ISSN 2319 - 5975

Volume 13, No.1, December 2023 – January 2024

International Journal of Networks and Systems

Available Online at http://www.warse.org/IJNS/static/pdf/file/ijns101312024.pdf

https://doi.org/10.30534/ijns/2024/101312024

Violence Detection Using Deep Learning

Krishna Sapagale[1], Manoj Sanikam[2], Nikitha[3], Prajwal M Shetty[4], Kiran B V[5],

[1] Alva's Institute of Engineering and Technology, Mijar, krishnasapagalevl30@gmail.com

[2] Alva's Institute of Engineering and Technology, Mijar, manojsanikam01@gmail.com

[3] Alva's Institute of Engineering and Technology, Mijar, nikithaswamy@gmail.com

[4] Alva's Institute of Engineering and Technology, Mijar, prajwalshetty63615@gmail.com

[5] Alva's Institute of Engineering and Technology, Mijar, kiranbv@aiet.org.in

Received Date : November 23, 2023 Accepted Date : December 22, 2023 Published Date : January 07, 2024

ABSTRACT

Due to the increased risk of exposure to violent and harmful content brought about by the spread of online video content, robust systems for automatic detection and filtering have to be developed. This research suggests a novel method for deep learning-based violent content detection in videos. Our model examines both temporal and spatial characteristics in video frames by utilizing the power of recurrent neural networks (RNNs) and convolutional neural networks (CNNs).The suggested system uses a two-stream architecture, where one stream is used for temporal information using bidirectional LSTM (Long Short-Term Memory) networks to capture sequential dependencies, and the other stream is devoted to spatial analysis using 3D CNNs for frame-level understanding [1]. To ensure strong generalization, the model is additionally trained on a varied dataset that includes both violent and non-violent content. Transfer learning is used with pre- trained deep learning models on large-scale datasets to improve the model's performance [5]. Comprehensive tests show how well the suggested method works to reliably identify violent content in videos of different genres and settings. The system demonstrates its potential for incorporation into online video platforms to give viewers a safer and more secure experience by achieving state-of-the-art outcomes in terms of precision, recall, and F1 score [4]. The suggested deep learning-based approach supports further initiatives to lessen the negative impacts of violent content in digital media and promote a safe and healthy online community [1]. Using Deep Learning to Address the Problem of Violent Video Detection: A Bright Future for Security and Safety.

The proliferation of violent content is a key concern posed by the ever-increasing abundance of online video content. This puts personal safety, public safety, and platforms' capacity to properly filter information at risk. Presenting deep learning, a potent technique that presents a viable way to automatically identify violent content in videos [2]. To sum up, deep learning presents a potent and exciting way to address the pressing problem of violent video content. We can create a more secure online environment for everyone by utilizing this technology properly and resolving the issues it raises [5]. Further investigation into cross-modality learning and real-time detection shows promise for even higher efficiency and accuracy.

Key words: Deep Learning Methods, Multi Model Feature Extraction, Machine Learning, Fight, Violent Flow, Motion feature extraction, Feature fusion baseline.

1. INTRODUCTION

Due to the ongoing rise in abnormal behaviour in different contexts, human behaviours detection in general and violence detection in particular have recently gained significant attention in Computer Vision (CV) research. Additionally, because of the complexity of the environment (i.e., social interaction) and the challenge associated with extracting a particular characteristic that is associated with a particular occurrence, violence detection is one of the most challenging problems in CV [3].

To put it another way, accurately detecting a violent situation requires two main feature extraction methods: 1) Spatial or shape feature extraction, and 2) Temporal or time features extractions. The spatial features represent the relationships or interactions between single frame pixels, but they are insufficient to identify the violence.

In the meanwhile, the most well-liked study in violence detection uses surveillance footage to extract spatiotemporal elements that aid in the clear identification of violent cases. In order to improve overall classification performance, this paper proposed various architectures based on extracting spatiotemporal features using various techniques (e.g., 3D Convolutional Neural Network (CNN) Convolutional Long Short-Term Memory (Conv-LSTM) networks integrating transfer learning with LSTM or Conv-LSTM). Additionally, the architectures included a combination of attention modules (i.e., channel attention and spatial attention).

Based on the UBI-Fights video data, a great deal of important work has been done recently in the area of violence detection. For instance, in order to provide weak/self- supervised learning, Bruno Manuel Degrading in suggests a complex iterative learning framework based on Bayesian filtration for the instances of unlabeled input. Further more, the author employed the late decision fusion ensemble technique to improve the overall performance of three models using the random forest algorithm, which has fifty decision trees [2].

The results showed that this framework performs 0.819 for the Area Under the Curve (AUC) metric and 0.284 for the Equal Error Rate (EER) measure on the UBI-fights data. Proposing different architectures based on integrating the Convolutional Block Attention Modules (CBAM) with various layers such as ConvLSTM2D or Conv2d&LSTM layers; to catch the spatiotemporal features, and increase the focus on the important ones.

Furthermore, the using for Categorical Focal Loss function (CFL) through the training, increases the focus on the important features, and overcomes the drawback of class imbalance data. Making two Comparisons to declare the significance of the proposed work results, by comparing the simply proposed architectures with other complicated ones; and also with respect to the state-of-the-art on the same data.

Nevertheless, we uncover a feature that many violent video identification algorithms in use today have overlooked: thereare instances in which the audio-visual data's semantic information does not match up within the same violent film. Some videographers, for instance, use the contrast of audio-visual semantics to artistically enhance their videos. They might play calming music during a combat scenario. Even if these movies are still categorized as violent, there is a clear inconsistency between the semantics of the two modalities because the auditory signal is non-violent and the visual signal [1]. The only way for the model to fully utilize complementarity between multimodal features through fusion is if they share the same semantics. Direct merger of multimodal characteristics is not appropriate in the aforementioned scenario.

The complete application of multimodal data will be hampered by an issue known as the heterogeneity gap, which may emerge because the auditory and visual signals represent data of two distinct modalities. Shared subspace learning, which attempts to incorporate data of many modalities into an intermediate common space where the heterogeneity can be viewed as having been eradicated, is the mainstream approach to deal with it [5]. The model might pick up some pertinent information regarding the correlation between audio-visual data implicitly throughout this process. However, we think that explicitly introducing correlation knowledge during the training phase benefits the model more, particularly when working with data that is semantically non-corresponding.

2 . DEEP LEARNING METHODS

In recent days, deep learning methodology have reached remarkable results in computer vision. Deep neural networkshave also been utilized for violent video detection.

In the Mediaeval affective task 2015, Fudan- Huawei designed a violent video detection system consisting of two stream networks and LSTM networks. Some conventional motion and audio features were used as complementary information The approach combined visual, movement and audio elements in a late fusion manner, producing the most impressive outcomes of the year. Zhou et al devised a model called Fight Net to identify visual violence interaction [2], the method relies on a classic action recognition model that operates on temporal segments. Given the extensive application of 3D Conv Nets in comprehending video content, certain researchers have begun employing them for the purpose of detecting violence in videos. In 3D Conv Net was used to extracted spatiotemporal features, whereas Song et al. built an end-to-end violent video detection system basedon 3D Conv Net.

While the methods mentioned earlier have shown reasonably positive outcomes, there remains ample room for enhancing the effectiveness of current violent video detection techniques. In this paper, we introduce a model designed for detecting violent content in videos. The utilization of the pseudo-3D model (P3D), as suggested in [3], is employed to capture short-term spatiotemporal features from the input video. The P3D model comprises pseudo 3D blocks, serving as substitutes for the original 3D Conv Net kernels to streamline computations. Following the P3D network, we integrate an LSTM network to extract long-term features from the video.

3. PROPOSED METHOD

The architecture of our model is depicted in initially, three distinct types of features namely, appearance, motion, and audio features are extracted from the video. Subsequently, we establish a feature fusion baseline utilizing shared subspace learning to combine these three features. Finally, the fusion network incorporates semantic correspondence information through a combination of multitask learning and semantic embedding learning.

3.1 Multimodal Feature Extraction

Violent videos commonly encompass the following elements Appearance Information: Includes items like firearms and cold arms. Involves gory scenes or situations where people are lyingdown. Motion Information: Encompasses activities such as fighting, chasing, and shooting [1]. Audio Information: Typical audio accompaniments in violent videos consist of screams, explosions, and gunshots. Given the analysis above, we opt touse three key features to characterize violent videos: appearance, motion, and audio features.

3.2 Appearance Feature Extraction

The current techniques for extracting appearance features have reached a relatively advanced stage. A widely adopted model for processing frame sequences is the 3D Conv Net, which excels in capturing spatiotemporal information compared to its2D counterpart. In this context, we employ the pseudo-3D model (P3D) as proposed in to extract short-term spatiotemporal features from the input video. The P3D model is composed of pseudo 3D blocks, strategically replacing the original 3D Conv Net kernel to streamline computations. To further enhance feature extraction, we incorporate an LSTM network after the P3D model, facilitating the capture of long-term features from the video.

3.3 Motion Feature Extraction

The frameworks employed for extracting appearance and motion features share a notable resemblance. However, the distinction lies in the fact that the former operates on individual video frames, while the latter works on stacked

optical flow displacement fields between consecutive frames. Opting for optical flow in motion feature extraction is grounded in its capacity to explicitly express motion information compared to video frames, making it a more suitable choice for this purpose.

3.4 Audio Feature Extraction

For audio feature extraction, we leverage the widely used VG Gish [4] network. This network is derived from the classic VGG network and has demonstrated superior performance compared to traditional sound processing methods, particularly on extensive voice datasets like Audio Set[4]. We have introduced modifications to the original VG Gish network, replacing the last three fully connected layers with a global average pooling layer to mitigate over fitting.
The audio signal extracted from the video undergoes processing to generate a mel spectrogram with a size of 96 × 64. This spectrogram is then fed into the VG Gish network, resulting in a 128-dimensional feature (Fau) that effectively represents the audio feature of the video.

4. FEATURE FUSION BASELINE

In contrast to late fusion, which operates on decision-level scores, feature-level fusion has the advantage of incorporating more information, leading to better results. Given that audio-visual data represent two distinct modalities that may exhibit a heterogeneity gap, shared subspace learning emerges as a widely adopted method to address this issue [2].
The fundamental principle of shared subspace learning involves mitigating the heterogeneity among different modal features through projection transformations and harnessing the complementarity of multimodal features. In the following sections, we will delve into the details of our shared subspacelearning approach.

5. DATASETS

Due to the challenges associated with collecting extensive violent data, there is currently a scarcity of large-scale public datasets specifically dedicated to violent videos. Our experiments address this limitation by utilizing three publicly available datasets: Hockey Fight [1], Violent Flow [3], and VSD2015. The videos in these datasets typically have durations ranging from 2 to 10 seconds. scene and semantic information remaining relatively consistent. In these datasets, the task of violent video detection is essentially transformed into a binary classification problem, distinguishing between violent and non-violent content.
Hockey Fight: This dataset features relatively simple scenes, primarily centered around one specific violent scenario: fights. However [1], it lacks audio data and is employed mainly to assess the effectiveness of appearance and motion features.
Violent Flow: Upon investigation, it is noted that nearly all audio data in this dataset does not contain violent audio events such as explosions and gunshots.

Consequently, the majority of audio data is classified as non-violent from an auditory perspective. The visual violence label of each video in this dataset closely aligns with the overall video violence label. Due to this strong correlation, experiments regarding semantic correspondence are not conducted on this dataset.

6. IMPLEMENTATION DETAILS

The extraction of three key video features— appearance, motion, and audio—utilizes deep learning methods in a separate manner.
Appearance Feature Extraction: For appearance features, all frames extracted from the input video are initially set to a size of 224 × 224, randomly cropped from the resized 240 × 320 video frames. Successive non-overlapping frames (16 in total) form a clip, sent through the P3D199 model pretrained on Kinetics-400 [2]. This process yields a temporal local feature with 2048 dimensions. These temporal local features are then processed through an LSTM network, and the final output of the LSTM (512 dimensions) is regarded as the temporal global feature of the input.

7. CONCLUSION

This paper conducts a thorough analysis of the violence detection in surveillance videos task using UBI-Fights as a reference dataset in order to assess the proposed work. The analysis is conducted in three steps:
1) Provide a thorough and lucid explanation of the problem case study and challenges by reviewing the most recent related work on the same data.
2) Develop various architectures that satisfy the requirements identified in the first step.
3) Assess the proposed work by contrasting it with the state-of-the-art work and each other.
The UBI-Fights dataset is used to implement and assess six distinct architectures. The Convolutional Block Attention Module(CBAM) was integrated with three basic architectures that were created from the ground up as a spatiotemporal feature extractor, the ConvLSTM2D or Conv2D&LSTM layers; the other three have a similar integration procedure based on ResNet50, VGG16, or Mobile Net. The channel attention module and the spatial attention module are the two primary attention modules found in the CBAM [3].
They are both built in a sequential manner to draw the architectures' attention to the most crucial elements, such as the character of human interactions, while ignoring the less crucial ones, such as environmental features. Furthermore, the application of Category Focal Loss (CFL) as a loss function during the training of the architectures, lessens the issues with imbalanced data, and sharpens the models' emphasis on the most key components.

The Equal Error Rate (EER) and Area Under the Curve (AUC) metrics serve as the foundation for the evaluation criteria and metrics. Additionally, the assessment is completed using two primary comparisons:

1) Comparison of ablation studies, which shows how well the simple suggested architectures perform compared to the othercomplex ones.
2) Comparison of state-of-the-art, which indicates how innovative the proposed work performs compared to the published papers on the UBI-Fights data.
The performance results of the comparison steps show that the Conv2d&LSTM-based architecture can obtain a high performance of 0.0507 for the AUC metric and 0.9493 for the AUC metric.

REFERENCES

1. S. Davila-Montero, J. A. Dana-Le, G. Bente, A. T. Hall, and A. J. **Review and problems of technology for real time human behavior**, Mason IEEE Transactions on Biomedical Circuits and Systems, vol. 15, no. 1, pp. 2– 28, Feb. 2021. monitoring.
2. GRU-FFN, B. Fan, P. Li, S. Jin, and Z. Wang, Proc. **Anomaly detection based on pose estimation** IEEE Sustain. Power Energy Conf. (I SPEC), Dec. 2021, pp. 3821–3825.
3. **A video abnormal behavior recognition system based on deep learning**. was presented by B. Cao, H. Xia, and Z. Liu in the IEEE 4th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC), vol. 4, June 2021, pp. 755–759.
4. R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, **Recognition of human activity and abnormal behavior using deep neural network**. In Proc. ELEKTRO (ELEKTRO), May 2022, pp. 1-4.
5. F. J. Rendón Segador, J. A. Álvarez-García, F. Enríquez, and O. Deniz. "**Violence Net: Dense multi-head self- attention with bidirectional convolutional LSTM for detecting violence**" Electronics, vol. 10, no. 13.