MINOR PROJECT REPORT

TWO STREAM NETWORK FOR VISION BASED VOILENCE DETECTION

Department of Computer Science and Engineering
National Institute of Technology Patna

Submitted by:
Sneh Kumar (2006052)
Arjun Singh (2006067)
Paras Punjabi (2006069)

Under the Supervision:
Dr. Piyush Kumar
(Project Supervisor)
Assistant Professor
Department of CSE

Table of Content

ABSTRACT

Detecting violence in video recordings through artificial intelligence is critical for law requirements and also keeping track of public security with the help of surveillance cameras. Additionally, it very well might be an incredible tool for protecting children from accessing inappropriate content and assist guardians with settling on a better choice with regards to what their children should watch. This is a difficult issue since the actual meaning of violence is expansive and exceptionally abstract. Subsequently, detecting such subtleties from recordings with no human management isn't just technical, yet in addition a theoretical issue. In light of this, in this work we will investigate how to more readily depict the idea of violence using a convolutional neural network.

At first by breaking, it into more even handed and concrete related ideas, like fights, explosions, blood, and so on, for later combining them in a meta-order to depict violence. We will likewise investigate approaches to address time-based events for the network, since numerous violent demonstrations are portrayed in terms of development. And at long last we will investigate how to localize violent events, since numerous video transfers don't contain just violence, yet are a combination of violent and non-violent scenes.

There are a lot of pretrained convolutional neural networks which can be used for image classification. These models have learned to extract powerful and informative features from natural images and use it as a starting point to learn a new task. These networks have been trained on more than a million images and can classify images into 1000 object categories. We will be using a pretrained network with transfer learning

because it's typically much faster and easier than training a network from scratch. For this project we will be exploring which model will give the best accuracy for detecting violence in videos using various deep learning techniques.

We have taken a dataset (Real Life Violence Situation Dataset) of 1000 videos of violence and non-violence each. We have tried video classification using two stream network model.
After running this dataset on our model we have got accuracy of 82%.

We have also taken a dataset (Movie Fight Detection Dataset) of 100 videos of Fights and no-Fights each. After running this dataset on our model we have got accuracy of 100%.

In future, to test the effectiveness of our model, we will create our custom dataset with labelled examples that are representative of the real-world data, ensuring that the model is trained on a diverse set of data and can generalize well to new and unseen examples.

INTRODUCTION

Human activity recognition is a widely investigated problem in the field of computer vision that has diverse applications in human-computer interaction, robotics, surveillance, etc. In recent years, large-scale video action recognition has gained impressive improvements mostly due to the availability of large datasets, deep neural network architectures, video representation techniques, etc. Many works, on the other hand, focused on specific sub-tasks of action recognition such as spatial-temporal localization of activity, anomaly detection, action quality analysis (AQA), egocentric activity recognition, etc. One such important subset is violence detection which is widely applicable in public monitoring, surveillance systems, internet video filtering, etc. As digital media technologies like surveillance cameras are getting more and more ubiquitous, detecting violence from captured footage using manual inspection seems increasingly difficult. To address this issue researchers have suggested different approaches that can detect violence from surveillance footage automatically without requiring any human interaction. Violence detection is a section of general action recognition task which specifically focuses on detecting aggressive human behaviors such as fighting, robbery, rioting, etc.

Earlier works on violence detection mostly focused on engineering various descriptors that could effectively capture violent motion present in the video. Later on, the performance of these handcrafted features was surpassed by several end-to-end trainable deep learning methods which require little to no pre-processing . To validate the effectiveness of these methods, commonly three standard benchmark datasets were used called Hockey, Movies, and Violent-Flows. Recently, a new dataset called RWF-2000 has been proposed which is substantially bigger and more diverse. For applying these deep learning models in real-life practical scenarios both computational efficiency and accuracy need to be considered. In this respect, we present a novel two-stream CNN-LSTM based network that can produce discriminative Spatio-temporal features

while requiring fewer parameters. In general action recognition tasks, surroundings or background information may serve as discriminative clues.

We can encapsulate our significant contributions in the following points:

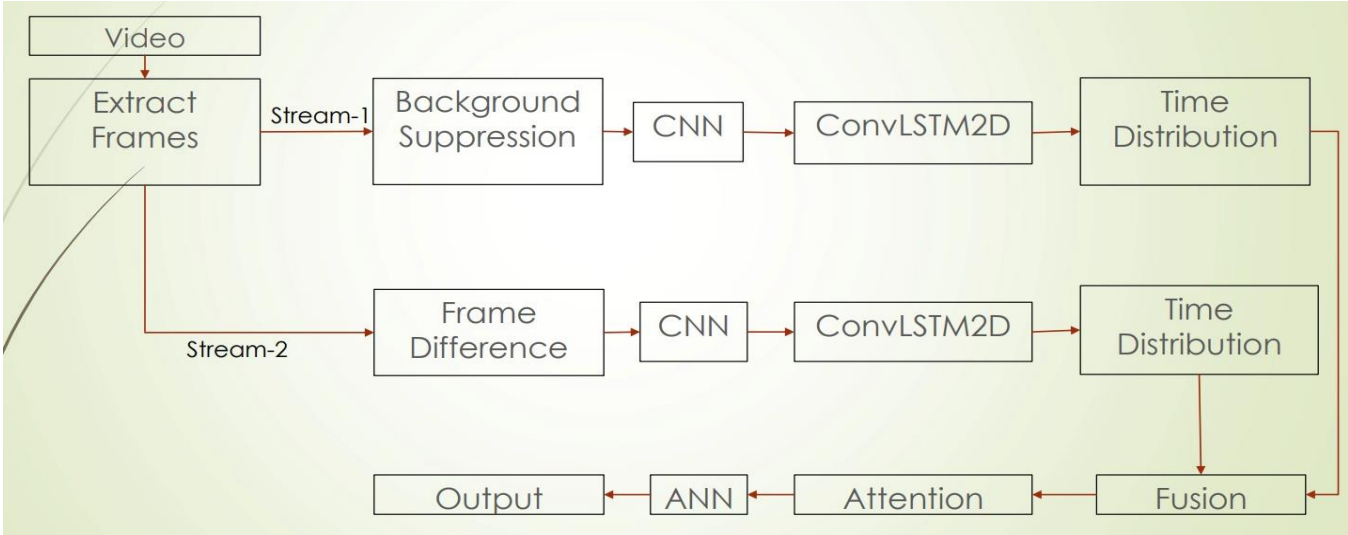
• We propose a two-stream deep learning architecture that leverages Convolutional LSTM (ConvLSTM) and Time Distribution

• We utilized simple and fast input pre-processing techniques that highlight the moving objects in the frames by suppressing non-moving backgrounds and capture the motion in-between frames.

LITERATURE SURVEY

Paper	Year	Dataset	Aim of Work	Method Used
Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM [2]	2021	RWF2000	To perform Vision Based Human Activity	Two Stream Network
FTCF: Full temporal cross fusion network for violence detection in videos [1]	2022	Real Life Violence Situations Dataset	To perform Vision Based Human Activity	Two Stream Networks using FTCF Blocks
Efficient violence detection using 3d convolutional neural networks[3]	2019	Movies Fight Detection Dataset	To perform Vision Based Human Activity	3D Convolutional Networks

METHODOLOGY



- Firstly, we are feeding videos in our model and here videos are converted to sequence of frames using opencv module in python. In this we are taking 15 frames per video and resizing each video frame to 64x64x3. Then after that our model is divided into two streams.

- In the first stream, we are passing all the frames through background suppression. We employed a simple technique to estimate the background to avoid adding computational overhead. We first calculate the average of all the frames. The average frame mostly contains the background information because they remain unvarying across multiple frames. Then we subtract this average from every frame which accentuates the moving objects in the frame by suppressing the background information. As violent actions are mostly characterized by body movements, not the non-moving background features, this promotes the model to focus more on relevant information.

- After that, we are passing these frames from CNN layers which comprises of Conv2D, Batch-Normalization and Dropout Layers.

- Then, we are passing this through ConvLSTM2D, which is a type of RNN layer that combines the properties of CNN and LSTM to process spatiotemporal data by applying convolutions across both space and time dimensions.

- Then, we are passing it through Time-Distribution layer, which is a wrapper layer in neural network that allows same layer to be applied to every time stamp of a sequence independently, treating each time step as a separate input, thereby enabling the network to learn temporal relationships between sequential data points.

- On the other hand in stream-2 the same frames are passed through frame difference. Frame difference is the difference between the matrices of consecutive frames. Frame difference actually promotes the model to encode temporal changes between the adjacent frames boosting the model to capture motion information.

- After that we are again passing these frames through CNN, ConvLSTM2D and Time Distribution Layers.

- Now we would fuse the data from these two streams by using Add Layer. Now the combined output is passed through MultiHeadAttention model. MultiHeadAttention model will enable our neural network to focus on different parts of input sequence to extract relevant information. This will made our model to capture more complex and diverse patterns in the data.

- After that, we are passing it through simple ANN which consists 2 dense layer to perform classification among the given classes.

RESULT AND DISCUSSION

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

Paper	Dataset	Accuracy	Our Accuracy
FTCF: Full temporal cross fusion network for violence detection in videos[1]	Real Life Violence Situation Dataset	98.5%	82%
Efficient violence detection using 3d convolutional neural networks[3]	Movies Fight Detection Dataset	100%	100%
Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM [2]	RWF2000	89.75%	73%

Classification metrics:

When performing classification predictions, there's four types of outcomes that could occur.

True positives are when you predict an observation belongs to a class and it actually does belong to that class.

True negatives are when you predict an observation does not belong to a class and it actually does not belong to that class.

False positives occur when you predict an observation belongs to a class when in reality it does not.

False negatives occur when you predict an observation does not belong to a class when in fact it does.

These four outcomes are often plotted on a confusion matrix.

❖ Real Life Violence Situation Dataset

Accuracy: 82%

Classes	Precision	Recall	F1-Score
Non-Violence	0.82	0.80	0.81
Violence	0.82	0.84	0.83

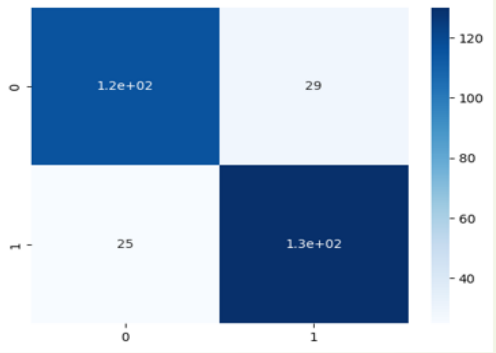


Table-1

❖ Movies Fight Detection Dataset

Accuracy: 100%

Classes	Precision	Recall	F1-Score
Fights	1.00	1.00	1.00
Non-Fights	1.00	1.00	1.00

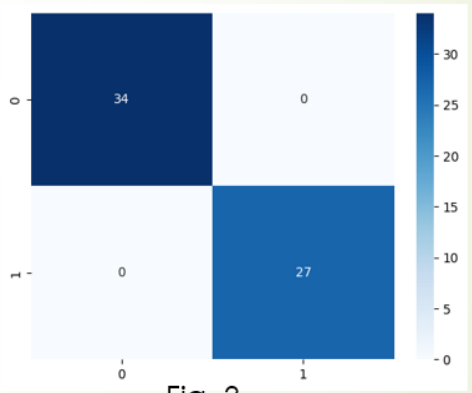


Fig. 3

Table-2

Precision is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

Precision = true positive / true positive + false positive

Recall is defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.

Recall = true positive / true positive + false negative

We can observe in Table-1 the recall for non-violence category is 80 percent and violence category is 84 percent in real life violence situation dataset. Here the violence category has more recall, which is good for the model as it increases scope for detecting violence.

Here in Table-1 we can observe that the F1 score for non-violence category is 81 percent and for violence category is 83 percent.

CONCLUSION AND FUTURE WORK

- The Two-Stream Network for Vision-Based Violence Detection paper introduces an innovative approach for detecting instances of violence in real-world scenarios using video data. The paper addresses the challenge of accurately classifying violent and non-violent scenes by utilizing both spatial and temporal information.

- Traditionally, violence detection methods have primarily focused on spatial information, analysing individual frames or images to identify visual cues associated with violence. However, this approach may overlook crucial temporal patterns and dynamics present in violent actions, leading to suboptimal performance.

- To overcome this limitation, the paper proposes a two-stream network architecture that incorporates both spatial and temporal information. The spatial stream processes individual frames, capturing visual features and patterns using convolutional neural networks (CNNs). This stream focuses on identifying spatial cues such as object appearances and their interactions.

- The temporal stream, on the other hand, aims to capture motion information and temporal dynamics by analysing sequences of frames. It utilizes optical flow estimation techniques to encode motion information between consecutive frames. This stream allows the network to discern temporal patterns associated with violent actions, enhancing the overall accuracy of the violence detection system.

- The combination of spatial and temporal streams enables the network to learn and exploit complementary information, leading to improved discrimination between violent and non-violent scenes. By considering both static appearance and dynamic motion

cues, the proposed approach achieves enhanced performance in violence detection tasks.

- One key aspect highlighted in the paper is the emphasis on utilizing temporal information for violence detection. By integrating temporal stream processing into the network architecture, the model can effectively capture and analyse temporal dynamics, which are crucial in identifying violent actions. This temporal-awareness contributes to better accuracy and robustness in detecting violence.

- Furthermore, the paper emphasizes the importance of creating a lightweight and efficient architecture suitable for real-time applications. By carefully designing the network, considering computational efficiency and reducing computational complexity, the proposed approach is capable of performing violence detection tasks in real-time scenarios.

- Overall, the Two-Stream Network for Vision-Based Violence Detection paper presents a promising approach to address the challenge of violence detection in real-world settings. By incorporating spatial and temporal information, the proposed architecture enhances performance, effectively leveraging both static appearance and dynamic motion cues. This work contributes to advancing violence detection systems and has implications in various domains such as security, surveillance, and public safety.

- To test the effectiveness of our model , we are thinking to create our custom dataset with labeled examples that are representative of the real-world data, ensuring that the model is trained on a diverse set of data and can generalize well to new and unseen examples.

REFERENCES

1. Tan Zhenhua, Xia Zhenche, Wang Pengfei, Ding Chang and Zhai Weichao, “FTCF: Full temporal cross fusion network for violence detection in videos” in Advances in neural information processing systems, 2022.

2. Zahidul Islam , Mohammad Rukonuzzaman, Raiyan Ahmed , Md. Hasanul Kabir , and Moshiur Farazi , “Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM” in Advances in neural information processing systems, 2021.

3. J. Li, X. Jiang, T. Sun, and K. Xu, “Efficient violence detection using 3d convolutional neural networks,” in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2019, pp. 1–8.

4. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in neural information processing systems, 2014, pp. 568–576.

5. S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017, pp. 1–6.