# CHAPTER I

# INTRODUCTION

## 1.1 HUMAN ACTIVITY RECOGNITION

Human Activity Recognition (HAR) is an approach to recognizing human behavior and intent based on a set of observations of human activity and its context. HAR is a subfield of computer science and engineering that aims to develop tools for automatically identifying and classifying human behavior in video footage. To recognize distinct human activities, HAR has been viewed as a common classification challenge in computer vision and pattern recognition. The increasing demand for HAR that makes use of visual and sensory data has aroused the curiosity of researchers. There is also a lot of debate over whether or not vision-based HAR techniques are more effective than sensor-based HAR approaches. Healthcare, surveillance, sports and event analysis, care for the aged, and Human-Computer Interaction (HCI) are just few of the current application sectors where HAR has been used (Dang et al. 2020). Lighting, backgrounds, crowded settings, camera viewpoint, and action complexity are some of the aspects that affect HAR's accuracy (Beddiar et al. 2020). Globally, the adoption of HAR applications has greatly increased people's security and prosperity (Chen et al. 2021).

## 1.2 HAR IN VIDEOS

Videos used by HAR systems may come from a variety of sources, including those recorded by static digital cameras, surveillance cameras, and television cameras that film from a variety of angles. (Paul et al. 2013) took a look at how and where humans might be spotted in surveillance footage. Any given sports video sequence's statistical features will change based on factors including lighting, camera settings, and more.

### 1.2.1 Basic Properties of Videos

In order to recognize events, a computer vision system needs to collect a set of shared characteristics from incoming visual data. The basic need is for the computer vision communities to agree on how to choose suitable parameters for event recognition. The following characteristics cause slight shifts in these parameters:

- **Spatial orientations**

It describes where something or someone is located in space are called "spatial orientation terms." Constructing spatial models from visual inputs is the focus of this effort, with the end result being a more complete understanding of the objects observed in video sequences. However, these are not trustworthy for shooting outside, as variations in lighting and atmosphere can cause significant distortions in the final product.

- **Shape**

The shape of an object is often fixed in computer vision programs. It is easy to find any analytical model that is less sensitive to variations in illumination if it is based on the external shape of the body, but this is not good enough for distinguishing between things and humans. Only in cases where the spatial information is not completely dependable in that situation shape information were most helpful.

- **Motions**

In most cases, actions or events span multiple consecutive frames, and their progression is often in a single direction. The acts of humans in sporting events are always a bit haphazard. Motion features should be used to estimate the human's position and orientation.

- **Scale and template transformations**

The scale variations over time in a successive frame sequence must be taken into account when refining the spatial feature space of the detected object. In addition, the fluctuations across a given activity necessitate careful feature selection.

## 1.3 IMPORTANCE OF HAR

The HAR is crucial in many different situations. Human operators have historically been responsible for monitoring and evaluating human behavior to ensure safety in surveillance settings, keep tabs on a patient's health to head off any potential complications, or develop new types of machine interaction games. One of the most important jobs for mobile and wearable sensing is human health monitoring because of

its significance. In many public places, human activity recognition is used for safety checks. Improvements to the HAR system have allowed for more precise and timely observations and analyses.

## 1.4 CHALLENGES IN HAR

One of the biggest obstacles to creating an action recognition system for humans is extracting features that can capture enough information to discriminate between different actions. When comparing features across activities in the same class, it is important to be able to ignore changes between tasks because the same activity might be completed in a variety of ways, even by the same individual. Both ambient noise and camera shake should have no effect on the functionality. Massive amounts of annotated videos are needed for training a classification system with these features. Large datasets with millions of annotated action videos are available, however the vast majority of these datasets are devoted to sports and so do not capture the complete spectrum of actions that occur in a typical day. In addition, high-priced hardware like graphics processing units (GPUs) is required for training on massive datasets. Human action recognition is difficult because some of the following challenges that arise during the various stages of the process:

### 1. Intraclass Variation and Interclass Similarity

Different students may approach the same activity in different ways, leading to intraclass differences. Performance velocities and muscularities also amplify existing class differences. Interclass similarity, also known as HAR, refers to the phenomena where two distinct classes of objects share the same shape, such as a laptop and a book. Figure 1.1 graphically displays the differences and similarities present in the feature space. To address these problems, we must develop and extract distinctive features from activity videos.

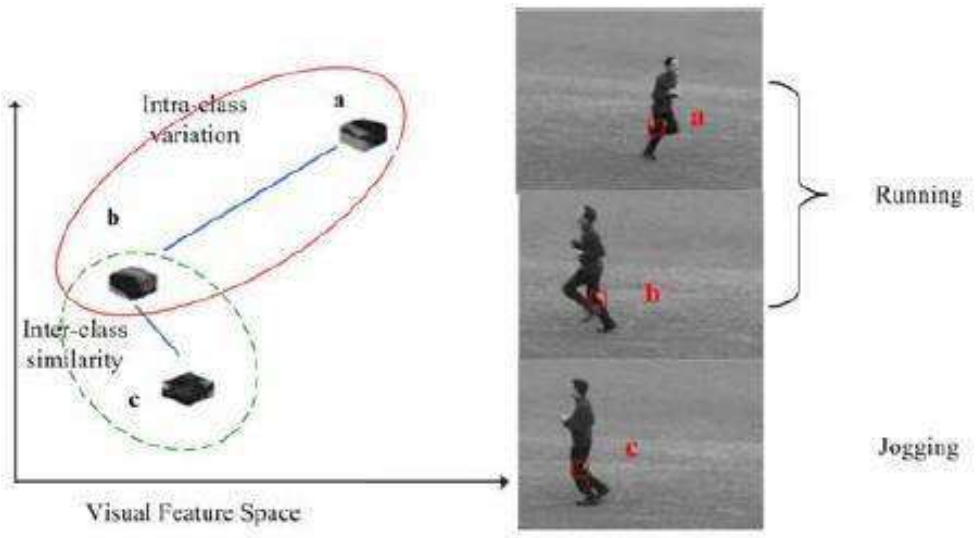


Figure 1.1 Variations and similarity in feature space

## 2. Complex and Various Backgrounds

Many modern applications rely on dynamic recording equipment, such as video surveillance and motion detection systems. Figure 1.2 depicts the example snapshots for different backgrounds. A growing trend is for people to use the built-in cameras on their wearable devices to shoot videos whenever they like. This is because of the widespread adoption of smart technologies like smartphones and wearable computers. These videos, captured with dynamic recording equipment, are authentic representations of the world as it actually is. It's split in two halves. One option is to use a number of various settings while filming the recordings or broadcasts. Second, occlusions, illumination fluctuation, and shifts in viewpoint make it more difficult to pick out actions in these videos.

**(a) Baseball swings** **(b) Golf swings** **(c) Tennis**

**Figure 1.2 Picture with different backgrounds**

### 3. Long-Distance and Low-Quality Videos

The field of video surveillance frequently faces the challenge of long-distance, low-quality, heavily-obstructed videos. Occlusions occur frequently in large, congested spaces like subway stations and airport terminals. Furthermore, surveillance cameras placed at great heights are unable to provide high-quality recordings, where the intended subject is easily discernible. Not everyone needs to be monitored, but the HAR system needs to be trained to identify suspicious or criminal actions.

Football broadcasts are another frequent example of long-distance interaction. The relatively low quality of long-distance camera placement, combined with the tiny size of the subject, makes it challenging to assess activities. Figure 1.3 displays long-distance, low-quality examples of action videos.

### 4. Temporal variations

Time is often considered to be neatly discrete when separating out events. While such an assumption relieves the recognition process of segmentation, it requires the use of a prior segmentation step. This is not always possible, and this challenge has been addressed in current action detection research. There can also be large differences in how quickly an action is completed. The time span covered by an action depends heavily on the recording rate, especially when motion characteristics are employed. An efficient algorithm for action recognition needs to be able to disregard individual differences in performance speed.

**(a)** **(b)**

**Figure 1.3 Action frames for (a) low- quality videos and (b) long-distance video**

### 5. Dynamic Illumination

Illumination shifts detract affect the visual appeal of video frames during processing. This is especially problematic in HAR methods, which rely heavily on human subjects in the foreground and on spatial and boundary relationships. Incorrect outcomes are produced for sports videos with erratic lighting. Figure 1.4 shows Example action videos with different illuminations.

**Figure 1.4 Action videos with different illuminations**

### 6. Execution rate

The time it takes to complete a task might vary greatly depending on the conditions. When the employed features are motion or temporal windows, this can have an effect on action recognition. It's important for an action recognition system to work regardless of how fast or slow an action is being performed.

## 1.5 DIFFERENT TYPES OF HUMAN ACTIVITIES

Any action carried out on a regular basis by a human being or humans is human activity. Indoor and outdoor activities are equally possible and can include strolling, lying down, and playing football or riding horses. Actions, behaviors, and gestures all fall under the umbrella term "human motion."

Figure 1.5 Demonstrates the Variety of Human Efforts. The complexity of human endeavors leads to the classification of those endeavors into,

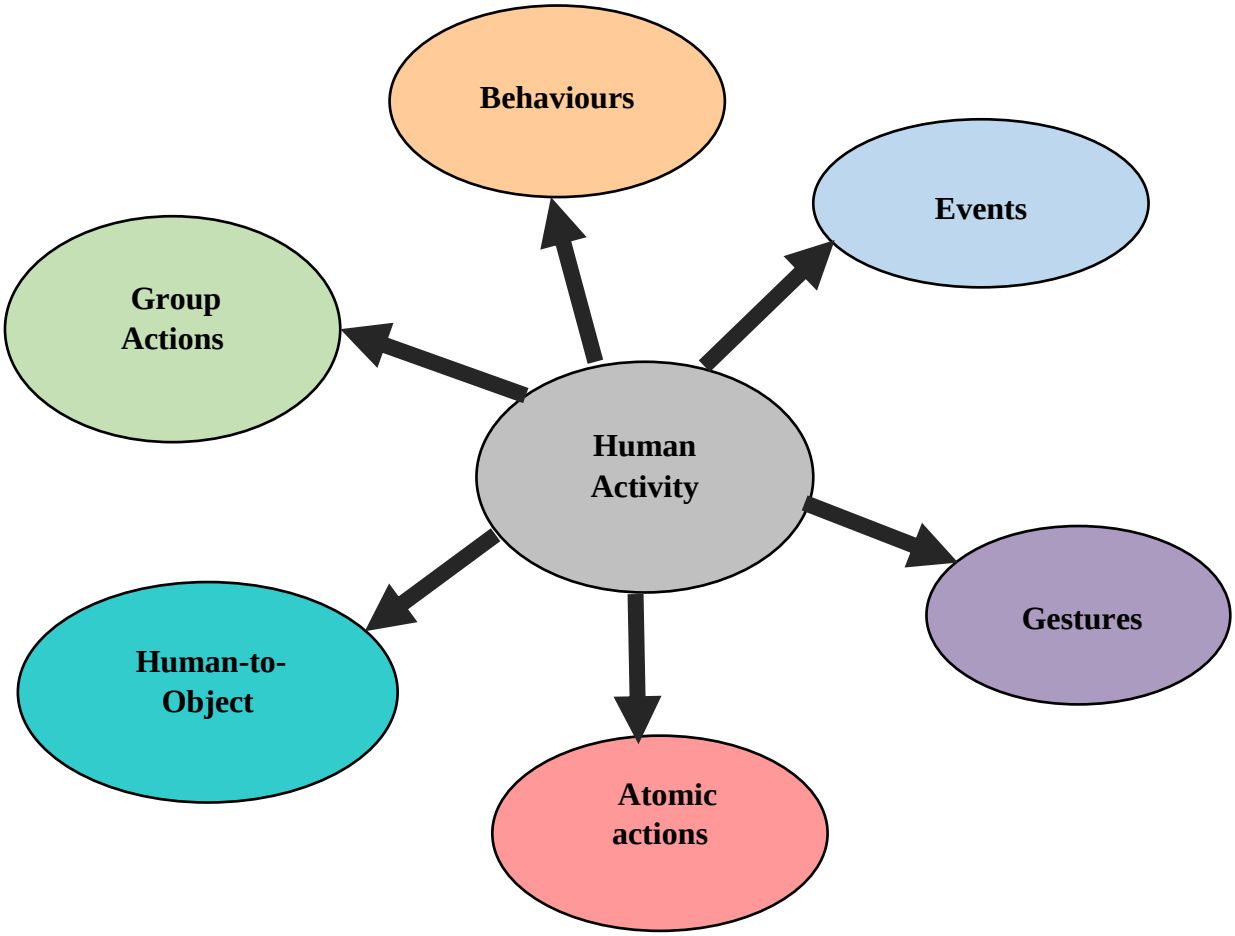


**Figure 1.5 Types of Human Activities**

**(i)** **Gestures** are interpreted as undeveloped actions that correspond to physical shifts in a person's body (Yang et al. 2013).

**(ii)** **Atomic actions** human actions describing a specific motion that can be applied into more involved activities (Ni et al., 2015).

**(iii)** **Human-to-object** are those carried out by humans that involve cooperation between multiple entities (Patron-Perez et al., 2012).

**(iv)** **Group actions** encompass collective endeavors (Tran et al., 2014b).

**(v)** **Human behaviours** are actions that reveal something about a person's character, mental health, and emotional make-up (Martinez et al., 2014).

**(vi)** **Events** are figurative deeds that tell us about the nature and purpose of human interactions and the roles they play in society (Lan et al. 2012).

## 1.6 TYPES OF HUMAN ACTION RECOGNITION

Recognizing the many kinds of human actions shown in Figure 1.6. Daily activity tracking has exploded in popularity and use as technology has improved and the price of monitoring devices has dropped. Activities such as cooking, eating, sleeping, and watching television, as well as the number of steps walked, are being recorded by many people. Various methods have been employed to record these events. Multiple strategies exist for tackling the problem of human activity recognition. Just to sum it up, they are,

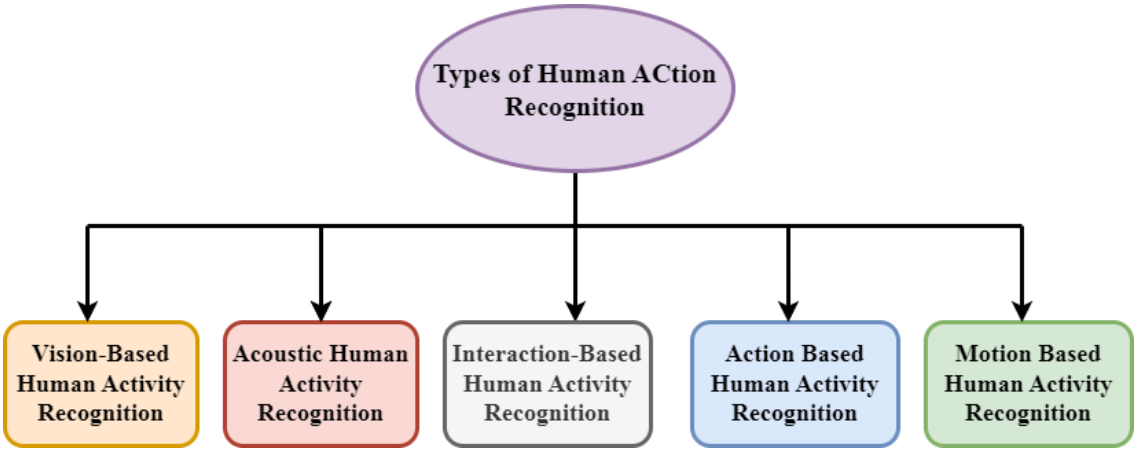


Figure 1.6 Types of human activity recognition

- Vision-based HAR
- Acoustic HAR
- Interaction-Based Sensor HAR
- Action Based HAR
- Motion Based HAR

### 1.6.1 Vision-based Human Activity Recognition

Many different configurations of cameras (monocular, stereo, infrared, thermal) and methods of describing the data (single-layer space-time, multi-layer description) have been used in systems for recognizing human activities based on machine learning techniques. The vision-based methodology was one of the first approaches taken in this area, and it entails recording information about people's actions using a camera. Activities can be identified in this collected data by employing computer vision

techniques. While there are many benefits to using computer vision-based techniques, there are also many problems that come along with this method. Secrecy is a top priority. The method also has the problem of being light-dependent.

In recent times, the advancement of cameras capable of capturing depth information has facilitated the progression of human activity detection using vision-based monitoring from a two-dimensional shapes to a three-dimensional context. Single- and direct-3D imaging technologies are now widely available at low cost, especially since the release of the Microsoft Kinect sensor (Shotton et al. 2013). Human activity detection is now feasible for in-home monitoring of the elderly because to the decreased prices and practical form factors of depth video sensors. (Jalal et al., 2014) proposed a life logging system that uses depth to learn about the habits of the elderly. In the beginning, depth images are obtained by a depth imaging sensor. These silhouettes form the basis for joint-informed human skeletons, which in turn are employed for activity recognition and the creation of life records. There are two distinct parts to the life-logging process. Initially, data was gathered using a depth camera, afterwards, distinctive attributes were recognized, and subsequently, a hidden Markov model (HMM) was constructed for each specific activity. The subsequent phase included the use of a recognition engine to categorize the incoming data and generate pertinent life records. The system was tested on the smart indoor activity datasets, where it performed admirably in terms of recognition rates thanks to an analysis of life recording features in comparison to main component and independent component features.

### 1.6.2 Acoustic Human Activity Recognition

Sounds in the background can provide clues about what goes on in a certain place. Several investigations have been conducted to determine whether or not auditory data may be used to identify activities. To categorize a large variety of human activities performed in the home, the web-based method of non-Markovian ensemble voting was developed by (Stork et al., 2012). Their algorithm does not rely on the detection of quiet or the separation of sounds in the audio stream. In addition, the approach can manage tasks that span indefinite amounts of time. The recognition was accomplished by using a scoring system applied to the votes provided by short audio frames in a

consistent manner with respect to the learnt model, which is based on sound books learned of activity courses. Real-world testing shows that the approach has an impressive 85% recognition rate for a continuous activity recognition setting, recognizing 22 distinct sounds that correlate to a wide range of human activities.

### 1.6.3 Interaction-Based Sensor Human Activity Recognition

The concept of using ambient sensors based on user interaction for smart home automation was first introduced in the late 1990s. Research utilizing such sensors for identifying patterns of behavior dates back to the turn of the millennium. The Gator Tech smart house was built at the University of Florida to investigate ambient assisted living (Helal et al., 2019). A number of sensor-equipped smart home appliances, including a smart refrigerator, were installed throughout the dwelling. Aware Home (Mano et al., 2016) was created by researchers at Georgia Tech to accomplish similar goals. In order to pinpoint their precise location, they utilized a number of in-ceiling cameras and radio-frequency identification (RFID) tags. Living laboratory experiments like these were among the earliest attempts to demonstrate an idea's viability.

House n, developed at Massachusetts Institute of Technology, was an early pioneer in the field of activity recognition. A total of more than 20 distinct occurrences were identified via the implementation of reed switch other piezoelectric switches across various household fixtures, including windows, doors, storage units, compartments, microwave ovens, refrigerators and freezers, cooking appliances, toilets, sinks, and bathrooms, shower switches for light lamps, boxes, and electronic devices, throughout two separate residential dwellings (Tapia et al., 2004). Subjects annotated their data using personal digital assistant (PDA) software; the data was then processed with a naïve Bayes classifier; and, depending on the evaluation metric, it showed a performance of 25% to 89%.

A comprehensive framework was devised that include the two processes of prediction as well as activity identification (Fatima et al., 2013). Researchers used a Support Vector Machine (SVM)-based kernel fusion activity detection approach to predict people's actions by identifying the most important sequential activities of the occupants. In order to forecast behavior, CRF was employed as a classifier. When comparing the linear kernel, the radial basis function kernel, the polynomial kernel, and

the multi-layer perceptron kernel, the fusion technique improved accuracy for detected activities by an average of 13.82%. When comparing CRF with HMM for action prediction, it was found that CRF yields an f-measure improvement of 6.61% to 6.76%.

### 1.6.4 Action Based Human Activity Recognition

(a)Kicking (b) Jumping (c) Pull ups (c) Squats (d) Tennis serves

Figure 1.7 Action based activities

Figure 1.7 depicts an example image of an activity that requires physical movement. Action-based activities are those that require participants to actively engage their bodies. This game can be played on any area of the body, or even the whole thing. Here, they compile the numerous solutions to the challenge of tagging human behavior. Figure 1.8 depicts an example of human action recognition using only visual cues.

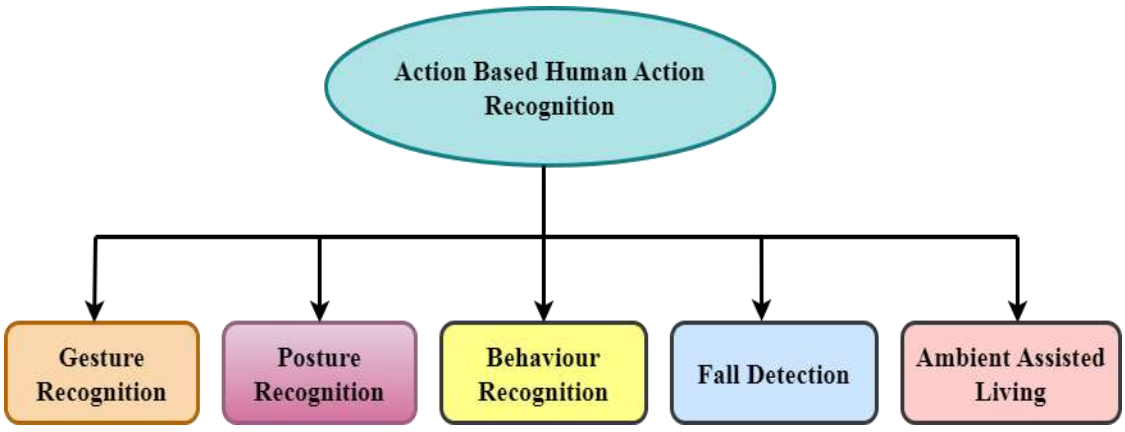


**Figure 1.8 Overview of Action Based Human Action Recognition**

**Gesture Recognition:** It is an essential part of the action recognition field. In recent years, its significance in the study of how humans interact with machines has come to be generally acknowledged. Until recently, humans communicating with computers required the use of input devices like mice, keyboards, and touch screens. However, that's not always possible. Disabled or elderly people, for instance, may struggle to

operate such gadgets. It may be difficult to install these input devices in heavily trafficked or high-priced public areas like parks, airports, and hospitals. Touch screens can spread germs and should be avoided in settings where the risk of infection is high, such as hospitals. Researchers have been working on new methods of human-machine interaction for the past decade. Wearable gadgets are being used in some systems for gesture recognition. Sensors to specialized gloves and bracelets are all included in this category. In order to recognize drinking motions, (Jayatilaka et al. 2017) employed A technologically advanced cup that is fitted alongside inactive Recognize Fluid Intake Gestures (RFIG) tags, while (Chen et al. 2017) attempted to do the same with a wrist-worn sensor. Link State Indicator (LSI) was presented as a device-free solution for gesture identification employing arrays of passive RFID tags (Ye et al., 2014). The number of tag reads serves as an indicator of connectivity. In a given time period, LSI calculates the percentage of correctly read tags in relation to the total number of references. The obstruction-free count is used as a reference. This study determines the gesture matrix for each gesture, which indicates whether or not all tags are completely obscured, slightly obscured, or not obscured at all. After all else has been tried, gesture recognition relies on Fisher's linear discriminant approach. The authors have conducted studies with a total of twelve different gestures to assess the efficacy of this strategy, including six squatting gestures and six standing gestures. For all twelve motions, the algorithm achieves an average accuracy of 94%. Since this method is not online, it cannot offer continuous recognition. The recognized movements are extremely dissimilar, and the performance is especially bad for gestures that are highly similar. The study does not address the issue of the proposed solution's complexity (in space and time). There haven't been nearly enough experiments done. Differences in execution, whether across individuals or within an individual's own performance of a gesture, are not addressed.

**Posture Recognition**: People engage in a wide variety of pursuits on a daily basis. These positions can be as basic as standing, sitting, lying down, or walking, or as sophisticated as sprinting, working out, and preparing elaborate meals. Recognizing even the most fundamental of these movements (postures) can be valuable in many settings. There are two broad categories that can be used to classify the many sensor-based solutions that researchers have explored for posture detection.

i. Utilizing sensors worn directly on the body
ii. Utilizing sensors placed throughout the outside world without the need for any additional hardware

In the first type of solution, various sensors are attached to a person or their clothing while they go about their daily activities. The inertial sensors in smartphones enable certain solutions, but this requires users to always have their device on them. (Torres-Huitzil et al., 2015) provides a technique that takes advantage of the accelerometer sensor found in most current cellphones. The Android smartphone is the first to fully use this strategy, increasing the variety of possible hand placements and viewing angles for mobile use. One major drawback of wearing sensors is that it's not always practical to use them while engaging in certain activities.

**Behaviour Recognition**: Human activity recognition relies heavily on this technique. From the information gathered by numerous sensors, a person's behavior can be inferred or recognized. Malls and other "smart" places like assisted living facilities and homes for the elderly are just two examples of where this technology could be put to good use (Chua et al., 2009). Due to the high cost of personnel, remote patient monitoring offers a major cost savings for senior care facilities. Those who need to be made aware of any concerning changes in an elderly person's behavior can be alerted immediately. client tracking in shopping malls can benefit business owners by revealing patterns of client behavior. Information about customers' purchasing habits, including their favorite products, brands, and interests, can be used to enhance the shopping experience for those customers. There has been a lot of effort put into understanding consumers' shopping habits recently.

**Fall Detection:** The term "fall" refers to the involuntary transition from an upright (or sitting, or walking) to a lying down position. Injuries from a fall can range from minor to severe. Merely a small proportion, around 3%, of falls lead to fractures. However, it is important to note that non-fracture injuries are still significant since they may impede the healing process and impose additional strain on the patient.

**Ambient Assisted Living:** Due to a combination of factors, including a declining birthrate and an expanding life expectancy, the world's population is getting older. The percentage of Australians 65 and up has doubled since 2006, and it is expected to reach 30% by 2056, as reported. Taking care of an aging population and the rising cost of

healthcare go hand in hand. Most persons over 65 either live independently or in residential care facilities for the elderly. They also necessitate care, which adds another burden on the working population. Extensive research has been done in recent years to try to understand and resolve such problems. Many technologies that improve the quality of life have been developed by scientists under the new paradigm of ambient intelligence. Remote monitoring, medicine administration, medication reminder, exercise management, and maintaining one's freedom are just a few of the ways in which Ambient Assisted Living (AAL) aids are transforming the lives of its users. Many strategies for assisting the elderly in maintaining their autonomy have been developed during the past decade under the umbrella of AAL (Queirós et al. 2017).

### 1.6.5 Motion Based Human Activity Recognition

In all of these endeavours, human motion is essential. In addition to actions, activities can also include the detection of things like presence or absence, motion, and so on in a monitored space. In the realms of security and monitoring, it is extremely helpful to be able to identify motion-based actions. Figure 1.9 shows the examples of motion based activities. Let's take a moment to review some of the most common types of motion based activities.

**Figure 1.9 Motion-based activities, such as path tracking, asset tracking, and recognizing movement in office environments.**

**Tracking:** Human activity recognition includes this as one of its key subfields. The Global Positioning System (GPS) makes tracking a breeze in the great outdoors, but it can't be used indoors. Augmented reality, occupancy detection, and indoor navigation are just a few of the many applications that might benefit from tracking data. There has been a lot of effort put into this because of how crucial it is becoming. The individual must also carry a device or tag, which is an inconvenience of the device-bound techniques. However, in other situations, including when tracking animals or unknown people, it is not practical to have them wear a gadget or tag. Users are not need to take any equipment around with them when using a device-free method (Li et al., 2017).

**Motion Detection**: It's a method for identifying the existence of anything that moves within a certain region. The significance of motion detection is highlighted by its use in surveillance and security applications, smart home and health monitoring systems, and other related fields. A smart house is one in which the environment is modified in response to the occupants' actions. The first and most fundamental requirement is awareness of whether or not there is a resident. Detecting the presence or movement of unauthorized individuals is a fundamental part of security and surveillance. Health care and remote patient monitoring, especially for the elderly, also benefit greatly from motion detection technology. Motion detection problems have been solved in a variety of ways. The Efficient motion detection (EMoD) method, presented by (Zhao et al. 2015), can not only pinpoint the location of an in-motion item, but also reveal its trajectory. Passive RFID tags are used in EMoD; therefore, no further hardware is needed. Multiple paired passive RFID tags are placed throughout the monitored area. Similar to Twins (Han et al., 2015), EMoD relies on a critical condition of the tags idea.

## 1.7 APPLICATIONS OF HUMAN ACTIVITY RECOGNITION

Recognizing human actions is a challenging and complex subject. The main goal of HAR is to explain human behavior by evaluating data collected from it. Given the importance of recognizing and making sense of human behavior, HAR has numerous applications. A study by Ranasinghe et al. (2016). The following are examples of where HAR has been put to use.

**Elder Health Care**

As medical knowledge and technology have progressed, so too has the average human lifespan. The global population is aging at an unprecedented rate. The percentage of some people aged 65 and up currently stands at 15% and is expected to treble by the year 2056. According to another estimate by Goldstone (Sonia et al. 2016), by 2050, 30% of the populations of China, Europe, Canada, and the United States would be 60 or older. Many issues, such as rising healthcare costs and a scarcity of caregivers, will arise as a result of the aging population. The need for healthcare professionals who can provide routine assistance to the elderly will rise as a result. In recent years, numerous proposals have been made to address this problem. The elderly can remain in their homes as they age. Various applications of this technology are now aiding individuals, including remote surveillance, detection of falls, managing medications, medication reminders, exercise management, as well as assistance for independent living (Queirós et al., 2017). The umbrella term for all of these devices is "Ambient Assisted Living" (AAL) technology. Human activity recognition (HAR) is an integral part of AAL. Thanks to recent developments in human activity recognition, elderly people can now be monitored remotely without requiring round-the-clock supervision. Human activity recognition technology allows the elderly to live independently. By monitoring human behavior and reporting any out-of-the-ordinary incidents like falls, HAR is helping the healthcare industry save money. Because of the world's aging population, there may soon be a severe scarcity of medical professionals, but HAR is helping to minimize this demand.

**Intelligent Environment**

Objects in an intelligent environment are able to perceive their surroundings and exchange information with one another. Sensors and other embedded devices in these settings can be used to gather and relay data. Nowadays the topic of "smart environment" construction is currently a very interesting and trending research topic. Smart environments, such as homes, offices, hospitals, and nursing homes, rely heavily on human activity recognition (Cumin et al. 2017). When its inhabitants are asleep, for example, the lights in a smart home will automatically shut off. Human activity recognition is useful because it provides insight into what people are up to in their communities. Smart environments can learn about the habits of its inhabitants through HAR methods and adjust to them automatically. If no one is home, a smart home might

shut down the lights, the HVAC, the windows, and other utilities. The moment its inhabitants step foot inside, the smart home will activate its lighting and other devices in preparation for their arrival. The behaviors of patients can be remotely observed in a context-aware smart health care center. Doctors may see how their patients are doing and verify their therapy and workout regimens. Seniors who live independently can get help from smart care centers. Most persons over 65 have many health issues, including mobility issues, and a smart care center can help them maintain their independence.

**Outdoor Navigation**

The use of GPS for outside navigation is mature and essential to the daily lives, but the same cannot be said for inside navigation. Most of a person's waking life also spent outside a structure, be it a house, an office, a store, or a restaurant. The weak GPS signal received outdoors (obstacles like concrete block the GPS signal) renders the technology useless indoors. An alternative to utilize GPS like one would for outside navigation is necessary for use. Important applications of navigation include localization and monitoring, helping the elderly and the disabled, supporting shoppers in a huge mall, helping people navigate vast buildings like airports and hospitals, assisting emergency response teams, and many more. Human activity identification research is being conducted to help with the challenge of outdoor navigation. A multitude of approaches have been used to tackle the challenges associated with the detection of human movement including presence, the localization and monitoring of their movements, and the identification of frequently traversed routes within a certain region (Khan et al., 2022). Human activity recognition research encompasses all of these techniques, which are all contributing to improved navigation for the general public.

**Human Computer Interaction**

**Figure 1.10 Human Computer Interaction**

Human-computer interaction is seen in Figure 1.6. Recognizing human actions is becoming increasingly important in human-computer interaction research. Input devices, such as keyboards and mouse, are the standard means of communicating with a computer or other equipment. This strategy does not work in most cases. It is impractical to install input devices in places frequented by the public such as hospitals, train stations, airports, and parks. Additionally, certain users may have difficulty operating these input devices, including the elderly and the sick. Recently, a number of techniques have emerged for communicating with machines without the need of conventional input mechanisms; this is due in large part to studies in human activity identification. The ability to communicate with machines through free-form gestures in the air or predetermined actions is now a reality. These instructions can be understood by machines. There have been numerous proposals for improving gesture recognition over the past few decades. Some of these options involve a wearable method (Siddiqui et al. 2017), while others don't require any additional hardware (Parada et al. 2016)The gaming and entertainment industries have been completely transformed by these methods. With the use of HAR, players are able to engage with games by really doing things, and their actions are picked up by the game. Human activity recognition (HAR) is also facilitating robot-human interaction.

## 1.8 OVERVIEW OF HAR APPROACHES

### 1.8.1 Machine Learning Based Human Activity Recognition

Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are the three machine learning techniques used in the HAR analysis. Due to mature areas of Machine Learning (ML) and Deep Learning (DL), as well as improvements in sensor technology and analytical prowess, wearable sensors have become increasingly popular and accurate for Human Activity Recognition (HAR). (Mehata et al., 2020) compare and contrast the three most essential aspects of a generic HAR and examine how they impact overall system performance. The process entails constructing several classification algorithms, assessing their efficacy, examining the impact of pre-processing, and extracting and selecting features from time-series data. Logistic Regression, K-Nearest Neighbors, SV M, and Artificial Neural Networks (ANN) are employed as classification methods in this investigation. SHNN-CAD increased classification performance with minimal parameter tweaking, but it couldn't keep up

with a training set that grew on the go. Multitask grouping framework for action analysis (Yan et al., 2015) was developed using visual input data collected from wearable cameras.

Many Machine Learning (ML) methods have been available in recent years for application to HAR datasets. Among them are RF, SVM, KNN, and a plethora of others. The unique wearable system developed by (Casale et al., 2011) was validated using an RF classifier. Additionally, they do admirably in human activity recognition tests. The hip accelerometer's ability to accurately classify different types of Physical Activity (PA) behavior and to predict Energy Expenditure (EE) was investigated by developing and testing two predictive models (Ellis et al., 2014). These models were a recursive feature (RF) classifier to predict activity type and a recursive feature (RF) of regression trees to estimate Metabolic Equivalents (METs). To improve activity recognition accuracy while reducing noise in the data, (Zeng et al. 2014) introduce a new approach. By integrating the Independent Components Analysis (ICA) algorithm with the wavelet transform algorithm, they attempt to eliminate the influence of noise during feature extraction. The SVM technique is used as the classifier for activity recognition due to its high level of performance.

### 1.8.2 DEEP LEARNING BASED HUMAN ACTIVITY RECOGNITION

In the areas of computer vision and action detection, deep learning has recently been applied with great success. As a result, many researchers are integrating inertial sensor data with deep learning to classify human activities. Many neural network-based algorithms, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Convolutional Networks, have been created for human activity recognition. The widespread usage of Convolutional Neural Networks (CNNs) may be attributed to their capability to successfully record the regional correlations present in sensor signals and their robustness in handling variations in motion intensity. For behavior classification without specialized expertise, (Zeng et al., 2014) presented a convolutional neural network (CNN) based technique. The proposed approach can record both the location and timing details of activity signals at different resolutions. Thus, the extracted features properly reflect variations in the same action.

Jiang and Yin (2015) describe a method for generating activity images that involves stacking the original sensor signals and then applying Discrete Fourier Transform (DFT) to the resulting signal image. Data from the gyroscope, total acceleration, and linear acceleration were utilized to train a Deep Convolutional Neural Network (DCNN) that was then used to categorize human activities. To boost activity identification accuracy, an SVM was given the DCNN's low-probability output. For the purpose of identifying actions in an order-picking scenario, (Rueda et al. 2018) presented an architecture known as a CNN- inertial measurement unit (CNN-IMU) network. The architectural design has parallel branches that use spatial convolution with maximum pooling procedures to handle the multiple channels of communication time-series information from the various IMUs. The raw sensor data is represented in an intermediate form by combining the results from each branch with those of the fully connected layers during the classification process. An efficient encoding strategy was proposed by (Hur et al. 2018) to construct an accurate image of the sensor signals, and then image-based CNN models were utilized to categorize the activities. To minimize distortion while preserving precision, the researchers opted to partition the signals generated by the sensors into three distinct components: the integer component, the first second decimal place, and the subsequent two decimal places. Subsequently, each of these components was assigned to a distinct color channel inside the picture. Seven distinct layers make up the image-centric CNN model: The proposed architecture consists of six convolutional layers, followed by 2 maximum-pooling layers, a fully linked layer, and a softmax layer (Yoon, Cho, & Cho, 2018) A suggested approach for activity identification involves a two-stage process and the use of test results for refinement. After separating the dynamic from the static actions, they employed a 1D CNN model to identify each one. In the first phase, a 1-dimensional convolutional neural network (CNN) model was used, whereas in the subsequent phase, two 1-dimensional CNN models were utilized, with each model having three distinct classes. The precision of activity recognition can be enhanced by employing data sharpening, a technique that places more emphasis on the data's high-frequency elements.

In the year 2016, Ravi and colleagues presented a methodology for the identification of human activities with low-power sensors. In order to mitigate the complexity of neural networks, the researchers devised a shallow network consisting of

just three layers of information: a temporal convolutional layer, a fully connected layer, and a softmax layer. The shallow neural network was trained by using the spectrogram of a sensor signal in order to effectively handle a wide range of sensor configurations. The study provided evidence that the suggested approach is capable of real-time recognition of human behaviors on low-power devices, while its accuracy is comparatively lower than the strategy relying on handmade characteristics. They created a neural network that uses both manually generated data and features collected from a shallow network to increase the accuracy with which it can recognize human behavior. Their methods were just as precise as those that relied solely on hand-crafted features, despite the fact that they blended characteristics with varying values. In order to lessen the load on the computer, they proposed utilizing a shallow neural network, which comes with some drawbacks, such as a lower activity identification rate.

(Ordonez and Roggen, 2016) created a Deep Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)-based activity recognition classifier using two publically available datasets consisting of data from seven inertial measurement units and twelve triaxial accelerometer wearable sensors. Using a combination of CNNs and long short-term memory (LSTMs), the authors were able to classify 27 hand gestures (including opening a door, cleaning dishes, clearing a table, etc.) and five movements (including standing, walking, sitting, laying, and nulling) from the sensory data transformed into sensor signal graphs. F1 scores of 0.93 and 0.958 were achieved in the simulations. The F1 score is the mathematical mean of the two subscores, "precision" and "recall." The Deep Appearance and Motion Learning (DAML) method was developed by (Wang et al., 2018) to enable individual-level activity recognition. Short-term egocentric activity recognition also improved, whereas long-term activity recognition using a basic classifier did not.

(Kim et al., 2018) proposed a fully convolutional network (FCN) -based approach to activity recognition that effectively balances recognition speed and memory consumption. To boost recognition rates and reduce the number of parameters, this method substituted Deep Neural Networks (DNNs) with five convolutional layers for shallow neural networks. Despite the fact that the classification engine of the HAR system has been the subject of numerous studies aimed at improving its energy efficiency, the segmentation process has not yet been perfected.

## 1.9 RESEARCH MOTIVATION

Nowadays, HAR can be learned automatically from thousands of videos and applied to all aspects of daily life. A rapid increase in video content on social media platforms has led to a need for a system that can accurately analyse these videos and provide solutions and suggestions. HAR systems are essential for many applications, like sports, etc. By accurately identifying and analyzing athletes' movements and actions, HAR systems contribute to performance optimization, injury prevention, skill assessment, and enhancing the overall sports experience. Therefore, HAR is one of the emerging research fields in computer vision attracting a lot of attention in recent years. Traditional methods for HAR can be time-consuming because those often rely on handcrafted features and shallow machine learning techniques, which might struggle to capture complex patterns in video sequences. They cannot learn spatiotemporal dependencies as human activities often have both temporal and spatial dependencies, resulting in less recognition accuracy. Also, these methods cannot capture features that are invariant to the noise and variations in the video sequences, leading to less robustness. This motivated to carry out the proposed research work, to learn complex, spatiotemporal dependencies from large-scale video sequences for effective HAR. The presented research focuses on exploiting deep learning models for HAR that can increase the recognition rate of human activities.

## 1.10 PROBLEM STATEMENT

The Research focus is to develop an accurate and robust human action recognition system capable of accurately detecting, classifying, and localizing human actions within video sequences. Given an input video clip, the system should be able to identify the specific action being performed by one or more individuals. Most of the deep learning models only capture body joints from the human action video sequences for recognition, while they lack motion information because of ignoring trajectory points, or optical flow fields from the video frames. The use of max-min pooling in those deep learning models mostly ignores the significant spatial dissimilarities among multiple video sequences of different human action classes from large-scale datasets. This is because the max-min pooling is very flexible to spatially smooth over the nearby kernels. Also, those deep learning models cannot be effective for long-range

video sequences, which leads to ineffective temporal feature learning. Though the deep learning models create video descriptors by capturing discriminative body joints and trajectories for HAR, the accuracy was still not satisfactory the captured features only express gesture information and do not define contour or geometrical relations. To increase the recognition accuracy, it is essential to capture geometrics from the video clips and learn the spatiotemporal dynamics among various geometrics.

## 1.11 OBJECTIVE OF THE THESIS

The proposed research work primarily aims to precisely recognize human activities with high accuracy. To accomplish this, the following interim objectives are considered:

- To enhance the HAR performance, both body joints and trajectory points can be learned together.
- To extract the relevant spatial dissimilarities among various action classes, a Positional Attention-based Bidirectional Recurrent Neural Network (PABRNN) model is adopted instead of max-min pooling.
- To avoid the vanishing gradient problem and learn the long-range joint correlations among multiple actions, a Positional Attention-based Hierarchical BRNN (PAHBRNN) model is proposed.
- To create a more informative video descriptor, different geometric features from the skeleton graph can be extracted and learned along with the body joints and trajectory points.
- To automatically learn the spatiotemporal dynamics among geometric features, a Graph Convolutional Network (GCN) is proposed.

## 1.12 CONTRIBUTION OF THE RESEARCH

- In the first phase of the research, an optical flow is suggested to be incorporated with the two-stream bilinear model in the Joints and Trajectory-pooled 3D-Deep Convolutional Descriptor (JTDD), which improves recognition accuracy.
- The second step involves incorporating the PABRNN model into a “two-stream Convolutional 3D network (C3D) network in order to extract meaningful

spatiotemporal features and boost the accuracy of recognizing individual activities." This network is called the Joints and Trajectory-pooled 3D-Deep Positional Attention-based Bidirectional Recurrent Convolutional Descriptor (JTPADBRD).

- "Joints and Trajectory-pooled 3D-Deep Positional Attention-based Hierarchical Recurrent Convolutional Descriptors (JTDPAHBRD)" is a method developed in the third stage of this study that use a PAHBRNN to improve feature aggregation.

- In the fourth phase of the research, a JTD-Geometric and PAHBRD (JTDGPAHBRD) is proposed, which explores the geometries of the graph structure, namely joints, edges, and surfaces, for HAR by considering skeletons as a sequence of graph joints.

- In the fifth phase of the research, GCN with the JTDGPAHBRD (JTDGPAHBRD-GCN) to create a video descriptor for HAR. The GCN can obtain complementary information, such as higher-level spatial-temporal features, between consecutive frames for enhancing end-to-end learning.

## 1.13 SCOPE OF THE RESEARCH WORK

The scope of the research is unique and distinct as follows:

- It can be used in sports to improve the performance of players and analyze their game plans.

- It is used to understand human behaviour and interactions with the environment.

- It can be used to monitor and prevent a person from becoming a duped of strange activities, burglaries, vandalism, fights, and harassment.

- It is useful in smart homes to monitor and assist elderly people and children.

## 1.14 ORGANIZATION OF THESIS

- **Chapter 1:** It provides the introduction of human action recognition, application used in human action recognition, along with the difficulties inherent in HAR and the difficulties of human action prediction.
- **Chapter 2:** The research's primary focus is on human action recognition by drawing on existing knowledge of Machine Learning and Deep Learning techniques.
- **Chapter 3:** The proposed approach, performance metrics, and study domain for human action recognition are all laid out here.
- **Chapter 4:** The initial result of this study's JTDD is presented in this chapter. In this model, an optical flow/trajectory point is derived between video frames using the body's joint locations as input to the proposed JTDD. Descriptors created from a C3D with two streams multiplied by a bilinear product are pooled and used to extract and capture spatiotemporal aspects of video sequences. The network is trained using a two-stream bilinear C3D model, which is then utilized to describe the videos. After that, they put the linear support vector machine to work to sort the video's description.
- **Chapter 5:** This chapter deals about the second contribution of this study, the JTPADBRD and its effectiveness are discussed. Extraction of important spatiotemporal information and improved activity recognition are achieved by incorporating into a two-channel C3D network using the PABRNN model.
- **Chapter 6:** In this chapter the third work is discussed. It describes the JTDPAHBRD method, which employs a PAHBRNN to improve feature aggregation.
- **Chapter 7:** This chapter explains the fourth contribution of the research work about the functionality of the JTDGPAHBRD. To acquire additional discriminative high-level features, a novel 3D-deep convolutional network is constructed using a PAHBRNN with view conversion and temporal dropout layers. The data is then processed through a fully linked layer to produce the

video description for that frame. Furthermore, an SVM classifier is used to separate out distinct aspects of human behavior based on the video description.

- **Chapter 8**: This chapter describes fifth major contribution of this study such as the JTDGPAHBRD-GCN and its performance.
- **Chapter 9:** Displays a summary of the proposed algorithms' findings.
- **Chapter 10:** Presents the thesis's concluding conclusions and discusses ways in which the research could be improved in the future.

## 1.15 CHAPTER SUMMARY

This chapter covers a wide range of human action recognition types, uses, and machine learning/deep learning approaches for making predictions about human action recognition. In addition, it also describes the research objectives, contribution and problems on predicting human action recognition.