Advances and Applications in Mathematical Sciences
Volume 21, Issue 11, September 2022, Pages 6627-6650
© 2022 Mili Publications, India

REAL TIME VIDEO BASED HUMAN SUSPICIOUS ACTIVITY RECOGNITION USING DEEP LEARNING

J. INDHUMATHI and M. BALASUBRAMANIAN

[1]Research Scholar
[2]Associate Professor
Department of Computer Science
and Engineering, Annamalai University
Annamalai Nagar, Chidambaram, India
E-mail: indhumathi20061996@gmail.com
balu_june1@yahoo.co.in

Abstract

Nowadays, the primary concern of any society is providing safety to an individual. It is very hard to recognize the human behaviour and identify whether it is suspicious or normal. In this proposed system, deep learning approach is used to classify human activities such as: (a) normal activity, (b) criminal activity and (c) suspicious activity in public environment. Novel 2D-CNN (8 layers, 10 layers, 12 layers, 14 layers) is trained using human activities video frames. Similarly, the proposed work uses Pre-trained VGG16 and ResNet50 techniques for recognizing the suspicious activities of humans. Trained models are used to recognize the human suspicious activities in testing video frames. The performance of system shows that novel 2D-CNN, VGG16, and ResNet50 achieve an accuracy of 98.96%, 97.84%, 99.01% respectively. The performance of system is evaluated using the Kaggle and real-time video datasets. The proposed system using 2D-CNN outperforms the pre-trained model VGG16 in Kaggle/real-time video.

1. Introduction

In recent times, Closed Circuit Television (CCTV) cameras are installed in many public places to monitor illegal activities. Monitoring human activities are highly sensitive in communal areas to prevent terrorism, theft or robbery, accident, riots, chain snatching, crime and other suspicious activities. [23] CCTV surveillance is very valuable when it combines with

2020 Mathematics Subject Classification: 68T07.

Keywords: Convolutional Neural Networks (CNN), Real-time, human Suspicious Activity Recognition, 2D-CNN, VGG- 16 and ResNet50.

Received March 28, 2022; Accepted June 28, 2022

image processing and object detection algorithms for the monitoring purpose. The object detection algorithms can be further extended to detect and recognize a person holding weapons like fire arms or sharp objects like knives. The human pose, face covered mask and weapon in the hands of a human are considered for human suspicious activity recognition. Human activities can be monitored from the following public places like Markets, Buildings, Railway stations, Courts, Automated Teller Machine (ATM) centers, Healthcare Sectors, Religious places, Hospitals, Airports, etc. In this system, human activities are classified into three stages, namely normal activities, suspicious activities and abnormal activities. Normal activities are the usual events that are not dangerous in the human world. Suspicious and abnormal activities are dangerous for everyone in the world. There are several abnormal measures like theft detection, unknown person entry to building or house, damaging (house windows, door, car windows or door), violence detection such as slapping, punching, hitting, shooting at public places like elevator and parking places and attack with weapons [11].

Currently, deep learning has been given particular attention by the computer vision community [19]. Deep learning algorithms learn gradually more about the image as it goes through several neural network layers. Early layers study how to detect low-level features like edges and subsequent layers blend features from earlier layers into a more holistic representation. A convolutional neural network (CNN/Conv Net) is a category of deep neural networks [25], mostly it is applied to analyse visual imagery. There are several variants of CNN architectures that have been developed to solve many image processing problems. In this paper, our proposed work has been implemented using three CNN architectures namely 2D-CNN, VGG16 and ResNet50. The main aim of this paper is to detect and recognize the suspicious activity using 2D-CNN, CNN-VGG16 and ResNet50. [2] model using human Dataset. This system is developed using Python language with the open-source library called Tensor Flow and Keras deep learning library to build CNN using Jupyter notebook environment.

1.1 Related Work

Several studies have been already made with Human Suspicious Activity Recognition approaches. Still, it remains a challenging task due to the complex human Activity or behavior patterns and identification with machine learning approaches. The Human Suspicious Activity recognition consists of four main steps namely: (a) identifying the region of activities (b) extracting the intrinsic values (features) from the region (c) Normalization phase to eliminate the wide changes of values in the dataset. (d) Classification techniques to identify behavior or activity. In dataset collection, the human activity is identified and preprocessed after that, it is fed to the feature extraction block. The features are given to any of the classification algorithms like Decision trees [9], Random Forest [5], Support Vector Machine [8], etc. After classification, the result of the proposed systems is analyzed. Until now comparing with other classification methods, CNN performance is more accurate in recognizing various features activity.

Adithyan Palaniappan et al., focused on identifying Abnormal Human Activity Recognition Using SVM [7]. The dataset contains 3-axis accelerometer and gyroscope data, which is mounted on 13 different individuals' body. The system recognizes 9 different activities of an individual using multi-class SVM. After designing the novel multi-class SVM, it has been tested. The feature dataset is given to the system for classification.

Tao Gu et al., proposed an approach using emerging patterns (EPs) and score functions [20] for recognizing sequential activities. The sliding window technique is used to separate the activities within time bound. To do this they also use trace segmentation algorithm to extract activity data within a specific time period. A sequential activity model is made to which EPs and activity correlation scores are used to recognize complex activities. This approach yields 91% accuracy for sequential activities and 82% accuracy for concurrent activity.

Amrutha et al., use footages obtained from CCTV camera for monitoring students' activities in a campus and send message to the corresponding authority when any suspicious event occurs. The system has three classes. There are students using mobile phone inside the campus-Suspicious class, Students fighting or fainting in campus-Suspicious class, Walking, running-Normal class. The architecture has different phases like video capture, video pre-processing, feature extraction, classification and prediction [1].

Srikanth et al., proposed transfer learning using Pre-trained VGG16 with Deep Convolutional Neural Network for Classifying Images. The main aim of transfer learning is to solve more complex problems easily by applying the technique already used with little modifications so that better results can be obtained more accurately [18].

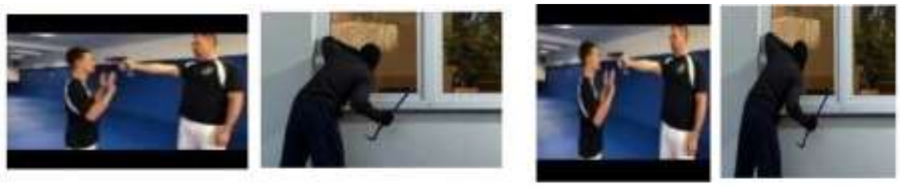
Rachana Gugale et al., explored and classified the five suspicious activities Shooting, punching, kicking, knife attack and sword fight. The non-suspicious activities were 6th class. Deep learning method is employed to notice suspicious activities from images and videos using Convolutional Neural Networks [10]. The total number of images used for dataset is 17, 716. They analyzed different CNN architectures and compared their accuracy. The Res Net architecture was used to build the CNN model. Then tried ResNet-18, ResNet-34 and ResNet-50 approaches. A learning rate in the range of $3 \times 10^{(-5)}$ to $3 \times 10^{(-4)}$ also works the best result [10].

2. Proposed Methodology

2.1 Description of Kaggle Dataset and Real-time Datasets

2.1.1 Description of Kaggle dataset.

This dataset contains 3 categories of color human activities images (criminal, normal and suspicious) with a total of 2160 images in the jpg format. The number of images vary on each category has at least 600 images. The images of the datasets are different size. Figure 1 illustrates the resizing process on a sample image.



(a) (b)

Figure 1. Kaggle dataset (a) original image (b) resized image.

2.1.2 Description of Real-time Dataset

This dataset contains 3 categories of color human activities video/images (criminal, suspicious and normal) with a total of 9000 images/video in the jpg format. The number of images vary on each category has at least 1800 images. The images of the datasets are different size. Figure 2 illustrates the resizing process on a sample image.

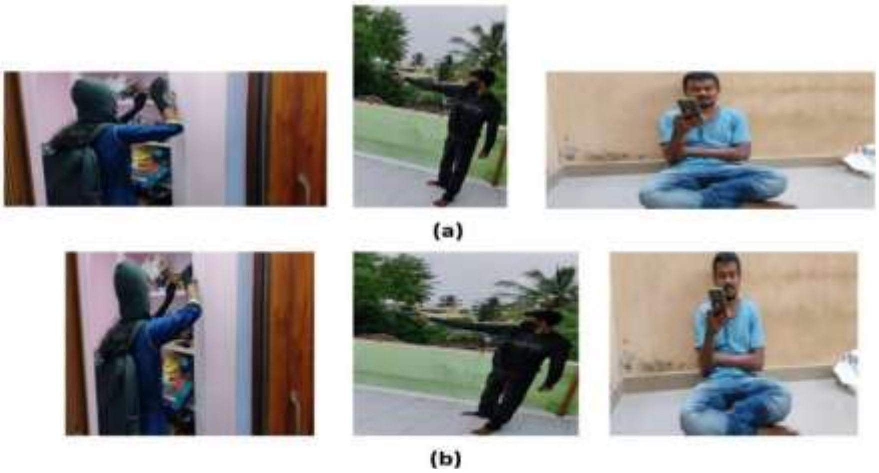


Figure 2. (a) Original Images (b) resized images.

2.2. Recognition of Human Activity

The traditional neural network consists of three layers, including input layer, hidden layer and output layer. Based on the traditional neural network, the convolutional neural network (CNN) [4] works as a feature extractor adding up the convolution layer and the sub-sampling layer. It is the combination of the artificial neural network and back propagation algorithm which simplifies the complexity of the model and reduces the parameters.

CNN eliminates the need for manual feature extraction and it works by extracting features from images. The features are learned automatically while the network trains on a set of images. This makes CNN models tremendously accurate for Image processing. CNN detects the features through tens or hundreds of hidden layers. Each layer of CNN increases the complexity of the learned features. Recognition of human suspicious activity block diagram is shown in Figure 3.

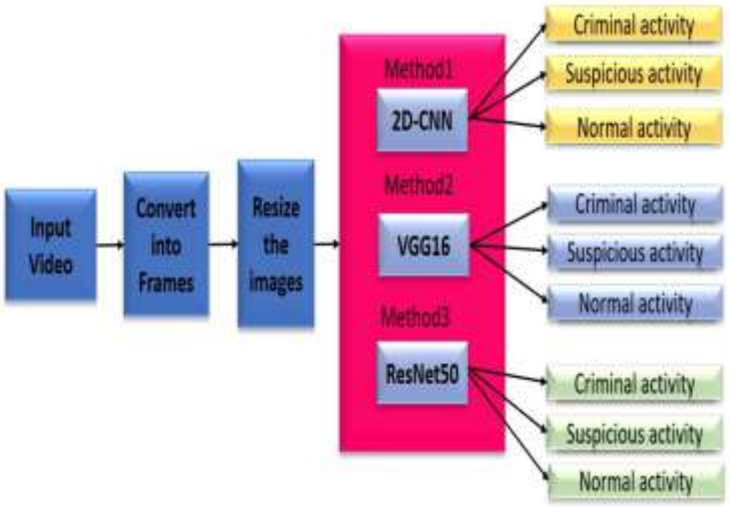


Figure 3. Block diagram of recognition of human suspicious activity using 2D-CNN, VGG16 and ResNet50.

2.2.1 2D-CNN Architecture

2D-CNN is the Standard CNN which can classify the two-dimensional inputs such as videos and images. The 2D CNN accepts the input layer of size 250×250. This CNN comprises varying convolution and sub-sampling layers followed by several multi-layer perceptron (MLP) layers. The MLP layers handle the outcomes from the convolution layers to generate an output vector that characterizes the input signal. The architecture of a 2D CNN hangs on the sizes of the convolution layers and the amounts of the MLP layers of the kernel size and the sub-sampling factors. The parameters of the 2D filter kernels in CNN are automatically optimized by back propagation. The architecture of 2D-CNN is shown in the Figure 4.

Convolution is a specific form of linear operation which is used for feature extraction. Convolution layer performs on the input data with the use of a filter or kernel to produce a feature map. In this layer, Kernel is a small array of numbers which is applied across the input which is also an array of numbers called a tensor. A component-wise product between each component of the kernel and the input tensor is calculated at each location of the tensor and add together to obtain the output value in the equivalent location of the output tensor, called a feature map. Max pooling is calculating the maximum or highest value in each patch of each feature map.

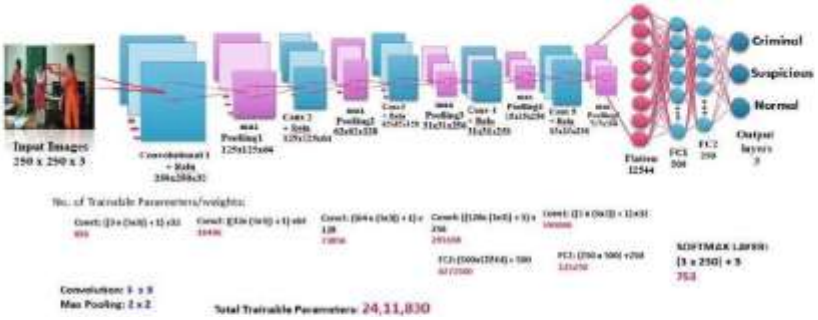


Figure 4. Architecture of 2D-CNN.

Numerous convolutions are performed on the input where each operation uses a different filter. This results in different feature maps. In the end, feature maps are collected and put them together as the final output of the convolution layer. ReLU Activation Function is responsible for converting the node's summed weighted input into the node's activation or output for that input. ReLU or rectified linear activation function $Y = \max(x, 0)$, if the input is positive, outputs the input directly; else, it outputs zero. It is quicker to train and produces higher performance. A pooling layer is added after the convolutional layer and it delivers the down sampling operation which reduces the dimensionality of the feature maps and decreases the number of subsequent learnable parameters. It provides the down sized feature maps by summarizing the features in patches of the feature map. The feature maps that deliver from final convolution or pooling layer are transformed into a one-dimensional (1D) array of numbers (or vector) called flattening process which are connected to one or more fully connected layers, so called dense layers. Flattening layer is fed to fully connected layer to classify the images. The final layer of CNN is the SoftMax activation function (instead of ReLU) which is used to get probabilities of the input being in a particular class (classification).

$$p_i = \frac{\exp(x_i)}{\sum_{j=1}^{j=c} \exp(x_i)} \text{ for } i = 1, 2, 3, \ldots, c \ldots > \qquad (1)$$

2.2.2 Architecture of Visual Geometry Group (VGG-16)

The second method practiced in this study is VGG-16 where the full name of VGG is the Visual Geometry Group. It is considered to be one of the excellent vision model architectures till date. VGG-16 is trained to a deeper structure of 16 layers consisting of 13 convolution layers with five max pooling layers and three fully connected layers. The architecture of VGG16 is depicted in Figure 5.

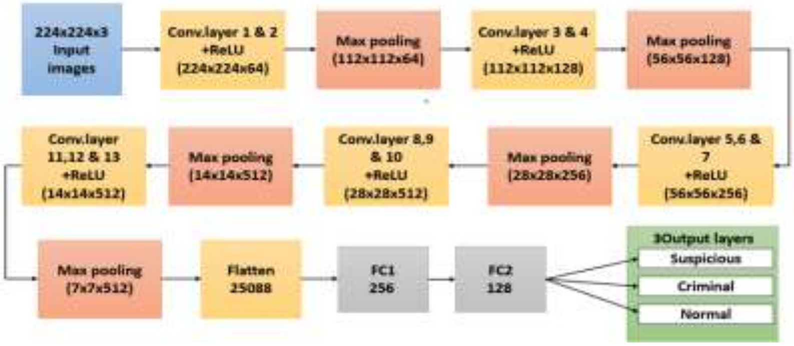


Figure 5. Architecture of VGG16.

The VGG architecture takes RGB image as the input with size of 224×244 [2]. The RGB values for the images in the training data set are composed and average values are intended. Then the image is fed as input to the VGG convolution network. The first convolutional layers in VGG-16 architecture are 3×3 convolutional layers having stride size as 1 and padding with one size, and the next pooling layers are 2×2 layer size with a stride size of 2. The pre-trained size of input image in the VGG-16 is 224×224. Pooling layer reduces the size of the feature map to half thus reducing the size of the image. The last feature map before the fully connected layers is 7×7 with 512 channels and it is expanded into a vector with $25, 088 (7 \times 7 \times 512)$ channels. The pooling layers are responsible for making the layers narrower and it has 3 VGG fully connected layers. A stack of convolutional layers is followed by three Fully-Connected (FC) layers [3]. The first two have 4096 channels each and the third performs classification thus contains 1000 channels (one for each class). The final layer is the soft-max layer. All networks have the same configuration for the fully connected layers.

2.2.3 ResNet50 Model

Residual network 50 model is used to locate the shortcuts in illustration of network functions in deep method [6]. The model is trained with various numbers of parameters and their weights are adjusted for 50 epochs. A very deep network's key advantage is that it can represent extremely complex functions. The problem of disappearing gradients is a significant impediment for training large neural networks. Gradient descent is slow in very deep networks because the gradient signal often decreases to zero soon. As the network back propagates from the final layer back to the first layer, weight matrix is multiplied with the value on each step and as a result, the gradient might soon drop to zero, obstructing the training process.

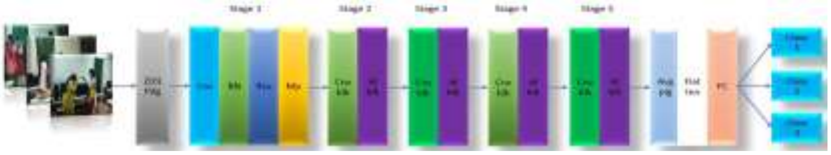


Figure 6. Architecture of ResNet-50.

The suspicious videos are converted into frames. The frames are resized to $250 \times 250,$ these images are fed to ResNet50. [12, 24] The input images of ResNet50 dimension are taken by $224 \times 224 \times 3.$ The next convolutional Block is activated when the input and output dimensions do not match, this sort of block is used. The shortcut path differs from the identity block in that it includes a CONV2D layer. The proposed model consists of various stages, each stage containing their own convolution and identity block. Each convolution block contains three levels and each identity block has three layers as well. There are around 23 million trainable parameters in the resnet-50. The architecture of ResNet-50 is depicted in Figure 6.

3. Experimental Results

3.1 Datasets

3.1.1 Kaggle Dataset

The dataset images were collected from Google Images, movie datasets, shutter stock Images etc. The dataset contains the human images related to terrorism (gun attack), theft or robbery, fighting, chain snatching, crime etc.

Figure 7. Datasets (a) criminal activity (b) normal activity (c) suspicious activity.

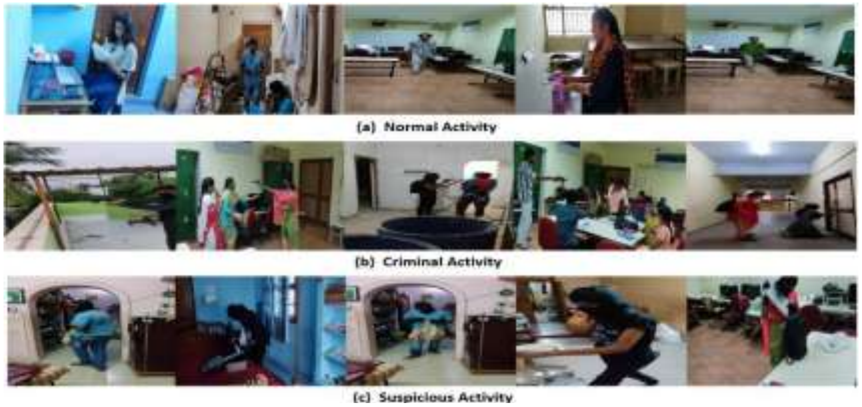
Table 1. Kaggle dataset.

Type of Images	Criminal	Normal	Suspicious	Total
Number of images for Training	520	520	520	1560
Number of images for Testing	200	200	200	600
Total	720	720	720	2160

Dataset for 2D-CNN, VGG16 and ResNet50 provides total of 2160 images datasets used from human (male and female) activities. They are Normal Activity, Criminal Activity and Suspicious Activity. All the Image datasets are converted into Array creation fed to the 2D-CNN. Dataset is divided into training and testing sets. For training the architectures, 1560 human activity images are used and 600 human activity-based images are used for testing. Description of Kaggle dataset given in the Table 1. Figure 7. shows the samples of the dataset collected for normal, suspicious and criminal activity.

3.1.2 Real-time Dataset

The dataset is collected using a camera with a resolution of 1280×720 in the home, unusable buildings, hotels and laboratory environment. The dataset is collected from persons of age group belonging to 18 to 30 years in



(a) Normal Activity

(b) Criminal Activity

(c) Suspicious Activity

Figure 8. (a) normal activity (b) criminal activity (c) suspicious activity.

laboratories having different locations (home, empty buildings, Collage laboratory and hotels) with different lighting and background conditions. 9,000 images are extracted for human activity (Criminal, Normal and Suspicious) identification from 25 subjects (10 male and 15 female). The dataset is allocated into training and testing sets. For training the models, 7,200 human suspicious activity images are used and 1,800 human suspicious activity images are used for testing. Figure 8 shows the three types like Criminal Activity, Normal Activity and Suspicious Activity used for the proposed human Suspicious Activity recognition experiments. Description of Real-time [17] dataset given in the Table 2.

Table 2. Real-time dataset.

Real-time dataset	Criminal	Normal	Suspicious	Total
No. of images for Training	2400	2400	2400	7200
No. of images for Testing	600	600	600	1800
Total	3000	3000	3000	9000

3.2 Performance of Human Suspicious Activity with 2D-CNN

In this section the training and testing process using 2D-CNN is discussed over the Kaggle dataset according to 81,012 and 14 layers. To work with 2D-CNN, 3-Dimensional input data is provided as input with patch size, filter size, number of filters, number of layers and number of epochs is fed as input to the model.

Table 3. 2D-CNN with 8, 10, 12, 14 layers.

2D-CNN					
	8 Layers	10 Layers		12 Layers	14 Layers
Input size	$250 \times 250 \times 3$		Input size	$250 \times 250 \times 3$	
Conv Layer1	250, 250, 64	250, 250, 32	Conv Layer1	250, 250, 64	250, 250, 32
Max_Pooling1	125, 125, 64	125, 125, 32	Conv Layer2	250, 250, 64	250, 250, 32
Conv Layer2	125,125,128	125, 125, 64	Max_Pooling1	125, 125, 64	125, 125, 32
Max_Pooling2	62,62,128	62, 62, 64	Conv Layer3	125, 125, 128	125, 125, 64
Conv Layer3	62,62,256	62, 62, 128	Conv Layer4	125, 125, 128	125, 125, 64
Max_Pooling3	31,31,256	31, 31, 128	Max_Pooling2	62, 62, 128	62, 62, 64
Conv Layer4	31,31,512	31, 31, 256	Conv Layer5	62, 62, 256	62, 62, 128
Max_Pooling4	15,15,512	15, 15, 256	Conv Layer6	62, 62, 256	62, 62, 128
Conv Layer5		15, 15, 512	Max_Pooling3	31, 31, 256	31, 31, 128
Max_pooling5		7, 7, 512	Conv Layer7	31, 31, 512	31, 31, 256
			Conv Layer8	31, 31, 512	31, 31, 256
			Max_Pooling4	15, 15, 512	15, 15, 256
			Conv Layer9		15, 15, 512
			Max_Pooling5		7, 7, 512
Flatten	115200	25088		115200	25088
Dense1	(500) 57600500	(500) 12544500		(500) 57600500	(500) 12544500

Dense2	(250) 125250	(250)125250		(250) 125250	(250) 125250
Dense2(SoftMax)	3 (753)	3 (753)		3 (753)	3 (753)
Trainable params	59, 277, 479	14, 239, 079		62, 411, 879	15, 022, 919

Network Structure of 14 Layers

2D-CNN with 14 layers consists of 9 Conv. layers and 5 max pooling layers. Conv. layer1 $(250 \times 250 \times 32)$ and Conv. Layer2 $(250 \times 250 \times 32)$ to Maxpooling1 $(125 \times 125 \times 32)$. Conv. Layer3 $(125 \times 125 \times 64)$ and Conv. Layer4 $(125 \times 125 \times 64)$ to Max pooling2 $(62 \times 62 \times 64)$. Conv. Layer5 $(62 \times 62 \times 128)$ and Conv. Layer6 $(62 \times 62 \times 128)$ to Max pooling3 $(31 \times 31 \times 128)$. Conv. Layer7 $(31 \times 31 \times 256)$ and Conv. Layer8 $(31 \times 31 \times 256)$ to Max pooling4 $(15 \times 15 \times 256)$. Conv. Layer 9 $(15 \times 15 \times 512)$ to Max pooling $5(7 \times 7 \times 512)$. Flatten (25088) parameters and Dense1(500), dense2(250), SoftMax (3) Outputs. The first, second, third, fourth, fifth, sixth, seventh, eighth and ninth convolutional layers are comprised of 512 feature kernel filters and size of the filter is 7×7. Last 3 layers Flatten (25088) units followed by a Dense layer 1(500), Dens2(250) and Dens3 Parameter 753 and 3output layer. Network parameters of 2D-CNN with 8 to 14 layers are given in the Table 4.

3.3 Performance of Human Suspicious Activity 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 $224 \times 224 \times 64$, conv. Layer2 is $224 \times 224 \times 64$, and Max_pooling1 is $112 \times 112 \times 64$. The following table shows the 16 layers structure and its dimensions. Network parameters are given in the Table 4.

Table 4. Network Parameters of VGG-16.

Layer	Patch Size	Input Size
Input Layer		$3 \times 224 \times 224$
convolutional x2	$3 \times 3/1$	$224 \times 224 \times 64$
Max_Pool	2×2	$112 \times 112 \times 64$
convolutional x 2	$3 \times 3/1$	$112 \times 112 \times 128$
Max_Pool	2×2	$56 \times 56 \times 128$
convolutional x 3	$3 \times 3/1$	$56 \times 56 \times 256$
Max_Pool	2×2	$28 \times 28 \times 256$
convolutional x 3	$3 \times 3/1$	$28 \times 28 \times 512$
Max_Pool	2×2	$14 \times 14 \times 512$
convolutional x 3	$3 \times 3/1$	$14 \times 14 \times 512$
Pool	2×2	$7 \times 7 \times 512$
Flattten	25088	25088
Fc1	256	256
Fc2	128	128
Softmax	Classifier	3

3.4 Performance of Human Suspicious Activity with ResNet-50

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. [15] The input dimension of Conv. Layer1 is $224 \times 224 \times 64$, conv. Layer2 is $224 \times 224 \times 64$, and Max_pooling1 is $55 \times 55 \times 64$. The Table 5 shows the 50 layers structure and its dimensions.

Table 5. Network Parameters of ResNet50.

Layers	50-Layers
Conv1	$7 \times 7,\ 64,\ \text{stride } 2$
	$3 \times 3 \times \max$ pool, stride 2
Conv2_x	$\begin{bmatrix} 1 \times 1,\ 64 \\ 3 \times 3,\ 64 \\ 1 \times 1,\ 256 \end{bmatrix} \times 3$
Conv3_x	$\begin{bmatrix} 1 \times 1,\ 128 \\ 3 \times 3,\ 128 \\ 1 \times 1,\ 512 \end{bmatrix} \times 4$
Conv4_x	$\begin{bmatrix} 1 \times 1,\ 256 \\ 3 \times 3,\ 256 \\ 1 \times 1,\ 1,\ 1024 \end{bmatrix} \times 6$
Conv5_x	$\begin{bmatrix} 1 \times 1,\ 512 \\ 3 \times 3,\ 512 \\ 1 \times 1,\ 10246 \end{bmatrix} \times 3$
	Average pool, 2048-d fc

3.5 Training and Validation

Description of Training and validation values for datasets using 2D-CNN, VGG16 and ResNet50 given in Table 6. Illustrated of Training and validation values for Real-time datasets using 2D-CNN, VGG16 and ResNet50 given in Table 7.

Table 6. Training and validation values for Kaggle datasets using 2D-CNN, VGG16 and ResNet50.

	Kaggle dataset					
model	Training Time	Tr loss	Tr accuracy (%)	Val loss	Val accuracy (%)	No of Epoch
2D-CNN 8L	3:01	0.2140	89.40	0.4165	80.40	25
2D-CNN 10L	3:25	0.2400	90.32	0.4570	82.83	25
2D-CNN 12L	6:47	0.4298	83.21	0.8630	67.00	25
2D-CNN 14L	4:01	0.2164	92.18	0.4076	85.83	25
VGG16	4:10	0.4232	85.41	0.6805	89.99	20
ResNet 50	4:08	0.0547	97.83	0.4200	93.00	20

The performance of human activity using 2D-CNN with 14 layers training validation accuracy 85.73% performance is superior in Kaggle datasets when compared to 2D-CNN (8, 10, and 12) layers. In Kaggle datasets, the ResNet50 training validation accuracy performance of 93.000% is superior to VGG16 and 2D-CNN.

Table 7. Training and validation values for Real-time datasets using 2D-CNN, VGG16 and ResNet50.

Real-time dataset						
model	Training	Tr loss	Tr	Val loss	Val	No of

	Time		accuracy		accuracy	Epoch
2D-CNN 8L	10:13	0.0220	0.6523	0.6523	0.9417	25
2D-CNN 10L	12:21	0.0091	0.9974	0.3181	0.9678	25
2D-CNN 12L	14:22	1.0989	0.3332	1.0986	0.3333	25
2D-CNN 14L	15:35	0.0072	0.9982	0.0629	0.9844	25
VGG16	6:30	0.0031	0.9959	0.0910	0.9784	20
ResNet 50	8:15	0.0258	0.9961	0.0021	0.9939	20

3.6 Performance Analysis

The computer configuration used in this experiment is Intel Core i5-6200U 11th gen CPU @ 2.40GHz speed, the operating system is Windows 10 and also uses 12GB memory for accelerate training. The software used for experiment is Python in Jupytor Notebook (Anaconda) tool. The platform used is the tensor flow 6.0 developed by Google which is mainly used for machine learning and artificial intelligence.

The training and testing are the important part of any application to analyse the performance of the trained models. Therefore 1680 testing samples are given to the trained model using 2D-CNN architecture. The results predicted from the proposed model are given as confusion matrix for the classification problem. It is the most efficient way to identify the True Positives (TP) [16], True Negatives (TN), False Positives (FP), False Negatives (FN) and Accuracy (ACC) of a classifier and it is used for classification problems [21] having binary or multi classes associated with the output. The true positives are the number of correctly classified samples as the labelled class. The False positives are the number of samples mistakenly classified as the labelled class. True negative is the samples correctly identified as labelled emotion in other classes. False negatives are the incorrectly classified emotion to other classes. The accuracy (ACC) is defined as the total number of corrected identified activities to the total number of samples.

The precision is calculated as the ratio of TP compared with TP along with FP. The recall is calculated as the ratio of TP with TP added with FN.

Table 8. Performance Metrix.

Performance Metrics	Accuracy	Precision	Re-call	F-score
Formulae	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{2*P*R}{P+R}$
TP, TN, FP, FN define as True Positive, True Negative and False Positive, False Negative.				

The precision and recall are combined to form the measure called F-score, which is the harmonic mean of P and R. Performance analysis for moving object detection using the proposed approach is calculated as shown in Table 8.

Table 9. Comparative performance of Recognition of Human Suspicious Activity with Kaggle dataset using 2D-CNN, VGG16 and ResNet50.

Model	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
2D-CNN (8 layers)	80.71	72.47	70.94	71.11
2D-CNN (10 layers)	88.55	84.95	82.83	83.13
2D-CNN (12 layers)	78.11	68.32	67.16	66.54
2D-CNN (14 layers)	90.88	86.72	86.33	86.36
VGG16	89.99	85.28	85.01	84.76
ResNet50	95.55	93.37	93.33	93.30

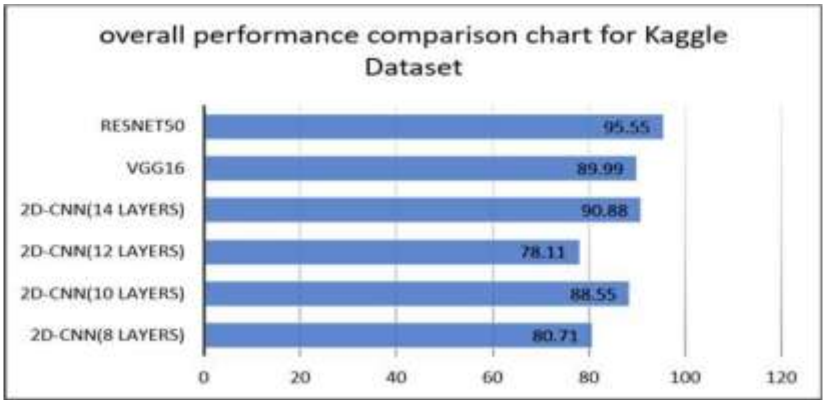


Figure 9. Overall performance accuracy chart for Kaggle Dataset using 2D-CNN, VGG16 and ResNet50.

The Two-Dimensional Convolutional Neural Network (2D-CNN) models are tested with Kaggle dataset and real-time video for varying numbers of Convolutional layers. The 2D-CNN model that gets trained with 14 convolutional layers performs well compared to 8 Layers, 10 Layers and 12 Layers convolution layers models and pre-trained Visual Geometry Group (VGG16) model. Comparative performance of human suspicious activity in Kaggle video given in Table 9. Comparative performance of human suspicious activity in real-time video given in Table 10.

Table 10. Comparative performance of human suspicious activity in Real-time video using 2D-CNN, VGG16 and ResNet50.

Real-time Video using Overall Performance compared with existing work					
	Model	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
Proposed work	2D-CNN (8 layers)	96.66	94.34	95.55	94.86
	2D-CNN (10 layers)	98.25	97.52	97.72	97.39
	2D-CNN (12 layers)	98.25	97.51	97.38	97.38
	2D-CNN	98.96	98.47	98.44	98.43

	(14 layers)				
	VGG16	97.84	96.94	96.61	96.77
	ResNet50	99.01	98.37	96.94	98.32
Existing work [syed and Marjan 13]	2D-CNN	95.69	-	-	-
Existing work [Wang Hao 22]	VGG16	75.6	-	-	-
Existing work [Sheldon 14]	ResNet50	97.33	-	-	-

The 14 layers convolutional model, when tested for every frame achieves better performance using Real-time videos compared with Kaggle dataset. ResNet50 is achieves better performance when compared with 2D-CNN and VGG16 using real-time video. Performance of Kaggle and Real-time video using 2D-CNN, VGG16 and ResNet50 shows in Figure 9 and Figure 10.

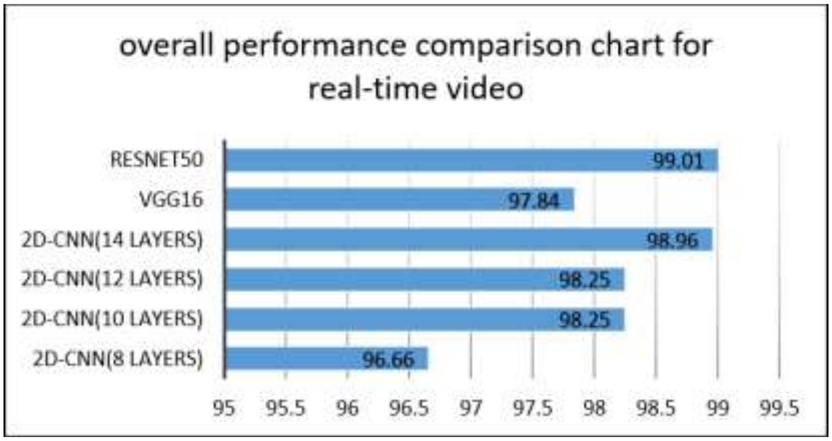


Figure 10. Overall performance accuracy chart with 2D-CNN, VGG16 and ResNet50 for Real-time video.

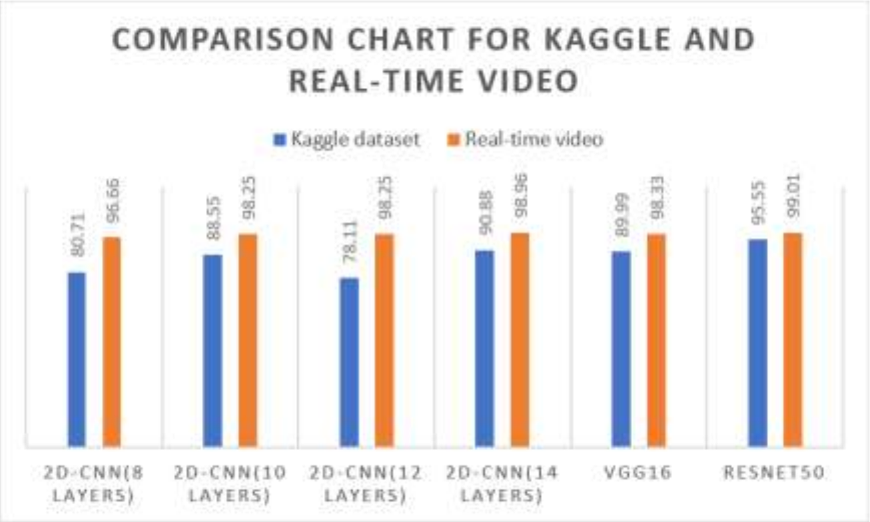


Figure 11. Overall performance comparison chart with 2D-CNN, VGG16 and ResNet50 for Kaggle dataset and Real-time datasets.

Performance of Recognition of human suspicious activity using 2D-CNN gives better accuracy rate of 90.88% when compared to VGG16 with Kaggle dataset. The performance of ResNet50 gives better accuracy 95.55 when compared to 2D-CNN and VGG16 with Kaggle dataset. Performance of Recognition of human suspicious activity using 2D-CNN gives better accuracy rate of 98.96% when compared to VGG16 with Real-time video. The performance of Recognition of human suspicious activity using ResNet50 gives better accuracy 99.01% when compared to 2D-CNN and VGG16 with Real-time video. Overall comparison chart for Kaggle dataset and Real-time datasets with 2D-CNN, VGG16 and ResNet50 shows in Figure 11.

4. Conclusion

Human Suspicious Activity Recognition is difficult task because of the inherent ambiguities of Activities during perception of the human behaviour. Violence, terrorism and other aggressive behaviours of human should be monitored through video surveillance camera and effective detection algorithms are in crucial need. Most of the violent behaviours are carried out using hand-held arms particularly guns (pistol, revolver), knife. In this system, deep learning approach is used to classify suspicious or normal activity in public environment. The proposed algorithm has been implemented using deep learning approach namely 2D-CNN, CNN-VGG16 and ResNet50 using real-time video Dataset as well as in Kaggle Dataset. The accuracy obtained for 2D-CNN in Kaggle dataset and real-time dataset are 90.88 and 98.96 respectively. Using VGG16, the accuracy obtained for the Kaggle Dataset and real-time Dataset are 89.99 and 97.84. In ResNet50 the accuracy abstained for Kaggle Dataset and real-time Dataset are 95.55 and 99.01. Based on the above comparison, ResNet50 using real-time video dataset processing performance is high when compared to VGG16 and 2D-CNN.

Acknowledgement

The authors are thankful to Annamalai University, Annamalainagar, for providing a good environment and well-equipped deep learning laboratory in the Department of Computer Science and Engineering to carry out research in an efficient manner.

5. References

[1] C. V. Amrutha, C. Jyotsna and J. Amudha, Deep Learning Approach for Suspicious Activity Detection from Surveillance Video, 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings, 335-339. https://doi.org/10.1109/ICIMIA48430.2020.9074920.

[2] Chaitanya Yeole, Hricha Singh, Hemal Waykole and Anagha Deshpande, Deep Neural Network Approaches for Video Based Human Activity Recognition, International Journal of Innovative Science and Research Technology 6(6) (2021).

[3] Chai C. Foong, Goh K. Meng and Lim L. Tze, Convolutional Neural Network based Rotten Fruit Detection using ResNet50, 12th Control and System Graduate Research Colloquium (ICSGRC) 2021 IEEE.

[4] Lu Xua, Weidu Yanga, Yueze Caob and Quanlong Li, Human, Activity Recognition Based on Random Forests, Institute of Electrical and Electronics Engineers, and IEEE Circuits and Systems Society, (n.d.), ICNC-FSKD 2017: 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery: Guilin, Guangxi, China, 29-31 July, 2017, 10.1109/FSKD.2017.8393329.

[5] Md Maruf Hossain Shuvo, Nafis Ahmed, Koundinya Nouduri and Kannappan Palaniappan, A hybrid approach for human activity recognition with support vector machine and 1D convolutional Neural Network, IEEE Applied Imagery Pattern Recognition Workshop, (AIPR). doi:10.1109/aipr50011.2020.9425332.

[6] Palaniappan Adithyan, R. Bhargavi and V. Vaidehi, Abnormal human activity recognition using SVM based approach IEEE 2012 International Conference on Recent Trends in Information Technology (ICRTIT) - Chennai, Tamil Nadu, India (2012.04.19-2012.04.21) International Conference on Recent Trends in Information Technology, 97-102. doi:10.1109/icrtit.2012.6206829.

[7] M. Patil Chandrashekar, B. Jagadeesh and M. N. Meghana, An Approach of Understanding Human Activity Recognition and Detection for Video Surveillance using HOG Descriptor and SVM Classifier, IEEE International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC) - Mysore (2017.9.8-2017.9.9), 481-485. doi:10.1109/CTCEEC.2017.8455046.

[8] Priyadarshini R K, Banu A, Bazila Nagamani T, Gradient Boosted Decision Tree based Classification for Recognizing Human Behavior, [IEEE 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE) -Sathyamangalam, Tamil Nadu, India (4.4-2019.4.6)], 1-4. doi:10.1109/icacce46606.2019.9080014.

[9] Rachana Gugale, Abhiruchi Shendkar, Arisha Chamadia, Swati Patra and Deepali Ahir, Human Suspicious Activity Detection using Deep Learning, International Research Journal of Engineering and Technology (IRJET), 2020, 07(06) 2020, e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[10] Rajesh Kumar, Anand Singh Jalal and Subhash Chand Agrawal, Suspicious human activity recognition: a review, Springer Science+Business Media Dordrecht 2017.

[11] A. Ramalingam, P. Aurchana, P. Dhanalakshmi, K. Vivekanandan and V. S. K. Venkatachalapathy, Analysis of Oral Squamous Cell Carcinoma into Various Stages using Pre-Trained Convolutional Neural Networks, IOP Conf. Series: Materials Science and Engineering, 2020, 012058 IOP Publishing DOI:10.1088/1757-899X/993/1/012058.

[12] Seyed Mohammad Taghi Almodarresi, Marjan Gholamrezaii, Human Activity Recognition Using 2D Convolutional Neural Networks, 27th Iranian Conference on Electrical Engineering (ICEE), IEEE-2019, 1682-1686. doi:10.1109/IranianCEE.2019.8786578

[13] Sheldon Mascarenhas and Mukul Agarwal, A Comparison between VGG16, VGG19 and ResNet50 Architecture framework for image classification, International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON) 2021(IEEE), 10.1109/CENTCON52345.2021.9687944.

[14] Shin Hoo-Chang, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, Transactions on Medical Imaging, IEEE 35(5) 2016. doi:10.1109/TMI.2016.2528162

[15] Showkat Ahmad Dar and S. Palanivel, Performance Evaluation of Convolutional Neural Networks (CNNs) and VGG on Real Time Face Recognition System, Advances in Science, Technology and Engineering Systems Journal 6(2) (2021), 956-964.

[16] Showkat A. Dar and S. Palanivel, Real Time Face Authentication System Using Stacked Deep Auto Encoder for Facial Reconstruction, International Journal of Thin Film Science and Technology, 2022, 11(1). https://digitalcommons.aaru.edu.jo/ijtfst/vol11/iss1/9/

[17] Srikanth Tammina, Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images, International Journal of Scientific and Research Publications (IJSRP), 9(10), October 2019, ISSN 2250-3153 (2019) pp. 9420. DOI-10.29322/IJSRP.9.10.

[18] J. Sujanaa and S. Palanivel, Real-time video based emotion recognition using convolutional neural network and transfer learning, Indian Journal of Science and Technology, (2020). https://doi.org/10.17485/IJST/v13i31.1118

[19] Tao Gu, Zhanqing Wu, Xianping Tao, Pung, Hung Keng and Jian Lu, An emerging patterns based approach to sequential, interleaved and concurrent activity recognition, IEEE International Conference on Pervasive Computing and Communications (PerCom) -Galveston, TX, USA (2009.03.9-2009.03.13), epSICAR:1-9. doi:10.1109/percom.2009.4912776.

[20] Tien Vo An, Hai Son Tran and Thai Hoang Le, Advertisement image classification using convolutional neural network, 9th International Conference on Knowledge and Systems Engineering (KSE), 2017, 197-202. doi:10.1109/KSE.2017.8119458.

[21] Wang Hao, Garbage recognition and classification system based on convolutional neural network VGG16, 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 2020, 252, 10.1109/AEMCSE50948.2020.00061

[22] A. Wiliem, V. Madasu, W. Boles and P. Yarlagadda, Adaptive unsupervised learning of human actions, IET 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009) - London, UK (3 Dec. 2009)] 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), P29-P29. doi:10.1049/ic.2009.0257.

[23] Xiuhua Hu, Yuan Chen, Xinyu Ma and Yingyu Liang, Research on Person Re-Identification Method Based on Fine-tune ResNet50 Network, 2020 International Conference on Computer Network, Electronic and Automation IEEE, 09 November 2020, (ICCNEA), 10.1109/ICCNEA50255.2020.00067.

[24] C. Yeole, H. Singh, H. Waykole and A. Deshpande, Deep Neural Network Approaches for Video Based Human Activity Recognition, In International Journal of Innovative Science and Research Technology 6(6) (2021).