International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

https://dx.doi.org/10.12785/ijcds/110105

Intelligent Video Analytics For Human Action Detection: A Deep Learning Approach With Transfer Learning

Saylee S. Begampure[1] **and Parul M. Jadhav**[2]

[1] *School Electronics and Communication Engineering, Dr.Vishwanath Karad MIT World Peace University,Pune, India*
[2]*School Electronics and Communication Engineering, Dr.Vishwanath Karad MIT World Peace University,Pune, India*
Received 1 Jan. 2021, Revised 11 Jul. 2021, Accepted 23 Jul. 2021, Published 9 Jan. 2022

Abstract: Human actions consist of a sequence of similar patterns which are difficult to classify using traditional image processing algorithms. Video analytics is a major research area that adds brains to eyes which means analytics to the camera. It monitors the video contents and extracts intelligent information from it. The human action analysis and its detection is a challenging task. The proposed method focuses on detection of normal human activity using Long-Short Term Memory (LSTM) as a deep neural architecture. The pre-processing technique of redundant frame detection along with pre-trained Convolutional Neural Network (CNN) is implemented for classifying the activities efficiently. Transfer learning approach is used followed by Long-Short Term Memory (LSTM) network to generate hybrid framework which further enhances the activity detection. Proposed method shows improvement in accuracy as compared to reference method. This method can be further implemented for on edge processing in embedded platforms for real time applications.

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

1. INTRODUCTION AND OVERVIEW

Detection and prediction of human activities is an important research area in many computer vision applications. It has wide application for the safety and security of the individual as it helps to identify normal as well as abnormal behavior of human beings. Enterprises and corporations put surveillance systems at every public place like the station, mall, airport, etc. for video surveillance. But the present surveillance cameras-based systems can do an analysis of normal or anomalous behavior but cannot predict the suspicious behavior of humans. This is needed to prevent hazards that will be useful for the safety and security of people.

An intelligent video analytics system that is the solution for all to predict normal and anomalous behavior and reduce the search time from weeks to hours to minutes will be useful for the safety and security of every individual, women, children, etc. Videos are getting captured every day and night with the different cameras, it is impossible task to monitor each video continuously to identify abnormal/normal activities as a normal individual has 22 minutes of attention span for monitoring any video. So, there is an absolute need for an intelligent video analytics in the surveillance system. A single surveillance camera typically produces 15 to 30 pictures a second, resulting in 1.3 to 2.6 million images per day. In series with thousands of camera's crucial pieces of information may exist as a handful of individual images from a single field. To maintain the secured information huge amount raw data is generated. Management of the video plays important role here as large storage space will be required for the same. With growing technology, quality of video is increasing and hence storing of such a huge data is a big question. With the growth of High-Definition video, the requirements for storage infrastructure just keep going up. Billions of videos are recorded each week which is a valuable tool for post-analysis and investigation of incidences and forensic evidence. The use of the manual method, by looking at hours of recording to identify few critical seconds of coverage that can make and break the case is difficult and takes more time. Video Analytics analyses video content in real-time, adds the "brains" (analytics) to the "eyes" (cameras). Video Analytics performs computational analysis of the contents of the video which eventually extracts intelligence from the video automatically. It is a software used to monitor video streams in near real-time and one of the booming research areas nowadays. In this paper, we will be focusing on normal activity detection from the surveillance videos. Problem domain knowledge rules help to distinguish activities. Video analytics will help to identify human activities with less or no human intervention.

Various deep learning architectures are being used to classify human actions. It uses layers of filters to extract features from the broader level to the granular level. Convolutional Neural Network (CNN) [1]is the basic architecture and has played a major role in enhancing the capabilities

E-mail address: sayalisbegampure@gmail.com parul.jadhav@mitwpu.edu.in

of deep learning networks. A Deep Neural Network[2]is a multi-layer network with input layers, several hidden layers, and a fully connected output layer. The hidden layers consist of filters used for feature extraction and feature engineering while the fully connected layer is a classifier. Inputs that are in the form of video frames act as an image and since video analytics is based on series of images from a video, Recurrent Neural Network (RNN) [3] or Long-Short Term Memory (LSTM) [4]is helpful to identify sequential information. The Pretrained networks which are readily available can be deployed for the existing dataset using the technique called transfer learning [5]. LSTM [3] and RNN [4] networks can be used under the transfer learning [5] approach to keep important features from the previous model and cascade the next model to it which saves training time drastically and improves the performance of the model. Traditional machine learning and deep learning algorithms are trained to solve a specific task. Models need to be retrained from scratch if feature space distribution changes. Transfer learning [6] overcomes this isolated learning paradigm by having information for one task to solve other related tasks. Hence transfer learning is used in the proposed model.

This paper focuses on only normal action detection and classification of human actions amongst the heterogeneous set of actions using transfer learning technique which can be further extended for abnormal action detection. The proposed model has shown promising results than the existing reference model. The overall organization of the paper is as follows: It starts with a Survey of related research articles under literature survey followed by a survey of the dataset under consideration, data pre-processing enhancement, system architecture, and finally transfer learning methodology followed by implementation and conclusion.

2. Literature Survey

Researchers have reviewed many generic object detection methods [7] starting from basics CNN [1], its different types like Mask-RCNN [8], Fast-RNN [9] up to newly developed efficient methods like SSD [10] or YOLO [11]. There can be two types of human action detection: Action detection based on features or video classification. At first, features are extracted from the image like a skeleton, ROI, HOG, etc. and the second classification is based on the content of the video. Most of the traditional deep neural architectures use 2D CNN or 3D CNN as a base Model. Use of 2D CNN on a single steam model [12] [13] and on multi-stream models[14] [15] using spatial and temporal features extraction and its fusion is carried out to get the desired output. Researchers from Google have used 3D CNN [16] as a multi-stream model [17] for large-scale video classification. It uses multi-resolution, foveated architecture which can train a larger dataset in lesser time, while 3D CNN single steam [18] models fuses the useful information of the frames having RGB data or other types of frames with the optical flow field to get desired results. There are different approaches based on the CNN model like frequency-based [19], motion-based [20], optical flow-based [14], color-based [21] developed by the researcher. The main drawback of this CNN-based approach is the training time required is very large. As principle work is here to identify different human activities from the videos, the dataset used for training consists of actions with different classes. CNN-based methods undergo high computations for these videos, hence the transfer learning approach [22] came into the picture. HAR-based transfer learning approach for activity detection. [23] At the first level common features are used using representation analysis and then user-specific features are transferred. feature-based transfer learning approach which uses CNN [1]and LSTM [2] combination shows the use of transfer learning to learn features from stationary activities[24]. The PCA-based approach using Shapley values under transfer learning has shown significant performance improvement[25]. The proposed architecture is based on a pre-trained convolutional neural network which is then further used to enhance performance metrics using transfer learning. Model taken as a reference model [26] is trained on KTH dataset with 3D CNN and results are compared with reference of it. The author of the same has achieved an accuracy of 68.98 %.

Various pre-trained models can also be used for action detection under transfer learning [22] [27] which gave better results than previous methods. To train any deep learning model, millions of records as a dataset is needed, but by using deep transfer learning a representation of data or transfer a representation of data to specific tasks is sufficient.[5] Pretrained models are used as a feature extractor. Pre-trained models have an advantage as its weighted layers extract features by keeping same weights while training on new data for another task. 4 pre-trained models are surveyed [28]. Comparison table for silent features, top 5 accuracy score, and the number of features used is explained in Table I. Out of 4 pre-trained networks surveyed, for normal action detection inception model is chosen as it has good accuracy with a comparatively fewer number of parameters.

The appropriate dataset is very important for training any deep learning model and plays a major role to measure performance indicators. The relevant datasets are surveyed and considered for comparative analysis. Prior art focuses on the use of datasets KTH, PETS, UCF, and CAVIAR for human action detection[33].

KTH Dataset: KTH dataset [34] consists of actions for human activity detection as shown in Figure 1 [33]The specifications as follows:

- Developed at: ICPR'04, Cambridge, UK
- Actions: Walking, Running, Jogging, Hand clapping, Handwaving and Boxing
- Length: 4 Seconds each
- Speed: 25 fps
- Background: Homogeneous

TABLE I. SURVEY OF EXISTING PRETRAINED MODELS

Network	Silent Feature	Top5 Accuracy	Parameters
AlexNet [29]	Deeper	84.70 %	62 Million
VGGNet [30]	Fixed Sized Kernel	92.30%	138 Million
ResNet 152 [31]	Skip- shortcut connections	95.51%	60.3 Million
Inception [32]	Parellel wider kernels	93.30 %	6.4 Million

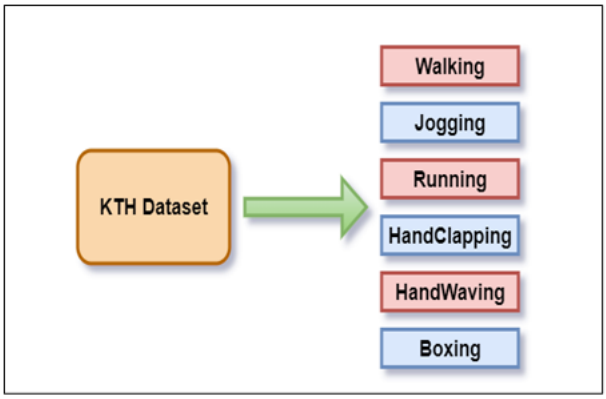

Figure 1. KTH Dataset

UCF crime dataset: UCF Crime [35] is the dataset for abnormal human activities as shown in Figure 2 [33] with specifications as:

- Developed by: University of Central Florida
- Actions: Abuse, Arson, Burglary, Assault, Explosion, Fighting, Vandalism, Road Accident, etc.
- Length: up to 1 min.
- Number of actions: 13
- Duration: 128 hours
- Background: Heterogeneous

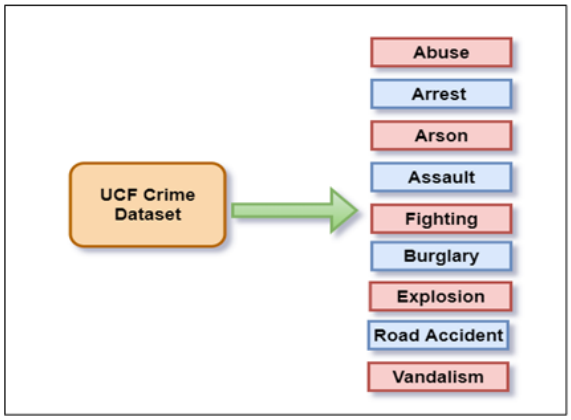

Figure 2. UCF Crime Dataset

PETS 2016: PETS 2016 [36] has 2 scenarios one related to coast boat attacks and the other related to human actions. Scenarios for Human actions are as shown in Figure 3 [33].

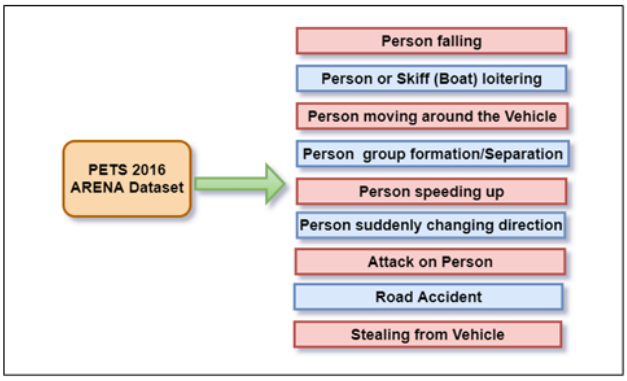

Figure 3. PETS Dataset

CAVIAR dataset: Context-Aware Vision using Image-based Active Recognition (CAVIAR) [37] dataset has normal and abnormal activities as shown in Figure.4 [33]

- Developed at: EC Funded CAVIAR project/IST 2001 37540
- Actions: walking, browning, resting, leaving a bag behind, people meeting together and splitting, etc.
- Length: up to 1 min.
- Number of actions: 1. City Centric 2. Market Specific
- Speed: 25 fps
- Background: entrance lobby of the INRIA Labs at Grenoble, France

This paper focuses on the use of the KTH dataset because of following reasons:

1) The background considered for creating this dataset is a constant per scenario which will help to measure the complexity across all the scenarios and in a way useful to measure the accuracy in terms of complexity.
2) There are three human actions related to hand gestures and three related to leg movements in KTH dataset.Actions considered while creating a dataset are very common and very similar, they have a very small difference that can be identified by a human,

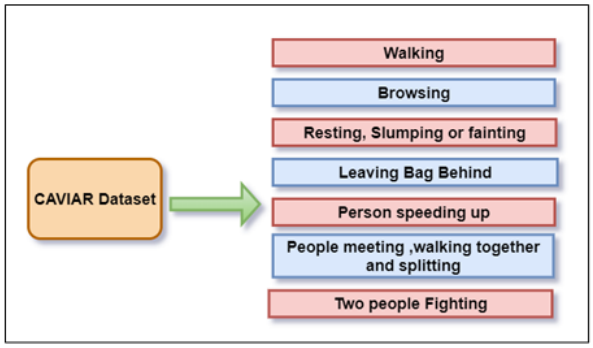

Figure 4. CAVIAR Dataset

but not by the machine. That is why it can train the model by giving similar action to the machine, hence confusing it while training to improve its prediction.

3. SYSTEM ARCHITECTURE

The system architecture consists following steps:

1) Appropriate event video dataset selection
2) Data pre-processing
3) Train, Test, Validation dataset creation
4) Designing of model architectures
5) Train the model with a pre-trained network
6) Use transfer learning to use pre-trained models' weights and apply it to LSTM train that model a certain number of epochs Test the model with the test dataset.
7) Calculation of performance metrics and plotting learning curve.

Important steps are as follows:

A. *Data Pre-processing*

The human action patterns consisting of normal behavior are to be learned so that it can predict the actions once unknown data is fed to it. The Reference Model [26] focuses on use of KTH dataset of 6 actions (walking, boxing, hand waving jogging, running, hand clapping) for activity detection. It uses for 4 scenarios such as Outdoors, outdoors with scale variation, outdoors with different cloths, indoors. [34] Frame rate is 25 fps as shown Figure5 with spatial resolution of 160*120.

Steps for Pre-processing:

1) Reading and splitting the Video:
   - Total number of videos: 598
   - Total number of videos for training: 398(66
   - Total number of videos for testing: 200(33
2) Extraction frames from the video: There are 2 ways for extracting the data while inputting the video frames one is extracting a particular number of frames from total number of frames or extracting a particular number of frames every second. The frame rate used is 25 frames per second. However, human actions will not change much in a span of 1 second, so there can be a large number of frames that are redundant and can be discarded. This technique for selecting frames with a fixed rate per second has been used as it can select frames uniformly from the entire video.
3) Resizing and Transforming: The extracted frames are resized to 128*128 pixels so that all the frames are of same spatial dimension, gray scale transformation is applied at the top of it to convert frames to black and white image.
4) Normalize the image: Min-max normalization is used to bring the pixel values normalized in between 0 and 1.

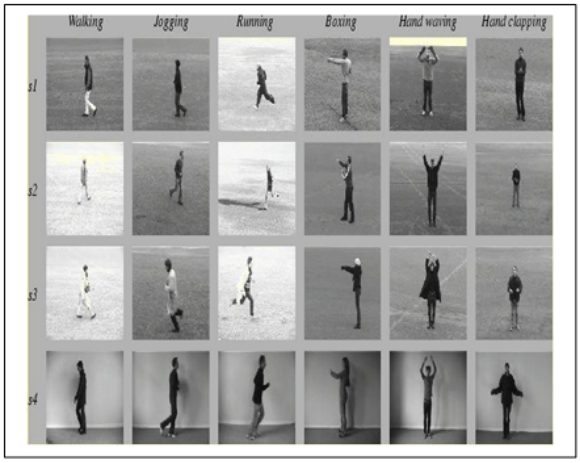

Figure 5. KTH Dataset Scenarios

Dimensionality of the tensors number of videos *number of frames *height*width* channel.

B. *Model Building*

After going through all pre-processing steps mentioned above model is trained on a pre-trained model using Inception V3 [32]. Transfer learning is applied to the above-trained model. The learned weights stored on the stack are transferred from the above model to the LSTM network [2]. Type of model used is Sequential as it is a linear stack of layers. Here dimensions of LSTM are 512 and the input shape used is (40, 2048). It is followed by the Dense layer of dimension 128. Sigmoid is used as an activation function in the next layer. Again, a dropout layer of 0.6 and a dense layer of dimension 6 is applied. Lastly, the SoftMax activation function is used. Model is trained for 50 epochs. The optimizer used is adam optimizer with categorical cross-entropy loss. The neuron which has the highest probability will be classified for particular action detection. Final system architecture is as shown in Figure.6

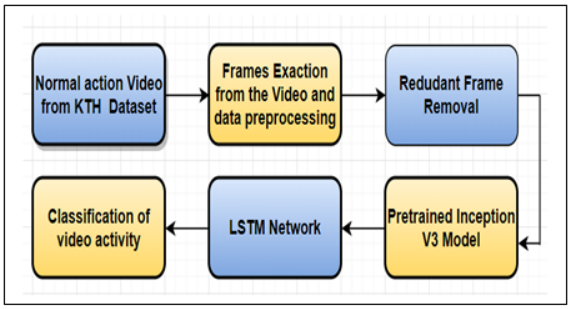

Figure 6. System Architecture

4. IMPLEMENTATION AND RESULTS

The pre-processing steps introduced in this paper uses the technique of removal of redundant frames (blank frames or frames without the person). In real-time scenarios 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 the training of deep learning architecture models. Such pre-processing technique has led to an improvement of accuracy. Steps in the proposed model[33]. The reference model uses the selection of some set of frames periodically from the center point in the sequence which may have red redundant frames. For any deep leaning based network dataset selection along with proper model architecture with suitable parameters selection is very important. In the proposed method, while pre-processing the dataset, frames without a person that is blank frames were identified and treated as the redundant frame which is further excluded from the dataset while training. Extraction of the redundant frames is implemented on the KTH dataset of 6 actions is used. The frame rate of 25 fps is there for each video in the KTH dataset and as discussed most of the frames are redundant.

In the previous art [33] the data pre-processing part is modified and balnk frames are removed. Originally extracted frames and frames remaining after removal of redundant frames is as shown in Figure.7 and 8.[33]

The Pre-processed data frames are given to a pre-trained Inception network, followed by LSTM,[2] dense layers as mentioned in network architecture above using transfer learning. Performance metrics like Loss, Accuracy, and Cross-Validation accuracy are calculated. Inception V3 [31] used as a pre-trained model acts as a deep classifier which helps reducing vanishing gradient problem. To prevent the network from dying out, auxiliary connections are added at the middle with batch normalization.

Main challenge faced was of High-speed parallel processing architecture as data is in video format. Due to memory constraints, the Google Colaboratory [38] and

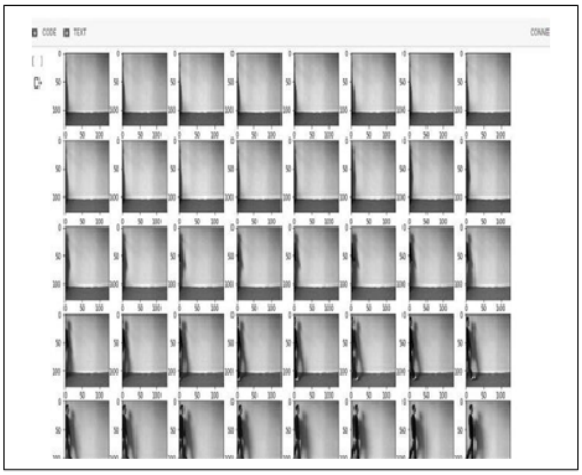

Figure 7. Sample video frame before redundant frame removal

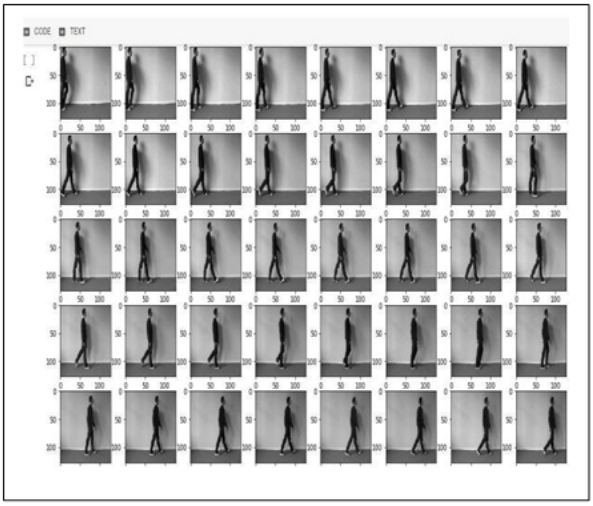

Figure 8. Sample video frame after redundant frame removal

Google Cloud Platform [39] are used. The best model amongst all after 50 epochs is considered for calculating the confusion matrix and another performance characteristic. Jupyter Notebook from Google Colaboratory is used along with TPU to run the model. Libraries used are sklearn, TensorFlow, matplotlib

Table II shows the comparative analysis of the reference model and the proposed work. It indicates that the accuracy of 68.98 % is achieved using the KTH dataset by the reference paper. The proposed technique using transfer learning shows improvement in accuracy of almost 20The use of transfer learning has also shown a reduction in

Code Snippet for benchmark model and modified model is shown in Figure 9. and Figure.10

TABLE II. COMPARATIVE ANALYSIS

Models	Accuracy
Reference Model [26]	68.98 %
Proposed model with Redundant frame removal and uses of transfer learning with LSTM	88.37%

```
# 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(bench_video, bench_target, batch_size=16, verbose=

print('Accuracy on test data: {:.2f}%'.format(accuracy * 100))
Accuracy on test data: 68.98%
```

Figure 9. Accuracy of reference model

```
Model.load_weights('Model_0.weights.best.hdf5')

# Testing the model on the Test data
(loss, accuracy) = Model.evaluate(X_test, y_test, batch_size=128, verbose=0)

print('Accuracy on test data: {:.2f}%'.format(accuracy * 100))
Accuracy on test data: 88.37%
```

Figure 10. Accuracy of modified model

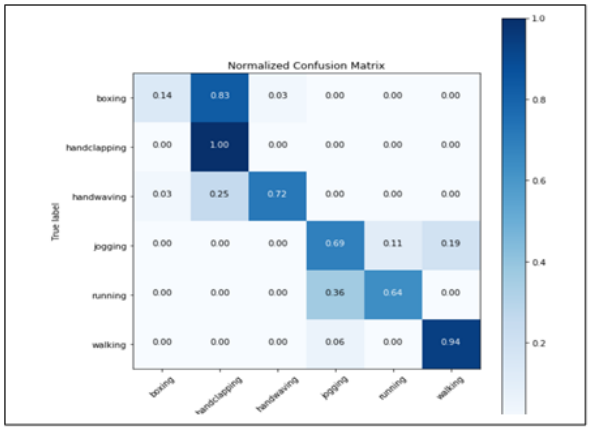

Figure 11. Confusion Matrix of a reference model

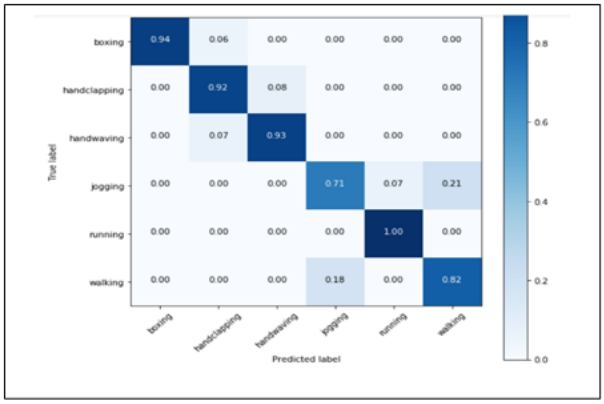

Figure 12. Confusion Matrix of a modified model

Comparative analysis is also done on Confusion matrix. Confusion matrix plotted after training the model against all 6 actions which as shown in Figure 12 and reference models confusion matrix is shown in Figure 11 [26] From the confusion matrix, it can be concluded that diagonal elements are having high values close to 1, which indicates the prediction accuracy of each action is good. Its shows-confusion of model between walking and jogging is more that is why false positive values are more, but other than that other actions classification accuracy is high as compared to reference model. All the diagonal values are improved as compared to original model.

Testing is done on unknown videos. Figure.13 the snap-shot for frames extracted from the video of a person walking on the road which is converted to the gray image first.

Optical flow vectors show how motion vectors are capturing the activity and thus predicting the activity from the video.

The output after predicting its motion with optical flow vectors at the start and once in motion is as shown below:

5. CONCLUSION

The proposed work focuses on normal human activity detection using a transfer learning approach on Inception V3 pre-trained model cascaded with LSTM. Dataset used is KTH for the classification of 6 actions. Accuracy is improved by 20The accuracy achieved is 88.37analysis of the confusion matrix shows an increase in a true positive index even in confusing actions. The use of Inception V3 which addresses the vanishing gradient issue with its auxiliary layer along with LSTM has shown improvement in the classification of similar actions. The use of LSTM model has contributed to memorizing similar sequential actions. The transfer learning approach has reduced the training time as it uses learned weights from pre-trained in the form of transferred values. The introduction of preprocessing step of removing/rejecting redundant or blank

Figure 13. Test video frame (Person Walking)

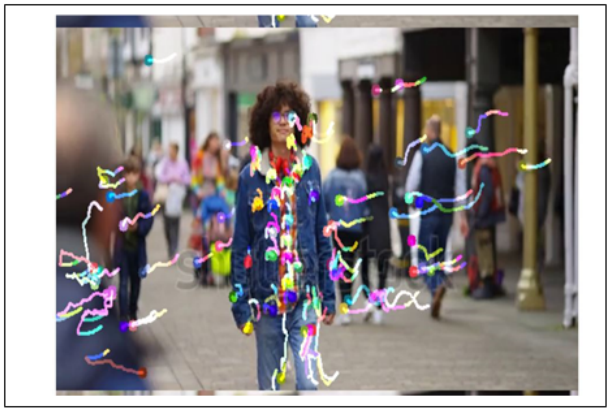

Figure 14. Optical flow of person walking at start

frames has played a significant role in the proposed model. This preprocessing can be taken care by on-edge devices in real-time applications.

Challenge faced was of high-speed parallel processing architecture as data was in video form, in future video data can be compressed, localized and unwanted part can be removed from it, before giving to the model so as to enhance the performance. This work can be extended with a region proposal network in which region-segmented images can be appended to the proposed model to improve its performance. Also, current work can be extended to anomalous activity detection for security at common places like airports, bus stops, shopping malls, residential areas, or any other public places.

Figure 15. Optical flow of person in motion

References

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[2] F. Altenberger and C. Lenz, "A non-technical survey on deep convolutional neural network architectures," 2018.

[3] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167278919305974

[4] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, 2002.

[5] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," 2018.

[6] R. Ribani and M. Marengoni, "A survey of transfer learning for convolutional neural networks," *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pp. 47–57, 2019.

[7] P. J. Saylee Begampure, "Comprehensive review of generic object detection frameworks using deep learning approach", international conference on contemporary," in *International conference on contemporary engineering and technology (ICCET-2019)*, April 2019, pp. 1–4.

[8] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[9] R. Girshick, "Fast r-cnn," 2015.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[12] H. Yang, C. Yuan, J. Xing, and W. Hu, "Scnn: Sequential convolutional neural network for human action recognition in videos," *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 355–359, 2017.

[13] A. Akula, A. K. Shah, and R. Ghosh, "Deep learning approach for human action recognition in infrared images," *Cognitive Systems Research*, vol. 50, pp. 146–154, 2018.

[14] M. Ma, N. Marturi, Y. Li, A. Leonardis, and R. Stolkin, "Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos," *Pattern Recognit.*, vol. 76, pp. 506–521, 2018.

[15] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, pp. 32–43, Jul. 2018, funding Information: This work is supported by the Singapore Ministry of Education Academic Research Fund Tier 2 MOE2015-T2-2-114 , and by grants from Office of Naval Research, US. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ONR. It was also supported by the Natural Science Foundation of China (61501198), Natural Science Foundation of Hubei Province (2014CFB461), Wuhan Youth Science and Technology Chenguang program (2014072704011248), the Dutch national program COMMIT and Dutch NWO TOP grant ARBITER. Publisher Copyright: © 2018 Elsevier Ltd.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[17] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018.

[18] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3d convolutional neural networks for action recognition," *Pattern Recognit.*, vol. 85, pp. 1–12, 2019.

[19] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 583–596, 2015.

[20] O. Elharrouss, A. Abbad, D. Moujahid, and H. Tairi, "Moving object detection zone using a block-based background model," *IET Computer Vision*, vol. 12, no. 1, pp. 86–94, 2018. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cvi.2017.0136

[21] J. Zhang, H. P. H. Shum, J. Han, and L. Shao, "Action recognition from arbitrary views using transferable dictionary learning," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4709–4723, 2018.

[22] A. B. Sargano, X. Wang, P. P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 463–469, 2017.

[23] S. An, G. Bhat, S. Gumussoy, and U. Ogras, "Transfer learning for human activity recognition using representational analysis of neural networks," 2021.

[24] R. Mutegeki and D. S. Han, "Feature-representation transfer learning for human activity recognition," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, 2019, pp. 18–20.

[25] P. C. Wen-Hui Chen and Y.-L. Jiang, "Activity recognition using transfer learning," in *Sensors and Materials*, vol. 29, no. 7, 2019, pp. 897–904.

[26] M. Jain, "Human activity recognition," in *Github repository by Mrinal Jain*, 2017.

[27] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, "Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos," *Procedia Computer Science*, vol. 133, pp. 471–477, 2018, international Conference on Robotics and Smart Manufacturing (RoSMa2018). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918310044

[28] G. Sapijaszko and W. B. Mikhael, "An overview of recent convolutional neural network algorithms for image recognition," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2018, pp. 743–746.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017. [Online]. Available: https://doi.org/10.1145/3065386

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[33] S. Begampure and P. Jadhav, "Enhanced video analysis framework for action detection using deep learning," *International Journal of Next-Generation Computing (IJNGC)*, vol. 12, no. 2, pp. 463–469, April 2021.

[34] C. Schüldt and C. Laptev, "Kth dataset," *ICPR'04, Cambridge, UK.*, vol. 18, January 2005. [Online]. Available: URL:http://www.nada.kth.se/cvap/actions/

[35] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[36] L. Patino, T. Cane, A. Vallee, and J. Ferryman, "Pets 2016: Dataset and challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[37] R. Fisher, "Ec funded caviar project," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2007.

[38] B. E., "Ec funded caviar project," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress, Berkeley, 2015.

[39] G. J. Krishnan S.P.T., "Ec funded caviar project," in *BuildingYourNext BigThing with Google Cloud Platform.*, 2015.

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