Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 1, pp. 233-242
DOI : 10.47974/JIOS-1867

Crime classification and synthesis leveraging CNN architectures for enhanced crime analysis

Udit Hasija *
Mahima Rohila †
Animesh Kumar §
Dharmender Saini ‡
Preeti Nagrath @
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering
New Delhi 110063
India

Sushil Kumar #
Department of Applied Sciences
Bharati Vidyapeeth's College of Engineering
New Delhi 110063
India

Mauro Zannetti$
Department of Industrial & Information Engineering and Economics
University of L'Aquila
I-67100 L'Aquila
Italy

Abstract

This paper proposes a novel deep learning approach for real-time activity classification in video surveillance applications. In this paper, we put forward a light mobile network-

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

* *E-mail:* udithasija1712@gmail.com (Corresponding Author)
† *E-mail:* mahimarohila600@gmail.com
§ *E-mail:* i.animesh001@gmail.com
‡ *E-mail:* dharmender.saini@bharatividyapeeth.edu
@ *E-mail:* preeti.nagrath@bharatividyapeeth.edu
# *E-mail:* sushilkumar16n@gmail.com
$ *E-mail:* mauro.zannetti@univaq.it

based LSTM called MobileSTM and an Inception V3 Network and combine their advantages through ensemble methods. Lightweight in architecture, the MobileNet can guarantee efficient processing on resource-constrained devices. LSTMs are powerful models in mining temporal information within sequences. Inception V3, known for its deep architecture and high accuracy, further enhances the model's capability to classify activities. MobileSTM and Inception V3 networks are combined and trained to recognize and classify these activities into pre-defined vital categories for public safety: violent acts and normal activities. The following approach brings forth the idea of strengthening the real-time video analysis for security personnel, so that they could focus their efforts and respond promptly in cases when some threats are observed or indicated.

Subject Classification: *68T07.*

Keywords: *MobileNet, Inception V3, Ensembling, LSTM, Activity classification, Video surveillance, Real-time, Public safety.*

1. Introduction

Video surveillance has definitely become the cornerstone of security in our homes, businesses, and public places today. Obviously, viewing hours of recordings to find those critical moments is not possible manually, and thus comes the increasing demand for automated systems of crime detection.

In this paper, we propose a deep learning video analysis approach towards the detection of crime activities in video surveillance. By leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs), we can classify video sequences into two essential categories: normal activities and violent acts. This method aims to close the gap between theoretical advancements in deep learning and their practical application in real-world crime prevention.

The main focus of our work is to distinguish between 2 categories, violent acts (like fighting, abuse etc) and normal acts in public places. By classifying any video sequences into these categories, we can successfully decrease the time taken to detect these crimes.

2. Literature Review

The applications of object detection, picture classification, and abnormal behavior detection in the intelligent video surveillance system and in the medical diagnosis are highlighted in this literature review. However, it has been work in progress with typical drawbacks such as handling of complex monitoring applications.

Models like MobileNet, ResNet, Ect. Several variants of DenseNet have been shown to be very efficient but more work on the aspect is

required [1]. Training performance has increased thanks to techniques like rectified linear units for optimisation and dropout for raising the degree of regularization. [2] The implication is the understanding of learning constraints on mobile devices. MobileNet-V2 performs superior in term of accuracy with less utilization of resources which justify its usage in mobile surveillance and ondevice analysis. [3]

To achieve efficient computation, models such as GoogLeNet [4] have been studied. The performance of the GoogLeNet architecture in picture categorisation has been impressive but certain difficulties. To mitigate these issues, Inception v3 [5] incorporates batch normalization and factorized convolutions.Other studies focus on novel deep learning architectures such as Models include ResNet50 [6], DenseNet [7][8], VGG16[9] and NasNet [10] although the last one has been proven effective in other works. Algorithms such as YOLOv3 have gained popularity and are well-suited for real-time video surveillance applications [11].

Intelligent Video Surveillance Systems (IVS): For capturing videos, such systems employ a combination of sensors such as thermal, infrared and visible light cameras The thermal, infrared and visible light cameras [12][13]. Object identification, identification of objects, tracking, background-foreground division and behavior description. to identify potential security risks are the main working steps of the ISS. [14] To be able to discover the contamination levels in overhead insulator IRT the following things must be considered; images, Current networks including CNN-BiLSTM are found to be more effective than simple CNN and CNN-LSTM networks [15].

Deep Learning in Healthcare: In one work, short-duration information from individuals' ECG mobile devices are utilized in the establishment of a MobileNet.detection [16]. This method is built upon utilizing MobileNet's efficiency. Human Activity Recognition (HAR) in healthcare uses deep learning a two-handed model of deep BiLSTM and pre-trained Things like gait analysis and other aspects of behavior that will require accurate CNN have to be carefully monitored. Thus, patient care features include potential for falls, risk assessment for falls, and video surveillance.[17]

Facial Emotion Recognition:

Automatic facial expression identification for emotion recognition is important in a variety of fields, including safety, healthcare, and human-computer interfaces [18]. Modern deep CNN architectures, such as Inception v3 [19], are utilized to address these problems by detecting emotions from masked faces on the FER-2013 dataset by applying artificial

masks. While reducing MobileNet-V2's computational load, reverse fusion, SELU activation, and merging SE-NET enhanced feature extraction and classification accuracy [20].

3. Methodology

This section details the methodology employed for developing a deep learning model to classify criminal activities and social behaviors in video surveillance footage.

Data Description: The dataset comprises a training set of 1,600 videos and a test set of 400 videos.

Data Preprocessing: Some picture frames were reduced to have constant input dimensions, such as 75 x 75 pixels. To enhance the similar concept detection, different data augmentation strategies were used to enrich the dataset.

Model Architecture Design

In this paper, we propose an average-based ensemble approach to classify video data.

MobileNetV2 employs depth-wise separable convolutions. The extracted features from MobileNetV2 are then fed into a BiLSTM layer which helps in capturing the temporal dependencies in the data. The features extracted by these layers are then processed by BiLSTM units to provide the final classification.

A standard convolution is used in InceptionV3 and inception modules are included in it. The features extracted by the InceptionV3 model are then fed into a BiLSTM layer to capture temporal dependencies.

Selecting the Base Learners: The base learners are MobileNetV2 [5] and InceptionV3[6], which is used for feature extraction and Bidirectional LSTM is used to detect the temporal nature of the video dataset.

Feature Extraction using MobileNetV2: MobileNetV2 employs depth wise separable convolutions, which break down the convolution into two separate steps to improve efficiency:

Depthwise Convolution:

$$Ydepth[i, j,cin] = sum(Wdepth[k, l, cin] * X[i + k, j + l, cin]) \quad \text{Eqn 1}$$

The depth wise convolution works specifically for each input channel without conducting multiply operations on all the channels *c* in. This

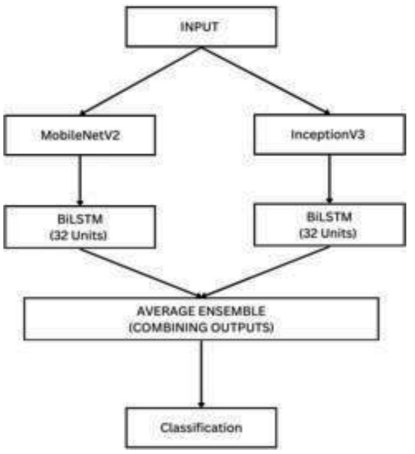

Figure 1

Average Ensemble Model Architecture

scans over the input tensor *X* and has a unique set of filters *W* depth for each channel to provide the respective channel's unique features.

Feature Extraction using InceptionV3: A standard convolution is used in InceptionV3 and inception modules are included in it.

$$Y[i, j] = \text{sum}(W[k, l, cin,cout] * X[i + k, j + l, cin] + b[cout]) \quad \text{Eqn 2}$$

Here, Y represents the output feature map, essentially a transformed version of the input video frame or patch (X). This transformation is achieved by applying a filter (weights), denoted by W.

Ensemble Aggregation of Models: Demonstrated in Figure 1. It consists of several convolutional neural stacked network models with a unidirectional LSTM. Ensemble Each prediction of the Model is calculated by:

Pi=modeli(Xtest) for

$$i = 1, 2, \ldots, M \quad \text{Eqn 3}$$

where Pi is the prediction from the i^{th} model. In this study, we use average ensemble on the base learners CNN-Bi-directional RNN model. Final prediction is obtained by the combination of all the predictions from these models by averaging:

$$P= 1/M*\text{sum}(Pi) \quad \text{Eqn 4}$$

And

$$Yj=\arg\max Pj,c \text{ for } j = 1 , 2, \ldots, N \quad \text{Eqn 5}$$

4. Model Evaluation and Explanation

A) *MobileNetV2 – BiLSTM*

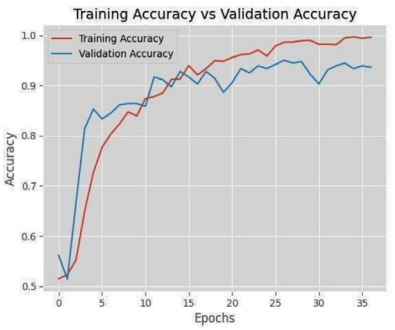

Figure 2

Training Accuracy vs Validation MobileNetV2-BiLSTM model

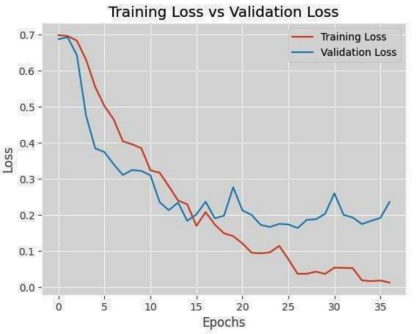

Figure 3

Training Loss vs Validation Loss MobileNetV2-BiLSTM model

The training process (Fig 2) depicts the model's learning behavior throughout training epochs. As expected, the training accuracy steadily increases, reaching a maximum of 0.99. Ideally, validation accuracy should follow a similar trend to training accuracy but at a slightly lower level. In this case, validation accuracy peaks at 0.93 on epoch 32 and then plateaus slightly.

Figure 3 illustrates the training process by comparing the training loss (red line) and validation loss (blue line) across epochs. The training loss signifies the model's performance generally decreases with each epoch as the model optimizes its parameters. The red line records its least value of 0.0165, indicating successful loss reduction during training.

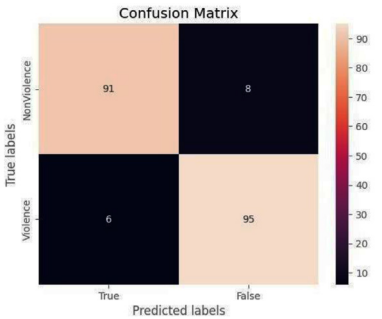

Figure 4

Confusion Matrix MobileNetV2-BiLSTM model

In Figure 4 the True Positive Rate is evaluated to be 91 and True Negative Rate is evaluated to be 95.

B) *InceptionV3 – BiLSTM*

The training process, as included in Fig. 5, shows the red line: It represents the training accuracy, which specifies how good the model performs on the training data which is steadily increasing and reaching a maximum of 0.99 whereas Finally, the blue line representing validation accuracy reaches it. It peaks slightly lower, 0.885 at epoch 13.

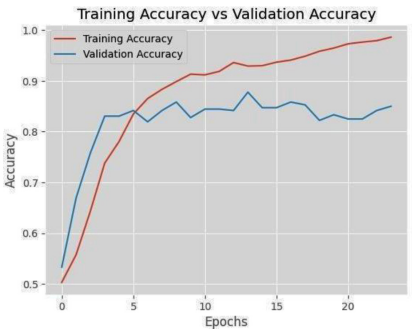

Figure 5

Training Accuracy vs Validation

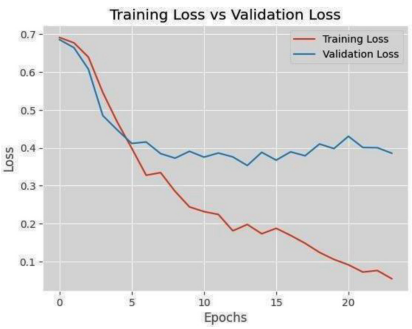

Figure 6

Training Loss vs Validation Loss

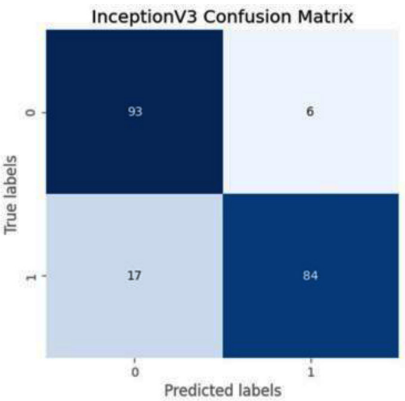

Figure 7

Confusion Matrix of Inception

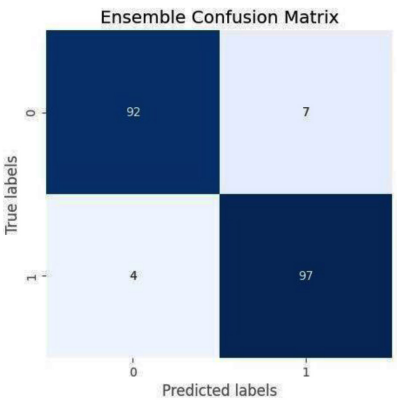

Figure 8

Confusion Matrix of Average Ensemble model

Figure 6 illustrates that training loss is decreasing with each epoch but the validation loss has plateaus at a very early epoch and is hardly decreasing. This states that InceptionV3 is not a very good choice for the chosen dataset. In Figure 7 the True Positive Rate is evaluated to be 93 and True Negative Rate is evaluated to be just 0.84. However, a concerning issue arises where Violence is classified as Non-Violence in 17 instances.

C) *Average Ensemble*

In Figure 8, the True Positive Rate is evaluated to be 0.92, and the True Negative Rate is evaluated to be only 0.97. Moreover, there has been a decrease in the False Positive rate and False Negative rate, which are now 4 and 7, respectively.

5. Results and Conclusion

Table 1

Classification Report Comparison Of the Various Models

Models	Classes	Precis ion	Rec all	F1-Score	Support
MobileNetV 2-BiLSTM	Non- Violence	0.94	0.92	0.93	99
	Violence	0.92	0.94	0.93	101
InceptionV3 -BiLSTM	Non- Violence	0.85	0.94	0.89	99
	Violence	0.93	0.83	0.88	101
Average Ensemble	Non- Violence	0.96	0.93	0.94	99
	Violence	0.93	0.96	0.95	101

The table 1 presents the classification performance of three models. MobileNetV2-BiLSTM performed well with 94% precision for Non-Violence and 92% for Violence, both achieving an F1-score of 93%, though further training enhancements could improve robustness. InceptionV3-BiLSTM had strong Violence detection precision (93%) but a lower recall (83%), with 85% precision for Non-Violence, indicating areas for improvement. The Average Ensemble method improved performance across most metrics, with 96% precision for Non-Violence and 93% for Violence, achieving the highest F1-scores (94% and 95%). Although the recall for Non- Violence dropped slightly, the ensemble approach enhanced overall accuracy and reliability.

Analyzing the training process of MobileNetV2-BiLSTM (Fig 2), we observed a high validation accuracy peaking at 93% but in contrast the InceptionV3-BiLSTM (Fig 5) has only achieved a validation accuracy of 88.5%. This highlights an area for improvement in future work, where techniques to mitigate overfitting can be explored. In MobileNetV2-BiLSTM, the validation loss (Fig 3) successfully decreased to 0.1861, indicating the model's ability to optimize its parameters during training. However, the training loss (Fig 6) shows that the validation loss in the InceptionV3-BiLSTM model remains stagnant after a certain point, without decreasing at all, which could indicate potential overfitting.

The Average Ensemble technique has demonstrated a very good classification report achieving a precision, recall and F2 score of 0.93 or above showing significant improvement from that of MobileNetV2-BiLSTM and InceptionV3-BiLSTM.

In conclusion, the proposed Average Ensemble model of MobileNetV2-BiLSTM and InceptionV3-BiLSTM has demonstrated effectiveness in classifying violence activities and social behaviors within video surveillance footage

References

[1] S. S. Bhandari, S. Dhoke, and O. Sarang, "Smart Video Surveillance Using YOLO Algorithm and OpenCV," *IJCSN - International Journal of Computer Science and Network*, vol. 9, no. 2, pp. 1–6, Apr. (2020).

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pp. 1–9.

[3] K. Dong and Y. Ruan, "MobileNetV2 Model for Image Classification," (2020).

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," Sep. 17 (2014).

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 11 (2015).

[6] M. H. Abid, R. Ashraf, T. Mahmood, and C. M. N. Faisal, "Multimodal medical image classification using deep residual network and genetic algorithm," Jun. 29 (2023).

[7] T. Chauhan, H. Palivela, and S. Tiwari, "Optimization and fine-tuning of DenseNet model for classification of COVID-19 cases in medical imaging," Apr. (2021).

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "*Densely Connected Convolutional Networks*," Jan. 8 (2018).

[9] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," Mar. 12 (2020).

[10] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," Apr. 1 (2018).

[11] P. Bhoyar, A. Butley, S. Dhole, N. Joshi, and A. Rumale, "Intelligent Video Surveillance using YOLO Object Detection," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 9, no. 11, Nov. (2021), ISSN: 2320-2882.

[12] Sutrisno Ibrahim, "A comprehensive review on intelligent surveillance systems", Communications in Science and technology 1 (2016)

[13] S. Ibrahim, "A comprehensive review on intelligent surveillance systems," *Communications in Science and Technology*, vol. 1 (2016).

[14] P. Singh and V. Pankajakshan, "A deep learning-based technique for anomaly detection in surveillance videos," in *2018 Twenty Fourth National Conference on Communications (NCC)*, IEEE (2018).

[15] A. K. Das, S. Deb, S. Chatterjee, B. Chatterjee, and S. Dalai, "Convolutional neural network and Bi-directional long short memory hybrid deep network aided infrared image classification framework for non-contact monitoring of overhead insulators."

[16] S. Shin, M. Kang, G. Zhang, J. Jung, and Y. T. Kim, "Lightweight Ensemble Network for Detecting Heart Disease Using ECG Signals," *Applied Sciences* (2022).

[17] N. Hassan, A. S. M. Miah, and J. Shin, "A Deep Bidirectional LSTM Model Enhanced by Transfer-Learning-Based Feature Extraction for Dynamic Human Activity Recognition," *Applied Sciences* (2024).

[18] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," Dec. (2017).

[19] Q. Zhu, H. Zhuang, M. Zhao, S. Xu, and R. Meng, "A Study on Expression Recognition Based on Improved MobileNetV2 Network," Jan. 29 (2024).

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," Apr. 17 (2017).

Received August, 2024