Circuit 43(2):566–576
349. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
350. Xu J, Yu Z, Ni B, Yang J, Yang X, Zhang W (2020) Deep kinematics analysis for monocular 3D human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 899–908
351. Nie Q, Liu Z, Liu Y (2020) Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In: European conference on computer vision, pp 102–118. Springer, Berlin
352. Deng J, Pan Y, Yao T, Zhou W, Li H, Mei T (2020) Single shot video object detector. IEEE Trans Multimed 23:846–858
353. Deng J, Yang Z, Liu D, Chen T, Zhou W, Zhang Y, Li H, Ouyang W (2022) Transvg++: end-to-end visual grounding with language conditioned vision transformer. arXiv:2206.06619
354. Rao H, Xu S, Hu X, Cheng J, Hu B (2021) Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Inf Sci 569:90–109
355. Purushwalkam S, Ye T, Gupta S, Gupta A (2020) Aligning videos in space and time. In: European conference on computer vision, pp 262–278. Springer, Berlin
356. Recasens A, Luc P, Alayrac J-B, Wang L, Strub F, Tallec C, Malinowski M, Pătrăucean V, Altché F, Valko M et al (2021) Broaden your views for self-supervised video learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1255–1265
357. Abdel-Basset M, Hawash H, Chang V, Chakrabortty RK, Ryan M (2020) Deep learning for heterogeneous human activity recognition in complex iot applications. IEEE Internet Things J
358. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D (2019) Invariant risk minimization. arXiv:1907.02893
359. Konečnỳ J, McMahan B, Ramage D (2015) Federated optimization: distributed optimization beyond the datacenter. arXiv:1511.03575
360. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
361. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
362. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929
363. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6836–6846
364. Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: ICML, vol 2, p 4

365. Liang J, Zhu H, Zhang E, Zhang J (2022) Stargazer: a transformer-based driver action detection system for intelligent transportation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3160–3167