International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 20 August 2023 **Revised:** 08 October 2023 **Accepted:** 19 October 2023*

A Survey on Different Deep Learning Model for Human Activity Recognition based on Application

Hetal Bhaidasna[1], Chirag Patel[*2], Zubin Bhaidasna[3]
[1]Research Scholar, Computer Engineering Department,
Parul University
Vadodara, India
hetal.bhaidasna@paruluniversity.ac.in
[2]Associate Professor , Computer Science &Engineering Department
CHANGA University
Kheda, India
Corresponding Author : chiragpatel.dce@charusat.ac.in
[3]Assistant Professor, Computer Engineering Department,
CVM University
zbhaidas@gmail.com

Abstract— The field of human activity recognition (HAR) seeks to identify and classify an individual's unique movements or activities. However, recognizing human activity from video is a challenging task that requires careful attention to individuals, their behaviors, and relevant body parts. Multimodal activity recognition systems are necessary for many applications, including video surveillance systems, human-computer interfaces, and robots that analyze human behavior. This study provides a comprehensive analysis of recent breakthroughs in human activity classification, including different approaches, methodologies, applications, and limitations. Additionally, the study identifies several challenges that require further investigation and improvements. The specifications for an ideal human activity recognition dataset are also discussed, along with a thorough examination of the publicly available human activity classification datasets.

Keywords- Recognizing human activity, Machine Learning, Computer Vision, Deep Learning, Video Surveillance, CNN, RNN, LSTM, GRU, Fast R-CNN

I. INTRODUCTION

Identification of human activities is one of the essential topics that has attracted a lot of interest from the computer vision community recently. From a video feed, Human Activity Recognition (HAR) attempts to identify actions such as kids playing ball, kids studying, kids playing games, kids strolling, kids sleeping, kids writing, kids singing, and kids cooking. These activities may be anything caught from the videos. Due of its effectiveness and extensive use in human-computer interface, sign language, artificial intelligence, and image processing, robotics, deep learning, and other fields, HAR is a popular study topic. Human activity detection and spotting is a key feature in some smart video control and tracking software, such as surveillance video, security video evaluation, sports video evaluation, detection of unusual activities, affected person tracking, site visitor tracking, sports activities, and industries. The goal of human activity recognition is to automatically analyze and interpret current behaviors using visual data. Human Activity Recognition in Video is a difficult problem that requires focusing on people and their body components that contribute to the activities. Data about human activity and interactions with the environment are being gathered as a part of a research effort called "Human Activity Recognition", enabling you to offer useful information. HAR is a challenging but crucial research area with various real-world applications. The development of accurate and efficient HAR systems can significantly improve our understanding of human behavior and interactions with the environment. Human activity is classified according to its level of complexity:

1. All acts involve a single actor such as bending, walking, and so on.
2. Interactions among individuals and between individuals and objects are examples of interactions such as lifting a bag, punching, etc.
3. Group activities are activities that take place inside a group. For example, group dancing, group robbing, and so forth.

Recognizing human activity from a video clip is a challenging task because of various factors such as backdrop clutter, partial occlusion, perspective, lighting, and appearance changes. The performance of HAR algorithms can be significantly affected by these factors, which can make it difficult to accurately classify human activities.

II. RELATED WORK

Finding the specific movement or activity that a person or individuals were engaged in when it was taken by a camera is the aim of the large field of study known as human activity recognition (HAR). Comparing frame baseline techniques with probabilistic reasoning, employing both motion characteristics and context information, Li Wei and Shishir K. Shah suggested a deep model to increase activity identification performance. The motion information contains low-level motion information such as spatial time and high-level mobility elements that capture human motions. The histogram of orientations serves as the space-time patch's descriptors and histogram of features are produced in this study. The context feature encloses a scenario that gathers information about the surroundings and human interactions. Othman O. Khalifa, Mohanad Babiker, Muhamed Zaharadeen, khyawHtike, Aisha Hassan [3] discovered common human behaviors in a video surveillance system. Background subtraction, binarization utilizing the Otsu technique, and morphological operations were suggested by a digital image processing system. To categorize activity models from the dataset, they use multi-layer feed forward perception. They undertake training, testing, and validation on activities after obtaining the dataset. The pictures that were recovered were subjected to the 2D median filter in order to decrease noise and distortion. Using CNN and LSTM convolutional approaches, Yongling FU, Jian Sun, Shengguang Li, Lin Tan Jie He and Cheng Xu [4] proposed a hybrid deep framework. In this technique, characteristics from video extracted using CNN, and LSTM units were used to simulate temporal dependence on the retrieved features, which was then utilized to solve a time series issue. For classification, an ELM (Extreme Learning Machine) classifier was utilized to improve generalization performance and reduce running time. Convolution layers, LSTM, and ELM are combined in the CnvLSTM-FC model. Guolong cui, Xiaobo yang, Mingyang wang, and Lingjiang Kong [5] presented a human body and limb identification system based on radar. They used a CNN, RNN, and LSTM-based stacked gated recurrent units' network (SGRUN), to do this. Yunhong wang, Longteng kong, di huang, jie qin, , and luc van gool [6] suggested utilizing a hierarchical framework for context modelling and attention based on LSTM networks to cope with concurrently attending both people and their body parts that are significant to the activity, as well as modelling contextual person structures in the group issues. Lee provided a model of a recurrent neural network with various spatiotemporal sizes. A standard convolution neural network model was combined with various timeframe recurrent dynamics to create the work of Minju Jung, Jun Tani, and others [7]. Wilton, Loius CW, and Carol Chen [8] proposed a deep learning network for identifying various sorts of activities and using a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to analyze dynamic video motion of sports motions. Dropout layer is added to RNN to alleviate the issue of overfitting simply eliminating a few nodes from the network at random. A dropout layer was placed in between the hidden and output layers. The LSTM is used with RNN to tackle the vanishing gradient issue. Using a smartphone's inertial sensor, Md. Zia Uddin, Mohammed Meahedi Hassan, Ahmad Almogren and Amr Mohamed, [9] offer a method for detecting human activity. First, efficient characteristics such as mean, median and auto-regression are retrieved from the raw data. The traits are then made stronger using kernel principal component analysis (KPCA) and linear discriminant analysis (LDA) (KPCA). At last, Deep Belief Network is used to train features (DBN). Masaya Inoue, SozoInoue, and Takeshi Nishida [10] advocated utilizing LSTM to construct a deep recurrent neural network for human activity detection and training it on raw time series data from mobile devices (Long Short-Term Memory). To generate RNN, the value of the weight gradient in each layer is altered using the back-propagation over time method. The number of layers, shortened time, gradient clipping parameter, and dropout rate are the best DRNN parameters. Jagadeesh B., Chandrashekar M. Patil, and Meghana M. N. centered on identifying activities and various people [11].Different human video datasets are tested for the recognition and tracking of many persons. Multiple human emotions are detected using the background subtraction approach. The noise component is then removed while maintaining the edges using the median filter. To extract features, the Histogram of Oriented Gradient (HOG) feature descriptor is utilized. The HOG technique is used to determine how frequently a gradient appears in a certain region of a photograph. The Support Vector Machine classifier was used to recognize human activities. Weria Khaksar, Md. Zia Uddin and Jim Torresan [13] proposed a human activity detection system based on recurrent neural networks. It uses solid scale-invariant body silhouette parameters. A silhouette of a human figure is created from a depth image after the backdrop has been removed. The matching body skeleton is then extracted from the picture by segmenting body parts 9" utilizing random forests. Furthermore, by describing the body joints in the spherical coordinate system, scale-invariant skeletal characteristics are restored. The motion aspects of the skeleton are then added to the skeleton's characteristics in a succession of frames. The depth silhouettes are altered using radon to get translation and scale invariant silhouette properties, and these characteristics are then coupled with the skeleton features. Using the robust characteristics extrapolated from the depth picture sequences, a deep recurrent neural network is trained to identify activities. RNN seems to be a particularly promising deep learning approach for modelling time-sequential information in this regard. Recurrent connections between hidden units are used by RNNs to connect the past with the present. RNN generally has a vanishing gradient problem since it processes long-term data. Long-term dependencies can be handled using Long Short-

Term Memory (LSTM). The table below includes summaries of several research publications.

TABLE I. SUMMARY OF DIFFERENT APPROACH FOR HUMAN ACTIVITY RECOGNITION

Paper Name	Method	Description	Dataset	Name of the Journal, Year
Human Activity Recognition using Deep Neural Network with contextual information [2]	HOG & HOF	Motion and context are two features.	CAVIR	IEEE, 2017.
Automated Daily Human Activity Recognize for video Surveillance using neural network[3]	Neural Network Multilayer Feed forward perceptron	With simple activity, detect and monitor the human body. Only work with static data and an indoor setting.	Real time Dataset (indoor Video)	IEEE, 2017
Sequential Human Activity Recognition Based on Deep Convolution Network and Extreme Learning Machine Using Wearable Sensors[4]	CNN & LST	They created a model to categorize the extracted characteristics and decrease the amount of time it takes. Feature extraction requires little or no specialist expertise.	OPPORTUNITY.	Journal of Sensors, 2018
Human body and limb motion recognition on via Stacked Gated Recurrent Units Network [5]	SGRUN (CNN + RNN +GRU)	Classify motion by extracting dynamic sequential characteristics. 10 GRU layers are added to the RNN, and features are retrieved using CNN.	Radar Data from Advanced Signal Processing LAB, Temple University	IET Radar Sonar Navig, 2018
Recognized on visually perceived compositional human actions by multiple Spatio-Temporal scaled Recurrent Neural Networks[7]	MSTRNN	Utilized are one conventional layer, one pooling layer, two context layers, and one softmax layer.	3ACWD CL1AD CL2AD	IEEE, 2017
A robust human activity recognition system using smartphone sensors and deep learning[9]	DBN	DBN is used as a NN, while KPCA is used to reduce dimension. Better than SVM in terms of results.	UCI Machine Learning Repository data set	Elsevier, 2018
Deep recurrent neural network for mobile human activity recognition with high throughput[10]	DRNN (RNN & LSTM)	RNN allows for trial and error. The DRNN has a high rate and throughput.	HAR Dataset, UCI Machine Learning Repository	Springer, 2017
An approach of understanding human activity recognition and detection for video surveillance using HOG descriptor and SVM classifier.[11]	HOG feature extraction SVM classifier recognition technique	HOG and SVM are better at detecting and recognizing human activities.	UT-Interaction Dataset Real Time dataset	IEEE, 2017
Hierarchical Attention And Context Modelling For Group Activity Recognition[6]	LSTM	A hierarchical attention and context modelling strategy was presented using LSTM for identifying group activities.	ImageNet GoogleNet	ICASSP, IEEE, 2018
Artificial Intelligence for Sport Action and Performance Analyzing using Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM)[8].	RNN + LSTM	A NN with RNN, LSTM, Dropout layer, and linear activation function was proposed.	Real Time Dataset	ACM, 2018
Structured RNN for Group Activity Analysis [15]	SRNN	LSTM layer, fusion layer, dropout layer, LSTM layer, and dropout layer are all part of the proposed architecture.	CAD-120 SBU-Interaction dataset M2I Dataset MTU RGB+D	IEEE, 2018

				Dataset	

III. MODELS USED FOR ACTIVITY RECOGNITION

A computer model learns to do categorization tasks directly from pictures, text, or voice using deep learning. Deep learning makes advantage of neural network functionality. Frank Rosenblatt designed the first Neural Network prototype in 1957. There are three distinct deep learning methodologies that is supervised, unsupervised, or semi-supervised.

All data and inputs are preset with certain labels in Supervised Learning. Following the application of models, the label must be finalized using the training dataset. For instance, to train a dataset of fruits with labels like as mango, apple, and banana. When a single fruit, such as mango, is supplied as an input, it will forecast mango. Data will be classified in Unsupervised Learning based on similarity or grouping of data. Classifying males, females, and children is similar. The goal of unsupervised learning is to infer a structure from a collection of unlabeled data. The algorithm is programmed to match action to circumstance in order to optimize the reward or feedback signal.

Deep neural networks may be divided into two groups.

a. Discriminative
b. Generative

Discriminative models are built from the ground up. In this paradigm, data travels from the input layer via a number of hidden levels before reaching the output layer. This model is often used in supervised learning applications like as classification and regression. For the top-down method, a generative model was utilized. It's most often employed for unsupervised learning. There are several deep neural network methods, including multilayer perceptron recurrent neural networks, neural networks, long short-term memories, convolution neural networks and deep belief networks. Each algorithm has a unique set of advantages and disadvantages. The figure below depicts the basic architecture for deep learning-based human activity detection.

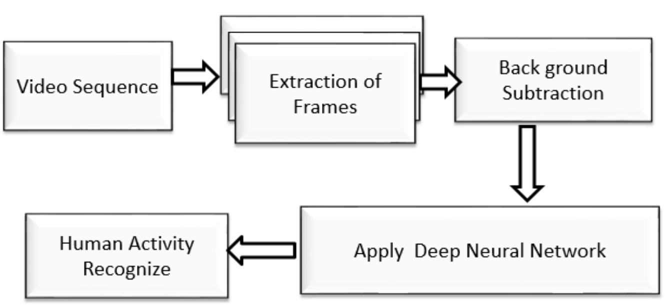


Figure 1. Basic Structure of HAR

The models discussed in this section for activity recognition range from Convolutional Neural Networks to Deep Neural Networks to Long Short-Term Memory, Recurrent Neural Networks and Gated Recurrent Units. These approaches are useful in detecting human activity.

A perceptron is a single neuron found in a deep learning neural network. Given a finite collection of m' inputs, multiply each by a weight, add a bias value, and lastly send the weighted combination of inputs through a non-linear activation function, which gives the output Y. Figure 3.2 depicts the neural networks' working model.

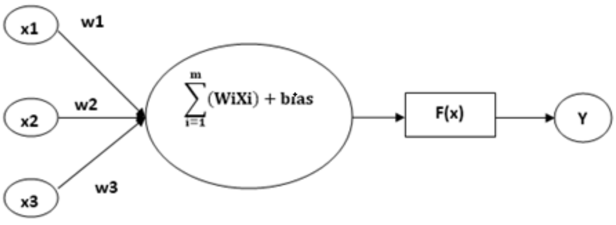


Figure 2. Neural Network

$$Y = f\left(\sum_{i=1}^{m} \quad xiwi + bias\right) \qquad (1)$$

A. Convolution Neural Network

Convolution Neural Networks (CNN), based on the Multi-Layer Perceptron architecture, are a frequently used approach for image processing issues in Deep Neural Networks. CNN is a supervised Deep Learning method utilized in areas such as Face Recognition, Image Retrieval, Object Detection, and Speech Recognition. CNN, sometimes referred to as ConvNet, is a feed-forward neural network that examines topology as input. Convolutional neural networks have neurons with trainable weights and biases, much like traditional neural networks do. Each neuron receives some inputs, does a dot product, and then, if required, performs a non-linearity.

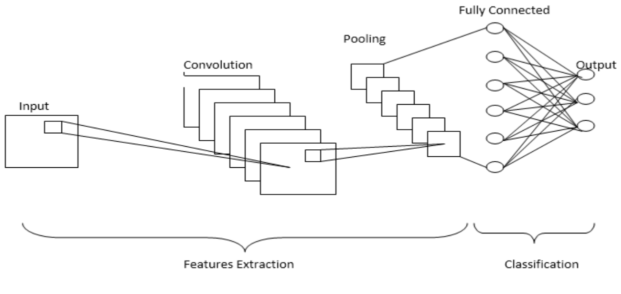


Figure 3. Convolution Neural Network

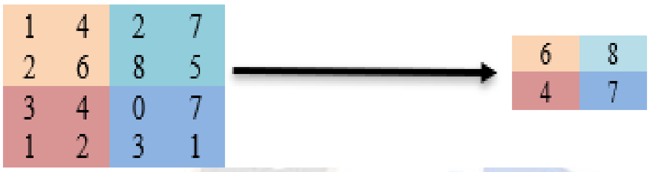
A simple ConvNet consists of a number of layers, each of which transfers activation from one region to another using a differential function. Convolution, pooling, and fully-connected layers are the three main types of layers used to construct ConvNet systems. Convolution layer performs convolution on the input layer, extracts the features, and passes the results to the next layer. Over the input volume, a convolution operation essentially computes the dot product between the weights and the tiny area to which they are related. Consider a 5*5 pixel input picture (X) and 3*3 filter (f), and the output matrix Z.

$$Z = X * f \tag{2}$$

For the non-linearity, any activation function should be used after extracting features. In general, the RELU activation function is the superior selection for the convolution layer.

$$Z1 = RELU(Z) \tag{3}$$

The pooling layer conducts a width-based down-sampling procedure, resulting in a decrease in dimensions. Pooling may be divided into two types: maximum pooling and average pooling. Pooling is mostly used to minimize spatial dimensions.



$$\begin{bmatrix} 1 & 4 & 2 & 7 \\ 2 & 6 & 8 & 5 \\ 3 & 4 & 0 & 7 \\ 1 & 2 & 3 & 1 \end{bmatrix} \longrightarrow \begin{bmatrix} 6 & 8 \\ 4 & 7 \end{bmatrix}$$

$$Z2 = Maxpool(Z1) \tag{4}$$

Convert the matrix to one dimension and send it on to the fully linked layer after pooling. The fully linked layer functions similarly to a simple deep neural network, with numerous neurons connected to one another. Each neuron has a particular weighting and bias. Finally, the non-linear activation function is used.

$$Z3 = \Omega * Z2 + \beta \tag{5}$$

Any activation function, such as sigmoid, sinusoidal, or other, may be used in a fully linked network. It's a straightforward feed-forward neural network. The final pooling-layer or convolution-layer output, which is compact, is fed into this layer. The final classification is made using the outputs from the convolution- and pooling-layers, which are fed into the fully connected neural network architecture.

CNNs have several advantages over traditional machine learning algorithms, including their ability to learn hierarchical representations of data, handle high-dimensional input data, and their flexibility in dealing with various types of data. One of the major advantages of CNNs is their ability to learn hierarchical representations of data. CNNs consist of multiple convolutional layers, each of which learns increasingly complex and abstract features of the input data. This allows the network to identify meaningful patterns in the data and capture the most relevant information for the classification task. Another advantage of CNNs is their ability to handle high-dimensional input data, such as images and videos. However, CNNs also have some limitations and disadvantages. One of the main limitations of CNNs is their requirement for large amounts of labeled data to train the model effectively. Without enough labeled data, the model may not be able to learn meaningful representations of the input data, leading to poor performance. Another disadvantage of CNNs is their sensitivity to adversarial attacks, which can cause the model to make incorrect predictions by adding small perturbations to the input data.

B. *Recurrent Neural Network*

Recurrent neural networks are a type of Deep Neural Network that can recognize patterns in sequential input. RNN is recurrent in nature since it uses the same function for each input data and depends on the results of the previous computation for each input. It evaluates the current input as well as the outcome learnt from past inputs while making judgments. A memory cell, also known as hidden state, is used to store intermediate values. The Hidden state, which remembers some information about the sequence, is the most essential element of RNN.

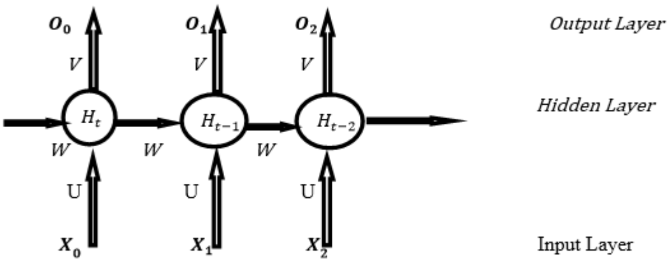


Figure 4. RNN Architecture

From hidden to hidden, the weight matrix W

x = input at time t,

O = network output at time t,

H = network internal (hidden) states vector at time t.

V = matrix of weights from hidden to output;

U = input to hidden weight matrix

In RNN, a hidden layer is made up of many nodes. Each node includes a function for creating the output state Ot, as shown in figure 3.2 and the current hidden state ht. With the assistance of the prior concealed state ht-1 and the input Xt.

$$Out(H_t) -> f\ ((U_t * X_t + W_{t-1} * Out(H_{t-1})) + b) \tag{6}$$

$$Out(O_t) -> f\left(\left(V_t * Out(H_t)\right) + b\right) \tag{7}$$

f is an activation function in this case. It is a non-linear function commonly selected from a set of functions that already exist, including ReLU, sigmoid, hyperbolic tangent, and others. Because RNN has memory cells in the form of hidden states, it works well for sequential data. However, because to the vanishing gradient and inflating gradient issue for big sequence data, training RNN is a difficult process. Recurrent Neural Networks (RNNs) are a type of neural network that is well-suited for processing sequential data, making them a popular

***Article Received:** 20 August 2023 **Revised:** 08 October 2023 **Accepted:** 19 October 2023*

choice for many natural language processing (NLP) and time series analysis tasks. RNNs have several advantages over traditional machine learning algorithms, including their ability to capture temporal dependencies, handle variable-length input sequences, and store memory. This makes RNNs well-suited for tasks where the order of input data is important, such as language modeling, speech recognition, and machine translation. Additionally, RNNs can handle input sequences of different lengths without the need for padding or truncation, making them more flexible and adaptable to a wider range of input data. RNNs also have some limitations like the vanishing gradient problem, where the gradient of the loss function with respect to the model parameters becomes extremely small, making it difficult to update the parameters during training.

C. Long short-term memory

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) architecture that is designed to overcome the limitations of traditional RNNs, such as the vanishing gradient problem and difficulty in capturing long-term dependencies. LSTMs were first introduced in 1997 by Hochreiter and Schmidhuber, and have since become one of the most widely used architectures for processing sequential data. In sequence prediction problems, it is capable of learning ordered dependency. This behavior is required for challenging problem domains like machine translation, speech recognition, and other areas. LSTM [21] effectively handles RNN limitations such as vanishing gradient and expanding gradient difficulties. At a high level, LSTMs work by maintaining a "memory" or "cell state" that can selectively store or discard information over time. This memory is updated at each time step based on the input data and the previous state of the memory. LSTMs use three types of "gates" - input gates, forget gates, and output gates - to control the flow of information in and out of the memory. These gates are implemented using sigmoid and element-wise multiplication operations, which allow the LSTM to selectively remember or forget information based on its relevance to the current task. The input gate determines which values from the input data should be added to the memory, while the forget gate determines which values should be discarded from the memory. The output gate determines which values from the memory should be output to the next layer or to the final output. Additionally, LSTMs have a "cell state" that allows them to store information over long periods of time, allowing them to capture long-term dependencies in sequential data.

LSTMs have been used for a wide range of applications, including natural language processing, speech recognition, machine translation, and image captioning. They have been shown to outperform traditional RNNs on many tasks, particularly those that involve long-term dependencies or require modeling complex patterns in sequential data.

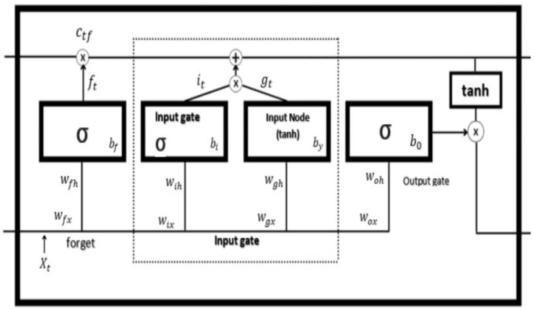


Figure 5. LSTM Structure

The cell's input gate regulates the flow of fresh information. When to forget stuff pertaining to the internal state, it is decided by the forget gate ft. The output gate ot regulates the data that is sent to the output. The primary input to the cell is input modulation gate gt. Cell internal recurrence is handled by internal state ct. Information from previously seen samples is contained in the hidden state h t in the context window:

Input Gate (i_t)

$$i_t = \sigma\,[(W_{ih} * h_{t-1}) + (W_{ix} * x_t) + b_i)] \quad (8)$$

$$g_t = tan\,tan\,h\,\left[\left(W_{gh} * h_{t-1}\right) + \left(W_{gx} * x_t\right) + b_g\right] \quad (9)$$

$$C_{ti} = i_t * g_t \quad (10)$$

Forget Gate (f_t)

$$f_n = \sigma\,[\left(W_{fh} * W_{t-1}\right) + \left(W_{fx} * x_t\right) + b_f] \quad (11)$$

$$C_{tf} = C_{t-1} * f_t \quad (12)$$

Output Gate

$$0_t = \sigma\,[(W_{on} * h_{t-1}) + (W_{ox} * x_t) + b_o] \quad (13)$$

$$C_t = C_{tf} + C_{ti} \quad (14)$$

$$h_t = tan\,tan\,h\,(C_t) * 0_t \quad (15)$$

The LSTM algorithm solves the issue of disappearing and expedient gradients. However, it falls short of completely eliminating the issue. Training the LSTM is a time-consuming and demanding task. As a result, it's tough to utilize for real-time applications. LSTMs have several advantages over traditional RNNs, including their ability to handle long-term dependencies, capture complex patterns in sequential data, and store memory. Additionally, LSTMs can handle variable-length input sequences without the need for padding or truncation. However, LSTMs can be computationally expensive, and may overfit on small datasets. Interpreting the inner workings of LSTMs can also be challenging, which can limit their explain ability and interpretability.

D. Gated Recurrent Unit

The Gated Recurrent unit combines the RNN with LSTM. It is frequently employed to address the problem of fading gradients. GRU employs update and reset gates to tackle the issue of disappearing gradients. The update gates are in charge of calculating how much prior data needs to be sent on to the next stage. It resembles the Output Gate of an LSTM recurrent unit. Reset Gates determines how much previous data will be erased. It is the end result of merging the Input Gate with the Forget Gate in a LSTM recurrent unit.

Update Gate:

$$z_t = \sigma \qquad (16)$$

Reset Gate:

$$r_t = \sigma \qquad (17)$$

Current Memory Content:

$$h_t = tanh\left(W_{x_t} + r_t U h_{t-1}\right) \qquad (18)$$

The memory capacity of RNNs will be increased, and training models will be simpler, thanks to the Gated Recurrent Unit. The hidden unit can also be used in recurrent neural networks to address the vanishing gradients problem. In GRU, the problem of growing and vanishing gradients has not yet been resolved. GRUs offer several advantages, including their ability to handle long-term dependencies and overcome the vanishing gradient problem. Compared to LSTMs, GRUs have fewer parameters, which makes them computationally less expensive.

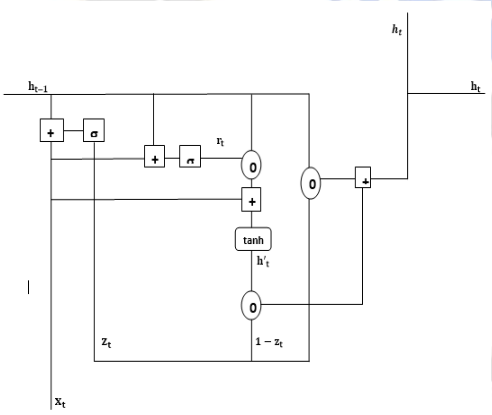


Figure 6. Gated Recurrent Unit

E. R-CNN

R-CNN was developed in 2014 by Ross Girshick et al. to recognize several items in a photograph. Computer vision applications frequently use R-CNNs (Region-based Convolutional Neural Networks), a family of Deep Learning models. The primary goal of the RCNN is to recognize objects and define borders around them. Three processes were used to create the R-CNN model. In the first stage, an input picture is fed into the R-CNN model, which extracts several areas (about 2000) using selection and rectangular bounds to indicate the region of interest. This area of interest is sent into CNN, which generates output features. The items displayed in an area of interest are then classified using these output attributes by an SVM (support vector machine) classifier. The RCNN model's key advantage is that it can extract all regions, but its main drawback is that computation takes a long time. The model's second flaw is that the selective search method is a fixed algorithm; therefore, no learning takes place at that point.

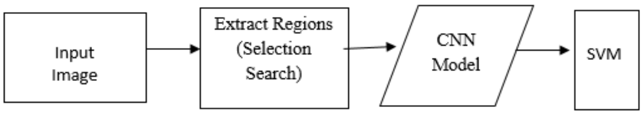


Figure 7. Region based CNN

While RCNN was a significant advancement in object detection, it had a major limitation of being computationally expensive due to the requirement of processing each region proposal separately. To overcome this limitation, the architecture was improved in the subsequent versions, such as Fast R-CNN and Faster R-CNN.

F. Fast R-CNN

A novel model dubbed Fast Region-based Convolution Neural Network was developed to overcome the RCNN issue. Selective search is addressed by replacing the slower R-CNN with a Region Proposal Network (RPN). Using a convolution neural network, extract feature maps from the input picture. Then, pass the maps to an area pooling network, which provides object recommendations. The suggestions are then transmitted to a fully connected layer, which categorizes any expected bounding boxes for the image, after a ROI pooling layer has been added to each proposal to make them all the same size. The classification in this strategy is done using a softmax classifier.

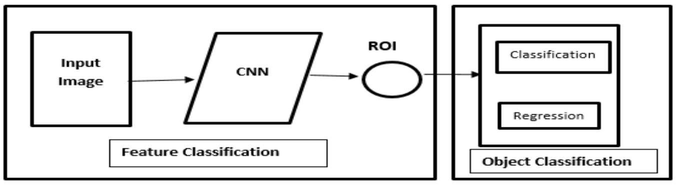


Figure 8. Fast Region based CNN

Advantages of Fast R-CNN include faster inference time compared to RCNN, end-to-end training for both region proposal and classification, and better accuracy in object detection. Additionally, Fast R-CNN eliminates the need for storing feature vectors for each region proposal, reducing the memory requirements. Limitations of Fast R-CNN include its dependency on selective search for region proposals, which may not always produce accurate proposals, and the need for pre-training on large datasets, which may be computationally expensive. Furthermore, Fast R-CNN may struggle with detecting small objects due to the ROI pooling layer.

***Article Received:** 20 August 2023 **Revised:** 08 October 2023 **Accepted:** 19 October 2023*

IV. METHODS FOR HAR BASED ON APPLICATION

METHODS FOR HAR BASED ON APPLICATION

Methods	Type	Application
Convolution Neural Network(CNN)	Discriminative	For the visual context, this model is often utilized. The categorization, face identification, flaw detection, and other applications are the most common.
R-CNN	Discriminative	This model use selective search to produce areas (ALexNet), extract region proposals for object identification, and then apply SVM to classify the results. It takes a long time to compute and does not provide excellent accuracy in a real-time environment.
Fast RCNN	Discriminative	This model use selective search to produce regions (ALexNet) and ROI extraction instead of Max Pooling, followed by Softmax classification. It's sluggish, therefore computation times are still long.
Recurrent Neural Network(RNN)	Discriminative	For sequential time series data, this approach is often employed. Speech recognition, voice recognition, activity recognition, and other applications are the most common. Vanishing or bursting slopes are difficult to teach.
Long short term memory(LSTM)	Discriminative	This model, which is a modified form of RNN, is also utilized for time series problems. It is used to overcome RNN limitations, although it requires a significant number of massive parameter changes.
BiLSTM	Discriminative	With two LSTM layers, it addresses the issue of fixed sequence to sequence prediction. Because BiLSTM includes two LSTM cells, it is more expensive and requires more memory to recognize specific input, which is why it takes a long time to train.
ConvLSTM	Discriminative	Convolution is used in place of matrix multiplication at each LSTM cell gate by ConvLSTM. In terms of computing, it is expensive.
Gated Recurrent Unit(GRU)	Discriminative	This model, which is a modified form of RNN and LSTM, is also utilized for time series problems.

In this section, discussed different models for the Activity Recognition based on structure and application. All models have their own advantage and disadvantage. Depending on requirement, researcher added new layer or modified the existing layer or hybrid the model. Here in below table will discuss the different approach from literature.

ACCURACY OF DIFFERENT HYBRID METHODS FOR HAR

Different Model	Name of Model	Structure	Dataset	Accuracy
CNN+ LSTM +ELM [4]	ConvLSTM-FC	4 Layers of CNN 2 Layers of LSTM 1 Layer of FC(ELM)	OPPORTUNITY	91.80%
GRU[5]	SGRUN	Input Layer(Pre-processing) 2 34-Neuron GRU Layer Output Layer	There are six different motions: walking, jogging, clapping and hand waving.	90.65%
Conv+ Pooling+ Context Layer[7]	MSTRNN	Convolutional layers, context layers, fully-connected layers, and pooling layers	CL1AD(Action)	93.56%
			CL2AD(Action)	81.43%
LSTM + CNN+GAP+BN +FC [23]	LSTM-CNN	2 LSTM (32) layer for extracting temporal features. 2 Convolution and Pooling Layer for extracting spatial features. Global Average Pooling for feature map Batch Normalization FULLY CONNECTED NETWORK	UCI	95.78%
			WISDOM	95.85%
			OPPORTUNITY	92.63%
Convolutin + LSTM + FC+BN+FC [14]	Hybrid Deep Learning	3Convolution and Pooling Layer for extracting spatial features. 2 LSTM layer for extracting temporal features.	UCI	95%

		Fully Connected layer. Batch Normalization to improve the training speed and accuracy. Fully Connected layer with softmax to check probability		

Algorithms like CNN, RNN, LSTM, and GRU were investigated in this research for the purpose of identifying human behavior, as it was covered in the literature review that came before it. Each algorithm has its own set of advantages and disadvantages. In studies comparing these models (for an overview, see Table 3), RNNs have consistently outperformed CNNs on benchmark computer vision datasets such as HAR, ImageNet, Google Net, and others. For the visual context, CNN is often employed. The categorization, face identification, flaw detection, classification, and other applications are the most common. CNNs have a number of advantages, including the ability to avoid handmade characteristics, which are required for other kinds of networks. CNNs, on the other hand, learn characteristics automatically. In terms of training process efficiency, CNN's training practices are time-consuming. The property that CNNs are invariant to transformations like rotation, translation and scaling is a benefit. One of the most important properties of CNNs is their invariance to translation, rotation, and scale, which is very important in computer vision issues like object identification. Because it permits the identification or category of an item to be separated from the details of the visual inputs. As a result, the network is able to detect a certain item even when the image's actual pixel values are greatly improved. However, for time series tasks such as voice recognition, activity detection, picture captioning, and so on, CNN yields less promising outcomes. For sequential time series data, RNN models are often utilized. All of the inputs and outputs of CNN are independent of one another, although in certain circumstances, such as when predicting information, prior data is necessary. As a result, it is necessary to recall the prior words. RNN was created to tackle this issue; it does so with the use of hidden layers that retain information from prior states. However, RNN takes longer to handle extremely lengthy sequences of input, and training the RNN is challenging due to its intricacies. Furthermore, the issue of gradient disappearing concerns about the use of RNN. An adapted form of RNN called as LSTM and GRU's are employed to solve this issue. The comprehensive solution of disappearing and bursting gradients is still a work in progress. Overfitting is an issue that both algorithms suffer. As per the observation of above literature, review, to improve the Human Activity Recognition, can combine CNN & LSTM algorithm which give the advantage of spatial and temporal domain both. Also, add some optimization techniques like the one batch normalization, dropout etc. will be used for the result improvement.

V. CHALLENGES

The acts that are to be identified are the foundation for the construction of a Human Activity Recognition system. The sort of activities and their complexity will have an impact on the recognition quality. The following are some of the difficulties that researchers regularly confront.

A. Complex & Various Background

Static cameras are often utilized in video surveillance and fall detection systems. There are scenes that take dynamic recordings from a variety of devices, such as sports, children playing in the yard, cooking in the kitchen, movies, and the entertainment business, among others. Smart gadgets have cameras implanted in them, and the majority of the captured movies include intricate dynamic backdrops. These films have a variety of backdrops that are continually changing. The rest of the realistic movies, which include occlusion, illumination changes, and angle changes, make it tough to recognize and detect actions in these complicated and continuously changing settings.

B. Occlusion

Clear sight of the motion carried out inside the video frames is one of the most basic needs of today's devices. This is not always possible with standard surveillance equipment. Films in which a large number of people are observed from inside the camera's field of view. Based on this, occlusions may be self-occlusions or occlusions caused by the obstruction of any object in the video.

C. Real Time Processing

Real-time processing is necessary for many HAR applications, such as in assistive technologies or robotics. However, the computational complexity of many HAR models can make real-time processing challenging, particularly on resource-constrained devices. This can limit the applicability of HAR systems in real-world scenarios where real-time processing is critical. Efficient model architectures and optimization techniques are necessary to overcome this challenge.

D. Cluttered Background and Camera Motion

It is fascinating to see how a variety of human sports and motion reputation systems perform in controlled conditions. However, due of the historical noise, it cannot perform as well outside and in out of control environments. The majority of popular hobby functions, such as histograms of oriented gradient and hobby factors, encode historical noise and decrease popularity performance. The motion of the digicam is another factor to consider when working with real-world worldwide applications. Extraction of motion functions is challenging because to digicam movement. Digicam movement must be represented and compensated in order to extract higher functions; other issues like, as lighting conditions, viewpoint shifts, and so on may also be challenging circumstances that prevent motion reputation set of rules from being used in realistic settings.

E. Variability in Human Movements

The variability in how people perform activities is a major challenge in HAR. People can perform the same activity in different ways, and the same person can perform an activity differently at different times. This makes it challenging to develop models that can accurately recognize different human activities. The variability can stem from differences in gender, age, physical fitness, cultural background, and other factors. HAR models must be robust enough to accommodate this variability and still produce accurate results.

F. Data Collection

Collecting human activity data can be challenging due to issues such as privacy concerns, ethical considerations, and variability in data quality. Collecting sufficient and representative data is critical for developing accurate models, but obtaining such data can be time-consuming and costly. Additionally, the data collection process may introduce biases, such as in the selection of participants or the environment in which the data is collected. Addressing these challenges requires careful planning and consideration of ethical and privacy issues.

VI. APPLICATION FOR HAR

In human-centric applications including patient monitoring, surveillance, biometrics, child monitoring, healthcare, human-computer interaction, sports and entertainment human activity detection is a crucial area of research.

A. Health Monitoring:

HAR can be used to monitor the physical activities of individuals and provide insights into their health status. For example, tracking activities such as walking, running, and cycling can help in monitoring and improving cardiovascular health.

B. Smart Homes:

HAR can be used to automate home appliances and provide a personalized experience for individuals. For example, turning on/off lights, AC, and other appliances based on activities such as sitting, sleeping, and walking.

C. Sports:

HAR can be used to track and analyze the movements of athletes during training and competition. For example, analyzing the movements of football players during a game can provide insights into their performance and help in improving their skills.

D. Security

HAR can be used for surveillance purposes to detect and prevent unauthorized access to restricted areas. For example, tracking the activities of individuals in a high-security area such as a bank vault or a military base can help in preventing security breaches.

E. Robotics

HAR can be used to develop intelligent robots that can interact with humans and perform tasks. For example, robots can be trained to recognize human activities such as waving, sitting, standing, and walking to assist individuals in daily tasks.

F. Gaming:

HAR can be used to develop interactive games that can respond to the movements and activities of players. For example, games such as Kinect Sports and Wii Fit use HAR to track the movements of players and provide a more immersive gaming experience.

G. Rehabilitation

HAR can be used in rehabilitation to monitor and track the progress of patients during therapy. For example, tracking activities such as walking and exercising can help in monitoring the recovery of patients from injuries.

H. Human-Robot Interaction

HAR can be used to enable robots to understand and respond to human activities and gestures. For example, recognizing hand gestures can help in controlling the movement of a robotic arm.

I. Automotive Industry

HAR can be used to develop intelligent cars that can recognize and respond to the activities of drivers and passengers. For example, tracking the activities of drivers such as steering, braking, and accelerating can help in improving the safety of vehicles.

J. Entertainment

HAR can be used to develop interactive entertainment systems that can respond to the activities of users. For example, interactive dance games can use HAR to track the movements of users and provide feedback on their performance.

VII. DATASET

Many public datasets are available for academics to verify their work and test the performance of various deep learning architectures.

A. UT-Interaction dataset:

The University of Texas [11] established the UT-Interaction dataset as a part of the Contest on Semantic Description of Human Behaviors (SDHA), a research competition aimed at recognizing human activities in realistic contexts. It includes 20 video sequences depicting the continuous execution of six different types of human-human interactions.

B. UCF-101 Action Recognition Dataset:

Central Florida University's Center for Computer Vision Research established it in 2012 [4]. It is made up of 13,320 videos from YouTube, divided into 101 realistic action categories.

C. HAR Dataset

This HAR[10] dataset contains information on 18 different activities, including sitting, chatting, standing, walking, jogging, and more, collected from 90 people using smartphone sensors (75 males and 15 women) (Accelerometer and Gyroscope). It contains 9185 subsamples derived from 1945 raw activity samples gathered directly from subjects.

D. OPPORTUNITY Dataset:

To evaluate human activity detection algorithms, In order to recognize human activity via wearable, object, and ambient sensors, the OPPORTUNITY Dataset [4] was created (classification, automated data segmentation, sensor fusion, feature extraction, and so on).

E. CAD-120

The RGB-D video sequences of people engaged in activities captured with the Microsoft Kinect sensor are part of the CAD-60 and CAD-120 data sets [15]. The CAD dataset includes numerous activities performed by four people in a variety of settings, including a kitchen, living room, and an office, among others. Robots were used to test the system, which reacted to the identified actions.

CONCLUSION

This paper delves into the topic of human activity recognition (HAR) using deep learning models. The study presents an overview of various deep learning models, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), long short-term memory (LSTM), and Gated Recurrent units (GRUs). These models have been applied to diverse applications and have shown significant potential in identifying and recognizing human behavior. However, recognizing human activities from video data is a challenging task due to various factors such as complicated backgrounds, camera motion, occlusion, and changes in views. These factors can pose problems for different models and affect their performance to varying degrees. To address these challenges, researchers have been exploring different deep learning models and their capabilities for HAR. The paper compares these models based on their architecture, applications, and advantages and limitations. For example, RNNs are useful for temporal modeling, while DNNs are better suited for image classification. CNNs can handle large datasets and complex features, and LSTMs are known for their ability to process long sequences of data. GRUs, on the other hand, offer similar capabilities to LSTMs but with fewer parameters and faster training times. The study also emphasizes the need for combining different deep learning models to enhance the accuracy of HAR. For instance, combining CNN and LSTM models can improve the integration of spatial and frequency features, leading to better performance. The paper also highlights the importance of selecting appropriate features and datasets for HAR, as well as the need for ongoing research in this field. However, further research is required to address the challenges and limitations in this field and to develop more robust and accurate models for HAR.

REFERENCES

[1] Sovan Biswas & Juergen Gall, "Structural Recurrent Neural Network (SRNN) for Group Activity Analysis", Winter Conference on Applications of Computer Vision-IEEE, 978-1-5386-4886-5, 2018.

[2] Li Wei &Shishir K. Shah, " Human Activity Recognition using Deep Neural Network with Contextual Information", In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications-IEEE, pages 34-43,ISBN: 978-989-758-226-4,2017.

[3] Mohanad Babiker, Othman O. Khalifa, khyawHtike, Aisha Hassan , Muhamed Zaharadeen " Automated daily human activity recognition for video surveillance using neural network", Proc. of the 4th IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Putrajaya, Malaysia ,November 2017.

[4] Jian sun, Yongling FU, Shengguang Li, Jie He, Cheng Xu, Lin Tan , "Sequential Human Activity Recognition Based on Deep Convolutional Network and Extreme Learning Machine Using Wearable Sensors" ,Journal of Sensors,2018.

[5] Mingyang wang, Guolong cui, Xiaobo yang, Lingjiang Kong, "Human body and limb motion recognition via

stacked gated recurrent units network" IET Radar Sonar Navig, Vol. 12 Iss. 9, pp. 1046-1051, 2018.

[6] Longteng Kong, jie Qin, di huang, Yunhong wang and luc van gool, "Hierarchical attention and context modelling for group activity recognition", ICASSP-IEEE, 978-1-5386-4658-8/18 2018.

[7] Haanvid Lee, Member, Minju Jung, and Jun Tani, "Recognition of visually perceived compositional human actions by multiple spatio-temporal scales recurrent neural network", IEEE Transactions on Cognitive and Developmental Systems, February 2017.

[8] Wilton, Loius CW, Carol Chen, "Artificial Intelligence for Sport Action and Performance Analyzing using Recurrent Neural Network (RNN) with Long Short Term Memory(LSTM)", ACM,2018, ISBN 978-1-4503-6584-0/11.

[9] Mohammed Mehedi Hassan, Md. Zia Uddin, Amr Mohamed, Ahmad Almogren , "A Robust Human Activity Recognition System using smart phone sensors and deep learning", Future Generation Computer Systems 81 (2018) 307–313-Elsevier,2018.

[10] Masaya Inoue, SozoInoue, Takeshi Nishida, "Deep Recurrent Neural Network for Mobile Human Activity Recognition with high Throughput", Artificial Life and Robotics-Springer, 2017.

[11] Chandrashekar M Patil, Jagadeesh B, Meghana M N , "An Approach of Understanding Human Activity Recognition and Detection for Video Surveillance using HOG Descriptor and SVM Classifier" , ICCTCEEC-IEEE;978-1-5386-3243-7,2017.

[12] Amin RasekhChien-An Chen Yan Lu , "Human Activity Recognition using Smartphone", 2014

[13] Md. Zia Uddin, Weria Khaksar, Jim Torresen, "Activity Recognition Using Deep Recurrent Neural Network on Translation and Scale-Invariant Features", IEEE, 2018.

[14] Xia K, Huang J, Wang H. "LSTM-CNN architecture for human activity recognition", IEEE Access, March 2020.

[15] Sowan Biswas ,Juergen Gall "Structural Recurrent Neural Network (SRNN) for Group Activity Analysis", Winter Conference on Applications of Computer Vision, IEEE, 2018.

[16] Ihianle, Isibor Kennedy, et al. "A deep learning approach for human activities recognition from multimodal sensing devices." IEEE Access 8 (2020): 179028-179038.

[17] Park, S. U., et al. "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services." Procedia Computer Science 100 (2016): 78-84.

[18] Jaouedi, Neziha, NoureddineBoujnah, and Med Salim Bouhlel. "A new hybrid deep learning model for human action recognition." Journal of King Saud University-Computer and Information Sciences 32.4 (2020): 447-453.

[19] Arifoglu, Damla, and AbdelhamidBouchachia. "Activity recognition and abnormal behavior detection with recurrent neural networks." Procedia Computer Science 110 (2017): 86-93.

[20] Murad, Abdulmajid, and Jae-Young Pyun. "Deep recurrent neural networks for human activity recognition." Sensors 17.11 (2017): 2556.

[21] Wang, Huaijun, et al. "Wearable Sensor-Based Human Activity Recognition Using Hybrid Deep Learning Techniques." Security and Communication Networks 2020 (2020).

[22] Xia, Kun, Jianguang Huang, and Hanyu Wang. "LSTM-CNN architecture for human activity recognition." IEEE Access 8 (2020): 56855-56866.

[23] Chen, Wen-Hui, Carlos Andrés Betancourt Baca, and Chih-HaoTou. "LSTM-RNNs combined with scene information for human activity recognition." 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). IEEE, 2017.

[24] Käse, Neslihan, MohammadrezaBabaee, and Gerhard Rigoll. "Multi-view human activity recognition using motion frequency." 2017 IEEE international conference on image processing (ICIP). IEEE, 2017.

[25] Xu, Cheng, et al. "InnoHAR: A deep neural network for complex human activity recognition." IEEE Access 7 (2019): 9893-9902.

[26] Osayamwen, Festus, and Jules-Raymond Tapamo. "Deep learning class discrimination based on prior probability for human activity recognition." IEEE Access 7 (2019): 14747-14756.

[27] Liu, Congcong, et al. "Abnormal human activity recognition using bayes classifier and convolutional neural network." 2018 IEEE 3rd international conference on signal and image processing (ICSIP). IEEE, 2018.

[28] Li, Meng, and Qiumei Sun. "3D Skeletal Human Action Recognition Using a CNN Fusion Model." Mathematical Problems in Engineering 2021 (2021).

[29] Tasnim, Nusrat, Mohammad Khairul Islam, and Joong-Hwan Baek. "Deep Learning Based Human Activity Recognition Using Spatio-Temporal Image Formation of Skeleton Joints." Applied Sciences 11.6 (2021): 2675.

[30] Deep, Samundra, and Xi Zheng. "Hybrid model featuring CNN and LSTM architecture for human activity recognition on smartphone sensor data." 2019 20th international conference on parallel and distributed computing, applications and technologies (PDCAT). IEEE, 2019.

[31] Abbaspour, Saedeh, et al. "A comparative analysis of hybrid deep learning models for human activity recognition." Sensors 20.19 (2020): 5707.

[32] C. Xu, J. He, X. Zhang et al., "Recurrent transformation of prior knowledge based model for human motion recognition," Computational Intelligence and Neuroscience, vol. 2018, Article ID 4160652, 12 pages, 2018.

[33] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," IEEE Transactions on Systems, Man, and

Cybernetics, Part B (Cybernetics), vol. 42, no. 2, pp. 513–529, 2012.

[34] C. Xu, J. He, X. Zhang, P. H. Tseng, and S. Duan, “Toward near-ground localization: modeling and applications for TOA ranging error,” IEEE Transactions on Antennas and Propagation, vol. 65, no. 10, pp. 5658–5662, 2017.

[35] M. Z. Uddin, W. Khaksar, and J. Torresen, "Human activity recognition using robust spatiotemporal features and convolutional neural network," 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Daegu, 2017, pp. 144-149.

[36] S. R. Ke, H. L. U. Thuc, Y. J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, “A review on video-based human activity recognition,” Computers, vol. 2, no. 2, pp. 88–131, 2013.

[37] C. Xu, J. He, X. Zhang et al., “Recurrent transformation of prior knowledge based model for human motion recognition,” Computational Intelligence and Neuroscience, vol. 2018, Article ID 4160652, 12 pages, 2018.

[38] B. F. Books and S. Haykin, Neural Networks a Comprehensive Foundation, Pearson Education, Singapore, 2010.

[39] C. A. Ronao and S. B. Cho, “Deep convolutional neural networks for human activity recognition with smartphone sensors,” in Neural Information Processing, pp. 46–53, Springer International Publishing, 2015.