Harm Anticipation : Action Recognition with C3D Convolutional Neural Network

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

[1,2,3,4,5]Department of Computer Science and Engineering, AIET, Mijar, Dakshina Kannada, India

krishnasapagalevl30@gmail.com[1], manojsanikam01@gmail.com[2], nikithaswamy@gmail.com[3], prajwalshetty63615@gmail.com[4], kiranbv@aiet.org.in[5]

1 INTRODUCTION

Violence detection in particular and human behavior detection in general have lately attracted a lot of attention in Computer Vision (CV) research due to the continuous growth in deviant behavior in many scenarios. Violence detection is also one of the most difficult challenges in CV because of the complexity of the environment (i.e., social interaction) and the difficulty of identifying a specific attribute linked to a specific occurrence [3].

Stated differently, two primary feature extraction techniques are necessary for the accurate detection of violent situations: 1) Extraction of spatial or form characteristics; 2) Extraction of temporal or time features. The linkages or interactions between single frame pixels are represented by the spatial characteristics, but they are not enough to pinpoint the violence.

Meanwhile, the most popular study on violence detection extracts spatiotemporal components from surveillance film to help distinguish violent situations from non-violent ones. This paper proposed several architectures based on spatiotemporal feature extraction using different methods (e.g., 3D Convolutional Neural Network (CNN) Convolutional Long Short-Term Memory (Conv-LSTM) networks integrating transfer learning with LSTM or Conv-LSTM) to improve overall classification performance. A mix of attention modules, such aschannel attention and spatial attention, was also incorporated in the structures.

Violence detection has seen a great lot of significant work recently, based on the UBI-Fights video data. For example, Bruno Manuel Degrading in proposes a complicated iterative learning framework based on Bayesian filtering for the instancesof unlabeled input in order to give weak/self-supervised learning. Additionally, the author used the random forest method, which includes fifty decision trees, to enhance the overall performance of three models by utilizing the late decision fusion ensemble technique [2].

The findings demonstrated that this framework's performance on the UBI-fights data is 0.819 for Area Under the Curve (AUC) metric and 0.284 for the Equal Error Rate (EER) measure.putting forth many designs that combine the Convolutional Block Attention Modules (CBAM) , including ConvLSTM2D or Conv2d&LSTM layers, in order to capture spatiotemporal characteristics and sharpen attention on the most significant on

Additionally, by utilizing the Categorical Focal Loss function (CFL) during training, the disadvantage of class imbalance data is overcome and the attention is increased on the key characteristics. Comparing the straightforwardly suggested architectures with more complex ones and the state-of-the-art onthe same data sets two comparisons that will help determine the importance of the work findings.

However, we also reveal a trait that many existing violent video identification algorithms have missed: there are cases where the semantic content in the audio-visual data does not correspond tothe same violent movie. For example, some videographers employ the artistic contrast of audio-visual semantics to enhance their recordings. They might put on soothing music in the event of a fight. There is a glaring discrepancy between the semantics of the two modalities because the visual signal is violent while the auditory signal is not, even if these films are nevertheless classified as violent [1]. The model can only take full advantage of complementarity between multimodal features through fusion when those features have similar meanings. In the aforementioned case, a direct merger of multimodal properties isinappropriate. The heterogeneity gap—a problem that may arise because the auditory and visual signals represent data of two distinct modalities—will impede the full application of multimodal data.The popular way to deal with it is shared subspace learning, which aims to merge data of several modalities into an intermediate common space where the heterogeneity can be considered as having been eliminated [5].

During this process, the model may implicitly learn some important information about the association between audio-visual data. However, we believe thatthe model gains more from explicitly integrating correlation knowledge during the training phase, especially when dealing with semantically non-corresponding data

2. DEEP LEARNING METHODS

Deep learning techniques have produced amazing improvements in computer vision recently. Additionally, deep neural networks have been applied to the detection of violent videos. Fudan-Huawei created a violent video detection system using twostream networks and LSTM networks for the Mediaeval Affective Task 2015. As supplemental information, a few standard motion and audio elements were included. The method created the most remarkable results of the year by fusing together visual, movement, and aural components in a late fusion fashion. To detect visual violence interaction, Zhou et al. developed a model called Fight Net [2]. This approach is based on a traditional action recognition model that functions on temporal segments. Considering how often used 3D Conv Nets are there. Some researchers have started using them to understand video material in order to identify violent scenes in videos. Spaciotemporal characteristics were extracted using 3D Conv Net, while Song et al. developed an end-to-end violent video detection system using 3D Conv Net. Although the previously indicated strategies have yielded somewhat favorable results, there is still much space for improvement. The different methods has to be implemented and the methods must follow-up with some of additional methods that include the basic and the most required things that posses along with the efficiency of the methods now used to detect violent videos. We present a methodology in this research that is intendedto identify violent material in videos. As recommended in [3], the pseudo-3D model (P3D) is used to extract short-term spatiotemporal information from the input video. In order to simplify computations, the P3D model consists of pseudo 3D blocks that act as stand-ins for the original 3D Conv Net kernels. Inorder to extract long-term features from the video, we add an LSTM network after the P3D network

3.PROPOSED METHOD

The first diagram shows the architecture of our model. Three different kinds of features—appearance, motion, and audio—are taken out of the video. We then use shared subspace learning to create a feature fusion baseline that combines these three features.Lastly, by combining multitask learning and semantic embedding learning, the fusion network adds semantic correspondence information.

3.1 MULTIMODAL FEATURE EXTRACTION

Typically, violent videos include the following components: Information about Appearance: Contains objects such as cold armsand weapons. involves graphic events or scenes in which people are lying down. Motion Information: Contains actions like chasing, firing, and fighting [1]. Audio Information: Gunshots, explosions, and cries are frequently heard as background music in violent videos. Based on the aforementioned investigation, we choose to employ look, motion, and audio aspects as the three primary features to describe violent videos.

3.2 APPEARANCE FEATURE EXTRACTION

The appearance feature extraction methods available now are at a reasonably sophisticated level. The 3D Conv Net is a popular model that is used to process frame sequences because it is better at capturing spatiotemporal information than its 2D equivalent. In order to extract short-term spatiotemporal information from the input video, we use the pseudo-3D model (P3D) as proposed in. Pseudo 3D blocks deliberately replace the original 3D Conv Net kernel in the P3D model to expedite computations. After the P3D model, we add an LSTM network to improve feature extraction even more by making it easier to extract long-term features from the video.

3.3 MOTION FEATURE EXTRACTION

There is a significant overlap in the frameworks used to extract appearance and motion data. The distinction, though, is that the latter works on stacked optical flow displacement fields between consecutive frames, whilst the former works on individual video frames. The rationale behind using optical flow instead of video frames for motion feature extraction is that the former may express motion information explicitly, which makes it a better option in this context.

3.4 AUDIO FEATURE EXTRACTION

To extract audio features, we make advantage of the popular Gish [4] network. This network, which is based on the well- known VGG network, has outperformed more conventional approaches to sound processing, especially when used on large- scale voice datasets such as Audio Set [4]. To reduce overfitting,we have made changes to the original VG Gish network by adding a global average pooling layer in place of the final three fully linked layers. A 96 × 64 mel spectrogram is produced by processing the audio signal that was taken out of the movie. Afterfeeding the spectrogram into the VG Gish network, a 128- dimensional feature (Fau) is produced that accurately captures the video's audio feature.

4.FEATURE FUSION BASELINE

Feature-level fusion has an advantage over late fusion, which uses decision-level scores; it incorporates more information and produces better results. As audio-visual data represent two different modalities that could show heterogeneity gap, shared subspace learning becomes a popular approach to deal with this problem [2].

The core idea behind shared subspace learning is to use projection transformations to reduce the heterogeneity between various modal data and to take advantage of the complementarity between multimodal features. We will go into more depth about our shared subspace learning strategy in the sections that follow.

5.DATASETS

There are currently few large-scale public datasets dedicated to

violent videos because of the difficulties in gathering comprehensive violent data. Our studies make use of three publicly accessible datasets—Hockey Fight [1], Violent Flow [3], and VSD2015—to overcome this issue. Scene and semantic information are reasonably stable throughout the 2 to 10 second videos that make up these datasets. Identifying violent and non-violent content in these datasets effectively turns the issue of violent video detection into a binary classification problem. Hockey Fight: The scenes in this collection are very straightforward and are mostly focused on one type of violent scenario: fights. Nevertheless, it is primarily used to evaluate the efficacy of visual and motion elements and lacks audio data [1]. Violent Flow: After looking into it, it is discovered 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

Three important video features—appearance, motion, and audio—are extracted separately using deep learning techniques. Extraction of Appearance Features: All frames taken from the input video are initially sized for appearance features.

arbitrarily cut from the resized 240 × 320 movie frames to a size of 224 × 224. A clip is created by sending 16 consecutive non-overlapping frames through the P3D199 model after it has been pretrained on Kinetics-400 [2]. A temporal-local feature with 2048 dimensions is produced by this approach. Following their processing by an LSTM network, these temporal local features are considered the temporal global feature of the input. The 512-

dimensional final output of the LSTM is used to determine this. 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

In order to evaluate the suggested method, this research thoroughly analyzes the task of violence identification in surveillance videos using UBI-Fights as a reference dataset. Three steps are taken in the analysis process:

1)Review the most current related work on the same data and provide a comprehensive and understandable explanation of the problem case study and challenges.

2)Create several architectures that fulfill the first step's requirements.

3)Evaluate the suggested work by comparing it to each other and the most recent work.

Six different architectures are implemented and evaluated using the UBI-Fights dataset. Three fundamental designs, the ConvLSTM2D or Conv2D&LSTM layers, that were built from the ground up as a spatiotemporal feature extractor were integrated with the Convolutional Block Attention Module (CBAM); the other three used a similar integration process based on ResNet50, VGG16, or Mobile Net. The two main attention modules in the CBAM are the channel attention module and the spatial attention module [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. Furt Both are constructed in a sequential fashion with the intention of highlighting the most important components—like the nature of human interactions—while downplaying the less important ones—like environmental aspects. Additionally, by using Category Focal Loss (CFL) as a loss function during the architectures' training, imbalanced data problems are lessened and the models' concentration on the most important elements is sharpened.

Furthermore, by using Category Focal Loss (CFL) as a loss function during the architectures' training, imbalanced data problems are mitigated and the models' focus on the most important elements is enhanced. The measurements and evaluation criteria are based on the Equal Error Rate (EER) and Area Under the Curve (AUC) metrics. In addition, two key comparisons are used to finish the assessment:

1) Comparison of ablation investigations, demonstrating the relative performance of the simple recommended structures and the more complex ones.

2) State-of-the-art comparison: This shows how inventively the suggested study stacks up against the published publications on the UBI-Fights data.

The Conv2d&LSTM-based architecture can achieve a high performance of 0.0507 for the AUC metric and 0.9493 for the AUC metric, according to the comparison step performance finding

8.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.Recognition of human activity and abnormal behavior using deep neural network, R. Vrskova, R. Hudec, P. Kamencay, and

P. Sykora. In Proceedings of Electrochemistry (ELEKTRO), May 2022, pp. 1-4.

5.F. J. Rendón Segador, O. Deniz, F. Enríquez, and J. A. Álvarez-García. "Violence Net: Bidirectional convolutional LSTM with dense multi-head self-attention for detecting violence" was published in Electronics, vol. 10, no