Violence Detection in Videos Using Deep Learning Techniques

John Vernon Baldeo
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
baldeojb@students.national-u.edu.ph

Neil Adrian Baltar
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
baltarnb@students.national-u.edu.ph

Carl Arvin Hipolito
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
hipolitocc@students.national-u.edu.ph

Rainnand Montaniel
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
montanielrp@students.national-u.edu.ph

Renaire Odarve
College of Computing and Information Technology(CCIT) National University(N.U.)
Manila, Philippines
odarverb@students.national-u.edu.ph

***Abstract*—The detection of violent behavior in video streams is a critical task with applications in public safety, surveillance, and law enforcement. Traditional methods relying on handcrafted features often fall short when dealing with diverse real-world scenarios. In this paper, we explore a deep learning-based approach for identifying violence in videos using Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal dynamics. Our method demonstrates promising accuracy and generalizability on benchmark datasets, highlighting the effectiveness of combining spatial and temporal modeling in video violence detection tasks.**

***Index Terms*—Violence detection, video analysis, deep learning, CNN, LSTM, computer vision, surveillance.**

I. Introduction

The increasing availability of video surveillance systems and online video content has intensified the demand for automatic analysis tools to detect harmful or violent behavior. Manual monitoring is often infeasible due to the sheer volume of video data and the potential for human error or fatigue. Automated violence detection systems can alleviate this issue by continuously analyzing video feeds and flagging suspicious or violent activity for review.

Conventional approaches to violence detection depend heavily on manually engineered features such as motion vectors, optical flow, and histogram-based methods. While effective in controlled settings, these methods often underperform in complex and unconstrained environments due to variability in camera angles, lighting, and occlusions.

Deep learning has shown significant promise in computer vision tasks, offering end-to-end learning capabilities that can learn abstract and discriminative features directly from raw data. In this study, we investigate a hybrid deep learning architecture that integrates CNNs for spatial analysis and LSTMs for temporal sequence modeling to identify violent behavior in videos.

II. Related Work

Violence detection in videos has attracted considerable research interest in recent years. Early methods such as those by Datta et al. [**?**] and Nievas et al. [**?**] relied on handcrafted descriptors and spatio-temporal interest points.

More recent approaches leverage deep learning models, particularly CNNs and Recurrent Neural Networks (RNNs), for automatic feature learning. For instance, Hassan et al. [**?**] proposed a two-stream CNN model to fuse spatial and motion features. Other studies, such as Sudhakaran et al. [**?**], utilized 3D CNNs and ConvLSTMs for end-to-end video classification.

While these methods demonstrate good performance, challenges remain in modeling long-term dependencies and maintaining efficiency for real-time applications. Our work builds upon these foundations by employing a CNN-LSTM hybrid model for improved performance on violence detection tasks.

III. Methodology

A. Overview

Our proposed methodology integrates both spatial and temporal information from video frames using a two-stage architecture:

1) A CNN backbone to extract spatial features from each frame.
2) An LSTM network to model temporal dependencies across the extracted features.

This pipeline enables the model to capture both the appearance and motion information necessary for detecting violent activities.

B. Dataset

We conduct our experiments using the *Hockey Fight* dataset and the *Violent-Flows* dataset. Each dataset contains video clips labeled as "violent" or "non-violent". Videos are pre-processed to a fixed resolution (e.g., 224×224 pixels) and uniformly sampled into frame sequences of fixed length (e.g., 30 frames per clip).

C. Preprocessing

The preprocessing stage includes:

- Frame extraction using OpenCV.
- Resizing frames to match CNN input dimensions.
- Normalizing pixel values to the range [0,1].
- Data augmentation techniques such as horizontal flipping and random cropping to increase robustness.

D. Spatial Feature Extraction

We use a pre-trained CNN such as ResNet-50 as the feature extractor. For each video frame, the CNN outputs a feature vector representing high-level visual characteristics. This step transforms each video clip into a sequence of feature vectors:

$$\{f_1, f_2, ..., f_T\}, \quad f_t \in \mathbb{R}^d$$

where T is the number of frames and d is the dimensionality of the CNN output.

E. Temporal Modeling

The sequence of feature vectors is passed to an LSTM network, which captures temporal dependencies across the frames. The LSTM outputs a final hidden state that is fed into a fully connected layer with a sigmoid activation to classify the video as violent or non-violent.

F. Training Details

We use the binary cross-entropy loss function:

$$\mathcal{L} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

where y is the ground truth label and $\hat{y}$ is the predicted probability.

The model is trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 16. Dropout and early stopping are employed to prevent overfitting.

G. Evaluation Metrics

Model performance is evaluated using accuracy, precision, recall, and F1-score. A confusion matrix is also generated to visualize classification outcomes.