INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Human Behavior Classification using 2D – Convolutional Neural Network, VGG16 and ResNet50

M Sowmya[1]*, **M Balasubramanian**[2], **K Vaidehi**[3]

1 Research Scholar, Department of CSE, Annamalai University, Annamalai Nagar, India
2 Associate Professor, Department of CSE, Annamalai University, Annamalai Nagar, India
3 Associate Professor, Department of CSE, Stanley College of Engineering and Technology for Women, Hyderabad, India

OPEN ACCESS

Received: 27-01-2023

Accepted: 30-03-2023

Published: 24-04-2023

Citation: Sowmya M, Balasubramanian M, Vaidehi K (2023) Human Behavior Classification using 2D – Convolutional Neural Network, VGG16 and ResNet50. Indian Journal of Science and Technology 16(16): 1221-1229. https://doi.org/10.17485/IJST/v16i16.199

* **Corresponding author**.

msoumya@stanley.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To develop a real-time application for human behavior classification using 2- Dimensional Convolution Neural Network, VGG16 and ResNet50. **Methods**: This study provides a novel system which considers sitting, standing and walking as normal human behaviors. It consists of three major steps: dataset collection, training, and testing. In this work real time images are used. In human behavior classification dataset there are 2271 trained images and 539 testing images. **Findings**: The Convolution Neural Network (CNN), VGG16 and ResNet50 are trained using human normal behavior images. **Novelty:** The dataset namely human behavior classification dataset is used in this work and the experimental results has shown that on human behavior classification ResNet50 has outperformed with accuracy of 99.72% compared to VGG16 and 2D-CNN. This work can detect the three normal behaviors of humans in an unconstrained laboratory environment.

Keywords: Deep Learning; 2D Convolution Neural Network (CNN); Human Behavior Classification; ADAM Optimizer; VGG16; ResNet50

1 Introduction

Human Behavior classification becomes a most populous and active research area for researchers from the past two decades. It is a categorization issue that considers identifying human movements and activities for monitoring purposes as well as the detection of anomalies in behavior. It plays an important role in human-to-human interaction and interpersonal relations. Examining human activity from still photos or video clips is the aim of understanding human activities recognition. In this work, 2D-CNN, VGG16 and ResNet50 are used to classify the human digital images into different classes like normal behaviors based on their behaviors. In [(1)], this research suggests a novel deep neural network for recognizing human behavior that blends LSTM and convolutional layers. The recurrent neural network (RNN), which is better suited for processing temporal sequences, has a variation called LSTM. The raw data gathered by mobile sensors was fed into a two-layer LSTM in the proposed architecture, which was then followed by convolutional layers. The model's effectiveness on the OPPORTUNITY dataset was assessed at 92.63%. In [(2)], one of the main research goals of this study is to identify human activities using data gathered from many gyroscopes and accelerometer sensors. Deep learning models are employed to perform HAR in the gathered data, yielding the best outcomes. They suggest combining gated recurrent units with hierarchical multi-resolution convolutional neural networks. They conducted an experiment using the UCI data sets, and the results show how effective the suggested model is because it obtained acceptable accuracy levels: 94.50% in the UCI data set. In[3], many artificial intelligence-based models are created for activity recognition; however, these algorithms perform poorly on long-term HAR in the real world because they are unable to extract spatial and temporal data. They create a hybrid model for activity recognition using a convolutional neural network (CNN) and a long short-term memory (LSTM) network, where CNN is used for extracting spatial characteristics and LSTM network is used for learning temporal information. The CNN-LSTM approach was used to produce an accuracy of 90.89%, demonstrating the suggested model's suitability for HAR applications. In[4], home care supervision has been more popular over the last several years as a result of the development of deep learning and the availability of depth sensors. This is becoming increasingly clear as a result of the ongoing COVID-19 epidemic, which has increased the health risks for the elderly and the disabled and generated a demand for online and remote-based alternatives to face-to-face connection. As a result, a quick method for detecting abnormal human activity in real time is suggested. The TST Fall Dataset shows that the approach is 91% accurate. In[5], in the field of computer vision, the study of human activity recognition (HAR) has gained a lot of attention. The primary goal of the paper is to identify six fundamental human behaviors walking, standing, sitting, and whether a person is moving upstairs or downstairs. Using a deep learning method called Convolutional Neural Network (CNN) with the accelerometer found in smart phones, the article focuses on predicting the activities. The evaluation of the two-dimensional CNN model's training and testing is supported by extra efforts. Finally, it was discovered that the model had an average accuracy of 89.67% and could accurately predict the activities. In[6], the study's objective is to determine whether two-stream approaches perform better than single-stream approaches when it comes to detecting human activities using deep learning. They initially used single-stream models that combined spatial and motion data. They discovered that MobileNetV2 CNN performs better when extracting motion features, while DenseNet201 CNN works better for extracting spatial features. On the UCF-101 dataset, the experimental results of the suggested model reached 85.21%Top-, outperforming the current state-of-the-art methodologies. In[7], for the HAR challenge, this study offers a pre-trained “CNN model VGG16” with the “SVM” classifier. The “VGG16 pre-trained CNN model” learns the deep features. The ”UniMiB” dataset contains 11771 examples of everyday human activity. VGG16 was used to categorize the images of human activity recorded by the mobile phone's accelerometer sensor. The evaluation metrics are classification accuracy and F-Score, and the proposed technique achieved 79.55 percent accuracy and 71.63 percent F-Score. In[8], in this study, a large dataset (500 video clips) collected from real elevator cabs with different backgrounds has been applied to ensure the robustness and generalizability of the proposed model. This study applies the two mainstream dangerous human behaviors, i.e., door blocking and door picking as case studies to test and evaluate the usability and availability. Experimental results show that the model has an 85% recognition rate of abnormal behavior. In[9], this investigation's goal was to identify human behaviors connected to a hypothetical synergistic task. To do this, a data gathering field experiment with five wearable sensors (integrated with tri-axial accelerometers, gyroscopes, and magnetometers) attached to twenty healthy people was created. The above mentioned activity included a number of load lifting and carrying-related sub-activities, which were completed by agricultural workers in actual field settings. A Long Short-Term Memory neural network, which is frequently used in deep learning for feature detection in time-dependent data sequences, was then fed the cleaned-up signals from the on-body sensors. With an average accuracy of 85.6%, the suggested methodology showed remarkable efficacy in forecasting the specified sub-activities. In[10], a method for accurately detecting violent behavior in surveillance situations was developed. A lightweight CNN model is trained on its own dataset of pilgrims as the first phase of the suggested framework to find pilgrims using security cameras. The second stage involves passing these previously processed salient frames to a CNN model for the extraction of spatial characteristics. A Long Short Term Memory network (LSTM) is created in the third step to extract temporal information. In the last step, the suggested system will generate an alarm in the event of violent activity or accidents to alert law enforcement agencies and enable them take the proper action, preventing accidents and stampedes. On datasets of violent behavior that are publicly available, like Surveillance, they have conducted numerous tests and achieved 81.05%. Human behavior classification dataset is created for this work and this work is focused on 2D-CNN, VGG16 and ResNet50 and results have shown the best accuracy compared to the existing work.

2 Methodology

The purpose of this work is to develop a classification system for different human behaviors using 2D CNN, VGG16 and ResNet50. The proposed system detects the normal behaviors of humans like sitting, walking and standing. For training, 2271 human behavior images are used, 539 human behavior images are used for testing for normal human behaviors like sitting, standing and walking. The steps followed in this work are as follows: Dataset collection, converting videos to frames, Training the 2D-CNN model, VGG16 and ResNet50 and testing the different human behavior images. The block diagram of the proposed system is shown in Figure 1.

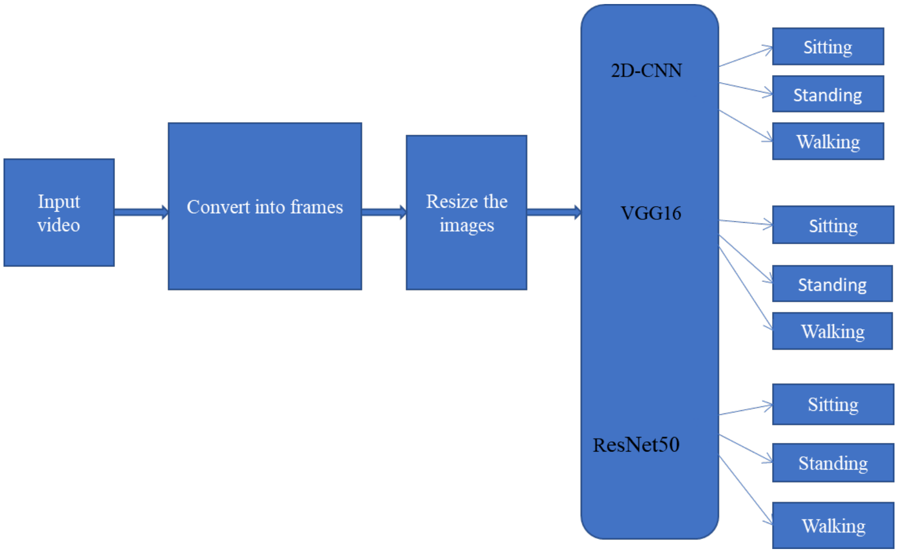

Fig 1. Block diagram of the proposed human behavior classification system

2.1 CNN Architecture

CNN can be used to play a key role in the fields like image processing, natural language processing (NLP), computer vision and other cognitive processing. A convolutional neural network is made up of an input layer, an output layer, and a numerous hidden layers. An image that is a matrix of pixel values provide as input. The matrices values for each pixel will range from 0 to 255 if the image is grayscale, the channel value will be 1. If a color image is present, the channel value is 3, which stands for Red, Green, and Blue. The main purpose of convolution layer is to extract features from input image and it contains with parameters like filter, stride and depth then it generates feature map. Filter size used for convolution is 3x3. The number of filters utilized is related to depth. The stride defines how many inputs steps the filter slides through. When the stride value is 1, the filters are moved one pixel at a time. When sliding the filters around with a stride of 2, they jump 2 pixels at a time. The activation function used in 2D-CNN are RELU, Softmax, tanH and the Sigmoid functions. RELU activation function is used in this work. The advantages of RELU are replacing negative values in the feature map by zero, Maximum Threshold values are infinite, so there is no issue of vanish gradient problem, the output prediction accuracy and their efficiency is maximum, speed is fast compare to other activation functions. When convolution is applied to an input, the produced output size of the matrix is reduced, resulting in information loss, for avoiding this padding concept is implemented. Padding is done through the input volume with zeros at the border. Valid and same are two popular padding options. The same padding indicates that output size remains the same as input size, and valid padding means no padding is added. After convolution layers, CNNs frequently employ the pooling layer operation, which has the goal of reducing the dimension, also known as down sampling. For pooling layer, max pooling is used which takes the maximum values from the feature map. Then features are fed into Fully Connected (FC) Layer which uses flattening. Flattening is used to convert all the resultant 2-Dimensional arrays from pooled feature maps into a single long continuous linear vector**.** Dropout and Activation Function are connected to the FC layer. Dropout is used to solve over fitting problem and improve generalization error. An Activation Function decides whether a neuron should be activated or not. It helps to normalize the output of each neuron to a range between 1 and 0.

In Figure 2, the convolution layers perform feature extraction and generate feature maps from the input image. The RELU activation function employed is a nonlinear function basically for image transformation. This accepts the feature maps as input and transforms them into the system to train and learn them properly. The transformed output will be sent as input to the next level of convolution. This is illustrated in the equation below:

$$Conv\_layer = (filter, RELU) \tag{1}$$

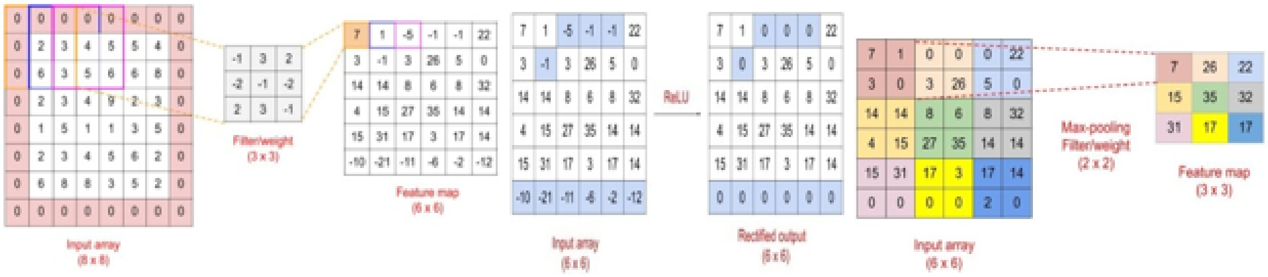

Fig 2. Sample of Feature map generated using 2D-CNN used in this work

The ReLU activation function uses the equation below:

$$Max(0,x) = \begin{matrix} x, & x \geq 0, \\ 0\ , & x < 0 \end{matrix} \tag{2}$$

$$Feature\ map\ Size\ = (N\ \ F\ +\ 2P)/S\ +\ 1 \tag{3}$$

2.2 Implementation of 2D-CNN on Human Behavior Classification Dataset

Convolution layers, Max-pooling layers, and a flattening layer are all included in a 2D CNN model. The variable number of fully connected dense layers get the flattened output. The first layer is made up of the unprocessed pixels from a 100x100 human behavior images with three color channels. The first convolution layer, which has 32 filters, then the size is 100x100x32, by performing a dot product of the weights of the filters and the input image pixel values. Max-pooling layer is applied along the spatial dimension (height x width), and this layer reduces the dimension to 50x50x32.down sampling operation is accomplished. Following the output of the first max-pooling layer of size 2x2 as input to next level and output dimension of 25x25x64 is down sampled, the CNN filter of size 3x3 and with 64 filters is applied in the second layer. The output of the third layer's CNN filter, which has a 3x3 filtering matrix, is 25x25x128. The output of the second 2x2 max-pooling layer is 12x12x128. Then, in layer 4, the images are "flattened." The output layer, which will be the final layer, have a softmax activation function and include six output neurons for classification. Standing, walking and sitting normal categories and hit, kick and punch abnormal activities are used for the proposed human behavior classification system.

2.3 VGG16 Architecture

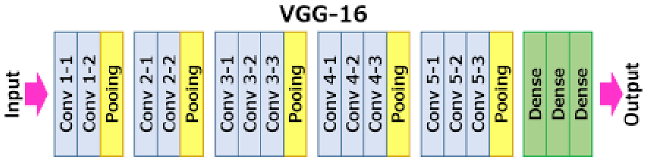

Fig 3. Architecture of VGG16

In Figure 3 the first two layers are convolutional layers with 3x3 filters. The first two layers contain 64 filters, resulting in a volume of 224x224x64 due to the utilization of the identical convolutions. The filters always have a 3x3 kernel size with stride value 1. This was followed by the use of a pooling layer with a max-pool of 2*2 size and stride 2, which reduces the volume's height and width from 224x224x64 to 112x112x64. Two additional convolution layers with 128 filters are added after this. The new dimension as a result is 112x112x128. Volume is decreased to 56x56x128 once the pooling layer is employed. The size is decreased to 28x28x256 by adding two additional convolution layers, each with 256 filters. A max-pool layer separates two more stacks, each having three convolution layers. 7x7x512 volume is flattened into a Fully Connected (FC) layer and a softmax output of 3 classes i.e. Sitting, Walking and Standing after the last pooling layer.

2.4 ResNet50 Architecture

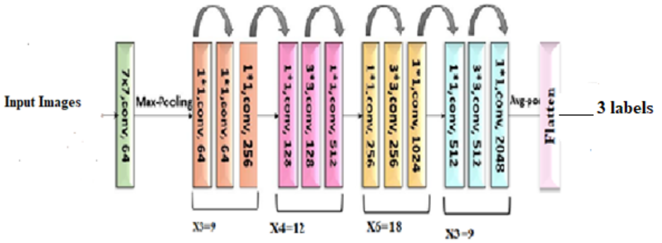

Fig 4. Architecture of ResNet50

Researchers at Microsoft Research first proposed ResNet in 2015 and introduced the residual network architecture, a new design. The network's performance decreases or becomes saturated as it gets deeper. Since gradients are vanishing, accuracy is reduced. The idea of a residual network provides a solution to vanishing gradient during back propagation. Skip connections is a method used by residual networks. A skip connection links straight to the output after skipping a few stages of training. Gradients can pass directly from later levels to starting layers through the skip connections which is shown in Figure 4. ResNet is a powerful tool which is used in many computer vision tasks. It uses Skip connections which adds previous layer output to next layer to prevent from vanish gradient problem. It contains two blocks, identity block and convolution block. If output and input is same then identity block is used and if output is not equal to input, then convolution block is inserted so that input will be equal to output.

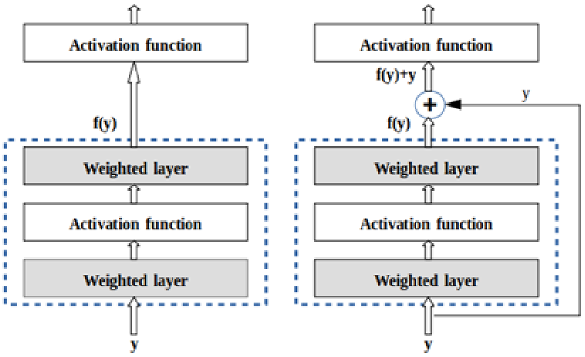

Fig 5. Basic block (left) and residual block (right) of ResNet50

'Skip connection' is a direct connection that skips over some layers of the model. The output is not the same due to this skip connection. Without the skip connection, input 'X gets multiplied by the weights of the layer followed by adding a bias term. The activation function, F() and the output is shown in Equation (4) as:

$$F(w*x + b) = (F(X)) \tag{4}$$

But with skip connection technique, the output is:

$$F(X)+x \tag{5}$$

In ResNet-50, there are two kinds of blocks — 1. Identity Block, 2. Convolutional Block.

The value of 'x' is added to the output layer if and only if the -

input size=output size

If this is not the case, then add a 'convolutional block' in the shortcut path to make the input size equal to output size.

There are 2 ways to make the input size equal to the output size -

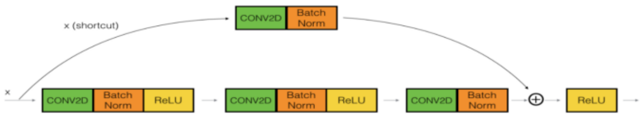

Fig 6. Convolution block

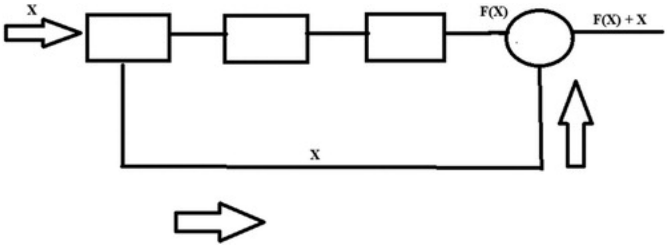

Fig 7. Identity block

3 Padding the input volume, 2 Performing 1*1 convolutions

To equal input and output size the equation used is

$$(n+2p-f)\div s+1 \quad x \quad (n+2p-f)\div s+1 \tag{6}$$

where, n= input image size, p=padding, s=stride, f=number of filters.

In CNNs, to reduce the size of the image, pooling is used. In Resnet 50, make use of stride=2 instead.

The ResNet 50 architecture contains the following element: a convolution with 64 distinct kernels, each having a stride of size 2, and a kernel size of 7 x 7, providing layer 1. Next max pooling with also a stride size of 2. The next convolution has three levels: 1x1,64 kernel, 3x3,64 kernel, and finally a 1x1,256 kernel. These three layers are repeated a total of three times, yielding nine layers in this step. The kernel of 1 x 1,128 is displayed next, followed by the kernel of 3 x 3,128 and, finally, the kernel of 1 x 1,512. This procedure was performed four times for a total of 12 layers. Following that, we have a kernel of size 1 x 1, 256, followed by two more kernels of size 3 x 3,256 and size 1 x 1,1024; this is repeated six times, giving us a total of 18 layers. After that, a 1x1,512 kernel was added, followed by two more kernels of 3x3,512 and 1x1,2048. This process was done three times, giving us a total of nine layers. Following that, add an average pool, finish it with a layer that has three fully linked nodes those are Sitting, Standing and Walking, and then add a softmax function to give one layer.

4 Result

The dataset is collected using real time video of human behaviors and then converted to frames as input images. All the RGB images are fed to the CNN. The dataset is divided into training set and test set 80:20 ratio. For training, 2271 human behavior images are used, 539 human behavior images are used for testing for normal human behaviors like sitting, standing and walking which is shown in Table 1.

The CNN model is trained with varying numbers of dense layers and then the model is trained with various numbers of parameters and their weights are adjusted. The model with six labels are trained with 3,999,079 training parameters gives good classification results.

Dataset URL: https://github.com/sowmyamudumba/human-behaviour-classification

Table 1. Normal images taken for training and testing in this proposed work.

Behaviors	Training Images	Testing Images
Normal Behaviours		
Sitting	310	201
Standing	990	32
Walking	971	306

4.1 Performance of human behaviors using 2D-CNN

Table 2. Trainable CNN Parameters of each layer

Name of the Layer	Calculation	Trainable Parameters
Convolution1 Number of filters are 32	RGB x Kernel size x Stride +1 x depth of the filters	3x((3x3)x1)+1x32 = 28x32=896
Convolution 2 Number of filters are 64	Kernel size x previous layer depth filters+1 x current layer depth of filters.	((3x3)x32+1)x64 289x64=18496
Convolution 3 Number of filters are 128	Kernel size x previous layer depth filters+1 x current layer depth of filters.	((3x3)x64+1)x128=73856
Convolution 4 Number of filters are 256	Kernel size x previous layer depth filters+1 x current layer depth of filters.	((3x3)x128+1)x256 =295168
Convolution 5 Number of filters are 512	Kernel size x previous layer depth filters+1 x current layer depth of filters.	((3x3)x256+1)x512=1180160
FC1	4609x500	2304500
FC2	501x250	125250
FC3	251x3	753
Total Trainable Parameters : 3999079		

First convolution is made up of kernel size with 3x3 and stride value 1 with 32 depth of the filters. As input image contains color image so channel value is 3 i.e. Red, Green and Blue. Second convolution is made up of previous depth of the filters i.e. 32 and kernel size 3x3 with 64 depth of the filters in that layer. Third convolution is made up of previous depth of the filters i.e. 64 and kernel size 3x3 with 128 depth of the filters in that layer. Fourth convolution is made up of previous depth of the filters i.e. 128 and kernel size 3x3 with 256 depth of the filters in that layer. Fifth convolution is made up of previous depth of the filters i.e. 256 and kernel size 3x3 with 512 depth of the filters in that layer. Total trainable parameters are 3999079.

In this work human behavior classification using 2D-CNN, VGG16 and ResNet50 are tested and the results are compared.

4.2 Performance of human behaviors with VGG-16

As discussed in previous sections VGG-16 is trained with 16 layers consisting of 13 convolution layers with five max pooling layers and 3 fully connected layers. The input dimension of Conv. Layer1 is 224 224 x 64, conv. Layer2 is 224 224x64, and Max_pooling1 is 112 112x64. The following table shows the 16 layers structure and its dimensions. Network parameters are given in the Table 3.

Table 3. Network Parameters of VGG-16.

Layer (type)	Output shape	Parameter
Conv2d_1	224x224x64	1792
Conv2d_2	224x224x64	36928
Max_Pooling2d_1	112x112x64	0
Conv2d_3	112 x 112 x 128	73856
Conv2d_4	112 x 112 x 128	147584
Max_pooling2d_2	56 x 56 x 128	0
Conv2d_5	56x56x256	295168
Conv2d_6	56x56x256	590080
Conv2d_7	56x56x256	590080
Max_pooling2d_3	28x28x256	0
Conv2d_8	28x28x512	1180160
Conv2d_9	28x28x512	2359808
Conv2d_10	28x28x512	2359808
Max_pooling2d_4	14x14x512	0
Conv2d_11	14x14x512	2359808
Conv2d_12	14x14x512	2359808
Conv2d_13	14x14x512	2359808
VGG16(Max Pooling2d)	7x7x512	0
Flatten	25088	0

Continued on next page

Table 3 continued

FC1(Dense)	256	6422784
Fc2(Dense	128	32896
Softmax	3	387
Trainable param		21,170,755

4.3 Performance of human behaviors with ResNet50

The back propagation method is used in this case. Convergence becomes more difficult as the network grows deeper. As discussed in previous sections ResNet-50 is trained with 50 layers. Consisting of convolution layers with zero padding, max pooling and activation function, batch normalization layers, average pooling and fully connected layers. The input dimension of Conv. Layer 1 is 224 224x64, Conv. Layer 2 is 224 224x64, and Max_pooling1 is 55 55x64. The Table 4 shows the 50 layers structure and its dimensions.

Table 4. Network Parameters of ResNet50

Layers	50 Layers
Conv1	7x7,64, stride 2
	3x3x max pool, stride 2
Conv2_x	[1 x 1,64 3 x 3,64 1 x 1,256] x 3
Conv3_x	[1 x 1,128 3 x 3,128 1 x 1,512] x 4
Conv4_x	[1 x 1,256 3 x 3,256 1 x 1, 1,1024] x 6
Conv5_x	[1 x 1,512 3 x 3,512 1 x 1,1024] x 3
	Average pool

Table 5. Comparative performance of Classification of human normal Behaviors with human classification dataset using 2D-CNN, VGG16 and ResNet50

	2D-CNN			VGG16			ResNet50		
Human Normal Behaviors	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Sitting	0.99	0.95	0.97	0.99	0.96	0.97	1.00	1.00	1.00
Standing	0.88	0.98	0.93	0.91	0.98	0.94	1.00	0.93	0.97
Walking	1.00	0.99	1.00	1.00	1.00	1.00	0.98	1.00	0.99

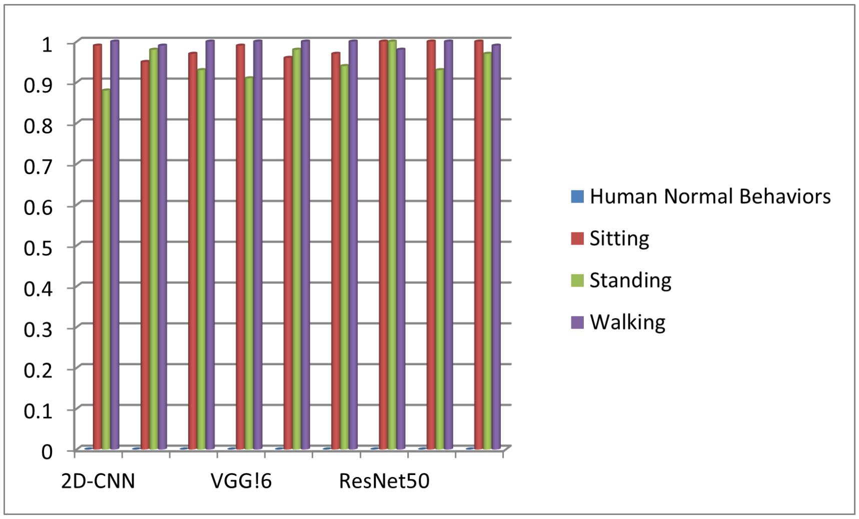

Chart 1: **Overall performance of human normal behavior classification on human behavior classification dataset using 2D-CNN, VGG16 and ResNet50 of different evaluation metrics**

Table 6. Comparative performance of Classification of human normal Behaviors with human classification dataset using 2D-CNN, VGG16 and ResNet50

Datasets	Model1	Model2	Model3	Accuracy
Human Behavior Classification dataset	2D-CNN	VGG16	ResNet50	2D-CNN-92.63% VGG16-94.56% ResNet50-99.72%

Table 6 Shows overall performance of human behavior classification dataset and it has shown that on human behavior classification dataset ResNet50 has outperformed compared to other models like VGG16 and 2D-CNN.

5 Discussion

Table 7. Existing accuracy with proposed method

Existing Methods	Methods	Accuracy
[3]	CNN+LSTM	90.89%
[7]	VGG16+-SVM model	79.55%
[9]	LSTM	85.6%
Proposed Method	**2D –CNN+VGG16+ResNet50**	**99.72%**

Table 7 Shows comparison of existing methods with that of proposed method. Our proposed exhibits the best accuracy of 99.72% using ResNet50.

6 Conclusion

This proposed research uses "convolutional neural networks", VGG16 and ResNet50 to build a system that can recognize the actions like sitting, standing, and walking. The human behavior classification dataset is created for this work and the experimental results has shown that ResNet50 has outperformed VGG16 and 2D-CNN. Multiple human activities can be done in future work with the use of LSTM Architecture.

References

1) Xia K, Huang J, Wang H. LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access*. 2020;8:56855–56866. Available from: https://doi.org/10.1109/ACCESS.2020.2982225.
2) Nafea O, Abdul W, Muhammad G. Multi-sensor human activity recognition using CNN and GRU. *International Journal of Multimedia Information Retrieval*. 2022;11:135–147. Available from: https://doi.org/10.1007/s13735-022-00234-9.
3) Khan IU, Afzal S, Lee JW. Human Activity Recognition via Hybrid Deep Learning Based Model. *Sensors*. 2022;22(1):323–323. Available from: https://doi.org/10.3390/s22010323.
4) Hristov P. Real-time Abnormal Human Activity Detection Using 1DCNN-LSTM for 3D Skeleton Data. In: and others, editor. 2021 12th National Conference with International Participation (ELECTRONICA). IEEE. 2021. Available from: https://doi.org/10.1109/ELECTRONICA52725.2021.9513696.
5) Prasad A, Tyagi AK, Althobaiti MM, Almulihi A, Mansour RF, Mahmoud AM. Human Activity Recognition Using Cell Phone-Based Accelerometer and Convolutional Neural Network. *Applied Sciences*. 2021;11(24):12099–12099. Available from: https://doi.org/10.3390/app112412099.
6) Chakraborty M, Pramanick A, Dhavale SV. Two-Stream Mid-Level Fusion Network for Human Activity Detection. In: International Conference on Innovative Computing and Communications, Singapore. Springer Singapore. 2020;p. 331–343. Available from: https://doi.org/10.1007/978-981-15-5148-2_30.
7) Athavale VA, Gupta SC, Kumar D, Savita. Human Action Recognition Using CNN-SVM Model. *Advances in Science and Technology*. 2021;105:282–290. Available from: https://doi.org/10.4028/www.scientific.net/AST.105.282.
8) Shi Y, Guo B, Xu Y, Xu Z, Huang J, Lu J, et al. Recognition of Abnormal Human Behavior in Elevators based on CNN. In: and others, editor. 2021 26th International Conference on Automation and Computing (ICAC). IEEE. 2021. Available from: https://doi.org/10.23919/ICAC50006.2021.9594189.
9) Anagnostis A, Benos L, Tsaopoulos D, Tagarakis A, Tsolakis N, Bochtis D. Human Activity Recognition through Recurrent Neural Networks for Human–Robot Interaction in Agriculture. *Applied Sciences*. 2021;11(5):2188–2188. Available from: https://doi.org/10.3390/app11052188.
10) Habib S, Hussain A, Albattah W, Islam M, Khan S, Khan RU, et al. Abnormal Activity Recognition from Surveillance Videos Using Convolutional Neural Network. *Sensors*. 2021;21(24):8291–8291. Available from: https://doi.org/10.3390/s21248291.