[1]Makhan Singh
[2]Nisha Tayal *

Performance Analysis of CNN Based Aggressive Human Behavior Detection Techniques

Journal of Electrical Systems

Abstract: - Aggressive Human Behavior is one of the most sophisticated concepts in social and situational contexts. A visible increase in aggressive human behavior can be seen over the news channels or in our surroundings every day. This work focuses on physical aggression in humans which includes hitting, kicking, punching etc. These aggressive events pose a direct threat to the public safety. Systems that can automatically monitor or detect surveillance videos and thus identify aggressive human activities in those videos will be of great help to the authorities. In this research, we used three different datasets i.e. hockey fights, movies and violent flow dataset. The video clips from these datasets are converted into pre-processed frame data sequences. The datasets are then divided into training dataset and validation dataset. The model through which training dataset is passed contains Convolution Neural Network (CNN) linked to convolutional LSTM (ConvLSTM) layer. The output of this model is binary classification of aggressive and non-aggressive flags. Further, the validation dataset is used to test the model efficiency. In case, the performance of the model is not satisfactory, the training of the model is re-evaluated until we achieve the desired performance. As depicted by the results, ResNet50 is the best performing CNN model with accuracy of 90%. The InceptionV3 CNN model yielded 89% of accuracy which is close to ResNet50. Further, VGG19 yielded very poor performance results of only 79% of accuracy. For future works, it is suggested to expand the model to more complex violence scenarios and appliances. Find creative solutions for data collection, advance generalization techniques and real-time optimizations.

Keywords: Aggressive Human Behavior Detection, CNN, ConvLSTM, datasets, performance analysis.

I. INTRODUCTION

Aggressive human behavior is a complex concept within social and situational contexts. It occurs when individuals intentionally inflict harm on others. The rise in such behavior is evident both in daily news reports and in our surroundings. This paper specifically examines physical aggression in humans, such as hitting, kicking, and punching. These acts of aggression pose a significant threat to public safety. To mitigate these risks, surveillance systems have been developed. Automated systems that can monitor surveillance footage and detect aggressive human activities would greatly assist authorities in maintaining safety [1]. This paper focuses on performance analysis of system that identifies aggressive and non aggressive human activities in surveillance videos with a help of pre-trained Convolution Neural Network (CNN). The pre-trained CNN is further linked to Convolutional LSTM layer (ConvLSTM) which gives output in binary form. The input to the system is raw videos captured by surveillance systems. Pre-trained CNN models (VGG19, IncepV3 and ResNet50) are implemented on benchmarks datasets. Performance analysis is done based on the results obtained for different datasets and best CNN based technique is analyzed.

It is highly demanding task when it comes to differentiate violent video from that of non violent videos. The videos generated by Fight detection systems are poor quality videos with lots of broken frames. Therefore, it is difficult or impossible to extract the intended features from theses frames. Violent videos particularly include aggressive moves or rapid moves (pushing, punching, slapping etc.) are shown in Figure 1. Whereas non violent videos specifically shows moving objects (hugging, clapping etc.) as shown in figure 2. In the aggressive human behavior detection system, sometimes false signals may be generated if the classification algorithms are not designed properly.

Figure 1. Violent Scenes including Human aggression

[1] Computer Science and Engineering, UIET, Panjab University, Chandigarh, India. singhmakhan@pu.ac.in
[2] Electrical and Electronics Engineering, UIET, Panjab University, Chandigarh, India. nisha.tayal@pu.ac.in
*Corresponding author: Nisha Tayal

Figure 2. Non-Aggressive Scenes

II. RELATED WORK

The incidents of human aggressive behavior, including crimes, offensive acts, and various unlawful killings, are now commonly heard worldwide. This behavior involves expressing anger in extreme forms, such as assault, harassment, and murder, all of which are harmful to others and violate basic human rights. This is a serious issue that continues to grow exponentially each year. To address the problem, surveillance cameras have proven to be very effective. Installing surveillance systems or CCTV cameras in public spaces has become a necessity [2, 3], as they have significantly contributed to reducing crime rates and other illegal activities.

However, ordinary surveillance systems may struggle to distinguish between nor-mal or non-violent behavior and abnormal or violent actions. To address this challenge, there is a need to develop advanced machine learning and deep learning techniques that can detect aggressive behavior in real time. These technologies can enable the system to identify various objects, analyze each image frame, and promptly alert the authorities when necessary.

A lot of research has been done in this field of computer vision since then. Abdali, et al. [4] proposed robust deep learning model that detects violence detection in real time. The research states a method that uses transfer learning. CNN model is used to extract the spatial features whereas for temporal relation learning method LSTM is used and the model gives 98% accuracy when implemented on three datasets (Movies, Hockey Fights and Violent Flow). This model needs to improve its performance with larger datasets. If the system can detect the type of violent action not just the existence of violence or non violence, it can be of greater usefulness.

A research paper from Mann B. Patel [5] produced a model that takes input of CCTV videos and draws inference to recognize where violent exists or not. The method is based on pose estimation using LSTM and hybrid CNN and LSTM. The datasets used are movies, hockey fights and violent flow giving accuracy 92.3% with Mean inference time 742. The paper suggests improving the accuracy of system while drawing inference from audio analysis.

Ullah, Fath U. Min, et al. [6] proposed a model to detect violence that consists of three stages. In the first stage, persons are detected using CNN model and unwanted data is eliminated. In second stage, the 3D CNN model is included to extract spatial temporal features and then softmax classifier is applied. In the end, the OpenVino toolkit is used to increase the performance of the model. Results over benchmark datasets justified that the method is good enough for violence detection in most of the cases. The future works indicated to apply the model on resource constraint systems to check the performance.

Zhang Xin et al. [7] proposed real time 6D pose estimation from a single RGB image. End to end deep learning model is used for detecting objects and recovering their 6D poses in an RGB image simultaneously. To guess the image coordinates of the object's 3D vertices, 2D pipeline is linked to pose estimation module. Then object's 6D pose can be predicted using Perspective-n-Point algorithm. The results are satisfactorily generated on two benchmark datasets (Linemod dataset and Occlusion dataset). The model is capable of real time processing and able to address texture less object analysis.

Ullah, Amin, et al. [8] proposed a lightweight deep learning assisted framework for activity recognition in which a person is detected using efficient CNN framework. The person is traced throughout the video stream using ultra fast object tracker called minimum object. The Deep Skip Connection Gated Recurrent Unit (DS-GRU) used to fetch the temporal changes. Datasets (HMDB51, UCF-101, UCF-50, YouTube Actions, Hollywood2 Actions) were used to test the efficiency of the system for real time surveillance applications. Future works suggest using embedded platforms to perform activity recognition over the edge.

Hammami, Samir Marwan et al. [9] presented vision based Model for detecting violence against children. The method is based on machine learning that uses skeleton joints data acquired by depth sensors. The dataset used is MMU VAAC to detect violent and non- violent actions. The system was able to give 99.03% accuracy.

Zaidenberg S et al. [10] contributed a generic framework for recognizing abnormal behavior in videos. The model keeps track of moving people by maintaining spatial-temporal group coherence. The trajectory is analyzed over the temporal window and clustered using Mean Shift Algorithm then formal event description language is introduced. The system yields satisfactory results on very challenging datasets (numerous occlusions and Long

duration sequences). The paper suggests working on the fight initiated in a group due to internal movements among people.

Sharma, Sarthak, et al. [11] presented a model based on deep learning CNN net-work and LSTM to detect and identify violent events. Datasets practiced Are Hockey fights, Movies and UCF crime dataset. The accuracy results obtained during the implementation of model were from 96.55% to 98.32%. The paper says to improve the model by specifying the severity of the violent action.

This paper uses pretrained CNN models like VGG19, IncepV3 and ResNet50 to differentiate violent and non violent videos. The performance of each pre trained model is evaluated on different data sets like hockey dataset, movies dataset and violent dataset one by one.

III. METHODOLOGY

The research paper shows comprehensive analysis of performance to underline the aspects and variables that directly or indirectly affect the fight detection systems. The main aim of this work is to recognize different CNN networks that are frame based systems using various available datasets in order to compare the performance of each CNN network. A basic model is followed to achieve our goal step by step which is shown in Figure 3.

Step 1: **Annotations:** The very first step is to apply the labels to the datasets that are to be used for model training purpose. It is a set of labels (aggressive or non aggressive behavior) which helps the model to learn specific features and recognize it. Here, we have used the label 1 as aggressive behavior and label 0 as non-aggressive behavior.

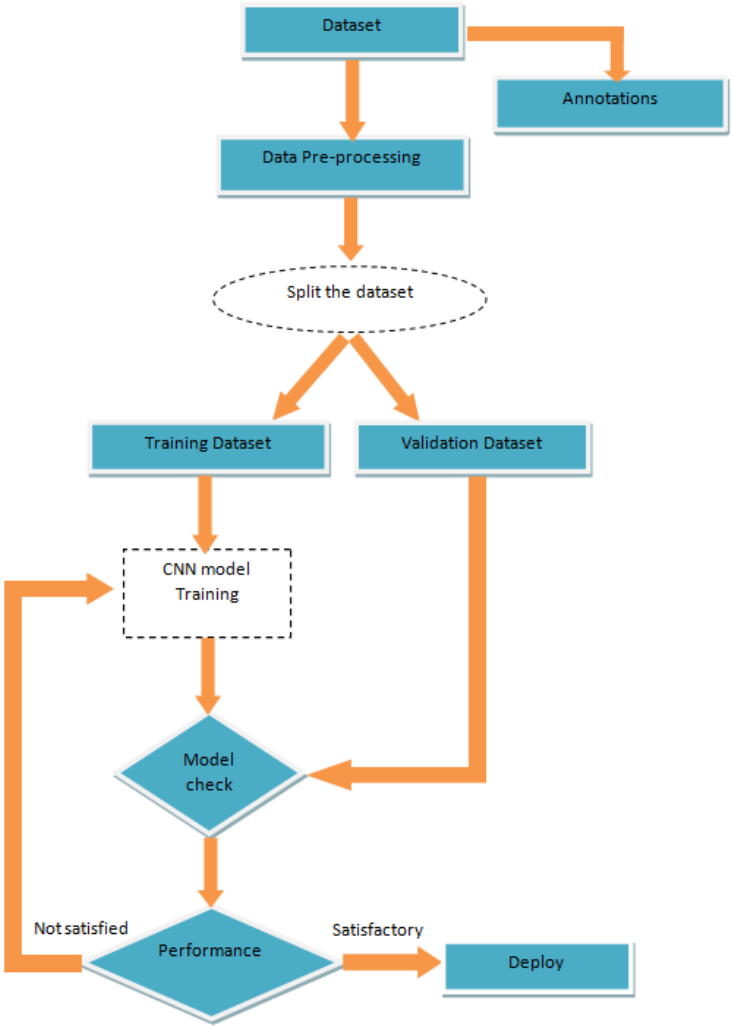

Figure 3. Flowchart of Proposed model

Step 2: Data Pre-processing: Few steps were taken into consideration for the graphical representation of the output. This is called the data preparation or pre-processing.

- Firstly, the videos were drawn into frames sequences keeping in mind the computational power of the devices.

- For all the datasets, combination of augmentation methods were implemented where as for some databases, dark edges of frames were taken out of frames.
- It is stated that the input to the model is subtraction of adjacent frames. This method ensures the inclusion of special movements in the input videos rather than crude pixels from each frame.
- Data augmentation is applied on the frames with the following transformations:
    - Image cropping: the image is cut with a different anchor corner each time.
    - Image transpose: after the image cropping, the image transpose is done. It is performed while fit generator processes.

Step 3: Splitting the dataset: While training the model, the most common problem is overfitting. This phenomenon takes place when the model gives better performance on the dataset used for its training but give poor results when it is given new or unseen dataset to learn. Most probably, the model could be able to learn only few inputs rather than learning other prediction parameters.

Second is the problem of underfitting. It arises when the model is not able to give better results even on the dataset that was used for its training. In order to deal with this problem, it is advised to split the dataset into training dataset and validation dataset. In this way, we can use the training data points to train the model and use the validation data points to check how the model performs when new or unseen data samples are fed to it.

Step 4: Performance Analysis: The training of the CNN model is required to see the performance of the model. During the step, Resnet50 CNN model is trained with dataset hockey, violent flow and movies clips one by one. The results obtained at the end of each epoch are recorded in csv files for further comparison. Similarly, incepV3 and VGG19 CNN models are trained. Then, each model is trained with the validation data samples of every dataset. These final results are compared to the results obtained while training the model with train data samples. The results obtained after training with validation dataset are also stored in CSV files. These results are compared to the results obtained while training the model with train data samples. The performance is evaluated by using the metrics of accuracy and loss. If the model is performing well with validation data samples then it's used for further deployments otherwise it is again trained by changing various hyper parameters such as learning rate, augmentation used, dropout values keeping the train type static.

In the next section, the results obtained by the implementation of various pre-trained CNN techniques on benchmark datasets such as hockey fight dataset, violent flow dataset and movies dataset will be discussed.

IV. DATASETS USED

In this research work, we have utilized three common datasets which are available openly for use i.e. Hockey Fights dataset [12], Movies dataset [13], and Violent Flow dataset [14]. The datasets contains data in the form of videos which has been with-drawn from closed-circuit-TV, from mobile cameras or from high-end recorders. The datasets vary in various other aspects such as quality, number of pixels and length. The datasets used for the implementation of the proposed model is shown in Table 1.

Table 1. Summary of Datasets used

Dataset	Size (MB)	Total Videos	Aggressive	Non-Aggressive	Types of Videos
Hockey	214	1000	500	500	Hockey fights players
Violent Flow	81	200	100	100	Crowded scenes from football match
Movies	159	246	123	123	Movies clips

V. RESULT

In order to compare the performance of different models, two graphs are plotted (validation accuracy versus train accuracy) and (validation loss versus train loss) for each dataset. Model overfitting and model underfitting are terms that we will use in the following section.

A. Comparison of Accuracy and Loss of all Models .

The behavior of ResNet50 model over three different datasets is shown in the figure 4. It is observed that the model is able to achieve 94.50% train accuracy where as the validation accuracy is 69.20%. For movies dataset, as indicated by Fig 4, the model achieved 94.11% train accuracy and 93.20% validation accuracy which is as good as compared to hockey data. The model was 90% able to recognize violent behavior of objects. For violent flow dataset, the train accuracy is 64.77% and validation accuracy is also 67.12%.

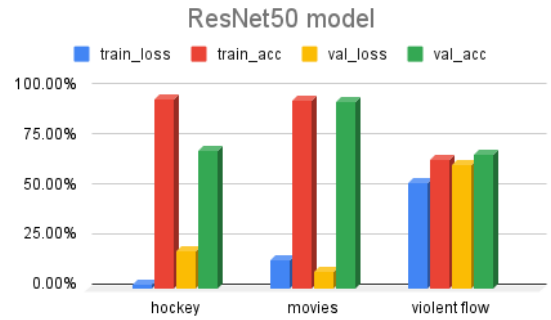

Figure 4. Comparison of performance of Resnet50 among different datasets

The behavior of InceptionV3 model over three different datasets is shown in the figure 5. As depicted by the results, the model gives 97.45% train accuracy and 98% validation accuracy for hockey dataset meaning that model is generalized well.

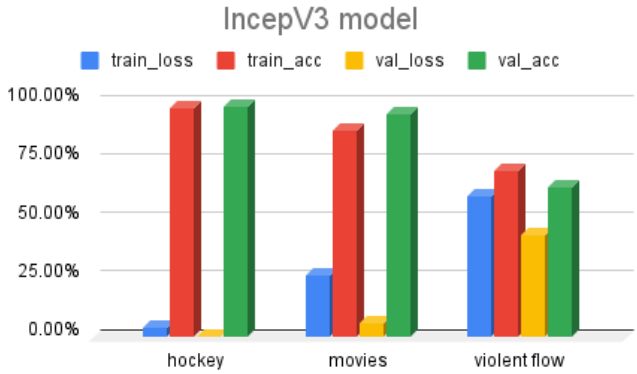

Figure 5. Comparison of performance of IncepV3 among different datasets

For movies dataset, the train accuracy is 88.24% and validation accuracy is 94.72% which is a good indicator. The model is learning well when new data is passed through it. For violent flow dataset, the train accuracy is 70.54% and validation accuracy is 63.70%.

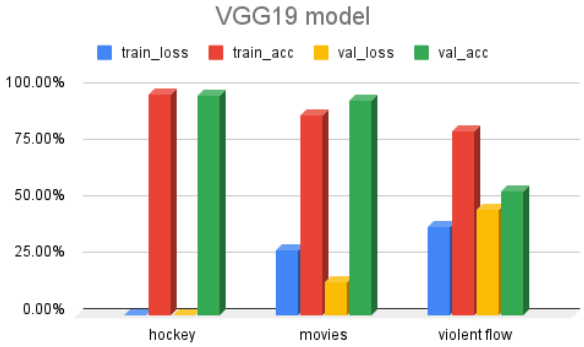

Figure 6. Comparison of performance of Vgg19 among different datasets

The behavior of VGG19 model over three different datasets is shown in the figure 6. For hockey dataset, the train accuracy and validation accuracy is 97.51% and 96.83% respectively. We need to add more data into the dataset to check its further generalizing ability.

B. Comparison of Models

Accuracy, validation accuracy, train loss and validation loss of Resnet50, IncepV3 and VGG19 is compared to check best performing model. The Figure 7 shows the plot which depicts that ResNet50 model is the best performing CNN model with accuracy of 90%. The InceptionV3 CNN model yielded 89% of accuracy which is close to ResNet50. Further, VGG19 yielded very poor performance results of only 79% of accuracy.

Figure 7. Comparison of performance of ResNet50, IncepV3 and Vgg19

VI. CONCLUSION AND FUTURE WORK

In this research, we used three different datasets i.e. hockey fights dataset, movies dataset and violent flow dataset on three different pre-trained CNN models. The video clips from datasets were converted into pre-processed frame data sequences. Then, the datasets were divided into training dataset and validation dataset. The models through which training dataset was passed were VGG19, ResNet50 and IncepV3. Each model contains Convolution Neural Network (CNN) linked to convolutional LSTM (ConvLSTM) layer. Further, the validation dataset was used to test the model efficiency. The results obtained are satisfactory and it was found that ResNet50 is the best performing CNN model with accuracy of 90%. The InceptionV3 CNN model yielded 89% of accuracy which is close to ResNet50. Further, VGG19 yielded very poor performance results of only 79% of accuracy.

In the future the efficiency of the proposed model can be checked on various other benchmark datasets or in real time surveillance. Smart data preprocessing of the video's frames can help the model to perform better in complex scenes. It is suggested to expand the model to more complex violence real time scenarios and appliances.

REFERENCES

[1] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

[2] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu. "Deep learning for sensor-based human activity recognition : Overview, challenges, and opportunities." ACM Computing Surveys. 2021 ; Vol. 54, No. 4. pp. 1-40.

[3] S. B. Atitallah, M. Driss, W. Boulila, and H. B. Ghezala. 2020. "Leveraging deep learning and IoT big data analytics to support the smart cities development: Review and future directions". Computer Science Review 38 (2020), 100303

[4] Abdali, Al-Maamoon R. and Rana F. Al-Tuma. "Robust Real-Time Violence Detection in Video Using CNN And LSTM," 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, 2019, pp. 104-108, doi: 10.1109/SCCS.2019.8852616.

[5] Patel, Mann. "Real-Time Violence Detection Using CNN-LSTM." arXiv preprint arXiv:2107.07578 (2021).

[6] Ullah, Fath U Min, Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. 2019. "Violence Detection Using Spatiotemporal Features with 3D Convolutional Neural Network" Sensors 19, no. 11: 2472. https://doi.org/10.3390/s19112472.

[7] Zhang Xin, Zhiguo Jiang, and Haopeng Zhang. "Real-time 6D pose estimation from a single RGB image." Image and Vision Computing Vol. 89, 2019: pp 1-11.

[8] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in IEEE Access, vol. 6, pp. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

[9] Hammami, Samir Marwan, and Muhammad Alhammami. "Vision-based system model for detecting violence against children." MethodsX, Vol. 7, (2020): 100744.

[10] Zaidenberg S, Boulay B, Bremond F. "A generic framework for video understanding applied to group behavior recognition." In: 2012 IEEE ninth international conference on advanced video and signal-based surveillance (AVSS), 2012, pp.136–142

[11] Sharma, Sarthak, et al. "A fully integrated violence detection system using CNN and LSTM." International Journal of Electrical & Computer Engineering (2088-8708) Vol. 11, Issue 4 (2021) p3374.

[12] Nievas, E.B., Suarez, O.D., Garciaand,G.B., Sukthankar, R. "Hockey fight detection dataset", pp. 332–339, Aug. 2011, [Online]. Available: http://visilab.etsii.uclm.es/personas/oscar/FightDetection/

[13] T. Hassner, Y. Itcher and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 2012, pp. 1-6, doi: 10.1109/CVPRW.2012.6239348.

[14] Nievas, E., Suarez, O., Garcia, G., & Sukthankar, R. “Movies Fight Detection Dataset”. In Computer Analysis of Images and Patterns. (2011) pp. 332–33.