Academic Science Journal

Detection of Violence Behavior using Deep Learning Technique

Mohamed Safaa Mohamed Shubber and Ziyad Tariq Mustafa Al-Ta'i

[1] Department of computer science – College of science – University of Diyala

msm.shubber@gmail.com

Received: 17 July 2022 **Accepted: 17 September 2022**

DOI: https://dx.doi.org/10.24237/ASJ.01.02.654B

Abstract

Deep learning-based violence detection approaches from video streams are a rapidly expanding subject of study, this is due to the necessity to develop appropriate and automated violence detection techniques based on visual data obtained from security cameras mounted in various locations. In this research, a modified pre-trained deep learning technique named Convolutional Neural Network-Visual Geometry Group16 (CNN-VGG16) are employed to implement a low complex and uncomplicated model for the detection of violence. The transfer learning technique is applied to take advantage of the pre-knowledge VGG16 in detecting shapes and edges. The final layers of the default VGG16 structure are replaced to accommodate the purpose of the research. The efficiency of this approach is evaluated using two datasets (Automatic Violence Detection Dataset (AvdDS) and Surveillance fight dataset (SfDS)). The experimental outcomes prove the efficiency of the proposed model against alternative methods. In experiment accuracy results for Dataset 1 94% and 91% after applying Canny filter, as for dataset 2, the accuracy is 99% and 92% using canny filter. Also In this paper, the effect of applying edges detector on classification accuracy results and processing time for training the model is observed.

Keywords: Violent Behavior Detection, Computer Vision, Deep Learning, Transfer Learning, Modified CNN-VGG16

Academic Science Journal

الكشف عن السلوك العنيف باستخدام تقنية التعلم العميق

محمد صفاء محمد شُبر وزياد طارق مصطفى الطائي

قسم الحاسبات – كلية العلوم – جامعة ديالى

الخلاصة

تعد طرق اكتشاف العنف القائمة على التعلم الآلي من تدفقات الفيديو موضوعًا سريعًا للدراسة نظرًا لضرورة تطوير تقنيات مناسبة واليات للكشف عن العنف استنادًا إلى البيانات المرئية التي تم الحصول عليها من الكاميرات الأمنية المثبتة في مواقع مختلفة. في هذا البحث، تم استخدام تقنية تعلم عميق معدلة ومدربة مسبقًا تسمى الشبكة العصبية التلافيفية مجموعة الهندسة المرئية (CNN-VGG16) لتنفيذ نموذج منخفض التعقيد للكشف عن العنف. يتم تطبيق تقنية (تحويل التعلم) للاستفادة من المعرفة المسبقة لـ VGG16 في اكتشاف الأشكال والحواف. تم استبدال الطبقات النهائية لهيكل VGG16 الافتراضي لتلائم الغرض من البحث. تم تقييم كفاءة النهج باستخدام مجموعتي بيانات (مجموعة بيانات الاكتشاف التلقائي للعنف (AvdDS) ومجموعة بيانات مكافحة المراقبة (SfDS)). أثبتت النتائج التجريبية كفاءة النموذج المقترح مقابل الطرق البديلة. أيضا في هذا البحث مراقبة تأثير تطبيق كاشف الحواف على نتائج دقة التصنيف ووقت المعالجة لتدريب النموذج. في نتائج دقة التجربة لمجموعة البيانات 1 94٪ و 91٪ بعد تطبيق مرشح Canny ، أما بالنسبة لمجموعة البيانات 2 ، فإن الدقة هي 99٪ و 92٪ باستخدام مرشح canny. أيضا في هذا البحث لوحظ تأثير تطبيق كاشف الحواف على نتائج دقة التصنيف ووقت المعالجة لتدريب النموذج.

كلمات مفتاحية: كشف السلوك العنيف ، الرؤية الحاسوبية ، التعلم العميق ، تحويل التعلم ، CNN-VGG16 المعدلة

Introduction

Although it might be challenging to outline abnormality in human behavior, it is generally simple to spot when it occurs. A psychological term describing acts that deviate from what is seen as normal in a human civilization or culture is called "abnormal behavior". There are four main ways to recognize unusual action in individuals: statistical rarity, societal norms being violated, personal distress, and unadaptable to societal behaviors[1].

Violence is considered one of the most dangerous violations of normal social habits and it may always be part of the human experience from the beginning of existence. Its impact can be seen in various forms worldwide[2].

Due to human exhaustion and inattention, it is harmful events like fights and aggressive actions won't be detected by security staff. Therefore, developing an intelligent video surveillance system that automatically identifies abnormalities is crucial. Given the significance of security, the research has been done in this area and several methods to identify anomalies in videos have been presented [3].

The domain of Computer Vision (CV) has lately seen a significant transformation in several applications like images categorization and human activities identification. Deep learning, is a type of machine learning, has also made its way into computer vision-related fields since its introduction and has shown to be quite effective. Since 2010, many researchers from the computer vision area have moved from conventional handcrafted features descriptor to the learned-based features descriptor, often referred to as data-driven algorithms[4].

In the areas of computer vision, Deep Learning (DL) has expanded the boundaries of what was previously thought to be achievable. A computing architecture that was inspired by the way the human brain operates is called an "artificial neural network (ANN)". (DL) is mostly based on ANNs. For instance, the human brain, it is made up of many different processing cells, often known as "neurons," that individually carry out a certain function and collaborate to produce an outcome[5].

This paper proposes a model for detecting violent behavior in videos on the frames level by preprocessing the frames and forwarding them to a modified VGG16 network for pattern detection and classification.

The rest sections in this paper are arranged as follows, Section 2. will go through some of the most related works, while, section 3. illustrate the materials and methods used in the module, as for section 4 it illustrates the proposed module methodology, section 5 shows the

experimental results and provides a discussion upon these results also, Finally, the conclusion is presented in section 6.

Related works

Detection of violence is a unique challenge in the area of action recognition. In the last decade, there has been critical progress in the field of computer vision, and numerous researchers succeeded in constructing reliable classifiers that conduct human activity recognition (HAR).

RNN was utilized in [6] to collect spatio-temporal characteristics using a 2D time distributed modified CNN. and employed two customized sub-networks, one for colored RGB images and the other one for optical flow, whose outcomes were combined together to encode the movement of the optical flow with characteristics.

[7] paper used deep neural networks that had already been trained to identify violence so that it provide an approach with lowest complexity level to this issue. To determine if a violent action had taken place, extracted characteristics using the pre-trained models were combined and fed to a fully-connected network. The ResNet-50 and VGG16 outputs were both examined in the suggested method.

[8] proposed a novel violence detection pipeline that can be combined with the conventional 2-dimensional Convolutional Neural Networks. On top of that, they presented temporal along with spatial attentions modules which are low complexity but persistently boost the efficiency of violence behavior recognition.

In [9] paper, the model comprises three phases: preprocessing, feature extraction, and violence sequence learning. Initially, a CNN model was utilized during the phase of feature extraction to collect the data and attributes. Afterward, the resulting feature map was constructed by concatenating remaining the optical flow CNN features with high-level features from the Darknet19 model. Finally, the LSTM network acquired and learnt the sequence characteristics for violence detection.

Using the Gated Recurrent Unit, researchers in [10] presented a novel Deep Convolutional Neural Network (DCNN) model called BrutNet (GRU). For every frame of the time-distributed layer, convolutional layers were used to obtain the image-feature set and pattern. Then, the GRU layer extracts the temporal character of these frames as a 1-dimensional vector, which is processed by many dense layers. Although these works achieved excellent results when being applied on the benchmark datasets, feature extraction was done to every aspect of the frame, and as a result this will lead to resources consumption problem in training and testing their network, which led to propose a classification module that reduces the extracted feature.

Material and methods

Visual Geometry Group (VGG16)

The architecture "VGG16 Convolutional neural network" was utilized to win the 2014 ILSVR (Imagenet) challenge[11]. It is considered one of the most effective vision model designs ever constructed. In place of a large number of hyper-parameters, VGG16 relies on (3x3) filter convolution layers with a stride value of one, while retaining the same padding and total max-pooling layer of (2x2) filter stride value of two. The convolution and max-pooling layers are structured similarly throughout the architecture. Finally, the top layers consist of two fully connected layers followed by a softmax for producing the outcome. The number 16 in VGG16 refers to the sixteen layers having weights [12], as shown in figure (1).

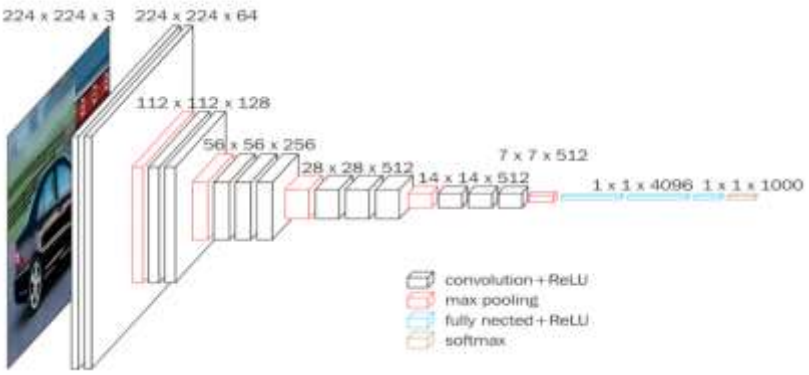

Figure 1:VGG-16 Model [12]

Image Enhancement

"Histogram Equalization" (HE) is an easy and simple method used for enhancing the contrast and improving image quality. During this correction, the intensities can be better distributed on the histogram [13]. "Contrast-Limited Adaptive Histogram Equalization (CLAHE)" is a special type of HE where excessive amplification and noise amplification will be overcome by cutting the spikes and improving the speed of calculation[14]. An example of the output of applying CLAHE on an image is illustrated in figure (2).

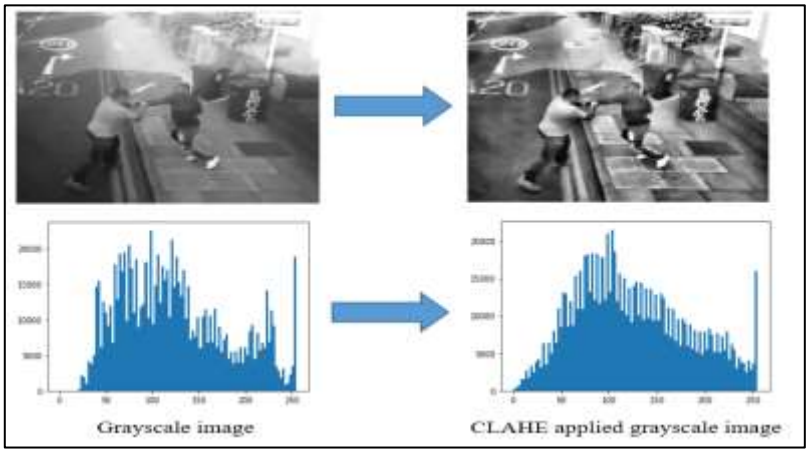

Figure 2: Samples of The Surveillance Fight Dataset

Edge detection

Representing only the edges of images considerably decreases the amount of data that should be processed while preserving critical information about the shapes of objects in the environment. This representation of an image is simple to implement in a wide variety of object identification and classification algorithms used in the computer vision and other image processing applications. The primary characteristic of the edge detection method is its ability to extract the exact edge line in a well-oriented form. Edge detection algorithms turn images into edge images that take advantage of the variations in grey tones within the images [15]. Algorithms for edge detection including Sobel, Canny, Prewitt, Roberts, Laplacian, and Zero Crossing are mostly employed. The Canny's method is the most effective for object extraction

in the majority of situations, according to analysis findings, since it produces fewer incorrect edges [16]. Figure (3) illustrates the effect of the edge detector.

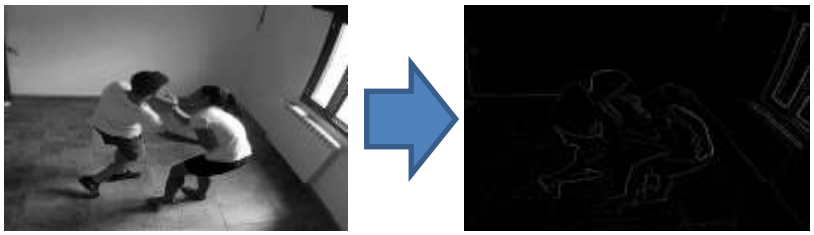

Figure 3: Edge Detector Effect

Canny edge detecting algorithm

The Canny edge detection algorithm adheres to a set of standards meant to enhance existing edge detection methods. First and foremost is a low miss rate. It is crucial that edges in images are not overlooked, and that actions to non-edges are not generated. The second standard is well-localized edge points. To put it in another way, the space between the detected edge pixels and the real edge must be as small as possible. The third standard is to have a single reaction to a single edge.[17]. This algorithm consists of four steps as illustrated in figure (4)

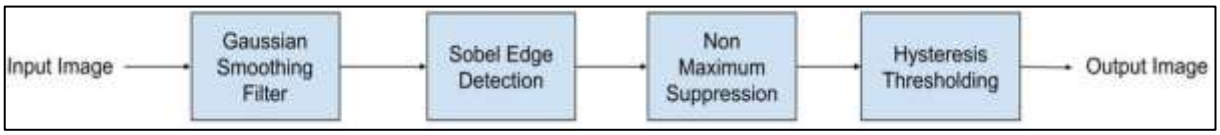

Figure 4: Canny Algorithm Steps

Methodology

The proposed model consists of four phases namely: data acquisition, data preprocessing, training, and finally testing as shown in the block diagram of figure (5).

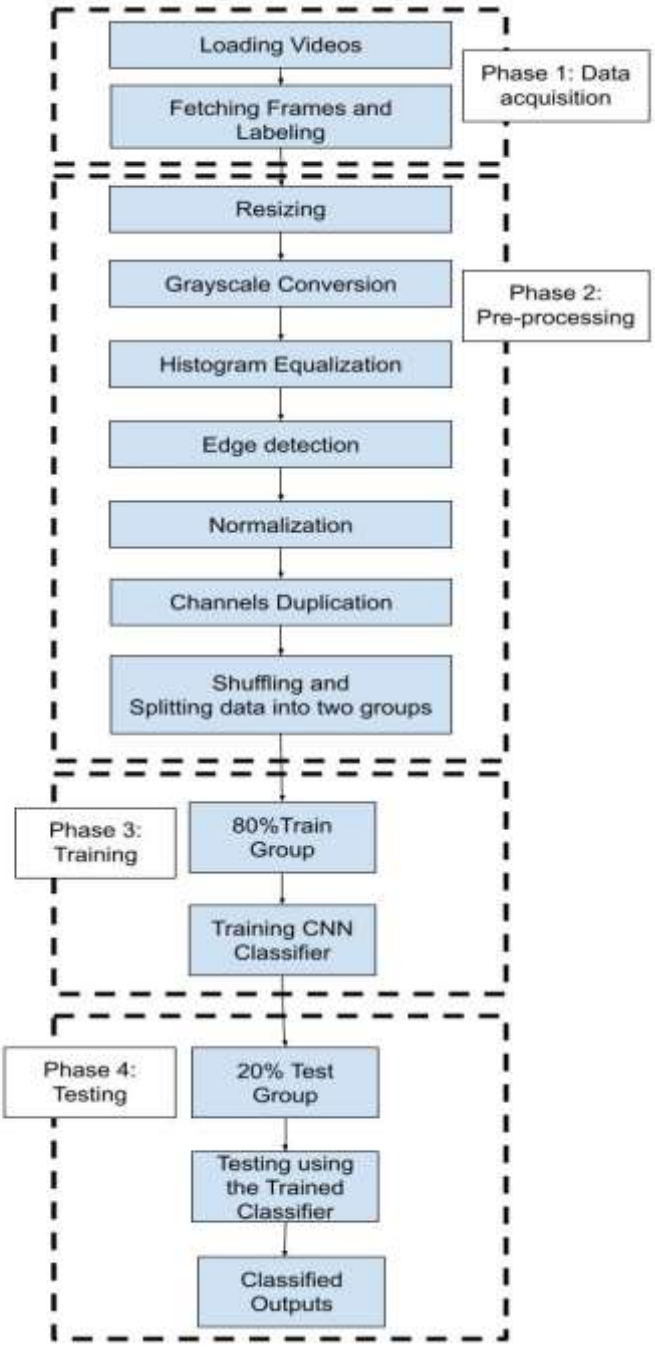

Figure 5: Block Diagram of the Proposed Model

Phase 1: Data Acquisition and Labeling

In this phase, the video files are converted into frames. The frames are gathered every (0.5) seconds for the first dataset and (0.2) seconds for the second dataset to prevent duplication. Also, reducing the total number of frames fetched from the video file. Then, each fetched frame will be labeled (0 for violent or 1 for non violent). The block diagram shown in figure (6) clarifies the above operation, and table (1) provides statistics for the acquisition operation.

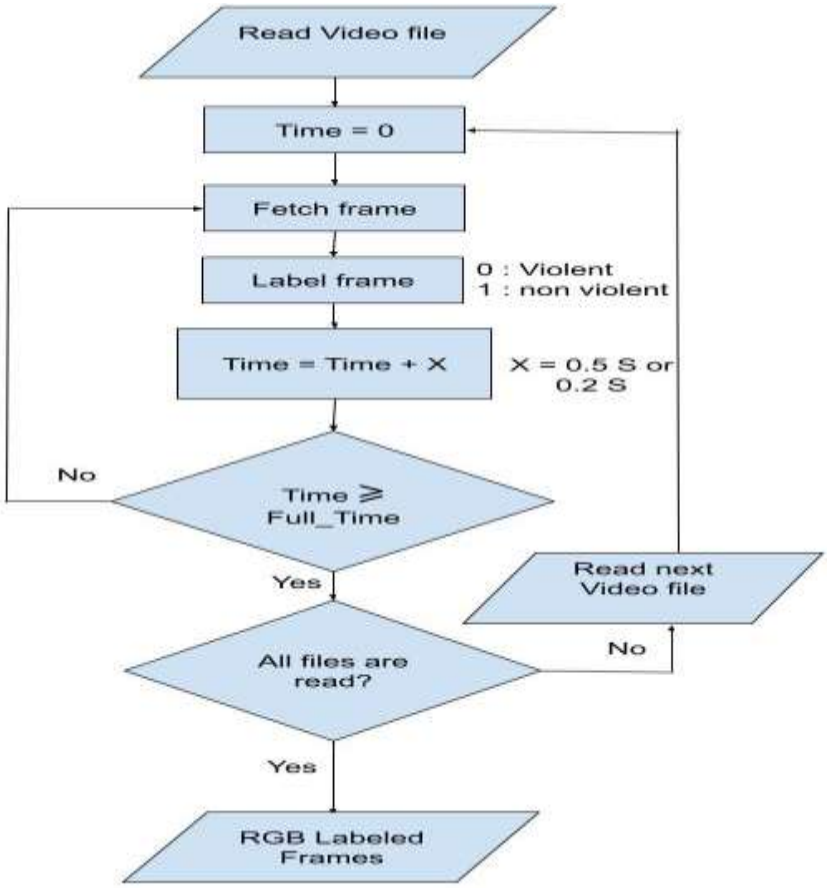

Figure 6: Data Acquisition and Labeling

Table 1: Frames Acquisition Statistics

Dataset	Behavior Type	No. of Videos	Time Gap Between Frames	No. of Fetched Frames for Each Class	Total Frames
Dataset1	Non violent	150	0.1 Sec	3095	6474
	Violent	150		3379	
Dataset2	Non violent	120	0.4 Sec	1856	4089
	Violent	230		2233	

Phase 2: Pre-Processing

In this phase, the fetch frames are going through the following steps:

- Resize the frames to (244 * 244) for faster processing time
- Convert the resized frames to a grayscale image because it's easier for processing than using colored frames.
- Apply Histogram equalization using CLAHE [14] on the grayscale frame to enhance the quality.
- Apply Canny edge detector to reduce the amount of information the frame.
- Normalize the pixel values from (0-255) rand to (0-1) range by using the equation 1:
  $Normalized\ pixel\ value = \frac{Pixel\ Value}{255}$ ..... (1)
- Apply the merge function to duplicate one channel(grayscale) three times so the frame can be processed by the CNN.
- Frames are shuffled and split into two groups: training and testing. The process of shuffling is used to prevent over fitting problem. The ratio of splitting is 80% training and 20% testing.

Phase 3: Training

In this phase, (80%) of the shuffled frames are used to train the CNN network. Since CNN-VGG16 is a pre-trained CNN, and transfer learning is implemented. So, it gives better accuracy in the feature extraction of shapes. CNN-VGG16 consists of 16 layers, but in this research, the last four layers (3 dense layers and the output layer) are removed because they are trained to classify 1000 general objects and replaced with these 3 layers:

1)**Flatten layer()**: used to convert all the 2-Dimensional arrays received from the previous layer into one long, continuous linear vector.

2) **Dense layer (size: 256)**: which contains 256 neurons using "RELU[18]" activation function.

3 **Output layer (sigmoid):** The output layer is using the Sigmoid activation function[19] which has only one neuron since it is binary classification (either 0 or 1); note that there is a dropout layer between the dense layer and the output layer value of (0.5).

The first 13 layers of CNN-VGG16 are using "RELU" activation function, and they are restrained from updating its Neurons weights (not trainable). Therefore, they are used as a feature extractor and only the new added layers are trainable. Furthermore, the learning rate is (0.2), the used optimizer is "Adam[20]" and the loss function is "binary cross-entropy".

Phase 4: Testing

In this phase, 20% of the data set frames are tested using the modified CNN-VGG16 network.

Experimental Results

Programming Environment

The proposed model is implemented using Google Colab (GC) . GC is a Python programming environment using the cloud technology offered by Google. GC Offers two versions of the services, free and paid versions. Since the paid version provides more RAM (25 GB instead of 12 GB) the paid version (which costs 9.99$ monthly) is selected, the hardware specifications of the selected service are:

- CPU: Intel(R) *Xeon*(R) CPU @ 2.30GHz
- GPU: K80(25GB),T4(16GB),P100(12GB) or V100(16GB) (which ever available during the time of execution)
- RAM: 25 GB
- HDD: 124 GB

Datasets

Two datasets are used to train and test the proposed model individually. Both datasets contain a collection of short video clips(*.mp4 files) which are divided into two folders. The first folder

is called “Violent” and as the name indicates, it consists of video clips with violent actions. The second is called “non-Violent” which holds video clips of normal activities:

First Dataset: Automatic Violence Detection Dataset(AvdDS)[21]: Total of 350 clips MP4 video files with a resolution of 1920 x 1080 and a frame rate of 30 FPS are included in this collection. The clips are divided as 120 clips as non-violent, and 230 as violent. All of them were recorded in an indoor environment using fixed camera as shown in figure (7).

Figure 7: Samples of the Automatic Violence Detection

Second Dataset:Surveillance fight dataset(SfDS)[22]: This dataset, which was compiled from YouTube, includes both violent and non-violent episodes that were recorded in the real world by security cameras in both indoor and outdoor situations. There are (300) videos in this dataset, having (150) clips for each incident category. The resolution dimension are ranges from (480 * 360) to (1280 * 720), and the frame rate ranges from (10 to 30)FPS . All the YouTube incidents were recorded by fixed camera with different angles and daytime as shown in figure(8)**.**

Figure 8: Samples of The Surveillance Fight Dataset

Results of Dataset1

After some experiments using batch size (32), the model needs (100 epochs and 162 steps) in each epoch to achieve higher accuracy results. The same parameters are used in the two cases. The results show:

Case 1: Without Edge Detection (Only Applying HE)

Every epoch lasted around 39 sec and the highest training accuracy was (**0.9345**) and the training loss: (**0.1626**) was achieved in epoch 100. The value of (Accuracy, loss, validation accuracy, and validation loss) according to the epoch number are presented in the line graphs in figure (9) and figure (10), also the confusion matrix is shown in figure (11).

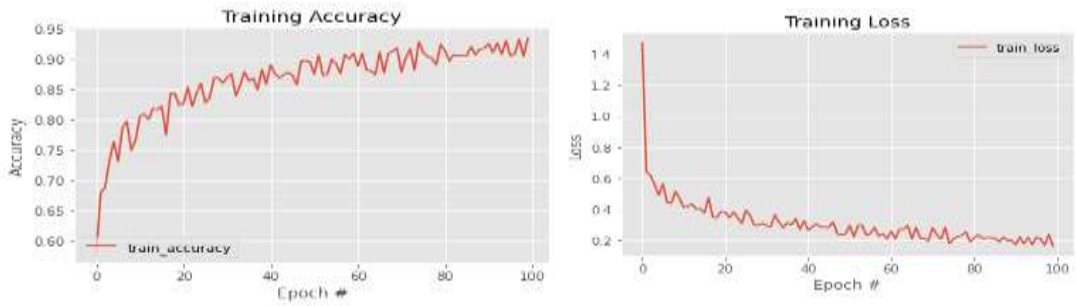

Figure 9: Line Graphs of Training Accuracy And Training Loss Against Epoch

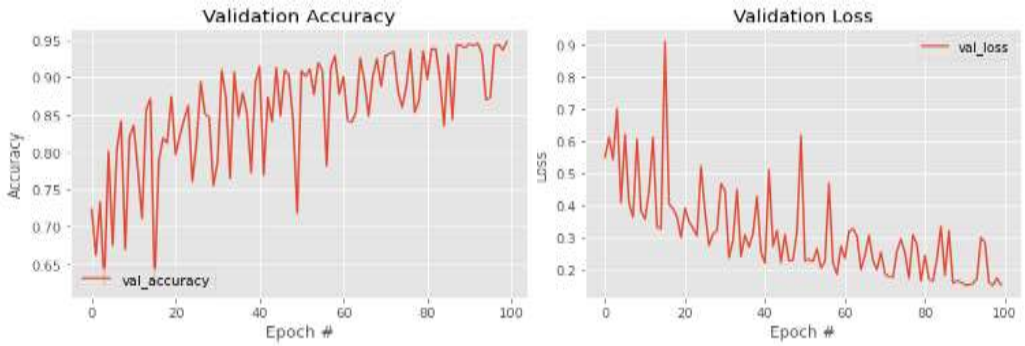

Figure 10: Line Graphs of Validation Accuracy And Validation Loss Against Epoch

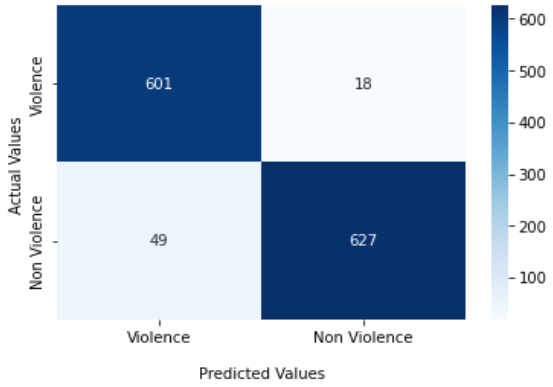

Figure 11: Confusion Matrix

Case 2: With Edge Detection (Using Canny Filter)

Every epoch lasted around 19 sec and the highest training accuracy was (**0.9102**) and the training loss: (**0.2172**) was achieved in epoch 100. The value of (Accuracy, loss, validation accuracy, and validation loss) according to the epoch number are presented in the line graphs in figure (12) and figure (13), also the confusion matrix is shown in figure (14).

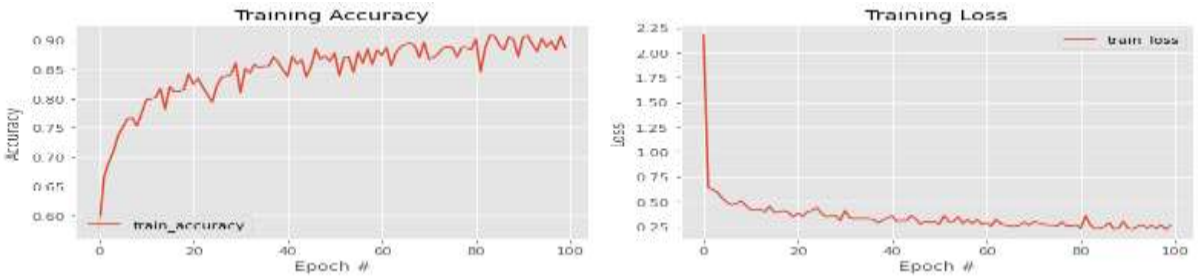

Figure 12: Line Graphs of Training Accuracy And Training Loss Against Epoch

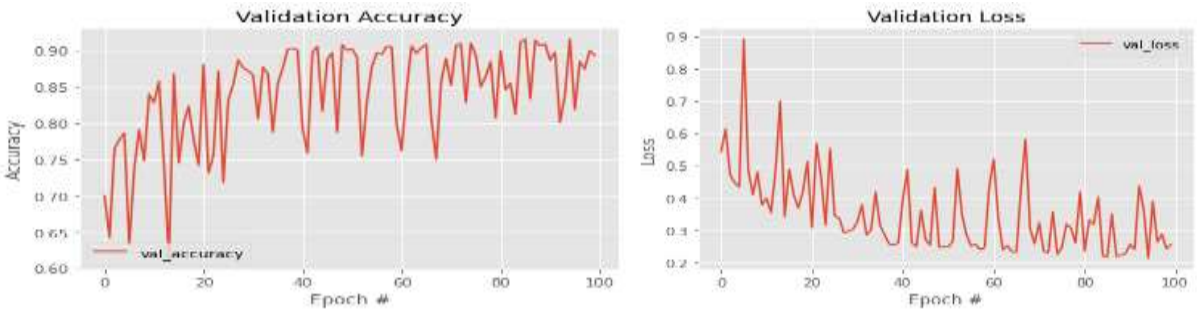

Figure 13: Line Graphs of Validation Accuracy And Validation Loss Against Epoch

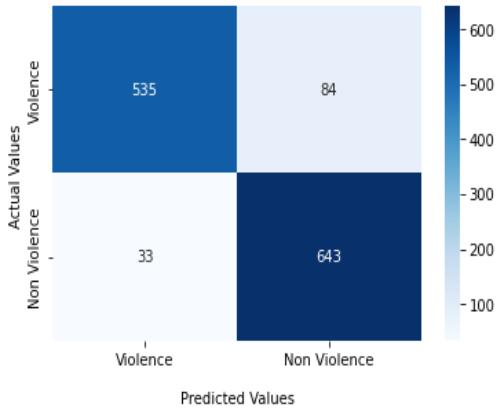

Figure 14: Confusion Matrix

Results of Dataset2

After some experiments using batch size (32), the model needs (50 epochs and 103 steps) in each epoch to achieve higher accuracy results. The same parameters are used in the two cases. The results show:

Case 1: Without Edge Detection (Only Applying HE)

Every epoch lasted around 26 sec and the highest training accuracy was (**0.9872**) and the training loss: (**0.1385**) was achieved in epoch 100. The value of (Accuracy, loss, validation accuracy, and validation loss) according to the epoch number are presented in the line graphs in figure (15) and figure (16), also the confusion matrix is shown in figure(17).

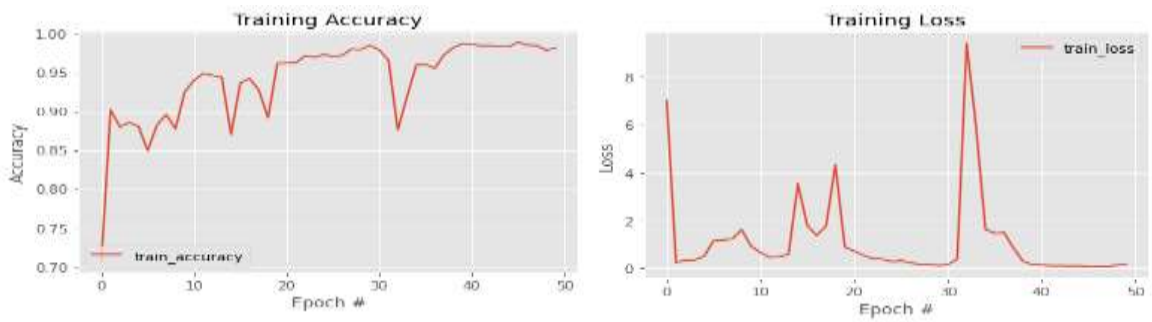

Figure 15: Line Graphs of Training Accuracy And Training Loss Against Epoch

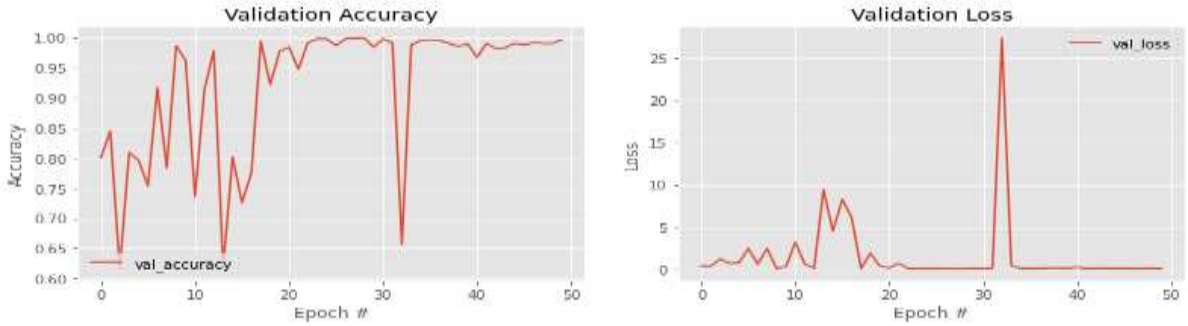

Figure 16: Line Graphs of Validation Accuracy And Validation Loss Against Epoch

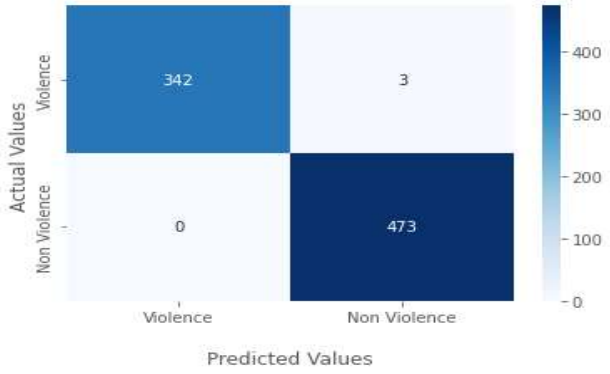

Figure 17: Confusion Matrix

Case 2: With Edge Detection (Using Canny Filter)

Every epoch lasted around 13 sec and the highest training accuracy was (0.8560) and the training loss: (0.3406) was achieved in epoch 45. The value of (Accuracy, loss, validation accuracy, and validation loss) according to the epoch number are presented in the line graphs in figure (18) and figure (19), also the confusion matrix is shown in (figure (20).

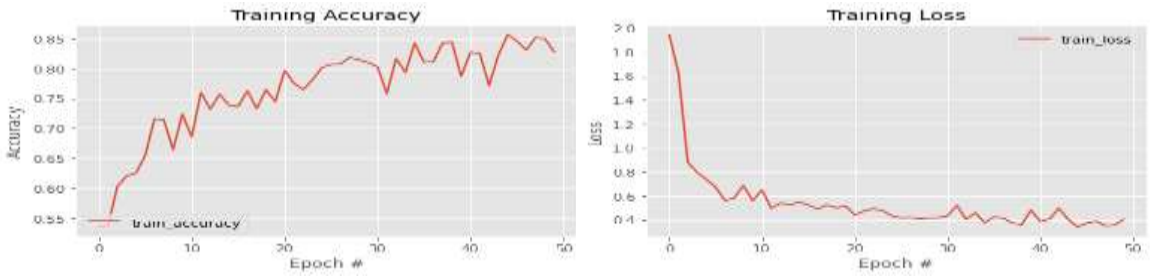

Figure 18: Line Graphs of Training Accuracy And Training Loss Against Epoch

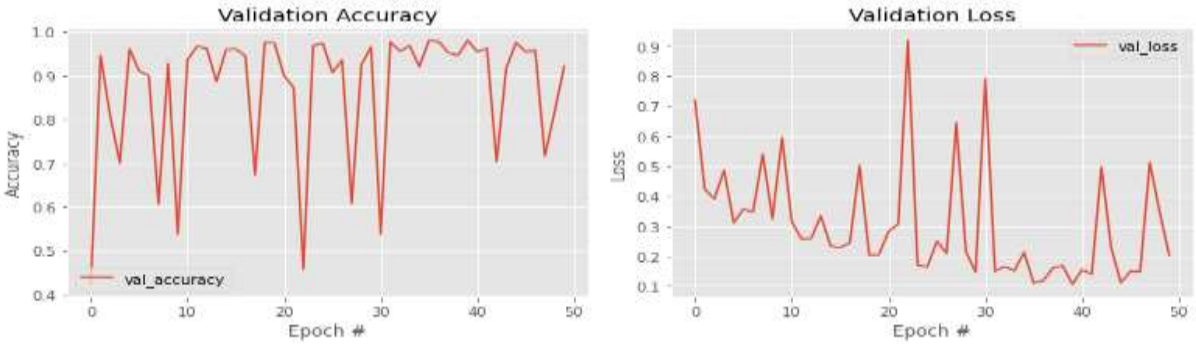

Figure 19: Line Graphs of Validation Accuracy And Validation Loss Against Epoch

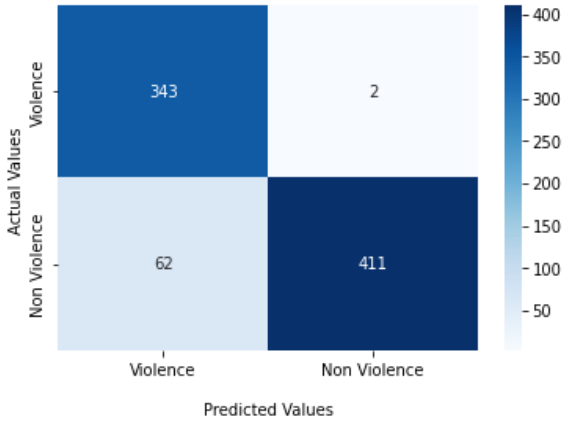

Figure 20: Confusion Matrix

Proposed Model vs. Previos Works

In Table (2) a comparison between the proposed model and the previos works is presented according to the accuracy results.

Table 2: proposed models against previous works comparison

Dataset	Researchers	Methodology	Accuracy
Dataset1 (SfDS)	F. U. M. Ullah *et al*[9]	Lightweight CNN for processing video stream acquired through vision sensor, and residential optical flow CNN used for extracting temporal optical flow features	74%
	M. S. Kang, R. H. Park, and H. M. Park [8]	MSM + EfficientNet-B0 with frame-grouping + TSE Block	92%
	proposed model	Modified VGG16: Only CLAHE	94%

		Modified VGG16: Using Canny filter	91%
Dataset2 (AvdDS)	M. Haque, S. Afsha, and H. Nyeem [10]	Deep CNN Model with GRU	90%
	proposed model	Modified VGG16: Only CLAHE Modified VGG16: Using Canny filter	99% 92%

Conclusions

Presently, the rate of violence is rising dramatically, posing a hazard to individuals, buildings, and institutions. There has always been a need for a more effective method to assist security in monitoring violence. In this paper, a low-complex network model is proposed that can detect violent actions in scene. This network uses the transfer learning method to implement the pre-trained modified CNN-VGG16 to extract features at the frame level. Two open datasets are used to train and evaluate the proposed model which are Automatic Violence Detection Dataset (AvdDS) and Surveillance fight dataset (SfDS), and this model achieved high accuracy results after applying preprocessing steps. Also, it has been shown that applying the Canny filter on the frames reduced the amount of processing time to half for training and testing the model, which is a key feature, yet the filter lowered the accuracy result slightly. In the experiment accuracy results for Dataset 1 were 94% and 91% after applying the Canny filter, in dataset 2, the accuracy was 99% and 92% using canny filter.

References

1. A. P. Association, Diagnostic and statistical manual of mental disorders (5th ed.) 2013
2. C. Mencacci, Quad. Ital. di Psichiatr., 30(1), 1–2(2002)
3. A. Ben Mabrouk, E. Zagrouba, Pattern Recognit. Lett., 92, 62–67(2017)
4. P. S. Arvindbhai, CNN and RNN based Deep Learning Models for Hand Gesture Recognition,(GUJARAT TECHNOLOGICAL UNIVERSITY, 2021)
5. N. O'Mahony,Adv. Intell. Syst. Comput., 943, 128–144(2020)
6. A. Traore, M. A. Akhloufi, Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks, In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 154–159(2020)
7. N. Honarjoo, A. Abdari, and A. Mansouri, Violence detection using pre-trained models, In: 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA),1–4(2021)
8. M. S. Kang, R. H. Park, H. M. Park, IEEE Access, 9, 76270–76285(2021)
9. F. U. M. Ullah, Int. J. Intell. Syst., 36, 1–23(2021)

10. M. Haque, S. Afsha, H. Nyeem, Developing BrutNet: A New Deep CNN Model with GRU for Realtime Violence Detection, In: 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), 390–395(2022)
11. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14(2015)
12. H. A. Khan, W. Jue, M. Mushtaq, M. U. Mushtaq, Math. Biosci. Eng,17(5), 6203–6216(2020)
13. R. Jaiswal, A. G. Rao, H. P. Shukla, Int. J. Electr. Electron. Eng., 1, 69–78(2010)
14. M. Siddhartha, A. Santra, arXiv Prepr. arXiv2006.13873(2020)
15. R. Muthukrishnan, M. Radha, Int. J. Comput. Sci. Inf. Technol., 3(6), 259(2011)
16. S. K. Katiyar, P. V. Arun, arXiv Prepr. arXiv1405.6132,(2014)
17. B. M. L. P. Vigil, ACCELERATING THE CANNY EDGE DETECTION ALGORITHM WITH CUDA/GPU.
18. N. Kulathunga, N. R. Ranasinghe, D. Vrinceanu, Z. Kinsman, L. Huang, Y. Wang, arXiv Prepr. arXiv2010.07359, (2020)
19. C. Nwankpa, W. Ijomah, A. Gachagan, S. Marshall, arXiv Prepr. arXiv1811.03378, (2018)
20. D. P. Kingma, J. Ba, arXiv Prepr. arXiv1412.6980, (2014)
21. M. Bianculli, Data Br., 33, 106587(2020)
22. S. Akti, G. A. Tataroglu, H. K. Ekenel, Vision-based Fight Detection from Surveillance Cameras, In: 2019 9th International Conference on Image Processing Theory, Tools and Applications, IPTA 2019, 1–6(2019)