Violence Detection in Surveillance videos using Deep Learning

Deepali Pawade
Matriculation ID: 7057787

Saanvi Khanna
Matriculation ID: 7046603

Pooja Kotresh Halannavar
Matriculation ID: 7058527

Abstract

In recent years, significant progress has been made in human action recognition, with violence recognition emerging as one of the challenging research topics in computer vision. This paper presents an approach to violence detection using deep learning methods, specifically leveraging transfer learning on three state-of-the-art architectures: DenseNet, SlowFast, and I3D. By fine-tuning these pre-trained models on a comprehensive video dataset containing four types of violence—abuse, assault, arrest, and arson, we aim to evaluate and compare their performance in terms of accuracy and efficiency. The experimental results demonstrate that each model shows unique strengths and weaknesses, with significant variations in precision, recall, and F1-score. This study provides an understanding of the effectiveness of transfer learning in violence detection and highlights the most promising approaches for future research in this domain.

1. Introduction

Human assistance is often required to monitor surveillance video screens, which can lead to mistakes and neglect in identifying violent events. This highlights the need for powerful automated violence detection systems. Deep learning offers a robust approach to detecting violent behavior. We propose a comprehensive deep learning solution designed to identify violent behavior in real-time video surveillance, classifying incidents into four categories: Abuse, Arrest, Arson, and Assault.

2. Related Work

Traditionally, hand-crafted-base methods extract visual features [1, 2, 3] using a manually predefined algorithm. These overcome specific problems such as occlusions and variable illumination. Alternatively, deep-learning methods have better results and train two-stream CNNs [4], RNNs [6], 3D-CNNs [7]. These are a combination of the neural network for spatial features and LSTMs [5, 13] for temporal features.

Nam et al. [6] present a method for detecting violent scenes in videos by combining audio and visual cues. The method employs a multi-step process, starting with the extraction of audio and visual features, followed by the fusion of these features to classify scenes as violent or non-violent. The authors demonstrate the effectiveness of their approach through experiments on a dataset of video clips, showing that the fusion of audio-visual information significantly improves the accuracy of violent scene detection.

Soliman et al. [5] utilizes a pre-trained VGG-16 model for spatial feature extraction, followed by Long Short-Term Memory (LSTM) for capturing temporal information. The fully- connected layers classify the video as violent or non-violent.

Li et al. [7] detect violence in crowded scenarios using two 3D-CNNs trained over RGB frames and optical flow images. Although their method achieves high accuracy, it is computationally expensive during training.

Freire-Obregón et al. [8], introduces a novel approach to automatic violence detection using an Inflated 3D ConvNet as the backbone. The authors focus on analyzing the impact of context on classifier performance. They demonstrate that context-free footage results in a 2% to 5% deterioration in classifier performance on publicly available datasets. However, performance stabilizes regardless of the level of context restriction applied.

3. Method

The model first pre-processes the video frames, which are then passed to pre-trained models to extract features. We then apply transfer learning and train a custom classifier on these extracted features.

3.1. Dataset

Our dataset-: “Violence classification for the 4 crime categories” [9] consists of 200 videos categorized into four different classes: arrest, assault, arson, and abuse, with each class containing 50 videos.

The videos are taken from various sources such as news channel recordings, YouTube, and videos posted online on multiple other social networking sites.

Figure 1. Samples of the dataset

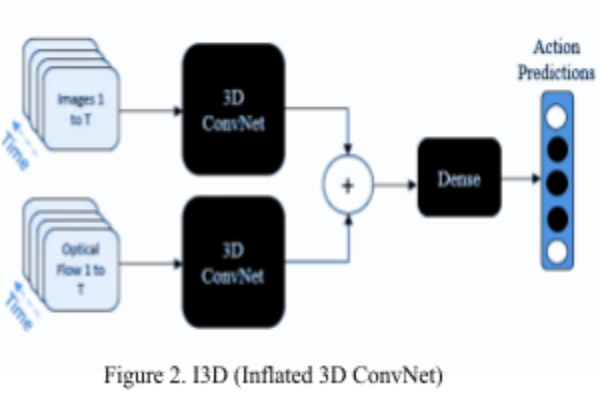

Figure 2. I3D (Inflated 3D ConvNet)

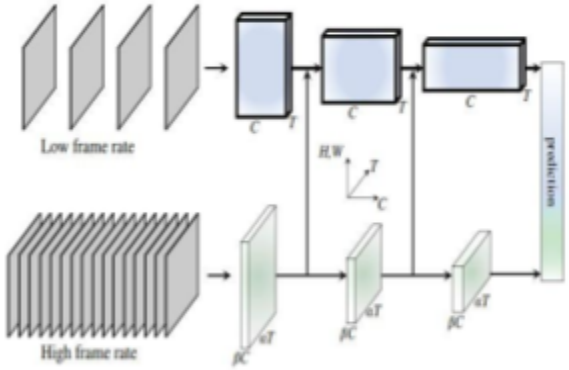

Figure 3. Slowfast Resnet-50

3.2. Pre- Processing

Initially, the dataset is acquired as a group of videos from which the frames are extracted. We extract up to 32 frames per video, resized them to 256x256 pixels and cropped them to 224x224 pixels to focus on the main content. We then applied normalization: mean (0.45, 0.45, 0.45), standard deviation (0.225, 0.225, 0.225) to match the pre- trained model's expected input format.

3.3. Feature Extraction

We utilize pre-trained models, specifically I3D (Inflated 3D ConvNets), SlowFast-ResNet50, and DenseNet121, to efficiently extract spatio-temporal features (embeddings) from videos. SlowFast and I3D, pre-trained on large action recognition datasets like Kinetics-400, enable robust feature extraction and improved performance, even when applied to violence detection with limited domain-specific data. By replacing the final classification layer with an identity layer, these models output high-dimensional feature vectors instead of class probabilities, allowing us to train a specialized classifier on top of these features.

I3D (Inflated 3D ConvNet) [10]-: I3D efficiently captures both spatial and temporal information by extending traditional 2D convolutional networks into the 3D domain. The model processes entire video segments to identify patterns in space and time, which are critical for detecting violent behavior. By inflating 2D convolutions, I3D effectively captures motion dynamics, object interactions, and scene context across multiple frames, allowing it to recognize both subtle and overt violent actions. Its ability to model complex spatio-temporal relationships makes it particularly well-suited for violence detection tasks.

SlowFast ResNet50 [11]-: Slowfast effectively captures both slow and fast temporal dynamics, which are crucial for identifying violent actions. The Slow pathway captures broader, slower movements, providing context, while the Fast pathway focuses on rapid, detailed motions typical of sudden violent acts. This dual-pathway approach allows the model to handle complex motion patterns in violent scenes. ResNet-50's balanced complexity offers good performance without excessive computational cost, making it practical for real-time applications.

DenseNet121 [12]-: DenseNets are known for their ability to extract features with less redundancy compared to other models. They perform well on identifying violence detection tasks where capturing subtle details in video frames is important. DenseNets require a low number of filters and parameters. In our method, we used DenseNet121 pre- trained on ImageNet to extract the features.

3.4. Classification

In this section, we detail the implementation of the Support Vector Machine (SVM) classifier and a Feed Forward Neural Network classifier.

To optimize the performance of the SVM model, we employed hyperparameter tuning through GridSearchCV.

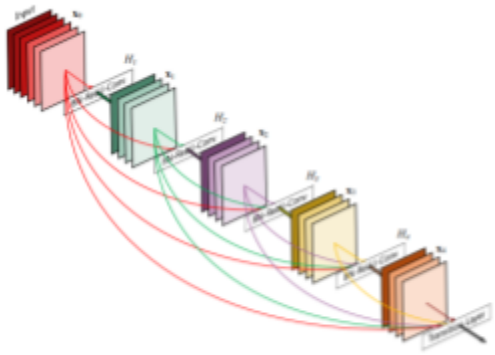

Figure 4. Dense Block

The best hyperparameters were then used to predict the test set labels. A 5-fold cross-validation approach was utilized within GridSearchCV to systematically test all possible combinations of the specified hyperparameters and to identify the optimal model.

The Forward Neural Network classifier was designed to incorporate several key components to enhance the model performance. Initially, Global Average Pooling is applied to reduce the spatial dimensions of the feature maps, simplifying the data while retaining essential information. The classifier then includes three dense layers, each with ReLU activation and dropout for regularization. Specifically, the first dense layer has 256 units with a 0.3 dropout rate, the second layer has 1024 units with a 0.5 dropout rate, and the third layer has 512 units with a 0.4 dropout rate. The final output layer utilizes SoftMax activation for multi-class classification. The model is optimized using Stochastic Gradient Descent (SGD) with a learning rate of 0.0003, ensuring efficient and effective training.

4. Experimental Results and Analysis

The dataset used in this study was divided into training (70%), validation (15%), and test (15%) sets. The training set included 140 videos (35 per label), while the validation and test sets each contained 30 videos (7 per label). This split ensured a balanced distribution, aiding effective model training, validation, and evaluation.

4.1. Evaluation Metrics

To assess the performance of the models in classification of violence, we employed several evaluation metrics such as Accuracy, Precision, Recall, F1- Score. These metrics provide a comprehensive understanding of the models' effectiveness and reliability. The final accuracy for the model fit with DenseNet121 was calculated by aggregating the results of all the four classes.

4.2. Results

The results are organized to highlight the performance of the models and the effectiveness of the methods employed.

The results indicate that the SlowFast model outperformed the others, achieving the highest accuracy demonstrating robust performance. This superior performance can be attributed to its dual-pathway architecture, which captures both fast and slow temporal dynamics allowing the model to handle the complex and varied motion patterns often present in violent scenes, leading to more accurate predictions.

In contrast, the I3D model exhibited signs of overfitting, achieving perfect accuracy on the training set (100%) but significantly lower accuracies on the test and validation sets. This indicates that the model struggled to generalize beyond the training data. The DenseNet-based custom classifier did not yield satisfactory results, with moderate training accuracy and even lower test and validation accuracies. This suggests that DenseNet may lack the necessary architecture to effectively capture spatio-temporal features in video data.

These findings emphasize the importance of model selection and the potential of the SlowFast model in accurately detecting violent actions. The results are as follows:

Model	Train Accuracy	Test Accuracy	Val Accuracy
I3D	100%	76.1%	43.33%
SlowFast	82.14%	76.67%	70.00%
DenseNet121	54.3%	43.33%	48%

Figure 5. Performance accuracy of pre-trained models

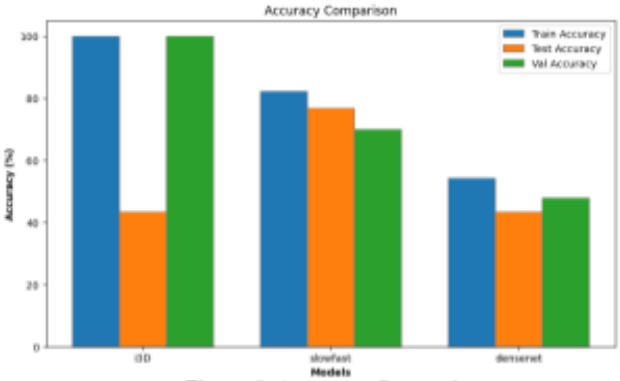

Figure 8. Accuracy Comparison

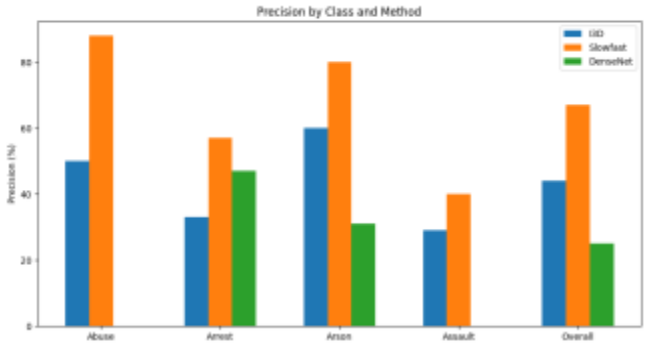

Figure 7. Precision Comparison

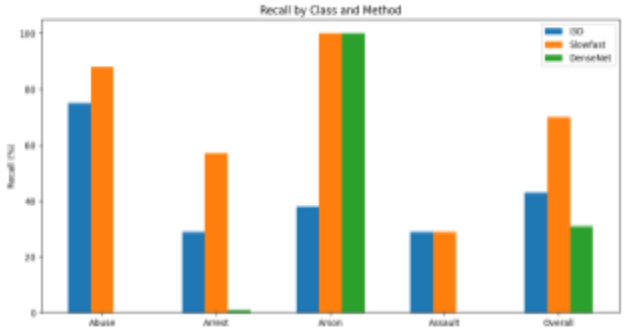

Figure 8. Recall Comparison

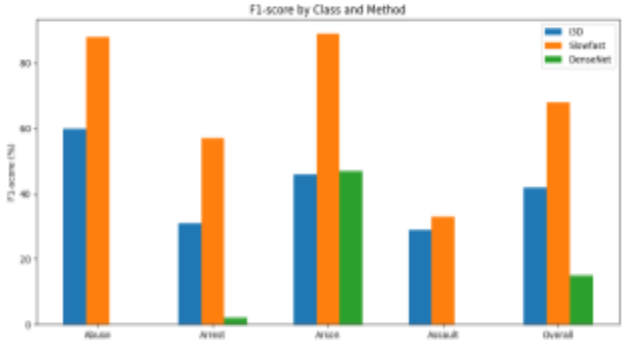

Figure 9. F1-Score Comparison

5. Conclusion

In this study, we focused on the classification of violent actions using various pre-trained models and transfer learning. Our comparative analysis revealed that the SlowFast model outperforms others, while the I3D model tends to overfit and DenseNet121 does not yield satisfactory results. We also employed Support Vector Machines (SVM) for classification, achieving promising results.

References

[1] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," Multimed. Tools Appl., vol. 75, no. 12, pp. 7327–7349, 2015.

[2] L. Lu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3535-3539.

[3] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei, "Real time violence detection based on deep spatio-temporal features," in Biometric Recognition, 2018, pp. 1-6.

[4] D. K. Ghosh and A. Chakrabarty, "Two-stream Multi-dimensional Convolutional Network for Real-time Violence Detection," arXiv preprint arXiv:2211.04255v1 [cs.CV], Nov. 2022.

[5] Soliman, M. M., Kamal, M. H., El-Massih Nashed, M. A., Mostafa, Y. M., Chawky, B. S., & Khattab, D. (2019). Violence Recognition from Videos using Deep Learning Techniques. In 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS). doi:10.1109/ICICIS46948.2019.9014714

[6] S. Vosta and K.-C. Yow, "A CNN-RNN Combined Structure for Real-World Violence Detection in Surveillance Cameras," Appl. Sci., vol. 12, no. 3, p. 1021, 2022. doi: 10.3390/app12031021.

[7] Li, C., Zhu, L., Zhu, D., Chen, J., Pan, Z., Li, X., & Wang, B. (2018). End-to-end Multiplayer Violence Detection Based on Deep 3D CNN. In Proceedings of the 2018 VII International Conference on Network, Communication and Computing (ICNCC 2018) (pp. 227-230). Taipei City, Taiwan.

[8] Freire-Obregón, David & Barra, Paola & Castrillón Santana, Modesto & De Marsico, Maria. (2022). Inflated 3D ConvNet context analysis for violence detection.Machine Vision and Applications. 33. 10.1007/s00138-021-01264-9

[9] Dataset: https://www.kaggle.com/datasets/ashwathbaskar/violence-classification-for-the-4-crime-categories

[10] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” arXiv preprint arXiv:1705.07750, 2017.

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “SlowFast Networks for Video Recognition,” Facebook AI Research (FAIR).

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269.

[13] Y. R. Elkhashab and W. H. El-Behaidy, “Violence Detection Enhancement in Video Sequences Based on Pre-trained Deep Models,” Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University, 5(1), 2023.