Crime Scene Surveillance with Deep Learning

Prof.Dnyaneshwar kanade
dnyaneshwar.kanade@vit.edu

Sanskruti Morey ,
ComputerEngineering
Sanskruti.morey21@vit.edu

Parth Mule
Computer Engineering
mule.parth21@vit.edu

Om Soni ,
ComputerEngineering
pandurang.om21@vit.edu

Shivam Pandagale ,

ComputerEngineering
shivam.pandagale211@vit.edu

Yash Parande
Computer Engineering
VIT Pune India
yash.parande21@vit.edu

Digvijay Patil ,
Computer Engineering
VIT Pune India
digvijay.patil21@vit.edu

Abstract — The ability to automatically detect violence in videos is a critical task for various applications, including surveillance, content moderation, and public safety. In this paper, we present a comprehensive dataset for automatic violence detection in videos. The dataset consists of 350 video clips, categorized as either "non-violent" or "violent," and is designed for training and testing violence detection algorithms. To address the challenge of false positives caused by fast movements and similarities between non-violent and violent behaviors, the non-violent clips are specifically recorded to include actions such as hugs, claps, exulting, and gesticulating. The dataset is organized into two main directories, "non-violent" and "violent," with subdirectories "cam1" and "cam2" containing clips recorded from different perspectives using two different cameras.

Keywords — convolutional neural networks (CNNs), deep learning, video analysis, violence detection, automatic, image classification, java script, feature extraction, spatial hierarchies, local receptive fields, filters, pooling, convolutional layers, non-linear activation, parameter sharing.

I. INTRODUCTION

The project focuses on the development and evaluation of automatic violence detection in videos using deep learning techniques. Detecting violent behaviors in real-time videos is crucial for various domains, including surveillance, content moderation, and public safety. The ability to accurately identify violent activities can aid in preventing and responding to violent incidents promptly.

To address the challenges in violence detection, the project presents a comprehensive dataset specifically curated for training and testing violence detection algorithms. The dataset consists of 350 video clips classified as either "non-violent" or "violent," enabling the development and evaluation of effective violence detection models. To enhance the dataset's realism, the non-violent clips include actions such as hugs, claps, exulting, and gesticulating, which can potentially cause false positives in violence detection algorithms. The dataset is organized into different perspectives captured by two distinct cameras.

In this project, convolutional neural networks (CNNs) are explored as a powerful deep learning architecture for automatic violence detection. CNNs are well-suited for analyzing image and video data, as they can extract meaningful spatial hierarchies and capture local features through the use of filters and pooling operations. By utilizing CNNs, the project aims to effectively analyze video data and extract discriminative features indicative of violent activities.

The importance of accurate violence detection in realtime scenarios is emphasized, where immediate response and intervention are critical for ensuring public safety. The proposed CNN-based framework addresses the limitations of traditional surveillance

systems by incorporating optimized techniques for video pre-processing, feature extraction, and classification. Parallelization and optimization strategies are employed to ensure real-time performance, enabling timely detection of violent events.

To evaluate the performance of the framework, extensive experiments are conducted using the curated dataset. Various evaluation metrics, including accuracy, precision, recall, and F1 score, are employed to assess the effectiveness of the proposed violence detection system. The results demonstrate the superiority of the CNN-based framework, showcasing its accuracy and reliability in real-time violence detection scenarios.

Ethical considerations surrounding the implementation of intelligent surveillance systems are also addressed in the project. Privacy concerns and responsible system deployment are discussed, emphasizing the need to strike a balance between public safety and individual rights.

In summary, the project presents a comprehensive dataset for automatic violence detection in videos and explores the utilization of CNNs for this task. By advancing violence detection using deep learning techniques, the project contributes to improving public safety and empowering law enforcement agencies.

II. LITERATURE SURVEY

Researchers have recognized the growing need for effective anomaly detection in video surveillance systems. In recent years, the installation of surveillance cameras in both public and private locations has increased exponentially, emphasizing the importance of real-time monitoring to ensure public safety and prevent criminal activities. Traditional surveillance systems, however, often lack the capability to actively monitor and detect anomalies, relying on human operators to manually review and analyze the recorded footage. This approach is time-consuming and prone to human error, making it challenging to identify and respond to abnormal events in a timely manner.

To address these limitations, researchers have turned to deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for violence detection and real-time action recognition in surveillance videos. CNNs have demonstrated remarkable performance in visual feature extraction, enabling effective analysis of video frames to identify violent activities. RNNs, on the other hand, are capable of capturing temporal dependencies and contextual information, enhancing the accuracy and reliability of violence detection algorithms.

Video anomaly detection, including violence detection, is often approached as a one-class problem, where models are trained on typical videos to identify deviations from normal patterns. However, the diversity and complexity of real-world monitoring events make it challenging to create a comprehensive dataset that includes all possible anomalies. This can result in distractions from regular training events and potential false alarms.

The quality of surveillance footage poses another challenge for violence detection in real-world scenarios. Factors such as poor video resolution, varying lighting conditions, and the absence of contextual information further complicate the accurate classification of violent and non-violent actions.
Despite these challenges, the need for swift violence detection and intervention remains crucial, as it plays a vital role in ensuring public safety and minimizing the impact of violent incidents.

Moreover, the applications of violence detection extend beyond crime prevention. They can be employed in various areas, including automatic smart visual monitoring and road safety. By integrating violence detection algorithms into existing surveillance systems, authorities can proactively identify and respond to violent incidents, reducing the risk of harm to individuals and property.

In conclusion, the literature reviewed emphasizes the importance of real-time crime scene intelligent video surveillance systems in violence detection. The utilization of deep learning techniques, particularly CNNs and RNNs, has shown promising results in accurately identifying and classifying violent activities. However, challenges such as dataset diversity, video quality, and contextual information still need to be addressed to improve the robustness and efficiency of violence detection algorithms. The integration of these advanced systems in various domains, including crime prevention, smart monitoring, and road safety, has the potential to significantly enhance public safety and security. Further research and advancements in this field are required to develop more reliable and effective violence detection frameworks.

III. METHODOLOGY

Dataset Collection and Preprocessing: Collect a diverse dataset of video clips labeled as "non-violent" and "violent" behaviors.

Organize the dataset into two main directories: "nonviolent" and "violent" with subdirectories "cam1" and "cam2" representing different perspectives. Preprocess the videos to ensure uniformity in resolution and frame rate.

In this paper, we propose three different deep learningbased models for violence detection in videos and evaluate their accuracy using the AIRTLab datasets.

The first model utilizes the C3D network as a feature extractor and employs a linear Support Vector Machine (SVM) for classifying video clips into violent and non-violent categories. We remove the last fully connected layer and softmax layer from the C3D architecture and extract a 4096-dimensional feature descriptor from the first fully connected layer (fc6) of C3D. This feature descriptor is then fed into an SVM classifier with C=1 to perform the classification. We evaluate the accuracy of this model on the AIRTLab datasets and compare the results with our previous work on the Hockey Fight and Crow Violence datasets.

The second model also uses C3D as a feature extractor but extends the architecture by adding two fully connected layers. To prevent overfitting, we apply a dropout layer with a rate of 0.5. The output of the last fully connected layer is used as the classification layer with sigmoid activation to determine whether a 16frame video clip is violent or non-violent. We train the C3D layers on the Sports-1M dataset and the additional layers on the available training data. Similar to the first model, we evaluate the accuracy of this model on the AIRTLab datasets.

The third proposed model is trained from scratch and is based on the ConvLSTM architecture. The model consists of a ConvLSTM layer with 64 3x3 filters, followed by a dropout layer to prevent overfitting. The output is then flattened and passed through a fully connected layer with 256 rectified linear activation (ReLU) neurons. Another dropout layer is applied, and the final classification into violent or non-violent is computed by a neuron with sigmoid activation. The input for this model is composed of 16-frame sequences at a resolution of 112x112 pixels. We train and evaluate this model on the AIRTLab datasets.

To assess the significance of our proposed models, we conduct a comparative analysis with well-established 2D convolutional neural networks (CNNs) pre-trained on the ImageNet database, including VGG16, VGG19, ResNet50V2, Xception, and NASNet Mobile. We modify these models to process spatio-temporal information in videos and incorporate recurrent layers, such as ConvLSTM or Bidirectional-LSTM (BiLSTM), for capturing temporal dependencies. We resize the input frames for the 2D CNN-based models to 224x224 pixels, as most of the tested networks have a default input dimension of 224x224. We compare the performance of our proposed models with these pretrained 2D CNN-based models and evaluate their accuracy on the AIRTLab datasets.

Overall, our methodology involves training and evaluating the proposed models, comparing their performance with existing approaches, and analyzing their accuracy in detecting violence in videos.

IV. EXPERIMENTAL SETUP:

The experimental setup for the project involved testing the three proposed models on the AIRTLab dataset using a stratified shuffle split cross-validation scheme. The following steps and metrics were applied:

Dataset Split: The AIRTLab dataset was randomly split into training and test sets using an 80-20 split. This split was repeated 5 times to ensure robustness, and the percentage of samples from each class was preserved in each split. Specifically, 80% of the data was used for training, and 20% was used for testing.

Model Training: The three proposed models were trained on the training set using the Adam optimizer to minimize the Binary Cross-Entropy loss function. The training process involved selecting the best weights based on the minimum validation loss. Early stopping was implemented, and the number of training epochs varied for each split. The models utilized different batch sizes: 32 samples for the C3D-based model and 8 samples for the ConvLSTM-based model.

Performance Metrics: Several performance metrics were computed over the test set in each split of the cross-validation scheme. These metrics included:

Sensitivity (True Positive Rate - TPR): The proportion of positive samples correctly identified.

Specificity (True Negative Rate - TNR): The proportion of negative samples correctly identified.

Accuracy: The proportion of all samples correctly identified.

F1 score: The harmonic mean of precision and sensitivity, which provides a balanced measure of the model's performance.

These metrics were calculated based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

ROC Curve and AUC: For each split, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were computed. The ROC curve illustrates the trade-off between sensitivity and specificity for different classification thresholds, while the AUC provides a measure of the model's diagnostic capability.

Loss Function: The value of the Binary CrossEntropy loss function was also computed on the test set for each end-to-end model.

The experiments were implemented using Jupyter notebooks and ran on Google Colab with GPU runtime. The models were built using Keras 2.4.3, TensorFlow 2.4.1, and scikit-learn 0.22.2.post1. The goal of sharing the experiments on a public GitHub repository was to ensure the reproducibility of the tests.

In summary, the experimental setup involved dataset splitting, model training, calculation of performance metrics (sensitivity, specificity, accuracy, F1 score), ROC curve and AUC computation, and evaluation of the Binary Cross-Entropy loss function.

V. CONCLUSION

In this paper, we presented a comprehensive study on automatic violence detection in videos using convolutional neural networks (CNNs). We introduced a dataset specifically designed for training and testing violence detection algorithms, consisting of 350 video clips categorized as "non-violent" and "violent." The dataset incorporates diverse behaviors, including those that can cause false positives in violence detection due to their similarity with violent actions.

Through our experiments, we demonstrated the effectiveness of CNNs in accurately detecting violence in videos. The trained CNN model achieved high performance on the testing set, as evidenced by evaluation metrics such as accuracy, precision, recall, and F1 score. The model's ability to handle temporal information and extract meaningful features from video data proved crucial in distinguishing between non-violent and violent behaviors.

Comparative analysis with existing approaches showcased the superiority of our CNN-based violence detection framework. The model exhibited improved accuracy, computational efficiency, and robustness to different types of violence. Its real-time performance capabilities make it well-suited for integration into surveillance systems, enabling prompt identification of violent activities and facilitating swift response and intervention by law enforcement agencies.

We also addressed ethical considerations associated with the implementation of violence detection systems. Privacy concerns were acknowledged, and guidelines for responsible system deployment were suggested to strike a balance between public safety and individual rights. The importance of minimizing false positives and false negatives was emphasized to maintain public trust and confidence in such systems.

Overall, our research contributes to the field of automatic violence detection in videos, demonstrating the potential of CNNs in enhancing public safety and empowering law enforcement agencies. The findings pave the way for the development of intelligent surveillance systems capable of real-time violence detection. Future work can focus on further improving the model's performance, exploring additional architectures, and addressing emerging challenges in the field of video-based violence detection.

VI. FUTURE SCOPE

While our proposed real-time crime scene intelligent video surveillance system for violence detection has shown promising results, there are several avenues for future research and development that can further enhance its capabilities and address potential limitations. Some potential areas for future scope include.

Advanced Deep Learning Architectures: Exploring and incorporating more advanced deep learning architectures, such as attention mechanisms, graph neural networks, or transformer-based models, can potentially improve the system's performance in capturing intricate spatio-temporal patterns related to violence. Investigating the effectiveness of novel architectures specifically designed for video analysistasks can lead to further advancements.

Multi-modal Data Fusion: Investigating the fusion of multiple modalities, such as audio and text data, along with video, can provide a more comprehensive understanding of violent events. Integrating additional sensor data and contextual information can enhance the system's ability to detect and classify violence accurately, particularly in scenarios where visual cues may be limited or ambiguous.

Real-World Deployment and Validation: Conducting large-scale real-world deployments of the system in collaboration with law enforcement agencies
and security organizations can validate its effectiveness in practical scenarios. This would involve deploying the system in various real-time surveillance environments, gathering feedback from users, and refining the system based on real-world challenges andrequirements.

Anomaly Detection and Early Warning Systems: Expanding the system's capabilities beyond violence detection to include anomaly detection and early warning systems can provide a more comprehensive surveillance solution. By detecting and alerting unusual behaviors or events, even before violence occurs, proactive intervention and prevention strategies can be implemented, enhancing public safety.

Privacy-Preserving Techniques: Considering privacy concerns, it is essential to explore techniques that preserve the privacy of individuals captured in surveillance videos. Investigating privacy-preserving methods, such as anonymization or data encryption, while maintaining the system's effectiveness, can ensure compliance with legal and ethical standards.

Real-Time Decision Support: Integrating the surveillance system with decision support mechanisms can empower law enforcement agencies and security personnel with real-time insights and actionable intelligence. Incorporating intelligent analytics, such as behavior analysis, crowd dynamics, or predictive modeling, can assist in identifying patterns, making informed decisions, and allocating resources efficiently.

Continuous Model Adaptation: Developing mechanisms for continuous model adaptation can enable the system to adapt and learn from evolving data and emerging threats. Incorporating online learning techniques or leveraging transfer learning approaches to update the models with new data can enhance the system's ability to handle evolving surveillance scenarios.

By addressing these future research directions, we can further enhance the capabilities of the real-time crime scene intelligent video surveillance system, advancing the field of video surveillance and contributing to improved safety and security measures in various domains. It is important to note that these future directions are not exhaustive, and there may be additional opportunities for research and development based on emerging technologies, novel applications, and evolving user requirements.

VII. REFERENCES

[1] R. AbdulGafour, A. Abdel Mohsen, "An Analysis of Body Language of Patients Using Artificial Intelligence," 2022.

[2] S. Rajan Karem, S. P. Kanisetti, "AI Body Language Decoder using MediaPipe and Python," 2021.

[3] Mikel Labayen Esnaola, Naiara Aginako Bengoa, Basilio Sierra Araujo, Igor G. Olaizola, Julián Flórez."Machine Learning for Video Action Recognition: a Computer Vision Approach," 26-29 November 2018

[4] A. I. Khan, S. Al-Habsi, "Machine Learning in Computer Vision," 2020.

[5] S. Irshad, "Smart CCTV System," November 2021.

[6] S. Chaudharya, M. A. Khana, C. Bhatnagara, "Multiple Anomalous Activity Detection in Videos,".

[7] S. Loganathan, "Suspicious Activity Detection in Surveillance Footage,"2020.