Dhirendra Tripathi, 2024, 12:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Deep Learning Techniques for Enhanced Violence Detection in Surveillance Systems

M. Tech Scholar Dhirendra Tripathi, HoD Nagendra Patel
Department of Computer Science Engineering
Rewa Institute of Technology, Rewa, India

Abstract- In the field of video content analysis, accurately distinguishing between 'Violence' and 'NonViolence' presents a significant challenge due to the dynamic and complex nature of video data. While traditional models like ResNet50 and MobileNetV2 have demonstrated strong performance in image classification, they often fall short in handling the temporal dependencies present in video sequences. To overcome this limitation, we propose a hybrid deep learning approach that leverages the spatial feature extraction capabilities of InceptionV3 along with the temporal pattern recognition strengths of Long Short-Term Memory (LSTM) networks. Our method begins with rigorous data preprocessing, which includes noise reduction and effective feature extraction. This is followed by a systematic model training process that optimally combines the features extracted by InceptionV3 with LSTM's sequence modeling capabilities. Performance evaluations indicate an impressive accuracy of 99.86% and a validation accuracy of 92.48%, significantly outperforming the other models tested. These results not only affirm the effectiveness of the hybrid model in video classification tasks but also highlight its potential for broader real-world applications that require nuanced content analysis.

Keywords- Video Content Analysis, Violence Detection, InceptionV3, LSTM Networks, Deep Learning, Hybrid Model.

I. INTRODUCTION

The enormous impact on safety, amusement, and social media has propelled automatic video categorisation to the forefront of AI and ML research. Compared to older computer vision approaches, deep learning has made it possible to acquire and extract complex characteristics from raw data. Sorting video elements into pre-defined categories is the backbone of video content classification. This approach, however, is difficult. The geographical and temporal data included in videos give them a high-dimensional quality. Everything within a frame, including people, places, and things, is considered spatial content. Videos capture the movement and motion across time, allowing viewers to get a feel for the environment. Traditional machine learning methods failed in this respect, since they need features that were hand-crafted and substantial subject expertise in order to get satisfying results. The field was turned upside down using Convolutional Neural Networks. These networks excel in image classification because they can learn spatial feature hierarchies on their own from input picture data. When it comes to video classification, Convolutional Neural Networks (CNNs) like ResNet50 and MobileNetV2 struggle, although they excel at analysing still photos. The main reason for this is because they failed to take temporal dependencies, namely changes from one frame to the next, into account. The difficulty or degree of difficulty.

© 2024 Dhirendra Tripathi. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

The complexity of human behaviours makes it more difficult to recognise them, including violent and nonviolent exchanges. The model's ability to understand complicated behaviours depends on its ability to correctly analyse each frame and understand the sequence of frames. While simple actions may be easily identified, automated systems struggle when faced with ambiguous contexts. To find patterns in data streams, recurrent neural networks (RNNs) with long short-term memory (LSTMs) are trained. Because they remember and make use of previous data, LSTMs work well for sequential prediction. In order to comprehend the material's evolution over time, LSTMs may examine the films' temporal aspects. Video categorisation is not a good fit for LSTMs because of their limitations in handling high-dimensional spatial data. Consequently, hybrid models that combine LSTM sequence modelling with CNN feature extraction have been developed. The models get their understanding of video material from the sequential analysis of data extracted at the frame level by Convolutional Neural Networks (CNNs) and stored in Long Short-Term Memory (LSTM) networks. Our hybrid approach combines LSTM networks for efficient temporal dynamics capture with InceptionV3, a powerful convolutional neural network well-known for picture recognition. In order to identify objects, shapes, and patterns in video frames, this model use the InceptionV3 architecture. The spatial context of these features allows LSTM layers to gradually comprehend the video's event sequence. Extensive data preparation is necessary for this method. In order to get the data ready for the model, the first step is to clean the video frames, lower the noise level, and standardise the data. Following preprocessing, the model is trained using samples that have labels. In order for the model to get reliable classification data, labelling is crucial.

It is necessary to optimise computer resources in order to train the model's complicated architecture efficiently. In order to do this, it is necessary to build an efficient network architecture, choose suitable hyperparameters, and guarantee that the training data adequately represents the complex and varied nature of the model's actual applications. Once trained, metrics assess the model's accuracy and generalisability. Metrics for accuracy show the proportion of right predictions, whereas loss measurements assess how well the model matches the real labels. Training and validation are carried out on these datasets, with the validation dataset acting as a stand-in for assessing the model's efficacy on unknown data. Our hybrid model outperformed both CNNs and LSTMs in our testing. It is clear that the InceptionV3 and LSTM combination has learnt the basic patterns without experiencing overfitting to the training data, as it has shown outstanding performance on the validation set and remarkable training accuracy.
When it comes to classifying video material, hybrid deep learning models work better. One possible solution to the problem with Convolutional Neural Networks (CNNs) when it comes to processing time-related data is the use of Long Short-Term Memory (LSTM) networks. We can now create systems that can comprehend video data with ease because of this. Automated video analysis might be taken to the next level with the help of these models, opening up new possibilities for data categorisation in a wide range of sectors and on social media.

II. REVIEW OF LITERATURE

Online violent entertainment was the focus of our content study. This study uses a dataset of 2,520 films hosted on YouTube to compare content created by amateurs and professionals. Sorting the films into categories is done using a combination of user ratings, popularity, and random sampling. Several criteria, including the characteristics of the perpetrator and victim, the reasoning for the violence, and its consequences, were considered in comparing the violent frequency and context of these YouTube video categories to studies on television violence. Compared to television, YouTube is more safer. This event occurred in a more disturbing setting and had real-life consequences, in contrast to violent television shows. Comparisons made after the fact showed that different video genres and producers portrayed violent content differently [1].

We must carefully examine the material that infants and toddlers see on YouTube since their use of the site is on the increase. Parents and teachers created YouTube channel indexes, so it's hard to tell what newborns and toddlers are watching. Using films found on YouTube, the study examined how infants' brains, emotions, and social development progressed. Individual measures of language impairment, aggressiveness (verbal and physical), empathy, emotional expression, management of emotions, depiction of prosocial and antisocial behaviour, and prosocial communication were evaluated in the research. This led to high levels of psychological suffering and minimal levels of physical violence. Emotional expressiveness, emotional awareness, and prosocial behaviour are more beneficial for early child development than destructive language, verbal hostility, or physical violence [2].

Careful management of both physical and digital security measures is required. One reliable way to keep dangerous locations safe is to install surveillance cameras. They are able to record vast amounts of video material and provide constant monitoring, not found in people. Quickly analysing this film will allow us to spot any irregularities and take necessary corrective measures. In order to solve modern security issues, automatic video and picture identification is essential. It is feasible to quickly detect and react to questionable behaviours shown in images and videos by using deep learning (DL) or machine learning (ML) approaches. Data and effort are needed for action recognition, which is an important part of violence detection. War, rioting, stone throwing, and general animosity are the only things that my work acknowledges [3].

Video analysis has been greatly improved thanks to computer vision technology and automated surveillance systems, which has greatly improved public and industrial security. This is especially true in the fields of detecting aggression, behavioural analysis, and human activity identification. Recent developments have not alleviated the processing limitations that prevent real-time surveillance systems from effectively detecting and assessing violent situations. Whether in a public or private setting, our IIoT-enabled VD-Net (Violence Detection Network) AI platform can detect hostile behaviour. The model zeroes down on critical input sequence properties by using lightweight ST-TCN blocks and bottleneck layers. Based on the qualities they have learnt, classifiers can discriminate between acts that are aggressive and those that are peaceful. In the event that our system identifies any type of aggression, it need to swiftly notify the proper authorities. We demonstrated that our strategy enhances the state-of-the-art accuracy by 1-4 percentage points by conducting extensive evaluations utilising both surveillance and non-surveillance datasets. [4].

The COVID-19 pandemic has coincided with a sharp rise in violent dating incidents. This article takes a look at how young people are being affected by the normalisation of dating violence on social media. By examining the "pretend to punch your girlfriend" fad on TikTok, this study seeks to understand how young persons interpret violence in online dating. How does their emotional response show that they have changed their mind on dating violence and relationship equality? Young people's "feeling rules" are affected by violent relationship circumstances, according to this mixed-method research. This study concludes by offering platform design methods to reduce violent dating on social media [5].

Extensive research and frequent discussions in rational choice theory focus on the relationship between the government and its opponents. This work aims to provide a thorough knowledge of dissident ideology by using computational tools to analyse ethnographic interviews. Dissidents, according to the research, are more likely to resort to violent means in response to violent persecution and more likely to resort to peaceful means in response to nonviolent persecution. Opting for government involvement or security measures is more popular than protesting. These findings add to the growing body of evidence demonstrating the two-way street between governments and dissidents, and they highlight the similarities in thought processes between political opposition and cooperation, a phenomenon often associated with

tit-for-tat dynamics. In addition, these results show that social contagion and perceived relative hardship are not the main factors that motivate resistance; rather, it is government persecution. Despite how repressive nations seem, heuristic reasoning suggests that dissidents might be more receptive to government change and cooperation than that [6].

One of the many uses for human aggression detection is in security cameras. One way to reduce crime rates and save lives is to be able to recognise violent situations in real-time. When it comes to ideas and studies, precision is usually valued more than efficiency and usefulness. Our technique is very accurate and efficient because it can identify aggressive human behaviour as it happens. Three modules make up the proposed model: the Spatial Motion Extractor (SME) for frame-level region extraction, the Short Temporal Extractor (STE) for frame-level temporal feature extraction of fast movements, and the Global Temporal Extractor (GTE) for frame-level long-lasting feature extraction and model tuning. To evaluate the plan's efficacy and efficiency, we ran an evaluation in real time. The RWF-2000, Movies, and Hockey datasets demonstrate the method's superiority. In order to evaluate the model's real-time capabilities, the VioPeru dataset was created utilising recordings of violent and non-aggressive episodes captured by Peruvian video security cameras. Using this dataset, our model achieved its best performance. While video violence detection (VVD) is crucial for security cameras, it becomes more challenging in densely populated areas due to the complexity and diversity of violent incidents. A quick, disorderly, and disorganised advance is a common trait of violent acts. When it comes to analysing violent video clips, we provide a novel method called Angle-level Co-occurrence. This technique makes use of a matrix that efficiently extracts the important details from these videos. The video volume model we're using has a rank-3 tensor with a single-plane fibre in it. The distribution of identical fibre pairs in one plane of the rank 3 tensor may be captured using ALCM. Two different quantised angle values that occur simultaneously between a fibre and its neighbouring fibres are measured. In order to build a complete TOP-ALCM that describes volume violence in detail, we compute three ALCMs for three planes that are perpendicular to each other. We also provide a DL-based system that classifies using TOP-ALCM features including entropy, homogeneity, and energy, in addition to a traditional VVD approach that uses CNN directly. In terms of experimental performance, TOP-ALCM surpasses state-of-the-art VVD algorithms [8].

The most recent occurrences of violence against children (defined here as viewers less than 17 years old) on primetime television are thoroughly examined in this research. Seven hundred sixty-five primetime episodes from twenty-one broadcast and cable networks were analysed using the same sample approach and codebook as the first National Television Violence Study in 2016 and 2017. Our goal was to compare and contrast the prevalence and kind of violent material by isolating children's and adult-oriented television shows. Although it is still higher than in adult television, the percentage of violent material in juvenile programming has dropped dramatically over the last 20 years. Kids' TV shows idealise violence, but they do it in a way that's more sanitised and trivialised than what viewers see in adult shows. Things to think about while teaching kids about aggression [9].

We analysed 540 news stories to find out how well the media portrays violence and how it affects viewers' aggression, according to scientific research. There has been mounting evidence over the last 30 years linking exposure to violent news stories with an increase in violent crime. After the year 2000, the tone quickly went back to neutral. Media type (e.g., television vs. video games), number of independent sources consulted, and gender of the journalist are all potential causes of this shift. Considering the potential effects on readers of this news article [10]. The public has different views about victims and demographic groups depending on how crimes are portrayed in the media. The media's power to perpetuate harmful stereotypes and downplay transgender people's experiences of assault only serves to amplify the enormous challenges that this group faces. The 316 news items that covered the

27 transgender fatalities in the United States in 2016 are analysed in this research. It examines the systemic organisation of transphobia, the use of terminology that affirms or denigrates transgender identities, and the positive and negative representation of transgender people. We cover a lot of ground, including results and potential directions for future research [11].

Recently, there has been a meteoric rise in scholarly interest in monitoring systems. Schools, hospitals, businesses, and roadways equipped with security cameras may record important events and movements, which can be utilised for a variety of purposes, including incident prediction, online activity surveillance, data analysis with targeted goals, and intrusion detection. Research in this area has led to the development of deep learning architectures that can detect instances of violent crime in progress using video surveillance in real-time. Using Deep Recurrent Neural Networks (DRNN) and spatio-temporal (ST) classification, the purpose is to collect video material from live surveillance systems at crime scenes and analyse the features. After the initial input was transformed into video frames, the attributes were extracted and organised. It may spot violent or aggressive actions in real time and spot out-of-the-ordinary occurrences. Our method is trained and tested on the massive UCF Crime anomaly dataset. With an F-1 score of78%, recall of80%, accuracy of98%, and precision of 96% on real-time datasets, the suggested method performs well [12].

Everyone, regardless of age, may safely peruse the web thanks to content screening. As content goes through moderation, human moderators mark it as violent or nonviolent. Graphic depictions of violence may have a significant psychological and emotional effect on content moderators. This research introduces a machine learning system that can distinguish between violent and non-violent films according to the level of violence they include. It uses data from both the senses of hearing and seeing. Separating the visual and auditory elements is the first step. In terms of loudness, an audio classifier can tell the difference between very loud and very soft sounds. In order to determine how violent a video is, video classifiers use the results of an audio classifier that labels the noises as peaceful [13].

III. PROPOSED METHOD

1. Pseudocode for Predict Video Content as 'Violence' or 'NonViolence' using InceptionV3 and LSTM

Load Models and Set Classification Threshold

- Load the pre-trained InceptionV3 and LSTM models.
- Define the threshold for categorizing videos as 'Violence' or 'NonViolence'.

Extract and Process Video Frames

- Extract frames from the input video.
- Resize the frames and normalize the pixel values for processing.

Generate Video Features

- Use InceptionV3 to extract features from each frame.
- Aggregate these features into a single feature vector representing the entire video.

Sequence Preparation and Prediction

- Prepare the feature vector as input for the LSTM model.
- Pass the sequence through LSTM to obtain a probability score for violence detection.

Classify and Output Results

- Compare the probability score against the predefined threshold.
- Classify the video as 'Violence' or 'NonViolence' based on the result.
- Output the classification label and, if needed, the probability scores.

IV. IMPLEMENTATION AND RESULT DISCUSSION

1. Dataset

The collection is organised into two distinct directories: 'NonViolence' and 'Violence'. A thousand films depicting various real-life situations, including eating, sports, and singing, all free of

violence, make up the 'NonViolence' category. The 'Violence' category, on the other hand, has a thousand films that depict violent actions in various contexts. This dataset is specifically designed to train and evaluate machine learning models that are tasked with distinguishing between violent content and non-violent acts, thanks to its accurate categorisation [14].

2. Illistrative Example

Figure 1: Displays a series of frames from a film where the word "Viole" is superimposed in a prominent red font, partly obstructing the picture.

Figure 1 shows a sequence of still images from a movie in which the word "Viole" is placed in a bold red script, partially blocking the scene's perspective. Viewed from a distance, the frames depict an outside setting like a park or street. The placement of the red text across the frames gives the impression that something is moving across the image, perhaps a person or an object, or that there is a focal point. With the identical background and lighting in every photo, it's likely that the series captures a fleeting moment in time. With the added text, it seems like it may be part of a video analysis or tracking presentation, maybe related to video editing or motion detection.

Figure 2: A succession of video frames, where each frame is superimposed with the phrase "NonViolence" in a green typeface.

Picture 2 shows a video sequence with the word "NonViolence" placed in green lettering on top of each shot. Outdoors on a track, we see a man working out, maybe performing some jogging exercises on his knees. As the word "NonViolence" is superimposed on every frame of the clip, it's possible that it's a tag or the outcome of a machine learning or video analysis system's attempt to categorise the footage. It seems like a sports or fitness video from the language, the setting, the man's attire, and his movements put together. The analysis of this film has shown that it does not contain any violent content.

3. Comprative Result

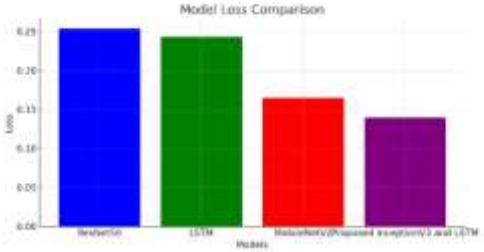

Figure 3: Comparison of model losses

"Model Loss Comparison" in figure 3 shows a graphical comparison of the loss values of four models: ResNet50, LSTM, MobileNetV2, and a suggested InceptionV3 LSTM hybrid. Among the models tested, the InceptionV3 and LSTM model produces the best results in terms of loss minimisation, followed by MobileNetV2, LSTM, and ResNet50.

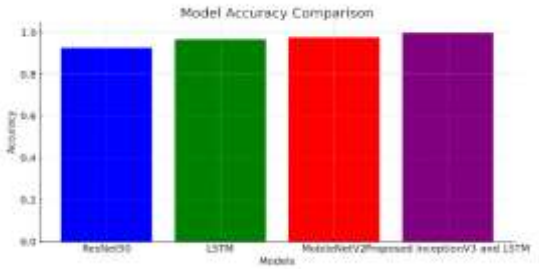

Figure 4: Comparison of model accuracy

Figure 4's "Model Accuracy Comparison" examines the performance of four different models: ResNet50, LSTM, MobileNetV2, and a hybrid model that combines InceptionV3 and LSTM. The accuracy of all the models is rather excellent, although the suggested InceptionV3 and LSTM model are a little better than the others. This shows that when

compared to ResNet50, LSTM, and MobileNetV2, the suggested model performs somewhat better in terms of accurate data classification.

V. CONCLUSION

The use of deep learning algorithms to the categorisation of video content has shown its capacity to identify intricate patterns and make predictions about the results. The performance of ResNet50, LSTM, MobileNetV2, and a hybrid model that was supposed to be a combination of InceptionV3 and LSTM was distinct, with the hybrid model outperforming the others in terms of training and validation indicators. Its validation result demonstrates that it is resilient in learning from training data and generalising to data that it has not before seen. Its improved accuracy and decreased loss metrics illustrate this, respectively. Cleaning, segmenting, and extracting features are the steps that we do to get the dataset ready for learning processes. Both the frame details and the context of the video sequence are captured by the hybrid model, which makes use of the spatial pattern recognition capabilities of InceptionV3 and the sequential data processing capabilities of LSTM. When it comes to data preprocessing and prediction, the deep learning method that modern AI systems use demonstrates their level of expertise. This method enables the sophisticated classification of video material as either "Violence" or "NonViolence," so indicating the potential for these models to be used in surveillance, content management, and media study.

REFERENCES

1. Weaver, Andrew J., Asta Zelenkauskaite, and Lelia Samson. "The (non) violent world of YouTube: Content trends in web video." Journal of Communication 62, no. 6 (2012): 1065-1083.
2. Choi, Yun Jung, and Changsook Kim. "A content analysis of cognitive, emotional, and social development in popular kid's YouTube." International Journal of Behavioral Development (2024): 01650254241239964.
3. Jain, Mahaveer, and Mukesh Kumar. "A Review of Violence Detection Techniques." In 2024 2nd International Conference on Computer, Communication and Control (IC4), pp. 1-6. IEEE, 2024.
4. Khan, Mustaqeem, Abdulmotaleb El Saddik, Wail Gueaieb, Giulia De Masi, and Fakhri Karray. "VD-Net: An Edge Vision-Based Surveillance System for Violence Detection." IEEE Access 12 (2024): 43796-43808.
5. Maddocks, Sophie, and Fallon Parfaite. ""Watch me pretend to punch my girlfriend": exploring youth responses to viral dating violence." Feminist Media Studies 24, no. 1 (2024): 103-118.
6. Dornschneider-Elkink, Stephanie, and Nick Henderson. "Repression and dissent: How tit-for-tat leads to violent and nonviolent resistance." Journal of Conflict Resolution 68, no. 4 (2024): 756-785.
7. Huillcen Baca, Herwin Alayn, Flor de Luz Palomino Valdivia, and Juan Carlos Gutierrez Caceres. "Efficient human violence recognition for surveillance in real time." Sensors 24, no. 2 (2024): 668.
8. Hu, Xing, Zhe Fan, Linhua Jiang, Jiawei Xu, Guoqiang Li, Wenming Chen, Xinhua Zeng, Genke Yang, and Dawei Zhang. "TOP-ALCM: A novel video analysis method for violence detection in crowded scenes." Information Sciences 606 (2022): 313-327.
9. Martins, Nicole, and Karyn Riddle. "Reassessing the risks: An updated content analysis of violence on US children's primetime television." Journal of Children and Media 16, no. 3 (2022): 368-386.
10. Ferguson, Christopher J., Anastasiia Gryshyna, Jung Soo Kim, Emma Knowles, Zainab Nadeem, Izabela Cardozo, Carolin Esser, Victoria Trebbi, and Emily Willis. "Video games, frustration, violence, and virtual reality: Two studies." British journal of social psychology 61, no. 1 (2022): 83-99.
11. Osborn, Max. "US news coverage of transgender victims of fatal violence: An exploratory content analysis." Violence against women 28, no. 9 (2022): 2033-2056.
12. Sahay, Kishan Bhushan, Bhuvaneswari Balachander, B. Jagadeesh, G. Anand Kumar, Ravi Kumar, and L. Rama Parvathy. "A real time

crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques." Computers and Electrical Engineering 103 (2022): 108319.

13. Rishab, K. S., P. Mayuravarsha, Yashwal S. Kanchan, M. R. Pranav, and Roopa Ravish. "Detection of Violent Content in Videos using Audio Visual Features." In 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), pp. 600-605. IEEE, 2023.
14. Cheng, Ming, Kunjing Cai, and Ming Li. "RWF-2000: an open large scale video database for violence detection." In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4183-4190. IEEE, 2021.