Enhanced video analysis framework for action detection using deep learning

Ms. Saylee Begampure
School of Electronics and Communication Engineering
Dr.Vishwanath Karad MIT World Peace University, Pune ,India
and
Dr. Parul Jadhav
School of Electronics and Communication Engineering
Dr.Vishwanath Karad MIT World Peace University,Pune ,India

Video Analytics analyzes the video content and adds brains to eyes that is analytics to camera. It extracts contents from the video by monitoring the video in real time. Normal and Abnormal human activity detection using deep learning models is a challenging task in computer vision. The detection of the same will help in detecting crime scenes which will help in preventing treacherous actions Proposed method focuses on classifying normal activities for humans in real time scenarios. The pre-processing technique for redundant frame detection, elimination and training the model efficiently using Convolutional Neural Network for classifying the activities is the main research contribution. Proposed method shows improvement in accuracy as compared to reference method which can be further implemented for on edge embedded platforms for real time applications

Keywords: Video Analytics, Deep Learning, KTH Dataset, Human Activity Detection

1. INTRODUCTION

Human action detection has wide applications in video analytics.There are two basic parts: Action Detection and Action Prediction. Human actions can be normal actions such as standing, sitting, jumping, and going up stairs, handclapping or some anomalous actions like abuse, assault, fighting, vandalism etc. Despite of many state of art algorithms, correct classification of human action from the video has always been a challenge. Deep learning can be used for analyzing the video data. Its layered filters enable the enhancement of an object detection rate and can be used to classify normal and abnormal actions using rule-violation.
Convolutional Neural Network (CNN) is one of the basic and important frameworks in deep learning. It is a special kind of multi-layer neural network, designed to recognize visual patterns directly from pixel images with minimal pre-processing.Albawi et al. [2017] It extracts high-level features of the images and reduces the complexity of the input thus helps in identifying normal human activities, which is then can be applied to Recurrent Neural Network (RNN) Sherstinsky [2020] or Long-Short Term Memory (LSTM) Sherstinsky [2020] which will help to identify abnormal human activities depending upon sequence of events. LSTM and RNN networks are used under transfer learning Tan and Sun [2018].
This paper focuses on only normal action detection and classification of human actions amongst heterogeneous set of actions. Main research contribution added here is in data pre-processing part. A technique of redundant frame detection and elimination is added before training the model, which removes the unwanted frames from the videos. Efficient Frames are then given to convolutional neural network to train the model and results are observed. Modified Model has shown promising results than existing reference paper. Organization of paper is as follows: It covers literature survey of related research articles along with the data-sets used, followed by system architecture. Here proposed method's details are explained along with data pre-processing

enhancement, model building, implementation details and finally experimental results and conclusion is described.

2. BACKGROUND AND RELATED WORK

Video contains the sequence of images having spatial and temporal characteristics. In Human activity detection normal or abnormal activities are analyzed and then detected using various deep learning models. Spatio-temporal relationship is the key feature in analyzing the human actions. Though deep learning-based architectures are useful for the same, however dataset also carries huge importance to get better accuracy and measuring the performance in terms of complexity. Comprehensive review on different object detection methods starting from basic CNN up to YOLO or SSD shows that there are various deep learning architectures used for general object detection purpose. Begampure and Jadhav [2019] Literature survey for existing models for recognizing human actions and dataset used for the same is as follows:

One of the initial works done in action detection uses temporal features with motion energy image and motion history image which further encodes all different human actions into a single frame.Bobick and Davis [2001] Drawback of this method was image are too much sensitive to view point. To overcome above drawback of view point, some authors used optical flow method Lucas and Kanade [1981]B.Horn and B.Schunck [1981]Sun et al. [2010] which captures the motion information. It performs action on horizontal and vertical axis and fins out motion pattern of in congestive frames.

A Local SVM approach is used for recognizing human actions by C. Schuldt, I. Laptev and B. Caputo. They used Local Space time features to recognize complex motion patterns by localizing. It captures local events in video and adapt velocity, frequency and size of moving pattern. Support vector machine-based classification is used for action recognition from the video.Schuldt et al. [2004] Some of the authors like Ivan Laptev, Tony Lindeberg used local descriptors approach for Spatio-Temporal Recognition. It represents video in local space time events. Optical flow method is used for comparison of motion representations. Several types of image descriptors are evaluated in context of activity recognition. Laptev and Lindeberg [2006]

Space time interest point method Laptev [2005]Laptev and Lindeberg [2003]is also as one of the approaches but drawback of this is space time interest points will work for only short temporal duration and not long duration. Recently Human activity Detection is done using various advanced methods. One of those method is using 3D CNN J. Arunnehrua and Bharathi [2018]Ji et al. [2010]Taylor et al. [2010]. Model taken as a reference model is trained on KTH dataset and 3D convolution neural network is used. Author has achieved accuracy of 64%.[1].Proposed architecture is based on 3D convolutional neural network which is then further used to enhance performance metric. Out of all above models, last model is considered as a reference model and results are compared with respect to it.

For training any deep learning model dataset is must. Dataset exploration is done for few relevant datasets and finalized one dataset for training. Following are the datasets considered Dataset: KTH Schüldt et al. [2005] [2].is basically a human activity recognition dataset. It is for normal human actions. It has 6 different actions:1.Walking 2. Running 3.Jogging 4.Hand clapping 5.Hand waving and 6.Boxing. Corresponding diagram for actions is as below: Each video is having length of 4 seconds. 25 subjects captured at four different scenarios, eventually having 2391 sequences in total. Frame rate is of 25 fps for all videos taken by static camera over homogeneous background. UCF crime dataset: UCF Crime Schüldt et al. [2005] dataset is developed by University of Central Florida. [3]. It has 13 abnormal human activities including Assault, Abuse, Arrest, Burglary, Robbery, Stealing, Road Accident, Explosion, shoplifting, Fighting, Arson and Vandalism. This dataset is designed as it will have great impact on safety of individual human being. It has all

[1] https://mrinaljain17.github.io/project/human activity recognition

[2] http://www.nada.kth.se/cvap/actions/

[3] https://www.crcv.ucf.edu/projects/real-world/

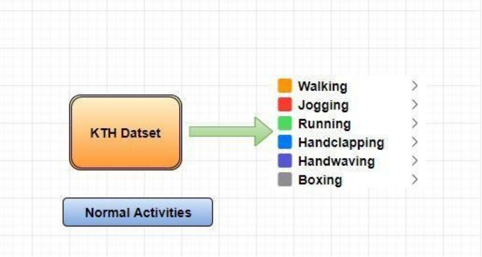


Figure 1.KTH Dataset

untrimmed real time video surveillance sequences for almost 128 hours. This dataset can be used for normal and abnormal human behavior activity detection. Following diagram illustrates the same. PETS 2016 dataset: PETS 2016 Sultani et al. [2018] dataset has 2 categories 1. IPATCH dataset 2. ARENA dataset. IPATCH is dataset created off the coast for scenarios of boat detection and attacks, anomaly near water, whereas [4].ARENA dataset is cretially criminal behavior which covers detection and tracking of norated for human behavior or activities scenario. It has multiple cameras with 22 different scenarios. ARENA dataset is further divided into 3 categories of human behavior. Normal behavior, criminal behavior and potenmal events Abnormal events and Threat events. Some of the abnormal activities considered under this dataset are as shown below: CAVIAR dataset: Cotext Aware Vision using Image-based Active Recognition (CAVIAR) Patino et al. [2016] mainly designed for considering 2 main scenarios 1. For surveillance at city center –to detect crime, vandalism, fights etc. 2. For marketers- to understand customer behavior pattern to increase profit. [5]. Out of which 1st scenario can be helpful for this project as its

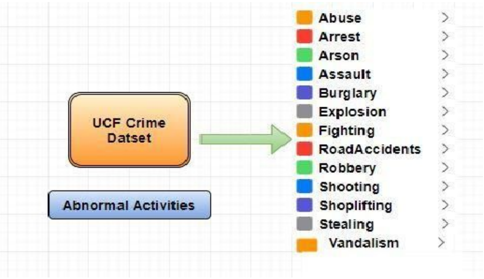


Figure 2.UCF Crime Dataset

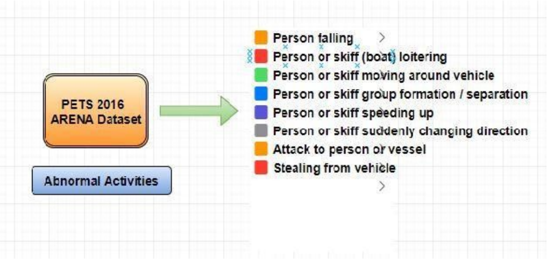

PETS 2016 ARENA Datset
Abnormal Activities
Person falling
Person or skiff (boat) loitering
Person or skiff moving around vehicle
Person or skiff group formation / separation
Person or skiff speeding up
Person or skiff suddenly changing direction
Attack to person or vessel
Stealing from vehicle

Figure 3.PETS 2016 Dataset

[4] http://www.cvg.reading.ac.uk/PETS2016/a.html

[5] https://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

considering important human behavior anomalous patterns as shown below: Walking, browsing, leaving bag behind, fighting between two people, people meeting in groups or spitting suddenly.

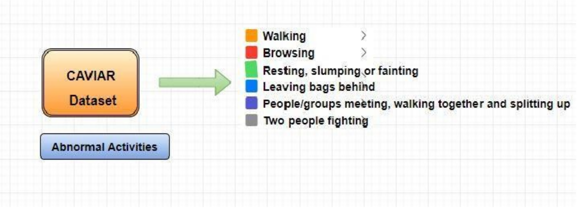


Figure 4.CAVIAR Dataset

Out of four datasets surveyed thoroughly, KTH dataset is finalized to go ahead with for normal activities detection and UCF crime will be used for abnormal activity detection in future.

3. SYSTEM ARCHITECTURE

(1) Selecting the relevant video dataset for human actions
(2) Preprocessing of data (converting into sequence of frames)
(3) Splitting the data into training, testing and validation dataset
(4) Design the model architecture (with number of layers and loss functions)
(5) Train the model with different layers of convolutional neural network for certain number of epochs Test the model with test dataset
(6) Calculation of performance metrics and plotting learning curve.

Flow Diagram for the system is as shown in Figure. 5

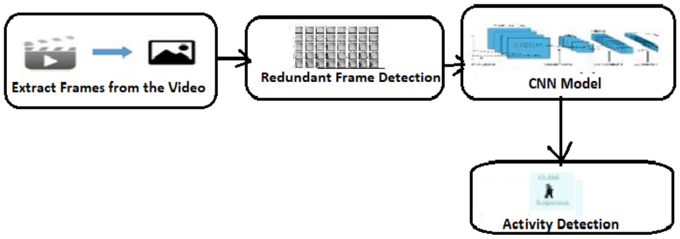


Figure 5.Flow Diagram of the System

Important steps are as follows:

3.1 Data Pre-processing

In detecting normal behavior of humans, there has to be learning about how person's movement pattern is. Human Activity Recognition using Deep Learning Ji et al. [2010] model was developed which considered 6 actions for training dataset like jogging, running, walking, boxing, hand waving and hand clapping is considered as base model for future enhancement. Dataset used here is KTH. In KTH dataset, as surveyed above has 25 subjects and each video is approximately 4 seconds. Four scenarios are considered. S1-Outdoors, S2-outdoors with scale variation, S3-outdoors with different cloths and S4-indoors. Frame rate is 25 fps. Spatial resolution of each video is maintained at 160*120 pixels by down sampling it. Sample image form all the 4 scenarios is as below:

First step will be to read videos from database. There are total 598 videos that is nearly 100 videos from each category (boxing, jogging, hand clapping, hand waving, running and walking). Out of which 33 % can be used for testing and 25 % of the training data can be used for validation, so the total number of video's for training data will be 398, total number of video's for testing data will be 200and total number of video's for validation =98.Frame rate is of 25 frames per second but human actions will not change much in period of 1 sec so there can be lots of frames which are redundant and can be discarded. Option for selecting a fixed number of frames per second is better option as it can take frames uniformly from the entire video.

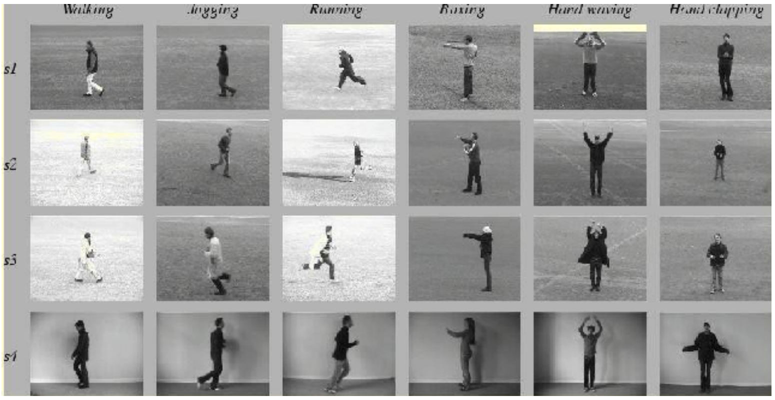


Figure 6.Samples from the KTH Dataset

Some data pre-processing is required before training the model using data. First step is to convert each frame into exactly same spatial dimension (same width and height) resizing videos to 128*128 pixels, which will take care of that. Second step is to convert the image into gray scale which will reduce the computational complexity and the last thing is to apply normalization. Normalization can be done by using 2 ways – One is Min–max normalization in which pixel values are normalized in 0 and 1 while second one is Z score normalization which determines standard deviation value from the mean. It is important to normalize pixel in the range 0 to 255 for better performance of network. 5 Dimension tensor will be obtained at end which will have a number of videos, number of frames, height, width and channel.

3.2 Model Building

Here 3D convolutional neural network J. Arunnehrua and Bharathi [2018]Ji et al. [2010]Taylor et al. [2010] is used which consists of repeated layers of convolution and max pooling and at the end connected to fully connected layer. Convolution layer used for filtering and pooling layer is used for sampling. These layers reduce spatial dimension form given to fully connected network which classify video frames. It uses Relu activation function, finally given to fully connected feedforward network which at the end consists of classifying neurons. Neuron which has the highest probability will be classified for particular action detection. Different combinations of number of layers are tried and finally combination which gave better accuracy has following specifications. It consists of 3 Convolution layer ,3 max pooling layers, followed by 1 global average pooling layer and finally 2 dense layers.

4. PROPOSED TECHNIQUE AND MODEL BUILDING

By considering the above model as a reference model, scope of improvement of existing model is analyzed by following ways. For any deep leaning based network dataset selection along with proper model architecture with suitable parameters selection is very important. In proposed method, while pre-processing dataset, frames without a person that is blank frames were identified and treating as redundant frame which is further excluded from dataset while training. In real time scenario's where inference is to be implemented then identification of redundant frame is very important and can be undertaken on the edge by the embedded platforms. This pre-processing technique is most applicable for training of deep learning architecture models. With such pre-processing technique has led to the improvement in accuracy.

Extraction of the redundant frames is implemented on KTH dataset of 6 actions is used. Frame rate of 25 fps is there for each video in KTH dataset and as discussed most of the frames are redundant. Author has used 2 approaches as discussed above

—Extracting fixed number of frames from video or

—Extracting fixed number of frames per second. content...

But problem in both the scenario is that there are still so many redundant frames. As author is calculating center point of frames and few frames before and few frames after are considered. Still there are some frames which are frames with on background and which will not be considered for our aim of detecting human actions. So if only frames with person is present in it will be extracted then accuracy will be improved as all redundant frames will be vanished. Going ahead with same approach, data preprocessing part is modified and frames with person are extracted. Following are the samples of video's before extracting and after extracting.

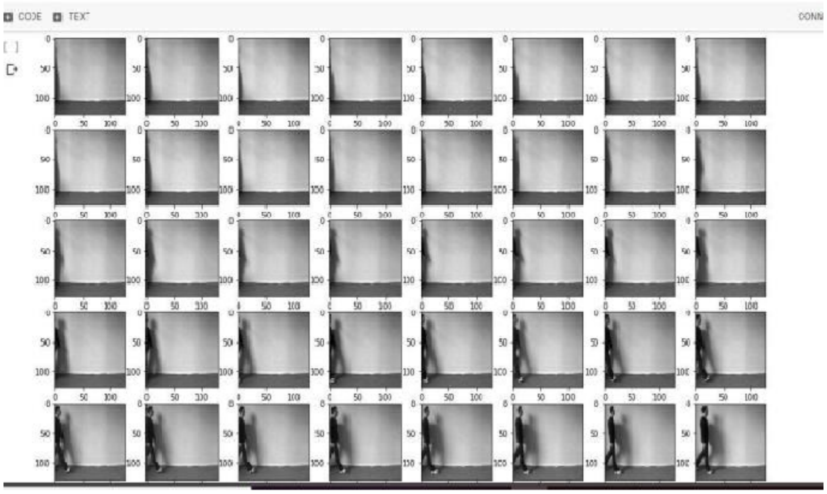


Figure 7.Sample video frames before extraction

After all blank frame are removed, all other preprocessing steps like resizing, converting to gray scale, and normalizing are kept same. 3D convolutional neural network is applied on newly extracted frames. Frames after removing redundant frames is as shown below

Convolutional Neural network-based architecture is as shown in Figure. 9.

Here, sequential model is used with 4 Convolution layer, 4 max pooling layers, followed by 1 global average pooling layer and finally 2 dense layers are used. No. of filters used in each layer are 16, 64, 256, 1054 respectively. Kernel size is kept as (5,3,3) and (2,3,3). No. of strides are kept as (1,1,1) for convolution layers while (2,2,2) for max pooling layers along with same padding. Activation function used is Relu and dropout is added of 0.5. Model is trained on the entire training dataset for 40 epochs with adam optimizer. Performance metrics like Loss, Accuracy and Cross Validation accuracy are calculated. Best model is saved every time depending upon metric. Finally, best model is calculated amongst all 40 epochs and accuracy and cross validation accuracy of that model is considered for plotting learning curve. Unknown video is given as an input to this system which predicts or classify the activity correctly as per the predicted accuracy.

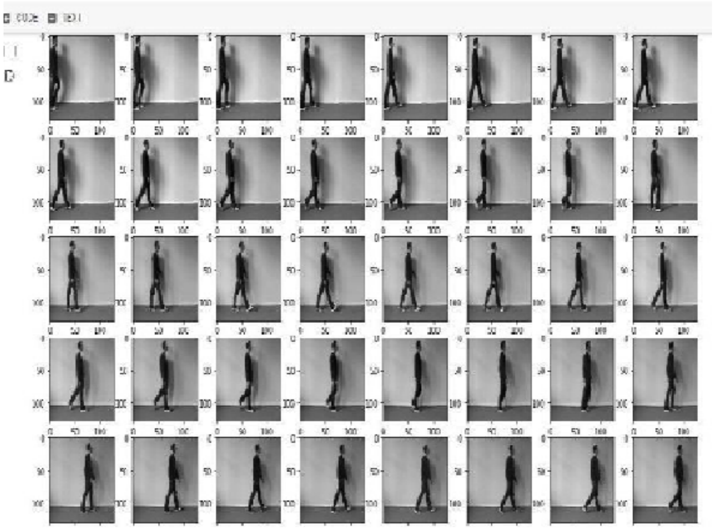
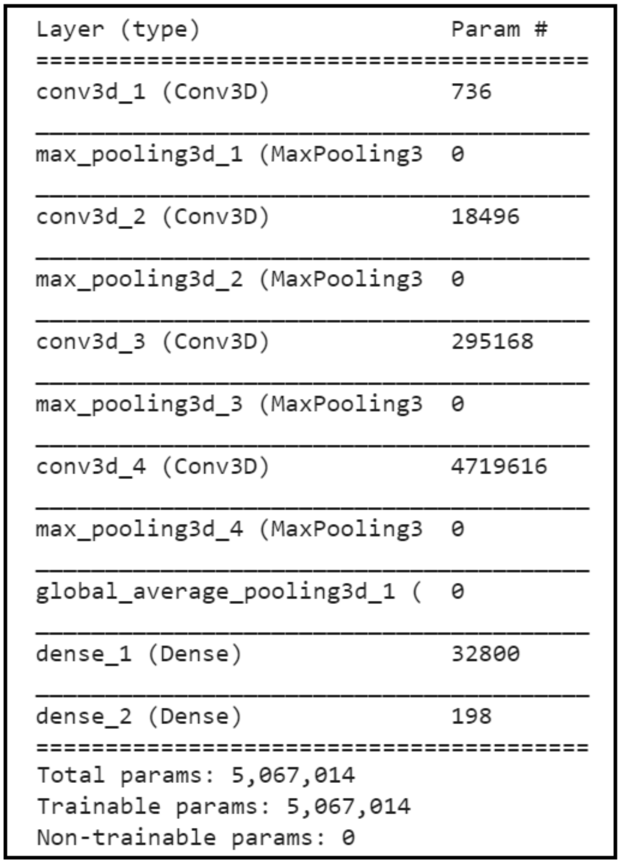


Figure 8.Sample video frames after extraction

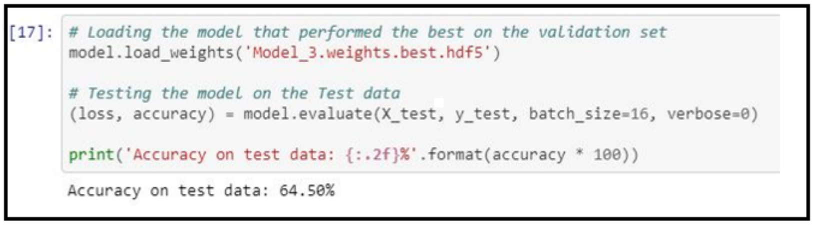


```
Layer (type)                 Param #
========================================
conv3d_1 (Conv3D)            736
________________________________________
max_pooling3d_1 (MaxPooling3 0
________________________________________
conv3d_2 (Conv3D)            18496
________________________________________
max_pooling3d_2 (MaxPooling3 0
________________________________________
conv3d_3 (Conv3D)            295168
________________________________________
max_pooling3d_3 (MaxPooling3 0
________________________________________
conv3d_4 (Conv3D)            4719616
________________________________________
max_pooling3d_4 (MaxPooling3 0
________________________________________
global_average_pooling3d_1 ( 0
________________________________________
dense_1 (Dense)              32800
________________________________________
dense_2 (Dense)              198
========================================
Total params: 5,067,014
Trainable params: 5,067,014
Non-trainable params: 0
```

Figure 9.Model Architecture

5. EXPERIMENTAL RESULTS

As its video dataset and each video is converting to frames large storing space is required, also model is trained on 40 epochs, so large processing power will be required. Normal CPU's will be insufficient for this. So, google Cloud platform and Google Collaboratory is used. Google Collaboratory E [2019] has functionality to use jupyter notebook along with it. We can upload entire dataset on drive and Collaboratory notebook will directly fetch data from it. It has CPU's, TPU's or GPU's. Important libraries in python like sklearn, TensorFlow, matplotlib are already installed can be easily imported. Google cloud platform S.P.T. and J.L.U [2019] has AI Platform for Deep Learning. It has facility to create virtual machine which can directly import Deep Learning Image, so that all the important libraries required for deep learning are already present. It has maximum limit of 24 CPU's and we can ask for required number of GPU's. For current implementation 24 cpu's and 6 GPU's are used form google cloud platform. Figure.10 shows result for base model. It indicates accuracy of 64.05 % is achieved using KTH dataset and selecting fixed 200 frames from center every time. With the use of proposed technique accuracy is improved to 86.23 % as shown in Figure.11.



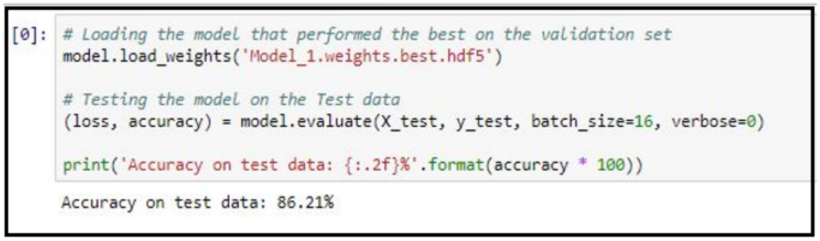
```
[17]: # Loading the model that performed the best on the validation set
      model.load_weights('Model_3.weights.best.hdf5')

      # Testing the model on the Test data
      (loss, accuracy) = model.evaluate(X_test, y_test, batch_size=16, verbose=0)

      print('Accuracy on test data: {:.2f}%'.format(accuracy * 100))

      Accuracy on test data: 64.50%
```

Figure 10.Base Paper Accuracy



```
[0]: # Loading the model that performed the best on the validation set
     model.load_weights('Model_1.weights.best.hdf5')

     # Testing the model on the Test data
     (loss, accuracy) = model.evaluate(X_test, y_test, batch_size=16, verbose=0)

     print('Accuracy on test data: {:.2f}%'.format(accuracy * 100))

     Accuracy on test data: 86.21%
```

Figure 11.Accuracy after Modification

Learning curve for the same can be plotted in Figure.12. It shows learning curve which is a plot of model learning performance experience over a time

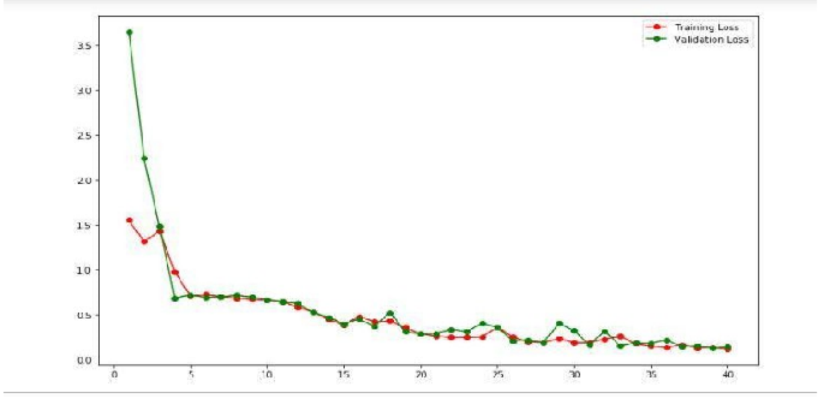


Figure 12.Learning Curve on test data

Here it is observed that model is not overfitting the data that is not learned statistical noise only learned from training dataset. Neither it is under fitting as its not showing high training and testing error. If input data size would have me more, it would have shown still better performance. Accuracy of model is improved after modification by almost 20%.

6. CONCLUSION

This work focuses on Normal activity detection using deep neural network. KTH dataset of 6 actions is used. In pre-processing of the data, frames are extracted and kept at the same resolution. Accuracy of reference model was 64.50 %. In proposed method, frames with only person in it are extracted and other redundant frames are removed, model is then trained again using convolutional neural network. Therefore, the size of the data got reduced by a large factor (5 times). This could be a great advantage for model training as model would now take less time to train. Also, since time is saved here, improvement was further made possible by making the model deeper (adding more layers). The model can still learn the patterns from the training data accurately because the data that removed was just redundant. The use of removal of blank frames shows improvement in accuracy to 86.21% after modification. This can be further extended with transfer learning approach for detecting abnormal activity detection using RNN and LSTM Networks. This approach for pre-processing on training datasets could be used with on edge embedded platforms for real time application.

References

Albawi, S., Mohammed, T. A., and Al-Zawi, S. 2017. Understanding of a convolutional neural network. *International Conference on Engineering and Technology (ICET), Antalya Vol.63,* No.1-6.

Begampure, S. and Jadhav, P. 2019. Comprehensive review of generic object detection frameworks using deep learning approach. *International conference on contemporary engineering and technology*.

B.Horn, J. and B.Schunck. 1981. Determining optical flow. *IEEE Trans Pattern Analysis and Machine Intelligence Vol.17,* No.185–203.

Bobick, J. A. and Davis, J. 2001. The recognition of human movement using temporal templates. *IEEE Transaction Pattern Analysis and Machine Intelligence Vol.23,* No.257–267.

E, B. 2019. Google colaboratory. in: Building machine learning and deep learning models on google cloud platform. *Apress, Berkeley, CA*.

J. Arunnehrua, G. and Bharathi, S. P. 2018. Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos. *International Conference on Robotics and Smart Manufacturing (RoSMa2018) ,Procedia Computer Science 133*, No.471–477.

Ji, S., Yang, W. X. M., and Yug, K. 2010. 3d convolutional neural networks for human action recognition. *ICML.*

Laptev, I. 2005. On space-time interest points. *IJCV Vol.64*, No.107–123.

Laptev, I. and Lindeberg, T. 2003. Space-time interest points. *ICCV*, No.432–439.

Laptev, I. and Lindeberg, T. 2006. Local descriptors for spatio-temporal recognition. *Spatial Coherence for Visual Motion Analysis Vol.3667.*

Lucas, B. D. and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. *Imaging Understanding Workshop*.

Patino, L., Cane, T., Vallee, A., and Ferryman, J. 2016. Pets 2016: Dataset and challenge. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1240–1247.

Schuldt, C., Laptev, I., and Caputo, B. 2004. Recognizing human actions: a local svm approach. Vol. Vol.3.

Schüldt, C., Laptev, and Caputo. 2005. ”kth dataset. *Proc. ICPR’04, Cambridge, UK*.

Sherstinsky, A. March 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) networks. *Elsevier journal "Physica D: Nonlinear Phenomena 404*, pp.181–199.

S.P.T., K. and J.L.U, G. 2019. Getting started with google cloud platform. in: Building your next big thing with google cloud platform. *Apress, Berkeley, CA.*

Sultani, W., Chen, C., and Shah, M. 2018. Real-world anomaly detection in surveillance videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Sun, J., S.Roth, and M.J.Black. 2010. Secrets of optical flow estimation and their principles. *CVPR.*

Tan, C. and Sun, F. 2018. A survey on deep transfer learning. *Artificial Neural Networks and Machine Learning ICANN 11141.*

Taylor, G. W., Fergus, R., LeCun, Y., and Bregler, C. 2010. Convolutional learning of spatio-temporal features. *ECCV.*

Ms.Saylee Begampure :Research Scholar at School of Electronics and Communication Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India

Dr.Parul Jadhav : Associate Professor at School of Electronics and Communication Engineering, Dr. Vishwanath Karad MIT World Peace University, Pune, India