A BRIEF INTRODUCTION TO HUMAN ACTIVITY RECOGNITION USING DEEP LEARNING

Abstract

Deep learning techniques for human activity recognition are gaining popularity due to their effectiveness in identifying intricate tasks and their cost-effectiveness compared to conventional machine learning methods. Human Activity Recognition (HAR) is a research domain concerned with detecting the everyday activities performed by individuals using time-series data captured by sensors. HAR encompasses a wide range of applications, including surveillance, baby monitoring, elderly healthcare, and smart driving. This article provides a brief introduction to the application of deep learning in HAR. It covers the fundamental concepts of CNNs and LSTMs, their strengths in capturing spatial and temporal features, and their integration for enhanced activity recognition. Different approaches are employed in HAR to address problems with efficiency and precision. Conventional human activity recognition (HAR) systems rely on wearable devices like IMUs and stretch sensors to identify different activities. These systems have proven to be effective in recognizing simple user actions like sitting, standing, and walking. However, when it comes to more intricate activities like running, jumping, wrestling, and swinging, sensor-based HAR systems encounter greater misclassification rates due to inaccuracies in sensor readings. These errors significantly impact the accuracy of the HAR system, resulting in suboptimal classification outcomes. In contrast, employing vision-based HAR systems enables improved accuracy in identifying complex activities, leading to enhanced overall performance.

I. INTRODUCTION

Human Activity Recognition also known as HAR is a process of correctly identifying the actions performed by a person based on sensor data [1].The goal of HAR is to automatically recognize and classify human activities, such as cycling, running, jumping, and other physical activities, from data collected from wearable sensors[2], cameras like Kinect [3], CCTV[4], smartphones [5] or other devices. Human activity recognition using deep learning methods are becoming increasingly popular due to its high efficiency in recognizing complex tasks and its relatively low cost compared to traditional machine learning methods.

Figure 1: Various actions performed by a person

Recognition of human activities plays a vital role in the advancement of wearable devices, health monitoring systems, smart residences, security systems, and human-computer interaction. The extensive adoption of wearable sensors like smartwatches, fitness trackers, and security cameras has led to a growing need for dependable and precise systems for identifying and categorizing activities. The information gathered from wearable sensors can be extremely diverse and intricate. For instance, individuals may demonstrate distinct movement patterns even while engaging in the same activity, and an individual's movements may fluctuate over time based on factors like age, physical well-being, and surroundings. Additionally, the data collected from wearable sensors may contain noise, artifacts, and other sources of interference, which can make it difficult to accurately recognize and classify human activities.

Conversely, human activity recognition (HAR) systems based on visual information offer enhanced precision and intricate details by capturing an individual's movements in three-dimensional space [6]. These systems have the capability to incorporate contextual elements like the surrounding environment and objects present, resulting in a more comprehensive understanding of the performed activity. However, implementing vision-based systems can be more intricate and necessitates specialized hardware like video cameras or depth sensors. To address these obstacles, human activity recognition systems commonly depend on deep learning algorithms, including neural networks, decision trees, and support vector machines. These algorithms are trained using extensive datasets comprising labeled activity information. By learning to identify patterns within the data that align with various activities, these algorithms can effectively recognize and categorize human actions [7].

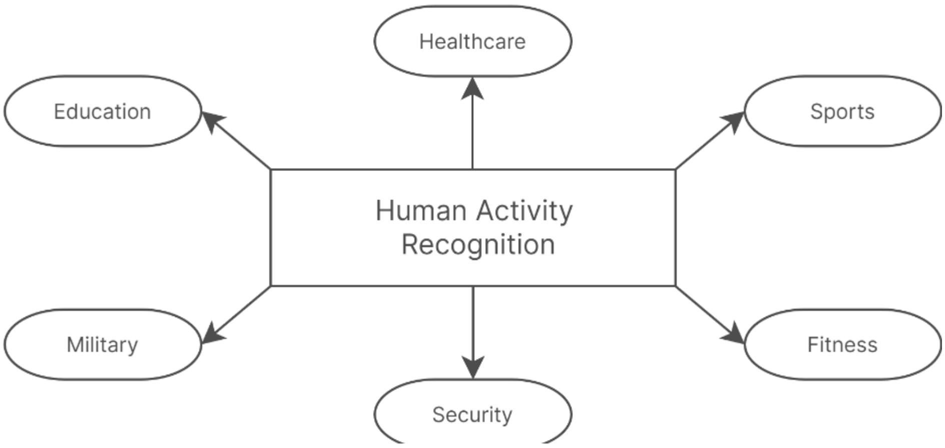

Figure 2: Human Activity Recognition in different fields

II. How A HAR system works?

In general, a human activity recognition system operates by employing diverse sensors and algorithms to detect and categorize distinct physical activities executed by a person. The system may use data from wearable sensors, such as accelerometers and gyroscopes, to collect information about the user's movements, body position, and orientation. The system then analyzes the data using machine learning algorithms to identify patterns and classify the activity. The system may use different techniques to analyze the data, such as feature extraction, pattern recognition, and statistical analysis. Once the system has determined the specific activity being carried out, it can offer the user feedback in the form of alerts or reminders, motivating them to uphold healthy habits or enhance their performance.

1. **Different stages of a HAR system: Human Activity Recognition (HAR) systems typically involve several stages, including:**
   - Data Collection
   - Data Pre-processing
   - Feature Extraction
   - Model Training
   - Model Testing and Validation
   - Deployment

2. **Types of HAR system:** Based on sensor type HAR systems can be divided into four categories [8].

   - **Sensor-based systems:** Sensor-driven human activity recognition (HAR) systems typically integrate a combination of various sensor types, including accelerometers (used to detect movements and postures), gyroscopes (used to detect rotation and orientation changes), pressure sensors (can be used to detect activities such as sitting, standing, and lying down), heart rate monitors (used to detect activities that involve physical exertion or stress) and magnetometers, to capture comprehensive data concerning an individual's movements and actions. This data is then processed using machine learning algorithms to classify the activities being performed.

A sensor-based HAR system has the capability to identify a range of activities, including walking, running, cycling, sitting, standing, and sleeping. These systems find utility across diverse applications, such as monitoring patients' physical activity levels, scrutinizing athletes' motions to enhance their performance, or controlling home automation systems in response to the occupants' activities.

- **Vision-based systems:** In a vision-based human activity recognition (HAR) system, visual data is captured using one or multiple cameras. This captured data is then analyzed to identify and categorize human activities. This approach has become increasingly popular due to the widespread availability of affordable and high-performance cameras in different devices.

  In vision-based HAR systems, there are multiple techniques to extract features from the visual data. Some commonly used methods include utilizing histograms of oriented gradients (HOG) [9], local binary patterns (LBP) [10], and employing deep learning-based approaches like convolutional neural networks (CNNs) [11].

- **RFID-based HAR system:** An HAR system based on RFID technology utilizes RFID tags, which can be affixed to objects or worn by individuals to monitor their activities. The RFID readers capture the data stored within the tags and transmit it to the processing unit, which then identifies the activities performed by the user. Both active and passive RFID technologies can be employed within HAR models [4],[12]. An advantage of RFID-based HAR systems is their non-contact nature, eliminating the need for physical interaction with the user, thereby enhancing comfort.However, the limitation is that the accuracy of the system may be affected by the distance between the RFID reader and the tag and cannot identify actions or behaviors that do not include the motion of objects equipped with RFID tags.

- **Wi-Fi-based HAR System:** A Wi-Fi-based Human Activity Recognition (HAR) system utilizes wireless signals emitted by Wi-Fi access points to identify and categorize human activities. By analyzing variations in signal strength and phase caused by human movements, the system can determine the type of activity taking place. The fundamental concept behind Wi-Fi-based HAR involves employing machine learning algorithms to analyze the data obtained from Wi-Fi signals and classify activities based on discernible patterns. To ensure accurate and reliable results, the system requires a Wi-Fi network equipped with access points strategically placed within the relevant area. Additionally, the system can be designed to function with multiple access points simultaneously, enhancing its accuracy and reliability. Various models have been developed utilizing Channel State Information (CSI) for purposes such as fall detection and gait recognition [13].

3. Different Methods used in a vision-based HAR system

- **Appearance-Based Methods:** Appearance-based methods in human activity recognition (HAR) rely on the visual appearance of body parts and their movements to classify different activities. These methods utilize various features like color, shape, texture, and motion to identify and categorize human actions. Several commonly employed appearance-based methods in HAR systems are:

- **Histograms of Oriented Gradients (HOG):** HOG computes a histogram that represents the orientations of gradients in an image. This information is utilized to detect human body parts and movements. HOG features have found extensive applications in activity recognition, particularly in tasks like pedestrian detection and tracking [13].

- **Scale-Invariant Feature Transform (SIFT):** SIFT identifies and extracts distinctive features from images that remain invariant to changes in scaling, rotation, and translation. These features can be used to recognize and track human body parts and movements. Recent studies have explored the use of SIFT in deaf sign language detection [14].

- **Local Binary Patterns (LBP):** LBP calculates a binary code for each pixel in an image based on the values of its neighboring pixels. This method has been widely used in activity recognition, especially for tasks like facial expression recognition [15].

- **Pose-Based Methods:** Pose-based methods in human activity recognition (HAR) rely on capturing and interpreting body poses and movements to identify and categorize activities. These methods utilize various features, including joint positions, angles, and velocities, to analyze the motion of the body and determine the nature of the activities performed. HAR systems commonly employ the following pose-based techniques:

4. Joint locations and Skeleton-based method:

- **Joint locations and Skeleton-based method:** The joint locations approach focuses on detecting and tracking the specific positions of body joints like elbows, wrists, and knees. On the other hand, the skeleton-based method constructs a skeletal representation of the human body using the detected joint locations. By utilizing either the joint positions or the skeletal model, it becomes possible to estimate the body's posture and movements, enabling the recognition and classification of different activities [16].

- **Optical Flow:** The optical flow method involves examining the motion of pixels between consecutive frames of a video sequence to estimate the body's movements. By analyzing the patterns of pixel motion, optical flow can determine the velocities and accelerations of various body parts. This information can then be utilized to recognize and categorize different activities [17].

- Kinematic analysis: Kinematic analysis involves applying the principles of kinematics to estimate the position, velocity, and acceleration of different body parts. By studying the movement patterns exhibited by the body, kinematic analysis enables the recognition and classification of activities based on these distinctive motion patterns.

- **Motion-based methods:** Motion-based techniques in Human Activity Recognition (HAR) systems employ various approaches to identify and categorize activities by monitoring the movement of body parts over time. These methods utilize characteristics like trajectories, velocity, and acceleration to analyze and classify motions. Some commonly employed motion-based techniques in HAR systems include:

- **Motion history images (MHI):** This approach entails generating a sequence of images that depict the historical movement of body parts over time. By capturing the overall motion patterns of the body, MHI enables the recognition and classification of activities.

- **Trajectory-based methods:** This method involves tracing the paths followed by body parts over time and leveraging this information to identify and categorize activities. Trajectory-based methods can effectively capture the trajectory's shape and route, facilitating activity recognition and classification[18].

- **Dynamic Time Warping (DTW):** This technique revolves around comparing the similarity between two time series by flexibly aligning their temporal structures. DTW can be employed to compare the movements of different body parts and identify and classify activities based on the similarity of their motion patterns [19].

- **Deep Learning Methods**: Deep learning methods excel in recognizing human activities in various systems due to their ability to automatically learn intricate patterns from raw sensor data, eliminating the need for manual feature engineering. By utilizing the back propagation algorithm, deep learning uncovers complex structures within extensive datasets. This algorithm guides the machine in adjusting its internal parameters, which compute the representation of each layer based on the representation from the preceding layer [20]. Deep learning models such as CNN, RNN, and LSTM play a crucial role in feature extraction and classification within a human activity recognition system [21]. This is particularly advantageous in HAR because the raw sensor data can be highly intricate and multidimensional, posing challenges in designing hand-crafted features that capture all the essential information.

5. Deep Learning in Human Activity Recognition

- **Role of Deep Learning:** Deep learning is of utmost importance in the field of vision-based human activity recognition (HAR) as it effectively extracts pertinent characteristics from visual data, categorizes activities, manages variations, enables comprehensive learning, and supports adaptability and transfer of knowledge. Deep learning models greatly enhance the precision, resilience, and scalability in detecting and comprehending human activities from visual input. Various roles of deep learning human activity recognition include:
  - ➢ Feature extraction
  - ➢ Activity classification
  - ➢ Multi-modal data integration
  - ➢ Anomaly detection
  - ➢ Continuous monitoring

- **Different Deep Learning methods:** Deep learning methods have shown great success in vision-based human activity recognition (HAR), and there are several commonly used techniques, including:

- **Convolutional Neural Networks (CNNs):** In HAR, Convolutional Neural Networks (CNNs) are frequently employed to extract spatial attributes from individual frames or image patches. The structure of the network typically involves multiple convolutional layers, followed by one or more fully connected layers. The convolutional layers learn spatial features by convolving filters across the input image or patch, and the fully connected layers learn to classify the learned features into different activities. CNNs can be trained end-to-end on large datasets of labeled videos or image sequences, and have been demonstrated exceptional performance in HAR.

- **Recurrent Neural Networks (RNNs):** RNNs belong to a category of advanced neural networks capable of understanding and representing patterns over time in sequential data. In the field of Human Activity Recognition (HAR), RNNs are often used to capture the temporal dynamics of activities by processing successive frames or image patches. Such networks usually comprise multiple recurrent layers, including Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) layers, which can selectively remember or forget information over time. The output of the final recurrent layer can be fed into a fully connected layer. The output of the final recurrent layer can be fed into a fully connected layer, allowing for accurate classification of activities [22].

- **Single-Frame Classification:** In HAR, single-frame classification refers to the process of determining the activity being performed by an individual using a single image or frame taken by a sensor. Although this method has its advantages for certain purposes, it is not without its limitations. Single-frame classification fails to encompass the complete sequence of actions performed by an individual, making it difficult to accurately classify intricate activities that involve a series of actions. Additionally, this approach may struggle to detect subtle variations in activities that happen within a brief timeframe.

- **Late Fusion:** Late fusion refers to a method where the outputs of the individual classifiers are combined at a later stage after they have been trained independently on different modalities or features. This approach allows for flexibility in the feature extraction and classification processes, as different modalities can be processed and classified separately [23].However, late fusion necessitates a cautious selection and design of the individual classifiers and feature extraction techniques to ensure that they work together synergistically and offer valuable information for the ultimate classification decision.

- **Early Fusion:** Early fusion in HAR refers to a technique which involves combining the features extracted from multiple modalities or sensors at an early stage before classification. The purpose of early fusion is to enhance the precision of activity recognition by allowing the classifiers to learn from the combined features, which can provide complementary information. In this approach, the features extracted from diverse sensors or modalities are consolidated into a solitary feature vector, which is used to train a single classifier [24].

- **3D Convolutional Neural Networks (3D CNNs):** Three-dimensional Convolutional Neural Networks (3D CNNs) are an advanced version of CNNs that possess the ability to comprehend both spatial and temporal characteristics present in video data. When applied to Human Activity Recognition (HAR), 3D CNNs can directly handle video sequences and extract features that capture both spatial and temporal aspects. The architecture of such networks typically comprises multiple layers of 3D convolutions, which enable the learning of spatial and temporal features by convolving filters throughout the video volume. These layers are often followed by one or more fully connected layers that specialize in classifying the acquired features into various activity categories [26].

- **Long Short-Term Memory (LSTM) Networks:** Long Short-Term Memory (LSTM) networks belong to the category of Recurrent Neural Networks (RNNs) and possess the ability to capture long-range relationships in sequential data by selectively retaining or disregarding information over time. In HAR, LSTMs can be used to model the temporal dynamics of activities by processing successive frames or image patches [27]. The typical architecture of an LSTM network encompasses input, forget, and output gates, which regulate the flow of information into and out of the LSTM cell. The LSTM cell's output can be directed to a fully connected layer for activity classification. LSTMs are capable of end-to-end training on extensive datasets comprising labeled videos or sequences of images, and have demonstrated enhanced accuracy in HAR tasks.

- **CNN + LSTM:** The CNN + LSTM technique is a popular method for human activity recognition (HAR) that combines the advantages of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This approach leverages CNNs to extract spatial characteristics from the sensor data and LSTM networks to capture the temporal relationships among these features. The CNN + LSTM method operates by initially applying CNNs to the sensor data in order to extract spatial features, which are subsequently inputted into LSTMs to capture the temporal dependencies between these features [28], [29]. The resulting characteristics are then employed to classify the activities performed by an individual.

- **Graph Convolutional Networks (GCNs):** Graph Convolutional Networks (GCNs) are a specific category of sophisticated neural networks capable of acquiring distinctive characteristics from data organized in a graph structure, such as the interconnections between human body positions or object components. In the context of Human Activity Recognition (HAR), GCNs can be employed to extract relational features from the correlations among human joints or object parts. Typically, the network is composed of multiple graph convolutional layers responsible for assimilating and transmitting information throughout the graph structure. Additionally, there are one or more fully connected layers that specialize in classifying the acquired features into different activity categories [30]. GCNs can be trained end-to-end using extensive datasets of labeled videos, and have demonstrated their ability to enhance the accuracy and resilience of HAR systems.

III. METHODOLOGY AND IMPLEMENTATION

The combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, has gained significant popularity as a deep learning technique in the field of human activity recognition (HAR). This approach, known as CNN + LSTM, leverages the unique advantages of both CNNs and LSTMs to effectively extract both spatial and temporal characteristics from video data.

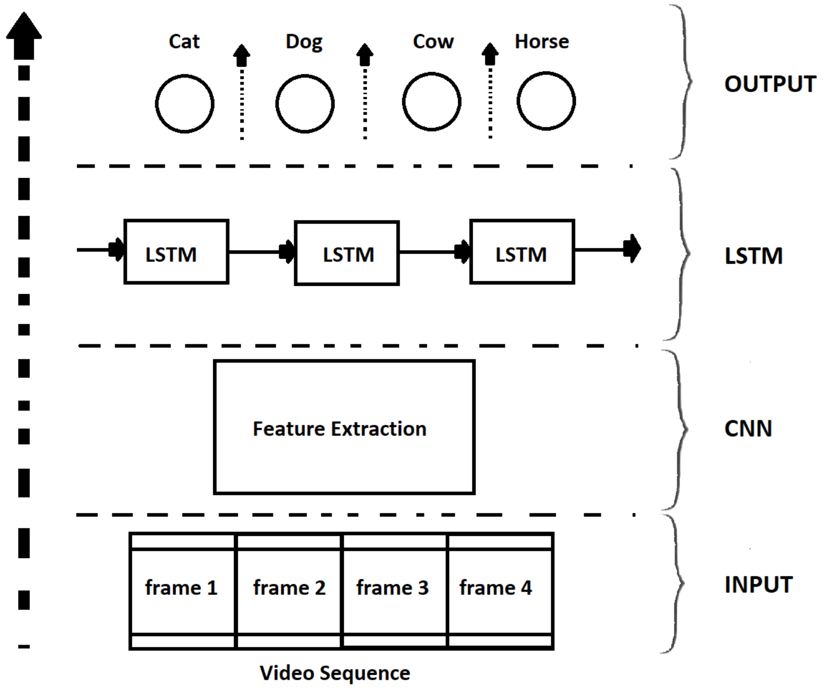

Figure 3: CNN + LSTM approach

1. **How CNN works:** Convolutional Neural Networks (CNNs) are a prevalent class of neural networks extensively applied in computer vision applications, including tasks like object detection and recognition. Their key strength lies in their ability to capture spatial characteristics from images or video frames. By employing convolutional filters, CNNs efficiently identify meaningful patterns and edges within the visual data, enabling them to extract informative spatial features [31], [32], [33], [34].

   The operation of a CNN can be summarized as follows:
   - The input data is fed into the convolutional layer, which applies a set of filters to extract features.
   - The output of the convolutional layer is fed into the pooling layer, which reduces the spatial dimensionality of the feature maps.
   - The output of the pooling layer is then fed into one or more fully connected layers, which perform classification or prediction based on the extracted features.
   - The output of the final fully connected layer represents the predicted class or output.

2. **Layers of CNN:** Convolutional Neural Networks (CNNs) generally comprise multiple layers designed to extract relevant features from input data for tasks such as classification or prediction. These layers are structured hierarchically, with each layer receiving input from the preceding layer and passing its output to the subsequent layer. The fundamental components of a CNN typically include the following layers:

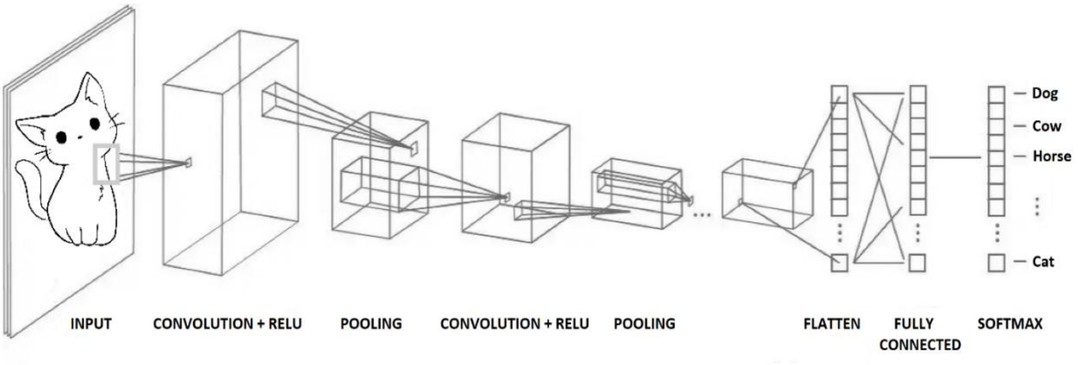

Figure 4: Layers of CNN

- **Convolutional Layer:** The Convolutional Layer serves as a fundamental component within Convolutional Neural Networks (CNNs). Its main function involves the application of a sliding filter or kernel (represented as anMxM matrix) across the input data, where a dot product is computed at each position.[35]. The convolution operation generates a feature map as the output, which highlights the presence of the learned feature at each location in the input data. These feature maps can be stacked together along a third dimension, forming a tensor that becomes the input for the subsequent layer. Deeper layers tend to generate more feature maps to capture additional details, but these maps are generally smaller in size compared to earlier layers[36].

- **Pooling Layer:** The Pooling Layer is commonly inserted between two convolutional layers in neural networks. Its purpose is to decrease the spatial dimensions of the feature maps produced by the Convolutional Layer. The pooling process involves dividing the feature map into non-overlapping regions, such as 2x2 or 3x3, and applying a function to each region to obtain a single output value. Various pooling techniques exist, including min pooling, max pooling, average pooling, and mixed pooling. These techniques provide different ways of summarizing the information within each region of the feature map.[37].

- **ReLU:** The Rectified Linear Unit (ReLU) serves as an activation function that brings non-linearity to a neural network. The activation is simply threshold at zero. The ReLU correction layer replaces all negative values received as inputs by zeros. ReLUis commonly applied between a convolutional layer and a pooling layer. $\mathrm{f}\,(\mathrm{u}) = \max(0, \mathrm{u})$

ReLU is advantageous due to its simplicity, computational efficiency, and ability to overcome the vanishing gradient problem. The vanishing gradient problem arises when the gradient signal diminishes to an insignificant magnitude, impeding the network from learning effectively. This results in sluggish convergence and subpar performance. However, ReLU mitigates this issue by allowing gradients to flow more freely during backpropagation, enabling faster and more effective learning.

- **Dropout Layer:** The Dropout Layer serves as a regularization method applied in neural networks, to prevent overfitting. It operates by randomly deactivating a fraction of neurons during training, with a probability defined by a hyper parameter called the dropout rate. Consequently, the network is compelled to acquire more resilient and independent features that are not excessively reliant on the behavior of particular neurons.

- Flatten Layer: The Flatten Layer serves the purpose of transforming the output generated by Convolutional and Pooling Layers into a condensed, one-dimensional feature vector. The flattening process involves taking the 2D feature maps produced by the convolutional and pooling layers and reshaping them into a one-dimensional array. For instance, if a pooling layer generates a tensor with dimensions of 4x4x64 (4x4 spatial dimensions and 64 feature maps), the Flatten Layer would convert it into a 1024 element array.

- **Fully Connected Layer:** Fully Connected Layer is commonly appended towards the end of the network, following the Convolutional, Pooling, and Flatten Layers. Fully Connected Layer, also referred to as the Dense Layer, plays a crucial role in conducting classification or regression tasks based on the extracted features from preceding layers.It consists of a set of neurons that are fully connected to the neurons in the previous layer. The output produced by the Fully Connected Layer is then forwarded to the output layer for the purpose of classificationor regression analysis.

- **Output Layer:** The Output Layer is the final layer of a CNN responsible for generating predictions. The configuration of the output layer varies depending on the specific problem at hand and can take on different forms such as a single neuron, multiple neurons, or a soft max layer. For instance, in a binary classification scenario, the output layer may consist of a single neuron employing a sigmoid activation function. This neuron produces a probability value ranging between 0 and 1, indicating the likelihood that the input belongs to the positive class. Conversely, in a multi-class classification scenario, the output layer can comprise several neurons, each representing a distinct class, accompanied by a softmax activation function, which outputs a probability distribution over the classes.

- **Sofmax and Sigmoid functions:** The Softmax or Sigmoid function is commonly employed to convert the output from the fully connected layer into a probability distribution across the available classes.
  - **Softmax Function:** The Softmax function is frequently utilized in scenarios involving multi-class classification, where the objective is to determine the probability of each class. By employing the Softmax function, a vector of inputs

can be transformed into a vector of probabilities, ensuring that the sum of all probabilities amounts to 1. The output of the Softmax function is calculated as:

$$S(z)i = \frac{e^{z_j}}{\sum_{j=1}^{C} e^{z_j}}$$

- Sigmoid Function: The Sigmoid function is commonly used in binary classification tasks, such as whether an email is spam or not. The Sigmoid function outputs a value between 0 and 1, allowing us to interpret it as a probability. The output of the Sigmoid function is calculated as:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

- **How LSTM works:** LSTM (Long Short-Term Memory) is a variant of recurrent neural networks (RNNs) frequently applied in tasks involving natural language processing, speech recognition, and sequence prediction. Its purpose is to address the challenge of the vanishing gradient problem encountered in conventional RNNs, where the network struggles to capture long-range dependencies within the data. LSTM is particularly effective for processing, predicting, and classification based on time series data [38].

Once trained, an LSTM network can classify real-time, unseen sensor data. The network receives a sliding window of sensor data as its input and generates a prediction for the ongoing activity. The sliding window mechanism allows the network to make predictions over a period of time, which proves valuable for monitoring the evolution of an activity or identifying transitions between different activities. LSTM accomplishes this by introducing a set of gates (input gate, the forget gate, and the output gate) that control the flow of information through the network.

- The input gate determines which information from the current input should be added to the cell state, which is the internal memory of the LSTM.
- The forget gate determines which information from the previous cell state should be discarded.
- The output gate determines which information from the current cell state should be output to the next layer of the network.

During the training process, each gate in the Long Short-Term Memory (LSTM) network possesses its unique set of weights that are learned. Sigmoid functions activate the input, forget, and output gates, controlling the flow of information. The cell state is altered using the input and forget gates and subsequently transformed by a tanh function, which compresses the values within the range of -1 to 1. The output gate determines which modified portions of the cell state should be transmitted to the subsequent network layer.

By leveraging these gates to govern the flow of information within the network, LSTM has the ability to selectively retain or discard information based on the input it receives, which makes it well-suited for tasks that involve long-term dependencies in the data.

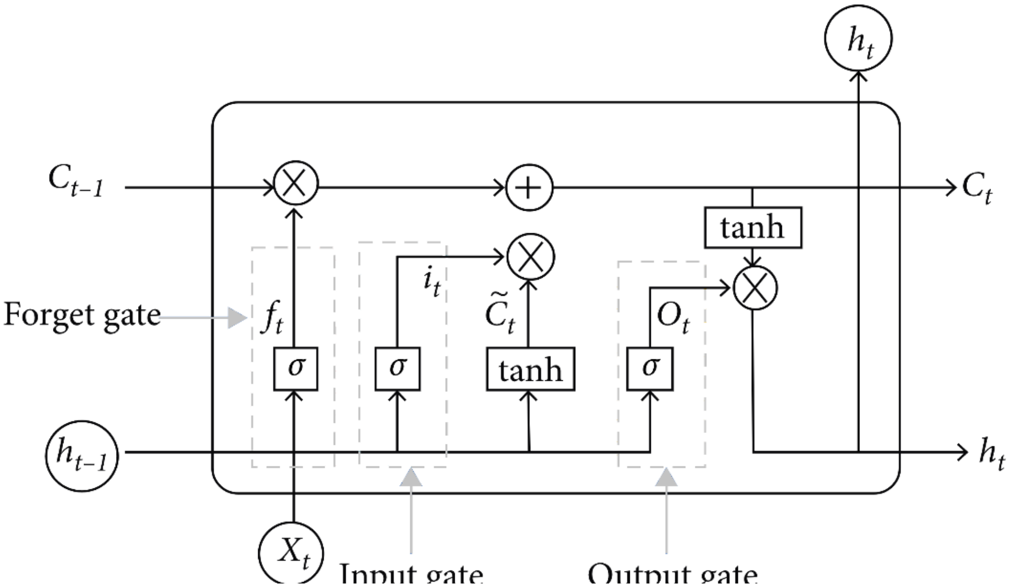

Figure 5: Architecture of LSTM

IV. CONCLUSION

The application of deep learning has brought about a significant transformation in Human Activity Recognition (HAR), leading to the creation of highly precise and effective systems for diverse purposes. Deep learning models, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) and hybrid models such as ConvLSTM, LRCNmodelshave shown remarkable performance in HAR tasks. The combination of CNNs and LSTMs enables the capture of both spatial and temporal features, allowing for more robust recognition of complex human activities. By employing a blend of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, it is possible to create a proficient system capable of accurately recognizing diverse human activities. The fundamental concepts and strengths of CNNs and LSTMs in activity recognition have been explored through this article. While deep learning has shown significant advancements in HAR, challenges still exist, including the need for large labeled datasets, model interpretability, and generalization to different environments. The future tasks involve developing a potent model and assessing its performance compared to numerous readily accessible models and algorithms.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, May 2015, doi: 10.1038/nature14539.

[2] C. Pham et al., "SensCapsNet: Deep Neural Network for Non-obtrusive Sensing Based Human Activity Recognition," IEEE Access, vol. 8, pp. 86934-86946, 2020, doi: 10.1109/ACCESS.2020.2991731.

[3] C. N. Phyo, T. T. Zin, and P. Tin, "Deep Learning for Recognizing Human Activities Using Motions of Skeletal Joints," IEEE Transactions on Consumer Electronics, vol. 65, no. 2, pp. 243-252, May 2019, doi: 10.1109/TCE.2019.2908986.

[4] Y. Du, Y. Lim, and Y. Tan, "A Novel Human Activity Recognition and Prediction in Smart Home Based on Interaction," Sensors, vol. 19, no. 20, article no. 4474, Oct. 2019, doi: 10.3390/s19204474.

[5] J. Qi, Z. Wang, X. Lin, and C. Li, "Learning Complex Spatio-Temporal Configurations of Body Joints for Online Activity Recognition," IEEE Transactions on Human-Machine Systems, vol. 48, no. 6, pp. 637-647, Dec. 2018, doi: 10.1109/THMS.2018.2850301.

[6] D. R. Beddiar, B. Nini, M. Sabokrou, et al., "Vision-Based Human Activity Recognition: A Survey," Multimedia Tools and Applications, vol. 79, no. 43, pp. 30509-30555, Dec. 2020, doi: 10.1007/s11042-020-09004-3.

[7] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A Survey on Deep Learning for Human Activity Recognition," ACM Computing Surveys (CSUR), vol. 54, no. 8, article no. 177, Nov. 2022, pp. 1-34, doi: 10.1145/3472290.

[8] N. Gupta, S. K. Gupta, R. K. Pathak, et al., "Human activity recognition in artificial intelligence framework: a narrative review," Artificial Intelligence Review, vol. 55, pp. 4755-4808, 2022. DOI: 10.1007/s10462-021-10116-x.

[9] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in Proceedings of the 20th ACM International Conference on Multimedia (MM '12), New York, NY, USA, 2012, pp. 1057-1060, doi: 10.1145/2393347.2396382.

[10] M. Z. Uddin, D.-H. Kim, J. T. Kim, and T.-S. Kim, “An Indoor Human Activity Recognition System for Smart Home Using Local Binary Pattern Features with Hidden Markov Models,” Indoor and Built Environment, vol. 22, no. 1, pp. 289-298, 2013. DOI: 10.1177/1420326X12469734.

[11] L. O. Chua and T. Roska, "The CNN paradigm," in IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 40, no. 3, pp. 147-156, March 1993, doi: 10.1109/81.222795.

[12] H. Ding et al., "FEMO: A Platform for Free-Weight Exercise Monitoring with RFIDs," in SenSys 2015—Proceedings of 13th ACM Conference on Embedded Networked Sensor Systems, 2015, pp. 141-154, doi: 10.1145/2809695.2809708.

[13] H. Zou, Y. Zhou, R. Arghandeh, and C. J. Spanos, "Multiple kernel semi-representation learning with its application to device-free human activity recognition," IEEE Internet Things J., vol. 6, no. 5, pp. 7670-7680, 2019, doi: 10.1109/JIOT.2019.2901927.

[14] C. I. Patel, D. Labana, S. Pandya, K. Modi, H. Ghayvat, and M. Awais, “Histogram of Oriented Gradient-Based Fusion of Features for Human Action Recognition in Action Video Sequences,” Sensors, vol. 20, no. 24, p. 7299, Dec. 2020, doi: 10.3390/s20247299.

[15] S. B. Patil and G. R. Sinha, "Distinctive Feature Extraction for Indian Sign Language (ISL) Gesture using Scale Invariant Feature Transform (SIFT)," J. Inst. Eng. India Ser. B, vol. 98, pp. 19-26, 2017. doi: 10.1007/s40031-016-0250-8.

[16] F. Kuncan, Y. Kaya, and M. Kuncan, "A novel approach for activity recognition with down-sampling 1D local binary pattern," Advances in Electrical and Computer Engineering, vol. 19, no. 1, pp. 35-44, 2019.

[17] J. Chen, W. Yang, C. Liu, and L. Yao, “A Data Augmentation Method for Skeleton-Based Action Recognition with Relative Features,” Applied Sciences, vol. 11, no. 23, p. 11481, Dec. 2021, doi: 10.3390/app112311481.

[18] G. Hua, G. Hemantha Kumar, and V.N. Manjunath Aradhya, "A Hybrid Speed and Radial Distance Feature Descriptor Using Optical Flow Approach in HAR," in Applied Intelligence and Informatics, M. Mahmud, C. Ieracitano, M.S. Kaiser, N. Mammone, and F.C. Morabito, Eds. Springer, Cham, 2022, vol. 1724, pp. 3-14. doi: 10.1007/978-3-031-24801-61.

[19] B. Boufama, P. Habashi and I. S. Ahmad, "Trajectory-based human activity recognition from videos," 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, Morocco, 2017, pp. 1-5, doi: 10.1109/ATSIP.2017.8075536.

[20] S. Seto, W. Zhang and Y. Zhou, "Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 2015, pp. 1399-1406, doi: 10.1109/SSCI.2015.199.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, May 2015, doi: 10.1038/nature14539.

[22] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," Pattern Recognition Letters, vol. 119, pp. 3-11, 2019, doi: 10.1016/j.patrec.2018.02.010.

[23] M. Atikuzzaman, T. R. Rahman, E. Wazed, M. P. Hossain and M. Z. Islam, "Human Activity Recognition System from Different Poses with CNN," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2020, pp. 1-5, doi: 10.1109/STI50764.2020.9350508.

[24] A. Tsanousa, G. Meditskos, S. Vrochidis and I. Kompatsiaris, "A Weighted Late Fusion Framework for Recognizing Human Activity from Wearable Sensors," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019, pp. 1-8, doi: 10.1109/IISA.2019.8900725.

[25] K. Gadzicki, R. Khamsehashari and C. Zetzsche, "Early vs Late Fusion in Multimodal Convolutional Neural Networks," 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 2020, pp. 1-6, doi: 10.23919/FUSION45008.2020.9190246.

[26] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[27] F. Hernández, L. F. Suárez, J. Villamizar and M. Altuve, "Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network," 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), Bucaramanga, Colombia, 2019, pp. 1-5, doi: 10.1109/STSIVA.2019.8730249.

[28] R. Mutegeki and D. S. Han, "A CNN-LSTM Approach to Human Activity Recognition," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 362-366, doi: 10.1109/ICAIIC48513.2020.9065078.

[29] K. Xia, J. Huang and H. Wang, "LSTM-CNN Architecture for Human Activity Recognition," in IEEE Access, vol. 8, pp. 56855-56866, 2020, doi: 10.1109/ACCESS.2020.2982225.

[30] W. Peng, J. Shi, T. Varanka and G. Zhao, "Rethinking the ST-GCNs for 3D skeleton-based human action recognition," in Neurocomputing, vol. 454, pp. 45-53, 2021, doi: 10.1016/j.neucom.2021.05.004.

[31] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," arXiv:1511.08458 [cs.NE], Nov. 2015, doi: 10.48550/ARXIV.1511.08458.

[32] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[33] P. Kim, "Convolutional Neural Network," in MATLAB Deep Learning. Berkeley, CA: Apress, 2017, ch. 6, doi: 10.1007/978-1-4842-2845-6_6.

[34] J. Wu, "Introduction to convolutional neural networks," National Key Lab for Novel Software Technology. Nanjing University. China, vol. 5, no. 23, pp. 495, 2017.

[35] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in Proceedings of the International Conference on Learning Representations (ICLR), 2015, Available: https://openreview.net/pdf?id=By-7Kxqeg.

[36] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov et al., "Devise: A deep visual-semantic embedding model," in Proceedings of the Neural Information Processing Systems (NIPS), 2013, pp. 2121-2129.

[37] M.U.G. Khan, L. Zhang, and Y. Gotoh, "Human focused video description," in Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, 2011, pp. 654-661.

[38] S. Herath, M.T. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," CoRR, vol. abs/1605.04988, 2016.