I.J. Image, Graphics and Signal Processing, **2023, 1, 47-62**
Published Online on February 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2023.01.05

Real-Time Video based Human Suspicious Activity Recognition with Transfer Learning for Deep Learning

Indhumathi .J*
Research Scholar, Annamalai University Department of Computer Science and Engineering, Chidambaram, Annamalai Nagar - 608002, India
Email: indhumathi20061996@gmail.com
ORCID iD: https://orcid.org/0000-0001-6522-7896
*Corresponding Author

Balasubramanian .M
Associate professor, Annamalai University Department of Computer Science and Engineering, Chidambaram, Annamalai Nagar - 608002, India
Email: balu_june1@yahoo.co.in
ORCID iD: https://orcid.org/0000-0003-4251-5144

Balasaigayathri .B
PG Scholar, Annamalai University/ Department of Computer Science and Engineering, Chidambaram, Annamalai Nagar - 608002, India
Email: balasaigayathri6154@gmail.com
ORCID iD: https://orcid.org/0000-0003-3344-6052

Received: 11 May, 2022; Revised: 13 July, 2022; Accepted: 17 September, 2022; Published: 08 February, 2023

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. Deep learning approaches paved the way for the development of various machine learning and artificial intelligence. The proposed system detects real-time human activity using a convolutional neural network. The objective of the study is to develop a real-time application for Activity recognition using with and without transfer learning methods. The proposed system considers criminal, suspicious and normal categories of activities. Differentiate suspicious behaviour videos are collected from different peoples(men/women). This proposed system is used to detect suspicious activities of a person. The novel 2D-CNN, pre-trained VGG-16 and ResNet50 is trained on video frames of human activities such as normal and suspicious behaviour. Similarly, the transfer learning in VGG16 and ResNet50 is trained using human suspicious activity datasets. The results show that the novel 2D-CNN, VGG16, and ResNet50 without transfer learning achieve accuracy of 98.96%, 97.84%, and 99.03%, respectively. In Kaggle/real-time video, the proposed system employing 2D-CNN outperforms the pre-trained model VGG16. The trained model is used to classify the activity in the real-time captured video. The performance obtained on ResNet50 with transfer learning accuracy of 99.18% is higher than VGG16 transfer learning accuracy of 98.36%.

Index Terms: Convolutional neural network, human suspicious activity recognition, pre-trained models, transfer learning, real-time human activity recognition, VGG16 and ResNet50.

1. Introduction

The human criminal activity classification places a very indispensable role in inspecting human behaviours and physical activity more precisely. Behaviour, clothing, and weapons can be used to identify a person's activity whether it is normal or criminal activity.In the normal or safe activities through, humans can interact with society through emotions. The basic types of normal activities in a person are reading books, drinking, eating, walking, using mobiles phones, communicate with each other, writing, sleeping. Activities like eating, writing and sleeping are easy to find out

whereas other activities like fighting, theft, gun and knife attack, fully covered suspicious person and chain snatching are very hard to find in human through activities. There are varieties of applications for human activity[1] classification like public places, religious places, school, colleges, airport/railway suspicious person classification , real-time crime activity monitoring systems, etc. In this work, the activities like criminal, normal, suspicious are considered in the face, appearance, hand-held arm and weapon regions of persons to classify the activity efficiently. The main issues for constructing a human activity recognition[2] system using deep learning needs a larger amount of dataset, normal and criminal behaviour classification in varying lightening conditions and identifying the human activity in real-time under different scenarios. In this paper, we employ five CNN architectures: ResNet50, VGG16, 2D-CNN, Transfer Learning[3] using CNN (VGG16 and ResNet50). The primary goal of this paper is to detect and recognise suspicious activity using 2D-CNN, CNN-VGG16 [4], and ResNet50 models on human data. Several architecture models have been discovered in previous ImageNet competitions, which are useful in redefining new models for each problem when their layers are properly fine-tuned and frozen, resulting in improved accuracy. It will minimize the workload to develop a completely new architecture for each problem, thereby reducing the complex task by reusing those pre-trained models instead of training from scratch.

2. Literature Review

This paper[5] proposed a TL-HAR framework based on transfer learning techniques. It consists of three main phases, namely, pre-training, preprocessing, and recognition. In the pre-training, three models are trained on a generic dataset to adjust network weights. This pre-trained network is used to recognize human activities in a realistic dataset. In preprocessing[6], a certain number of frames are extracted from the whole video. The segmentation technique ensures fifixed computational cost with long-range temporal coverage. The extracted frames are used to feed spatial-streams in the proposed architectures. TV-L1 is used to generate of frames. Stacked of frames are used to feed the temporal-streams in the proposed architectures. Data augmentation techniques are applied to the training and validation of the model.

In this paper[7], we propose unsupervised transfer learning from source dataset to target dataset, which are completely different in terms of number of users and samples. They present (i) base model (ii) Maximum Mean Discrepancy (MMD) and (iii) transfer learning model 855 for human activity recognition. The MMD is used in the transfer learning model to measure the dissimilarity between source domain features and target domain features.Different users can perform the same activity in different ways. Therefore, the sensor value pattern for same activity in different dataset varies. It results in feature distribution between source and target datasets varies a lot that leads to poor recognition performance for the transfer learning model.

Deep learning[8] approach is used to detect suspicious or normal activity in an academic environment, and which sends an alert message to the corresponding authority, in case of predicting a suspicious activity. Monitoring is often performed through consecutive frames which are extracted from the video. The proposed model has the benefit of stopping the crime before it happens. The real time CCTV footages are being tracked and analyzed. The analysis result is a command to the appropriate authority to take action if the result indicates that an untoward incident is about to occur and thus this can be avoided. Even though the proposed system is limited to academic area, this can also be used to predict more suspicious behaviors at public or private places. The model can be improved by identifying the suspicious individual from the suspicious activity.

In this paper[9], a training strategy for classifying Dermatofibroma (DF), Keratosis-like Lesions (BKL), and Basal Cell Carcinoma (BCC) types of cancer was explained, demonstrated, and evaluated. VGG19-based CNN and TL proved to be powerful tools to aid skin cancer diagnosis with high accuracy. The overall accuracy and loss of the network indicate satisfactory outcome that can be improved further.

[10] Our results demonstrated that we can get greater than 0.9 (F1) accuracy using both our CNN pipeline as well the MobileNet model with transfer learning on over 9 activities. MobileNet proved to provide higher accuracy than the much larger and more complex Inception V3 model DNN pipeline and the MobileNet based model has the potential for generalization across multiple users with a sufficient sample size across subjects with multiple attributes in the training set.

3. Proposed Methodology

This work classifies the criminal activities of the person in real-time transfer learning of pre-trained models. Human criminal, suspicious and normal activity classification, training, validation, and real-time[11] testing are the main steps in the system

3.1 Description of Kaggle Dataset and Real-time Datasets

A. Narration of Kaggle dataset

This kaggle dataset and google open source images contains 2160 jpg images in three categories of human activities images (criminal, normal, and suspicious). Each category has 520 images for training and 200 images for

testing. The dataset images vary in size of 950x540 pixels and after re-sizing the image size are 250x250. Figure 1 depicts the resizing process on a sample image.

(a)

(b)

Fig. 1. Kaggle dataset (a) original image(suspicious,normal and Criminal activities) (b) resized image

B. Narration of Real-time Dataset

The dataset contains 3 categories of human activities RGB images (Criminal,Normal and Suspicious) with a total of 9000 images/videos in the jpg and Mp4 format. Each category has 2400 images for training and 600 images for testing. While resizing, the datasets are 250x250 in size. Real-time original and resized images are shown in Fig.2.

(a)

(b)

Fig. 2. (a) original images (b) resized images

3.2 Human activity Recognition:

The traditional neural network is made up of three layers: the input layer, the hidden layer, and the output layer. The convolutional neural network (CNN), which is based on the traditional neural network, functions as a feature extractor by combining the convolution layer and the sub-sampling layer. The combination of the artificial neural network and the back propagation algorithm simplifies the model's complexity and reduces the parameters.CNN works by extracting features[12] from images and eliminates the need for manual feature extraction. The features are automatically learned as the network trains on a set of images. As a result, CNN models are extremely accurate for image processing. CNN detects features by utilising tens or hundreds of hidden layers.

Pre-trained models such as Visual Geometry Group (VGG) have been made publicly available ConvNet models for all computer vision problems used in image and video recognition[13] tasks, with VGG16 and ResNet50 being used in this work to determine activities. Block diagram of recognition of human suspicious activity using 2D-CNN, VGG16 and ResNet50 shown in Fig.3.

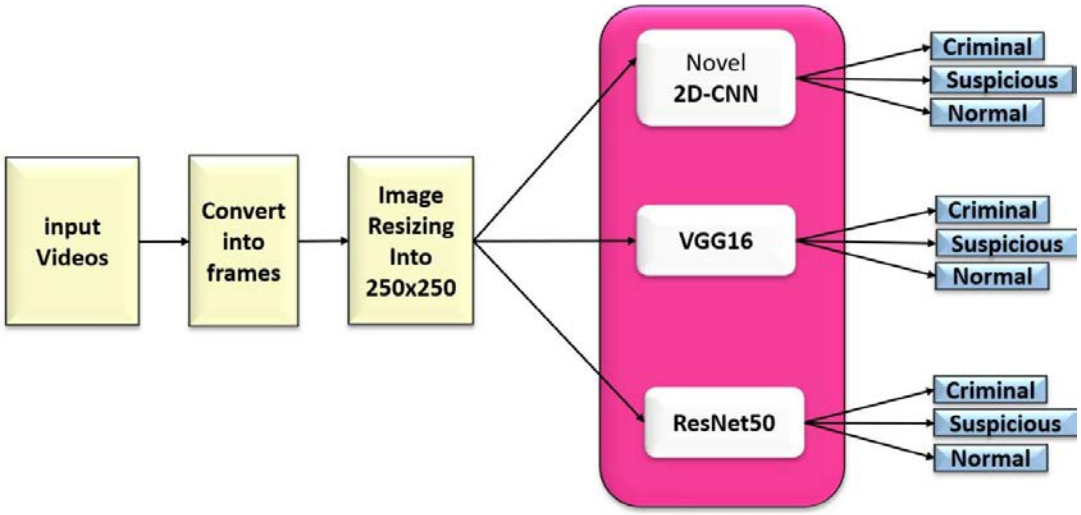

Fig. 3. Block diagram of recognition of human suspicious activity using 2D-CNN, VGG16 and ResNet50

A. 2D-CNN Architecture

The standard CNN is 2D-CNN, which can classify two-dimensional inputs like videos and images. The size of the input layer for the 2D CNN is 250x250. This CNN is made up of various convolution and sub-sampling layers, which are then followed by several multi-layer perceptron (MLP) layers. The MLP layers use the output vectors from the convolution layers to characterize the input signal. The architecture of a 2D CNN is determined by the convolution layer sizes, MLP layer amounts, kernel size, and sub-sampling factors. Back propagation is used to optimize the parameters of the 2D filter kernels in CNN. Convolution is a type of linear operation used to extract features. Convolution layer generates a feature map from input data by using a filter or kernel. Kernel is a small array of numbers in this layer that is applied across the input, which is also a small array of numbers known as a tensor. The maximum or highest value in each patch of each feature map is calculated by max pooling. The ReLU Activation Function is in-charge of converting the node's weighted summed input into the activation or output for that input. The architecture of 2D-CNN is shown in Fig.4.

To comprehend the operation of ReLU,outputs the input directly. otherwise, it outputs zero.

The rectified linear activation function is characterised as:

$$Y=\max(x,0) \tag{1}$$

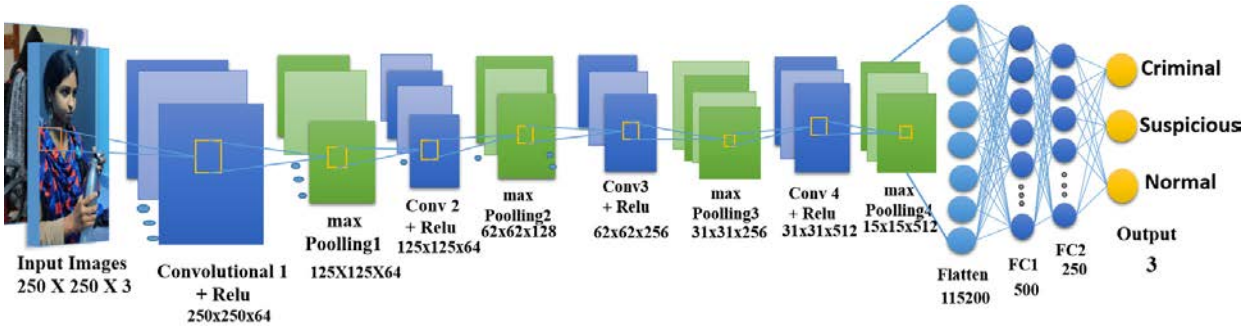

Fig. 4. Architecture of 2D-CNN

Pooling Layer

The size of a convolution process data image will be decreasing at this pooling layer. The pooling layer is made up of filters with a particular size and stride, which are then moved over the feature map area. The pooling method most commonly used for Convolutional Neural Network designs is max pooling[14]. The convolution output layer is divided into numerous grids by max pooling, and each filter shift will select the most significant value from each grid. In order to reduce the dimensions of the data and the number of parameters in the following stage, a tiny percentage of the image size generated by this method is useful. A feature map is the pooling layer's output. Process pooling layer in Fig.5.

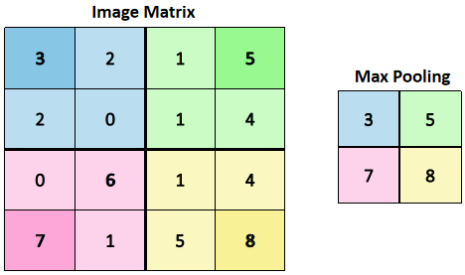

Fig. 5. Pooling Layer

Softmax Function

The softmax function converts a vector of K real values to a vector of K real values that sum to 1. The softmax transforms input values that are positive, negative, zero, or greater than one into values between 0 and 1, allowing them to be interpreted as probabilities.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\Sigma_{j=1}^{K} e^{z_j}} \quad (2)$$

Where

$\vec{z}$ = input vector
z_i = elements of the input vector
e^{z_i} = exponential function (standard)
K = multi-class classifier's number of classes
$\Sigma_{j=1}^{K} e^{z_j}$ = normalisation term

B. Architecture of VGG16

Conv.Layer1 and 2 (224x224x64) to Max_pooling1(112x112,64). Conv.Layer3 and 4 (112x112x128) to Max_pooling2 (56x56x128). Conv.Layer5,6 and 7 (56x56x256) to Max_pooling3 (28x28x256). Conv.Layer 8,9 and 10 (28x28x512) to Max_pooling4(14x14x512) 13 Conv.Layer11,12 and 13 (14x14x512),Max_pooling(7x7x512). Flatten(25088), Dense1,2 and 3,softmask output ,7526 parameters. The VGG16 architecture diagram is shown in Fig.6.

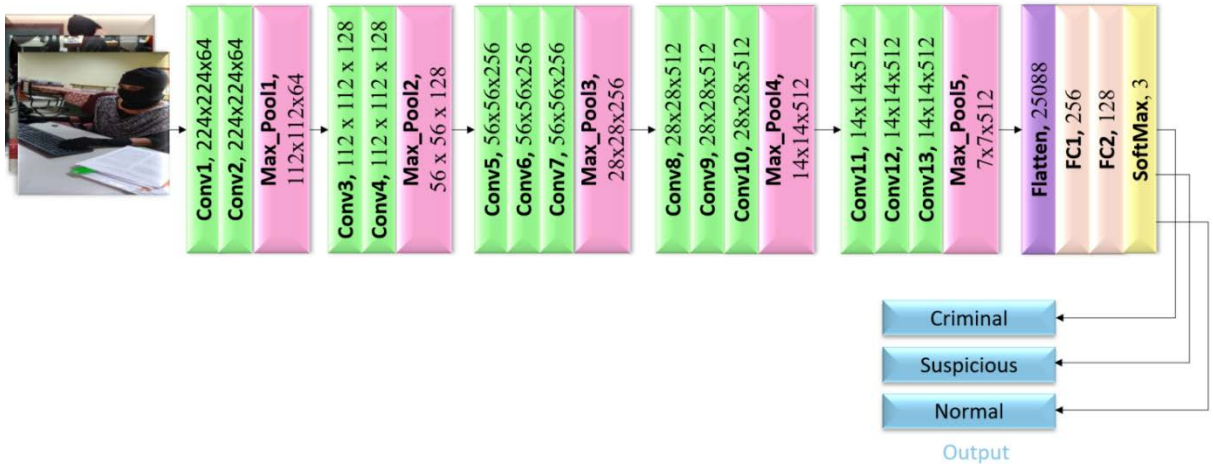

Fig. 6. Architecture of VGG16

C. ResNet50 Architecture

The ResNet 50 architecture contains convolutional layers, pooling layers, batch normalization, average pool ,fully connected layer and Softmax function. The model is trained with versatile number of parameters, and their weights are adjusted over 50 epochs. A block diagram of the ResNet model's architecture is shown Fig.7.

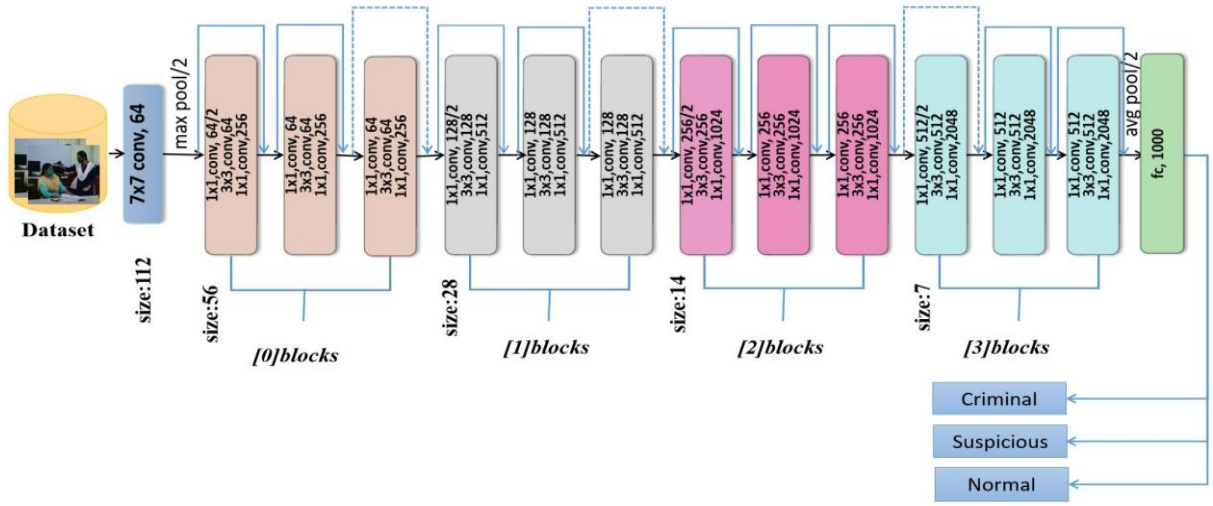

Fig. 7. Architecture of ResNet50

D. Transfer Learning

The pre-trained models in CNN architectures have learned the weights on a larger dataset after being trained with a subset of the ImageNet dataset during ImageNet Competitions. As a result, the weights of these models can be used to learn the proposed human activity recognition problem. Transfer learning occurs when knowledge from one domain can be applied to another. Transfer learning can be accomplished through two methods: fine-tuning and freezing[15]. The pre-trained model layers are fine-tuned with varying filters, layers, and hidden units to optimise their learning in current problems, thereby increasing the accuracy in learning the newly defined problem.The pre-trained model layer weights are frozen (locked) in the case of freezing, preventing them from being changed during the current training. The pre-trained models alleviate the burden of creating CNN architecture models from scratch. Convolution[16] block, max-pooling layer, rectified linear activation unit (ReLU), batch normalisation layer, separable convolution layer and so on comprise the pre-trained convolution base. These previously trained ImageNet blocks are frozen to prevent their weights from being trained for the proposed activity recognition task.

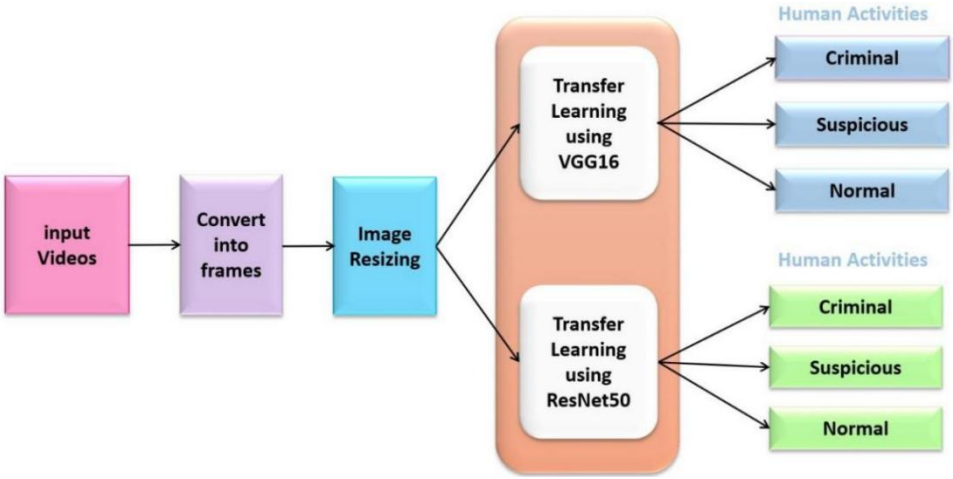

Fig. 8. Block diagram for the recognition of human Activities using Transfer Learning

The block diagram for the recognition of human Activities using Transfer Learning is shown in the Fig.8.Transfer Learning Architecture is shown in Fig.9.

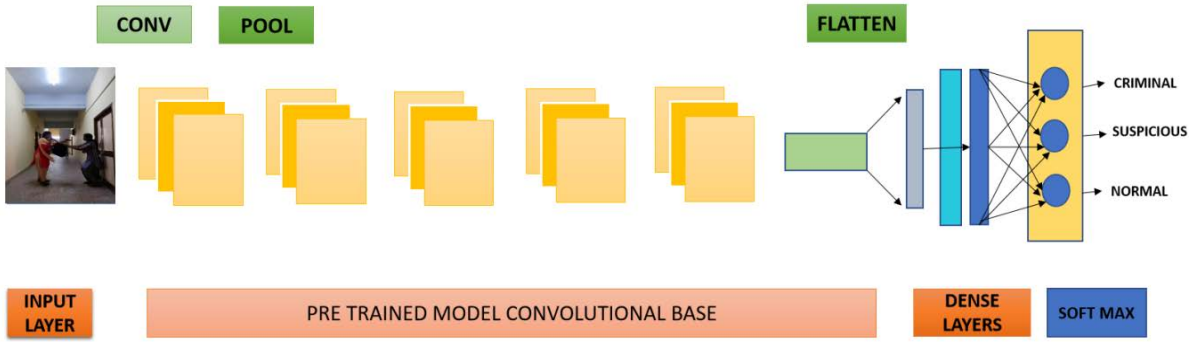

Fig. 9. Architecture of Transfer Learning

4. Experimental Results

A. Kaggle Dataset

The datasets are gathered from Google Images, movie datasets, shutter stock Images, etc. The dataset includes photographs of people involved in crimes such as robbery, fighting, chain snatching, and terrorism (gun attack). Normal activities are listening music, reading book, singing and eating.

(a)

(b)

(c)

Fig. 10. Datasets: (a) criminal activity (b) suspicious activity (c) normal 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. Fig.10. shows the samples of the dataset collected for normal, suspicious and criminal activity.

B. 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 46 years in 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.

(a) (b) (c)

Fig. 11. Real-time Datasets: (a) criminal activity (b) normal activity (c) suspicious activity

For training the models, 7,200 human suspicious activity images are used and 1,800 human suspicious activity images are used for testing. Fig.11. 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 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

4.1 Human Suspicious Activity Performance Using 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.

A. Network Structure of 14 Layers

2D-CNN with 14 layers consists of 9 Conv. layers and 5 max pooling layers. Conv. layer1 & 2 (250x250x32) to Maxpooling1 (125x125x32). Conv.Layer3 & 4 (125x125x64) to Max pooling2 (62x62x64). Conv. Layer5 &6 (62x62x128) to Max pooling3 (31x31x128). Conv. Layer7 & 8 (31x31x256) to Max pooling4 (15x15x256). Conv. Layer9 (15x15x512) to Max pooling5(7x7x512). Flatten(25088) parameters and Dense1(500) ,dense2(250), SoftMax (3)Outputs. Network parameters of 2D-CNN with 8 to 14 layers are given in the Table 3.

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

2D-CNN					
	8 Layers	10 Layers		12 Layers	14 Layers
Input size	250x250x3		Input size	250x250x3	
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
Dense 1	(500) 57600500	(500) 12544500		(500) 57600500	(500) 12544500
Dense 2	(250) 125250	(250)125250		(250) 125250	(250) 125250
Dense 3(SoftMax)	3 (753)	3 (753)		3 (753)	3 (753)
Trainable params	59,277,479	14,239,079		62,411,879	15,022,919

4.2 Human Suspicious Activity Performance Using VGG-16

Table 4. Network Parameters of VGG-16

Layer(type)	**Output shape**	**Parameter**
Conv2d_1	224x224x64	1792
Conv2d_2	224x224x64	36928
Max_Pooling2d_3	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
FC1(Dense)	256	6422784
Fc2(Dense)	128	32896
Softmax	3	387
Trainable param		21,170,755

4.3 Human Suspicious Activity Performance Using ResNet50

Table 5. Network Parameters of ResNet50

Layers	**50-Layers**
Conv1	7x7,64, stride 2
	3x3x 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

4.4 Training and validation

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

Table 6. Training and validation values for Kaggle and Real-time 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	6:39	0.4232	85.41	0.6805	89.99	20
ResNet50	7:08	0.0547	97.83	0.4200	93.00	20
Real-time dataset						
model	Training time	Tr loss	Tr accuracy	Val loss	Val accuracy	No of Epoch
2D-CNN 8L	10:13	0.0220	65.23	0.6523	94.17	25
2D-CNN 10L	12:21	0.0091	99.74	0.3181	96.78	25
2D-CNN 12L	14:22	1.0989	33.32	1.0986	33.33	25
2D-CNN 14L	15:35	0.0072	99.82	0.0629	98.44	25
VGG16	12:30	0.0031	99.59	0.0910	97.84	20
ResNet50	22:15	0.0325	99.46	0.0566	98.94	20

4.5 Human Suspicious Activity Performance Using Transfer Learning (VGG16 & ResNet50)

Various number of layers are frozen and trained on free cloud-based service, Google Colab network GPU's(23) for the pre-trained models. All models are trained with a total of 20 epochs. The training time and accuracy for validation data for different models are shown in Table 7, where model 1,2,3 and model 4 are obtained from VGG16 using transfer learning. In VGG16 model1 16 layers are frozen continued by model 2,3, and 4 with 14 layers, 12 layers and 10 layers respectively. High validation accuracy was obtained on VGG16 model 3 using Transfer learning with accuracy 97.55 during training and validation data.

Table 7. Number of frozen layer with training data for validation accuracy and training time using transfer learning in VGG16 model.

VGG16 using transfer learning	No.of frozen layer	Training time (in Time)	Validation accuracy (%)
Model 1	16	00:20:00	86.68
Model 2	14	01:25:00	95.59
Model 3	12	01:55:00	97.55
Model 4	10	02:45:00	96.00

In the ResNet50 using transfer learning model, having layers of Conv2D, ReLu activation function with number of MaxPooling[17], Batch normalization, ResNet_Conv2D,Flatten with two dense layer model and Softmax Activation function, the training time and Validation accuracy saturated at 10 epochs. In ResNet50 model5 50layers are frozen continued by model 6,7,8 and 9 with 45 layers, 40 layers, 35 layers and 30 layers respectively. High validation accuracy

was obtained on ResNet50 model 7 using Transfer learning with accuracy 98.99 during training and validation data. Table 8, shows value of model 5,6,7,8 and model 9 that are obtained from ResNet50 using transfer learning.

Table 8. Number of frozen layer with Accuracy for training and training time data using transfer learning in ResNet50 model.

ResNet50 using Transfer Learning	No. of Frozen Layer	Training Time (in Times)	Accuracy (%)
Model 5	50	00:30:01	88.76
Model 6	45	03:25:00	92.32
Model 7	40	04:50:00	98.99
Model 8	35	06:15:00	95.51
Model 9	30	07:06:00	90.93

In transfer learning[18] using VGG16 model 3, 12 layers are frozen and dense layer hidden neurons are changed from 25088 to 256 neurons in the first dense layer and 128 neurons in the second dense layer and given to 3 output classes with a dropout of 40% and it has 21,170,755 trainable parameters. In transfer learning using VGG16 model 7, 40 layers are frozen and dense layer hidden neurons are changed from 32768 to 256 neurons in the first dense layer and 128 neurons in the second dense layer and given to 3 output classes with a dropout of 40% and it has 31,956,739 trainable parameters. To retrieve the VGG-16 model trained on the imagenet dataset, we must include weights = 'imagenet'. Include top should be set to False to avoid downloading the pretrained model's fully connected layers. We must add our own classifier because the pretrained model classifier has three classes and our goal is to classify the image into three classes (suspicious, criminal and normal). The fully connected layer classifies low level image features[19] such as edges, lines, and blobs after the pretrained model's convolutional layers extract them.

4.6 Performance analysis

The computer configuration used in this experiment includes a Windows 11 operating system, an Intel Core i5-6200U 11th generation CPU running at a speed of 2.40GHz, and 12GB of memory for accelerated training. Python is the programme used for the experiment, which is a tool called Jupytor Notebook (Anaconda). Any application's training and testing phases are crucial for analysing how well trained models perform. As a result, the trained model built using the 2D-CNN architecture is given 1680 testing samples. The proposed model's predictions for the classification[20] issue are provided as a confusion matrix.

This experiment comprises the use of real-time datasets for performance evaluation of transfer learning using VGG16 and ResNet50 architecture for image classification task. In real-time dataset, I have used for training 7200 image and testing 1800 images. The overall process of human activity classification is done. After applying transfer learning using VGG16 and ResNet50, the following Table 6 shows the Performance Metrics[21] With the values of true positive and false positive values are calculated. In the confusion matrix, true positives are the total number of criminal activities correctly identified as belonging to the respective classes, while true negatives are the number of criminal activities correctly identified as belonging to other classes.

The accuracy calculation (AC) compares the system's efficiency. It considers the classifier's total number of correct predictions. The following equation is used to compute it:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Precision is the number of positive cases correctly predicted by the classifier. It is calculated using the following equation:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

Precision is the number of positive cases correctly predicted by the classifier. It is calculated using the following equation:

$$\text{Re-call} = \frac{TP}{TP+FN} \tag{5}$$

The F1 score, also known as the F measure, is a measure of the accuracy of the test. It is defined as a weighted mean of precision and recall. Its best value is 1 and its worst value is 0.

$$\text{F1-score} = \frac{2*P*R}{P+R} \tag{6}$$

Where,

True Positive (TP) : Test Samples correctly identified
False Positive (FP) : Test Samples incorrectly identified
True Negative (TN) : Test Samples correctly rejected
False Negative (FN) : Test Samples incorrectly rejected

The performance of the classifier can be evaluated using measures like precision, recall, F-score, and accuracy. The performance of human suspicious activity recognition for kaggle dataset is carried out where the 14 layers gives maximum accuracy of 90.88% using 2D-CNNwhen compared to the Vgg16 accuracy of 89.99% .The precision, recall, F-score, and accuracy for Kaggle and real-time dataset testing images are shown in Table 9 and Table 10.

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

Kaggle Dataset using Overall Performance				
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

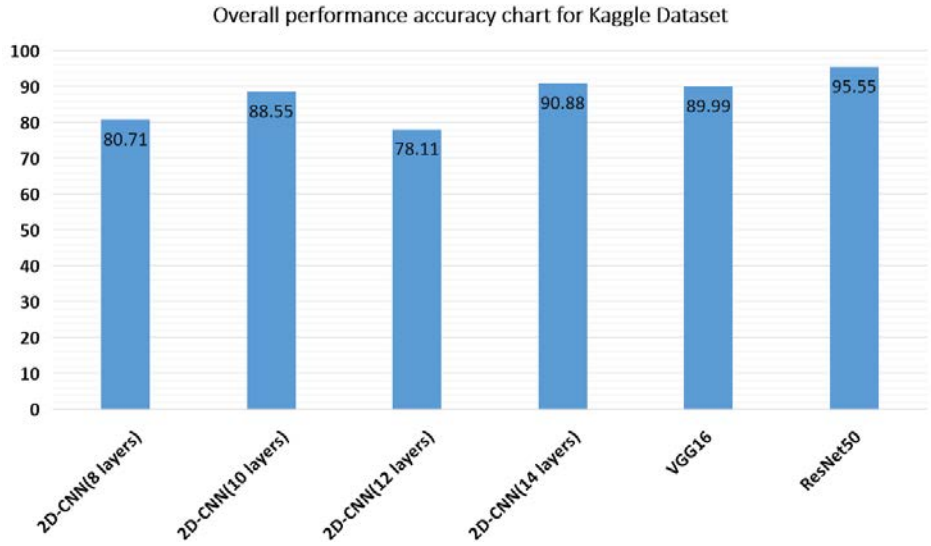

Fig. 12. Overall performance accuracy chart for Kaggle Dataset using 2D-CNN, VGG16 and ResNet50

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 Without transfer learning	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(14 layers)	98.96	98.47	98.44	98.43
	VGG16	97.84	96.94	96.61	96.77
	ResNet50	99.03	98.55	98.55	98.54
Existing work (Kamal Kant Verma [22])	2D-CNN	97.88	-	-	-

The Kaggle dataset and real-time video are used to test the 2D-CNN models with different numbers of convolutional layers. 2D-CNN with the 14 layers providing maximum accuracy of 98.96% when compared to the 8,10,12 and pre-trained VGG16 accuracy of 96.66%, 98.25% ,98.25 % and 97.84% respectively. Tables 10 and 11 compare real-time video and human suspicious activity performance in Kaggle. When tested for each frame, the 14-layer 2D-CNN model performs better using real-time videos than the Kaggle dataset. Using real-time video, ResNet50 outperforms 2D-CNN and VGG16 in terms of performance. Figures 10 and 11 depict the performance of Kaggle and Real-time video using 2D-CNN, VGG16, and ResNet50.

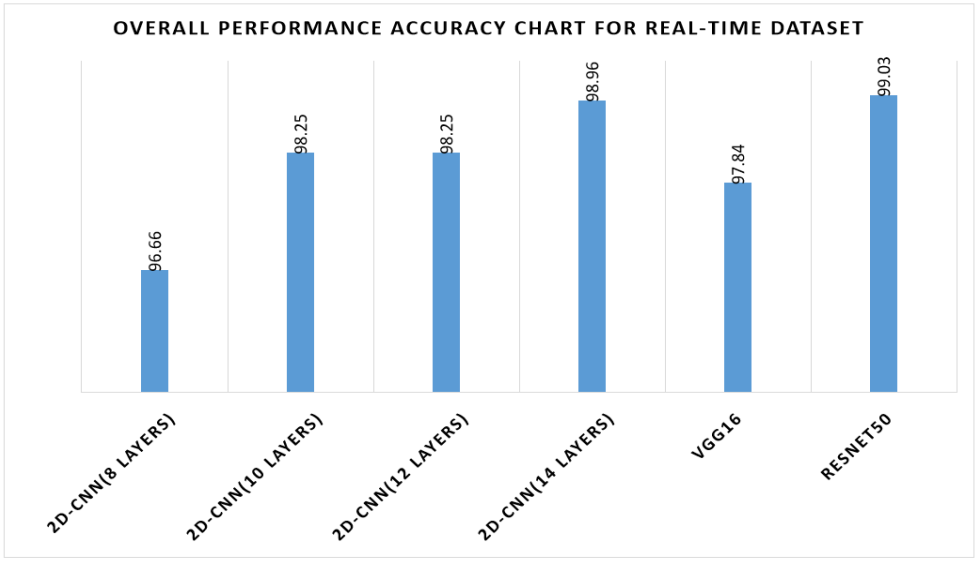

Fig. 13. Overall performance of 2D-CNN, VGG16 and ResNet50 for Real-time video

When compared to VGG16 with the Kaggle dataset, performance of human suspicious activity recognition using 2D-CNN has a better accuracy rate of 90.88%. When compared to 2D-CNN and VGG16 with the Kaggle dataset, ResNet50's performance for recognizing human suspicious activity has a better accuracy of 95.55%. When compared to VGG16 with real-time video, the performance of 2D-CNN for the recognition of human suspicious activity has a better accuracy rate of 98.96%. In comparison to 2D-CNN and VGG16 with Real-time video, ResNet50's Performance of Recognition of Human Suspicious Activity achieves better accuracy of 99.03%. As shown in Fig.13, a general comparison chart for the Real-time datasets with 2D-CNN, VGG16, and ResNet50 is available.

Table 11. Overall Performance of Human Suspicious Activity using VGG-16 With Transfer Learning

Real-time Video using Transfer Learning					
	Model	Accuracy	Precision	Recall	F-Score
Proposed work With transfer learning	Model 3 in VGG16 with transfer learning	98.36	97.55	97.55	97.55
	Model 7 ResNet50 with transfer learning	99.18	99.03	99.02	99.01
Existing work (Apri Junaidi [23])	VGG16 with transfer learning	90.00	-	-	-

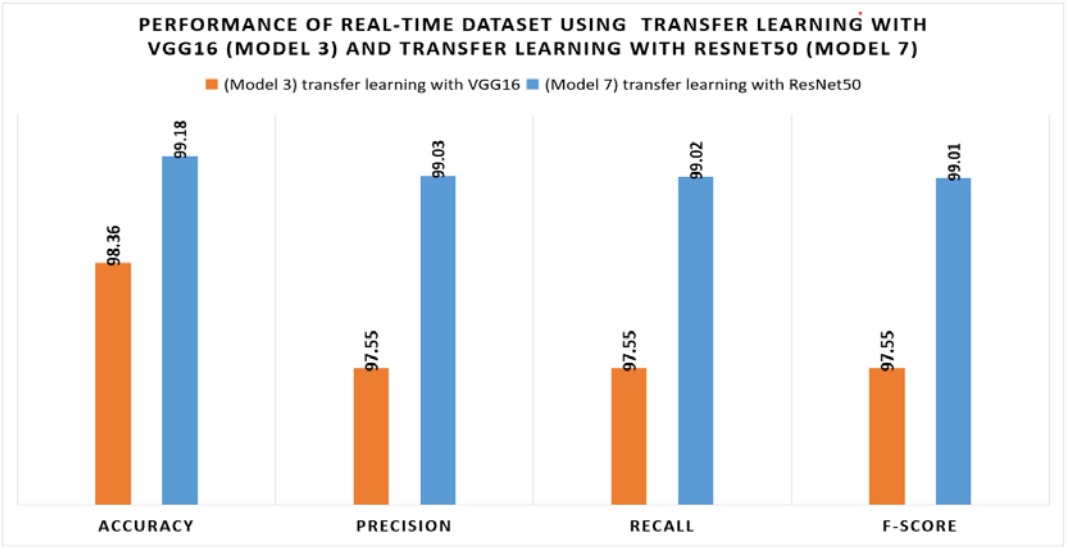

Fig. 14. Overall performance accuracy chart for real-time video using transfer learning

The performance of human suspicious activity recognition for real-time video is carried out where the Model 3 gives maximum accuracy of 98.36% using the transfer learning in VGG16 with 12 frozen layers when compared to the without using transfer learning in VGG16 accuracy of 97.84%. The performance of human suspicious activity recognition for real-time video is carried out where the Model 7 gives maximum accuracy of 99.18% using the transfer learning in ResNet50 method with 12 frozen layers when compared to the without using transfer learning in ResNet50 accuracy of 99.03%. Table 11. Fig.14. shows the overall performance accuracy chart for real-time video using transfer learning. shows the overall performance of suspicious human activity using VGG16 and ResNet-50 with transfer learning. Overall Performance recognition of human activities(suspicious, criminal and normal) using 2D-CNN, VGG16 and ResNet50 and transfer learning (with VGG16 & ResNet50) shown in Fig.15.

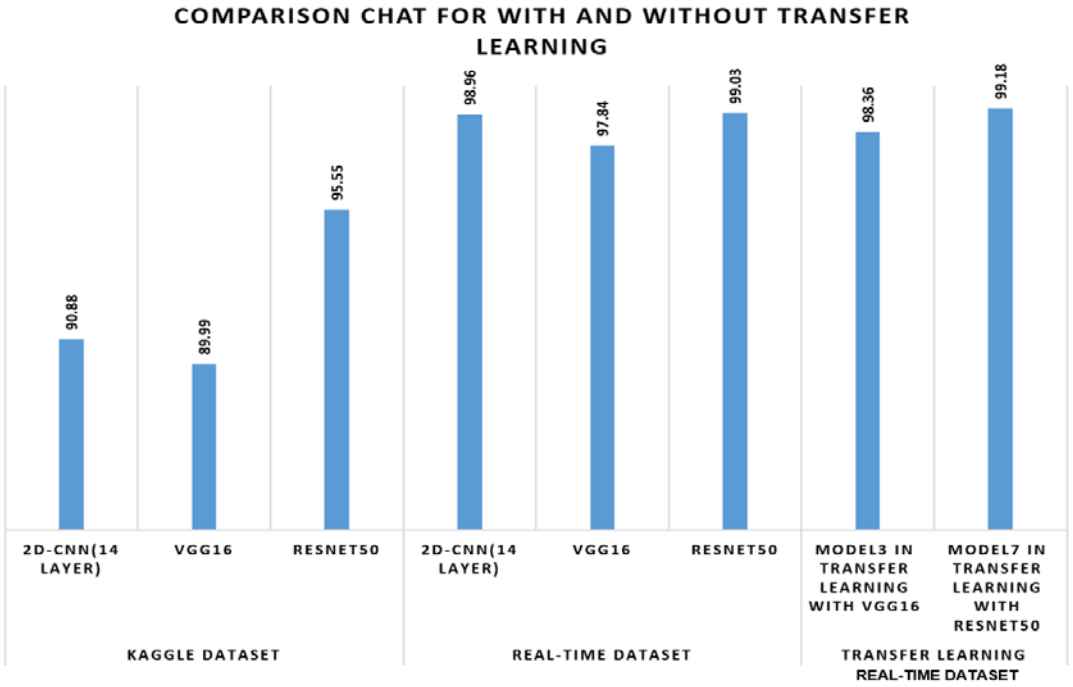

Fig. 15. Overall Performance recognition of human suspicious activity using 2D-CNN, VGG16 and ResNet50 and transfer learning using kaggle and real-time dataset

5. Conclusion

The majority of violent behaviours involve hand-held weapons, particularly a gun, knife, chain snatching, robbery, and everyday human activities, which are all included in the criminal, suspicious, and everyday normal activity datasets. To classify suspicious or common activity in a public place, this system uses deep learning. On real-time video datasets and Kaggle datasets, the proposed algorithm has been applied using CNN-VGG16, ResNet50 and 2D-CNN deep learning techniques. The criminal activity accuracy is significantly higher than the accuracy in the other two cases. This is due to more complex dense layers, higher dropout rates after convolutional layers, but perhaps most importantly, batch normalization's introduction. Dropouts and batch normalisation work together to significantly reduce overfitting and improve model performance. The accuracy for the 2D-CNN (14layer) is 90.88% in the Kaggle dataset and 98.96% in the real-time dataset, respectively. The accuracy for the Kaggle Dataset and real-time Dataset using pre-trained VGG16 is 89.99% and 97.84%, respectively. Kaggle Dataset and Real-Time Dataset accuracy abstained in ResNet50 are 95.55% and 99.03%, respectively. Based on the comparisons made above, ResNet50 outperforms VGG16 and 2D-CNN in terms of processing real-time video datasets. In transfer learning methods, 10 epochs are used and it gives an accuracy of 98.36% in VGG-16 with transfer learning and 99.03% in ResNet-50 with transfer learning.

References

[1] Allah Bux Sargano et..al., 2017 International Joint Conference on Neural Networks (IJCNN),“Human Action Recognition using Transfer Learning with Deep Representations”, IEEE,2017.

[2] Amrutha, C.V; Jyotsna, C. Amudha, J. (2020). [IEEE 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) - Bangalore, India] 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) - Deep Learning Approach for Suspicious Activity Detection from Surveillance Video, 335–339.

[3] Al Amin Biswas;Md. Mahbubur Rahman;Aditya Rajbongshi;Anup Majumder; (2021). Recognition of Local Birds using Different CNN Architectures with Transfer Learning. International Conference on Computer Communication and Informatics (ICCCI), 2021.

[4] Mutegeki, Ronald; Han, Dong Seog (2019). IEEE International Conference on Information and Communication Technology Convergence (ICTC) Feature-Representation Transfer Learning for Human Activity Recognition, 18–20,2019.
[5] Yousry Abdulazeem;Hossam Magdy Balaha;Waleed M. Bahgat;Mahmoud Badawy, IEEE Access, Human Action Recognition Based on Transfer Learning Approach,2021.
[6] Phan, Ha Tran Hong; Kumar, Ashnil; Kim, Jinman; Feng, Dagan. IEEE 2016 13th International Symposium on Biomedical Imaging (ISBI 2016) - Prague, Czech Republic" Transfer learning of a convolutional neural network for HEp-2 cell image classification, 1208–1211.
[7] Islam, Md Shafiqul; Okita, Tsuyoshi; Inoue, Sozo, 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech) - Fukuoka, Japan- Evaluation of Transfer Learning for Human Activity Recognition Among Different Datasets.854–859,2019.
[8] Yang Liu, Peng Sun, Max R. Highsmith, Nickolas M. Wergeles, Joel Sartwell, Andy Raedeke, et.al.,"Third International Conference on Data Science in Cyberspace", IEEE, Performance Comparison of Deep Learning Techniques for Recognizing Birds in Aerial Images,2018.
[9] Nour Abuared;Alavikunhu Panthakkan, Mina Al-Saad, Saad Ali Amin, Wathiq Mansoor. Skin Cancer Classification Model Based on VGG 19 and Transfer Learning. 2020 3rd International Conference on Signal Processing and Information Security (ICSPIS), 2020.
[10] Nutter, Mark; Crawford, Catherine H.; Ortiz, Jorge, "IEEE 2018 International Joint Conference on Neural Networks (IJCNN) - Rio de Janeiro", Brazil Design of Novel Deep Learning Models for Real- time Human Activity Recognition with Mobile Phones,1–8,2018.
[11] Miss. Sayali V,Mr.R.G.Mevekar,i "International Research Journal of Engineering and Technology (IRJET)",Developments of Deep Learning for Animal Classification: A Review, 2021,e-ISSN: 2395-0056.
[12] Priya Gupta, Nidhi Saxena, Meetika Sharma, Jagriti Tripathi, "Deep Neural Network for Human Face Recognition", International Journal of Engineering and Manufacturing,Vol.8, No.1, PP.63-71, 2018.
[13] Ahmad Ilham Gustisyaf, Ardiles Sinaga, "Implementation of Convolutional Neural Network to Classification Gender based on Fingerprint", International Journal of Modern Education and Computer Science, Vol.13, No.4, PP.55-67, 2021.
[14] Iorga, C., & Neagoe, V.-E.,"11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)" A Deep CNN Approach with Transfer Learning for Image Recognition, 2019.
[15] Kamal Kant Verma ET.AL., "International Journal of Interactive Multimedia and Artificial Intelligence", Two-Stage Human Activity Recognition Using 2D-ConvNet, Vol. 6,2.
[16] Yang Xing, Chen Lv, Member, Huaji Wang, Dongpu Cao, "Transaction of Vechicular Technology," Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach, IEEE 2019, Vol.68, Issue 6.
[17] Yang Liu, Peng Sun, Max R. Highsmith, Nickolas M. Wergeles, Joel Sartwell, Andy Raedeke, et.al.,"Third International Conference on Data Science in Cyberspace", IEEE, Performance Comparison of Deep Learning Techniques for Recognizing Birds in Aerial Images,2018.
[18] Afshar Shamsi, Brij Mohan Singh, H. L. Mandoria, Prachi Chauhan, "IEEE Transactions on Neural Networks and Learning System", An Uncertainty-Aware Transfer Learning-Based Framework for COVID-19 Diagnosis, Vol.32, NO. 4, April 2021.
[19] S.M. Mohidul Islam, Farhana Tazmim Pinki,"I. J. Engineering and Manufacturing",Colour, Texture, and Shape Features based Object Recognition Using Distance Measures,PP.42-50,August 2021.
[20] Tu, Xinyuan; Lai, Kenneth; Yanushkevich, Svetlana (2018). "[IEEE 2018 9th International Conference on Software Engineering and Service Science (ICSESS) - Beijing ", China - Transfer Learning on Convolutional Neural Networks for Dog Identification, 357–360,2018.
[21] Md. Rayhan Ahmed, Towhidul Islam Robin, Ashfaq Ali Shafin, "Automatic Environmental Sound Recognition (AESR) Using Convolutional Neural Network", I.J. Modern Education and Computer Science, 5, 41-54, 2020.
[22] Kamal Kant Verma1 et.al., "International Journal of Interactive Multimedia and Artificial Intelligence", Two-Stage Human Activity Recognition Using 2D-ConvNet,24 April 2020, 10.9781/ijimai.2020.04.002.
[23] Apri Junaidi et.al., IEEE International Conference on Communication, Networks and Satellite COMNETSAT)", Image Classification for Egg Inclubator using Transfer Learning of VGG16 and VGG19, IEEE, 17-18 July 2021.

Authors' Profiles

J. Indhumathi is a PhD (full-time) Research Scholar in the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram (India). She received her Master of Engineering (M.E) degrees in the year of 2019 and Bachelor of Engineering (B.E) degrees in the year of 2017 from the Department of Computer Science and Engineering, Annamalai University (India). Her area of research interest includes image and video processing, Machine and Deep Learning.

Dr. M. Balasubramanian is currently working as an Associate Professor in the Department of Computer Science and Engineering, Annamalai University (India). He is a member of Computer Society of India & ISTE. He awarded Ph.D in Computer Science and Engineering from Annamalai University in the year 2011. He received his M.Tech degree in Computer Applications from Indian Institute of Technology, Delhi in the year 2004. He received his B.E degree in Computer Science and Engineering from Government College of Engineering (GCE), Tirunelveli in the year 1996. He has published 45 papers in various reputed international journals and 19 papers in various national and international conferences. His area of research includes Image and Video processing, Machine and Deep Learning.

Balasaigayathri. B is a Master of Engineering (M.E) in the Department of Computer Science and Engineering, Annamalai University (India). She is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology (Deemed to be University) (India). She received her Bachelor of Engineering (B.E) degrees in the Department of Computer Science and Engineering, Annamalai University. Her area of work is image and video processing and Deep Learning.

How to cite this paper: Indhumathi .J, Balasubramanian .M, Balasaigayathri .B, "Real-Time Video based Human Suspicious Activity Recognition with Transfer Learning for Deep Learning", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.15, No.1, pp. 47-62, 2023. DOI:10.5815/ijigsp.2023.01.05