

Optical Flow based Violence Detection from Video Footage using Hybrid MobileNet and Bi-LSTM

by

Tashfia Haque
18201140

Farhan Fuad Ahmed
19101549

S. M. Irfan Ahmed
19101390

Mohammad Siam
23141065

A thesis submitted to the School of Data and Sciences
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science & Engineering

School of Data and Sciences
Brac University
September 2023

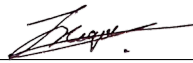
© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Tashfia Haque

18201140



Farhan Fuad Ahmed

19101549



S. M. Irfan Ahmed

19101390



Mohammad Siam

23141065

Approval

The thesis titled “Optical Flow based Violence Detection from Video Footage using Hybrid MobileNet and Bi-LSTM” submitted by

1. Mohammad Siam (23141065)
2. S. M Irfan Ahmed (19101390)
3. Farhan Fuad Ahmed (19101549)
4. Tashfia Haque (18201140)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science on September 21, 2023.

Examining Committee:

Supervisor



Dr. Md. Golam Rabiul Alam, PhD

Professor
Computer Science & Engineering
Brac University

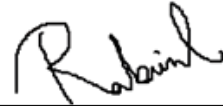
Co-Supervisor:



Md. Faisal Ahmed

Lecturer
Computer Science & Engineering
Brac University

Program Coordinator:



Dr. Md. Golam Rabiul Alam, PhD

Professor
Computer Science & Engineering
Brac University

Head of Department:
(Dean)

Mahbubul Alam Majumdar, PhD

Professor
School of Data and Sciences
Brac University

Ethics Statement

As the authors of the thesis paper entitled “**Optical Flow-based Violence Detection from Video Footage using Hybrid MobileNet and Bi-LSTM,**” we collectively declare our firm commitment to upholding ethical principles during all stages of this research project. Our collective effort is guided by the following basic ethical principles:

Privacy and Consent: We commit to abiding by all relevant legal statutes and regulations pertaining to privacy and informed consent in the process of obtaining video data. Our group is dedicated to ensuring the protection of the private rights of individuals portrayed in the recorded visual content. In instances where it is deemed necessary, the process of obtaining consent shall be diligently pursued, and the practice of data anonymization will be rigorously implemented in order to safeguard the identity of those involved.

Data Security: Both parties share the duty of establishing strong data security measures to guarantee the confidentiality and integrity of the video data utilized in this research. To mitigate the risk of unauthorized access or data breaches, we shall implement robust safe storage methods, stringent access controls, and encryption mechanisms.

Bias and Fairness: The primary objective of our group is to actively acknowledge and reduce bias, as well as uphold principles of fairness, during the entire process of developing the violence detection model. The selection of datasets will prioritize the representation of various populations, and a collaborative effort will be made to detect and address any potential biases that may arise during the processes of data collection, preprocessing, and model development.

Transparency and Accountability: In order to uphold transparency and accountability, it is imperative that we collaboratively document the research methods, algorithms, and approaches employed. Our unwavering dedication is in the resolution of ethical issues and the response to enquiries raised by various stakeholders, encompassing participants, coworkers, and the general public.

Compliance: Collectively, we commit to strictly abiding by the ethical research rules and regulations set forth by institutional, national, and international bodies. This entails the strict adherence to the ethical review procedures set by our academic institution.

Through the combined efforts of our collaborative team, we aim to abide by these ethical principles in order to ensure that our study makes a valuable contribution to the advancement of knowledge, while simultaneously safeguarding the rights, privacy, and dignity of the individuals portrayed in the video data being analyzed.

S. M Irfan Ahmed
Mohammad Siam
Farhan Fuad Ahmed
Tashfia Haque

17 September, 2023

Abstract

This thesis presents a novel approach for the automatic detection, categorization, and sub-categorization of violent and nonviolent behaviors in video footage. This research addresses the growing necessity for enhanced security protocols in both public and private sectors. Surveillance cameras are commonly accessible and easily affordable; however, their utilization is frequently inefficient due to boundaries related to human real-time monitoring. This occurrence may lead to delayed responses to unanticipated events, hence highlighting the need for enhanced and efficient monitoring measures. Our thesis presents a novel approach for the automation of violence detection by utilizing machine learning and deep learning techniques. The techniques applied in this study integrate object and motion detection through the utilization of optical flow analysis and a MobileNet-Bi-LSTM fusion architecture. This methodology exceeds conventional methods by incorporating both temporal dynamics and spatial features. We have invested notable efforts in enhancing our dataset acknowledging the significance of training an efficient violence detection system. In addition to the existing dataset, we have systematically compiled an adequate number of video footage. The compiled videos contain a diverse array of circumstances, effectively representing a variety of environments, lighting conditions, and situations. The inclusion of this range is crucial in facilitating our model's ability to generalize and adapt to real-world scenarios seamlessly. A thorough annotation procedure of meticulous labeling of 'violent' 'non-violent' actions, along with specific subcategories of violence like 'Beating,' 'Use of Weapons,' and 'Burning' was done to uphold the standards of quality and precision in the enhanced dataset. For an in-depth review, a comparison study was undertaken to examine two unique methodologies. The first approach centers on the categorization of actions into two distinct categories: 'Non-Violence' and 'Violence,' based on a binary classification system. The second approach entails the categorization of behaviors of the videos of our unique dataset named 'Beating-Burning-Weapon (BBW) Violence' Dataset into two main groups, namely 'Non-Violence' and 'Violence,' which further subdivided into three sub-categories of violence, which are 'Beating,' 'Burning,' and 'Use of Weapons.' In our comprehensive evaluation of violence detection methods, we tested two violence detection methods on the two previously mentioned datasets. The 'Frame Selection at Equal Intervals' method achieved higher accuracy, 90.16% in the 'Real Life Violence Situations (RLVS)' Dataset and 85.32% in the BBW Violence Dataset, making it a precise choice. On the other hand, the 'Merged Frame Stacking' method, offering computational efficiency, achieved respectable accuracies of 85% and 74% in the RLVS and BBW Violence Datasets respectively. This provides a foundational baseline for violence detection, thus highlighting method-specific advantages and trade-offs. Our research holds significant potential for proactive security management by promptly detecting and responding to possible threats.

Keywords: Surveillance Camera, Violence Detection, Machine Learning, Deep Learning, Motion Detection, Beating, Use of Weapons, Burning, Optical Flow, Bidirectional Long Short-Term Memory (Bi-LSTM), MobileNet V2, Crime Detection, Real-Time Monitoring, Proactive Security Management, Image Classification

Acknowledgement

Firstly, all praise to the Great Almighty for whom our thesis has been completed without any major interruption. Secondly, to our co-advisor, Md. Faisal Ahmed sir for his kind support and advice in our work. He helped us whenever we needed help. Thirdly, the whole judging panel for all the reviews they gave. It helped us a lot in our later work. And finally, to our parents without their thorough support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
List of Tables	x
1 Introduction	1
1.1 Motivation	2
1.2 Problem Statement	2
1.3 Research Objective	3
1.4 Research Contributions	3
1.5 Thesis Organization	4
2 Related Works	5
3 Dataset	11
3.1 Dataset Description	12
3.2 Dataset Collection	14
3.3 Data Pre-Processing	15
4 Methodology	18
4.1 Model Specifications	19
4.1.1 Dense Optical Flow	19
4.1.2 Bi-directional Long Short-Term Memory (BiLSTM)	20
4.1.3 MobileNetV2	22
4.1.4 Proposed OptiMobBi-LSTM Model	23
5 Result and Discussion	24
5.1 Performance Evaluation Measures	24
5.2 Experimental Results	25

5.2.1	OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on Real Life Violence Situations Dataset	25
5.2.2	OptiMobBi-LSTM Model Based Merged Frame Stacking on Real Life Violence Situations Dataset	26
5.2.3	OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset	28
5.2.4	OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset	29
5.3	Discussion	31
6	Conclusion	33
6.1	Future Works	33
	Bibliography	37

List of Figures

3.1	RLVS Dataset	12
3.2	BBW Violence Dataset	13
3.3	Original Frames	16
3.4	Processed Frames (first method)	16
3.5	Merged Frame (second method)	17
4.1	Workflow	18
4.2	Bi-LSTM	21
4.3	MobileNetV2	22
4.4	Proposed OptiMobBi-LSTM Model	23
5.1	OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on Real Life Violence Situations Dataset	26
5.2	OptiMobBi-LSTM Model Based Confusion Matrix of Frame Selection at Equal Intervals on Real Life Violence Situations Dataset	26
5.3	OptiMobBi-LSTM Model Based Merged Frame Stacking on Real Life Violence Situations Dataset	27
5.4	Confusion Matrix of OptiMobBi-LSTM Model Based Frame Stacking on Real Life Violence Situations Dataset	27
5.5	OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset	28
5.6	Confusion Matrix of OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset	29
5.7	OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset	30
5.8	Confusion Matrix of OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset	30
5.9	Overall Comparison	32

List of Tables

3.1	RLVS Dataset (Kaggle)	14
3.2	BBW Violence Dataset	14
5.1	OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on Real Life Violence Situations Dataset	25
5.2	OptiMobBi-LSTM Model Based Merged Frame Stacking on Real Life Violence Situations Dataset	27
5.3	OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset	28
5.4	OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset	29
5.5	Overall Comparison	31

Chapter 1

Introduction

In an era of technological progress, video cameras are essential to our commitment to safeguarding public spaces and ensuring security. Surveillance systems offer an effective method for monitoring human actions and identifying potentially unlawful or criminal conduct [11]. Despite their extensive deployment, current surveillance systems encounter certain constraints to reach their full potential [5]. The reliance on human operators to monitor video footage often results in the oversight of suspicious activities, undermining the system's prompt identification of such behaviour. Furthermore, the post-incident analysis of video recordings proves to be a time-consuming and inefficient endeavour, doing little to suppress rising crime rates [20]. Addressing these limitations, the integration of automated software and intelligent algorithms emerges as a promising avenue to unlock the full potential of surveillance systems. [9]. By doing so, the detection of unusual behaviours and violent activities can be substantially improved [6], [7]. Advanced software solutions have emerged that efficiently organize digital video footage into searchable databases, streamlining analysis procedures and enabling quicker and more accurate detection [8]. Harnessing these technologies optimally positions surveillance systems as powerful solutions, particularly by enhancing their effectiveness and efficiency in real-time monitoring and detection. The landscape of violence detection research has witnessed a notable shift, prompting our study's evolution [17].

This thesis aims to transcend the constraints of existing surveillance systems, introducing an innovative methodology for detecting both violent and non-violent behaviours. Furthermore, it endeavours to differentiate and classify specific types of violence, including 'Beating', 'Burning', and 'Use of Weapons'. These sub-sectors represent distinct manifestations of violence, demanding fine detection approaches. Recognizing the need for a comprehensive evaluation of our methodology's effectiveness, we have incorporated a comparative analysis between two distinct methods. The first method focuses on binary classification of 'Non-Violence' and 'Violence'. The second method involves the classification of behaviours into 'Non-Violence' and 'Violence' with sub-categories encompassing 'Beating,' 'Burning,' and 'Use of Weapons'. This meticulous categorization enables a refined understanding of the nature of detected violence. This comparative examination empowers us to measure the performance of our approach across these dimensions. In response to the shortage of appropriate and relevant datasets within contemporary research communities, our work undertakes a proactive measure. The creation of the new dataset in

conjunction with the existing dataset [23], addresses our novel approach. This augmentation effort, supported by detailed annotation, enables a rich repository capable of refining our model’s understanding of diverse events. The predicted outcome is a comprehensive framework capable of distinguishing between violent and non-violent actions and further classifying specific types of violence. This advancement bears transformative potential for proactive security management, enhancing surveillance systems’ effectiveness in swift threat detection and response. This thesis seeks to advance the realm of violence detection in surveillance videos by deploying optical flow [30] analysis alongside a Hybrid MobileNet-Bi-LSTM architecture. Through enhanced classification techniques and an enriched dataset, the study endeavours to create a paradigm shift in automated surveillance systems, trained to deliver precise, efficient responses to a range of security threats.

1.1 Motivation

The rising demand for enhanced surveillance systems that are able to detect suspicious behaviour [29] and acts of violence in an accurate and timely manner is what motivated us to write this thesis. The limitations of human operators in the context of real-time monitoring underscore the necessity for automated solutions. This thesis aims to enhance the effectiveness and efficiency of surveillance systems by implementing machine learning and deep learning approaches. This will be achieved by creating a dataset with a mixture of more relevant videos, categorizing specific violent acts (‘Beating,’ ‘Burning,’ ‘Use of Weapons’) for greater precision. Moreover, a novel approach was taken by including optical flow [30] in the pre-processing stage. The primary objective of this study is to contribute towards a worldwide problem by enhancing public safety and security procedures through the utilization of advanced algorithms for automated detection processes [17].

1.2 Problem Statement

The global rise in terrorism and violent incidents necessitates the widespread implementation of CCTV and IP surveillance systems to ensure security [29]. However, the reliance on human operators in these systems compromises their effectiveness, leading to underreported instances of violence and suspicious behaviour [20]. Furthermore, the post-incident investigation process is often protracted and fails to effectively reduce crime rates. Prior research has explored various methodologies, emphasizing the importance of analyzing appearance and motion features for precise violence detection [15], [16]. Techniques like key frame extraction and sampling have shown promise in achieving heightened precision [19]. This thesis addresses these critical challenges by developing a new crime detection system. Our main motivation is to enhance the efficiency and accuracy of surveillance systems. To achieve this, we created a new dataset and included more relevant videos. Following that, by incorporating advanced algorithms for automated detection processes have been used. In essence, this thesis seeks to bridge the gap in surveillance systems, ensuring that they can accurately and promptly identify violent behaviour, thereby contributing to the improvement of public safety and security procedures on a global scale [29].

1.3 Research Objective

The main purpose of this thesis is to create an advanced violence detection system through the utilization of machine learning and deep learning techniques [28] including object and motion detection, along with posture estimation, and other significant methodologies, to ensure a thorough analysis. Our primary objective is to significantly enhance the identification and classification of violent incidents in order to get an increased degree of precision in detection. Furthermore, we undertake a comprehensive examination of current methodologies for detecting violence, with the objective of identifying their respective merits, constraints, and prospects for improvement [21]. Additionally, this work aims to solve the existing lack of diverse datasets within the field [23]. Finally, we conduct thorough testing and evaluation of the performance of the violence detection system that has been established. This involves comparing it with established methods and benchmarks in order to assess its accuracy, efficiency, and reliability. The research has the potential to significantly improve security and safety on a larger scale.

1.4 Research Contributions

Our thesis contributions encompass dataset refinement, novel methodological innovation, enhanced preprocessing techniques, and the introduction of a sub categorization schema. These collective contributions substantially elevate the accuracy, comprehensiveness, and sophistication of violence detection systems. Additionally, our work establishes a robust foundation for ongoing exploration in the domain of event categorization, offering a pathway for the incorporation of additional categories in future research endeavors. The datasets were efficiently processed using optical flow [30] and afterwards inputted into a hybrid model consisting of BiLSTM [26], [27] and MobilenetV2 [22]. This approach yielded a more accurate outcome in the identification of violent incident.

The key contributions are concisely described as follows:

1. This research commenced with the meticulous creation of an enriched dataset named BBW Violence Dataset. This new dataset was thoughtfully structured to encompass a diverse spectrum of 1,150 violent videos which includes categorical violence manifestation of 625 videos as ‘Beating’, 252 videos as ‘Burning’, and 273 videos as ‘Use of Weapon’. Notably, the augmentation of this dataset which is based on numerous features such as pixel quality, fps rates, and duration sourced from various online platforms yielded a notable advancement in the field of violence identification in categorical classifications.
2. A pivotal contribution of our research was the introduction of a pioneering methodology tailored to detect non-violent and categorical violent events. This methodological innovation revolved around the concurrent utilization of both our original dataset named BBW Violence Dataset and existing RLVS Dataset [23]. We used a novel approach by integrating optical flow-based features into the data preprocessing pipeline. This integration underscored the pivotal role of optical flow-based features in augmenting the discriminatory prowess of

violence detection models. Subsequently, by harnessing the collective power of BiLSTM [26], [27] networks and MobileNetV2 [22] algorithms, this novel approach significantly improved accuracy measures in the realm of violence classification.

3. Another notable contribution of our research was the formulation of a label-based categorical violence detection schema. This scheme facilitated a more definitive identification of violent events across three distinct categories. More importantly, this meticulous categorization framework not only enhances the precision of violence classification but also offers a foundation for future research endeavors. Our thesis stands open to the prospect of introducing additional categories, both violent and non-violent, to further enrich the field of violence detection.

1.5 Thesis Organization

This thesis is organized into seven main sections. The introduction (Chapter 1) establishes the background by presenting the motivation, problem statement, research objectives, and research contributions made by the study. Chapter 2 explores the literature review, delivering significant background knowledge. Chapter 3 focuses on the dataset, including its description, collection methods, and data pre-processing. Chapter 4 provides a comprehensive analysis of the proposed OptiMobBi-LSTM Model in detail, encompassing Dense Optical Flow, Bi-directional Long Short-Term Memory (BiLSTM), MobileNetV2, and the Proposed OptiMobBi-LSTM Model. This chapter delves into an extensive review of these models. Chapter 5 presents the results and discussions, beginning with an exploration of performance evaluation measures, followed by an analysis of experimental results for frame selection and merged frame stacking techniques on both the BBW Violence Dataset and Real Life Violence Situations Dataset. This chapter also includes a series of graphs, charts, and tables that provide a visual representation of the findings, enhancing the reader's understanding and facilitating a deeper analysis of the results. Chapter 6 concludes the thesis, highlighting key findings and suggesting future research directions. The bibliography cites the sources referenced throughout the thesis.

Chapter 2

Related Works

This chapter aims to conduct an in-depth review of the existing literature, identifying areas where knowledge is lacking, offering valuable insights into the theoretical framework, and situating the study objectives within the wider academic discourse.

In recent years, the development of reliable automatic surveillance systems has attracted significant attention, particularly in areas prone to recurrent criminal activities. The necessity for real-time violence recognition has become paramount in enabling swift police responses during criminal incidents. Various techniques for action recognition have been explored, with 3D Convolutional Neural Networks (CNNs) emerging as a prominent method.

Traditional CNNs are capable of classifying individual image frames and can process 2D inputs. However, the dynamics of video data require models that capture both spatial and temporal information effectively. This is where 3D CNNs come into play. These models can extract information from multiple consecutive frames, allowing them to capture the motion encoded within these frames [10]. For instance, a deep 3D-CNN model, trained with 152 ResNets layers on the Kinetics database [18], achieved remarkable results, with accuracy rates of 78.4% on the Kinetics dataset and up to 94.5% on the UCF101 test set. However, a limitation of such approaches lies in the requirement for complete sequences of frames, making them less adaptable to scenarios where actions are not contextually bound. Consequently, some actions may be perceived as constrained or learned with irrelevant contextual data, potentially resulting in over-fitting.

The analysis of both the video's appearance and its motion are essential components in the process of determining whether or not it contains violent content. Researchers are increasingly utilizing audio features to aid in the detection of violent videos because the vast majority of videos contain data from both the visual and auditory modalities. The early works primarily concentrate on hand-crafted aspects of the design. Common appearance descriptors include SIFT, HOG, etc. Motion features that are frequently employed include space-time interest points (STIP), also known as improved dense trajectories (iDT). For auditory features, Mel-frequency cepstral coefficients (MFCC) emerged as a prevalent choice. Afterwards, a classifier is applied to the extracted features to determine an overall rating. Deep neural networks such as 2D ConvNet [13], 3D ConvNet [10], LSTMs [1] have been used by some

researchers to detect violent video in recent years. This improvement can be seen in both the publicly available dataset VSD2015 [12] and VCD. As data from two different modalities, audio and visual signals, there may be a heterogeneity gap issue that prevents the full use of multimodal data. Data from several modalities are to be integrated into a single intermediate common space using this method. Researchers have been exploring methods to integrate data from multiple modalities into a single intermediate common space to bridge this gap.

Efforts in video summarization have led to the development of key frame extraction methods. These techniques aim to efficiently represent a video's content. Key frames play a pivotal role in creating video summaries, allowing for quicker analysis and retrieval of crucial content. Using Key frame extraction Authors of [24], proposed a new sampling method. The video is first segmented based on the keyframes that are retrieved from it, and then the segmented movie is displayed. First, they transform the RGB frames into grayscale and then determine the centroid of the grayscale. The next step is to move on to the second frame, at which point we will be able to determine whether or not the frames share a visual similarity. The concept of visual similarity threshold refers to the minimum number of successive images that show similar visual characteristics to a certain extent. The currently shown frame and any frames that came before it can be thought of as being part of the same sequence of succeeding frames that have a visual similarity. Following this, the selection of the key frame within a sequence is achieved through the use of a filtering process, where the frame showing the shortest distance between the gray centroid and the average gray centroid is selected. This is a novel sampling method they introduced. In most cases, the uniform sampling approach samples every frame or sample frames at predetermined intervals. Videos are split into 16 consecutive frame chunks for 3D ConvNet using a unique uniform sampling technique. When sample movies are kept to a manageable length, the uniform sampling approach is an efficient and straightforward option. However, for longer films, the fixed sample technique introduces redundancy and motion discontinuities. They tested their method against traditional means which resulted in a new sampling method achieving 94.3 percent accuracy whereas the traditional method scored 93.5 percent.

In response to the increasing prevalence of terrorist attacks and social issues, video surveillance systems have become a focal point of academic research. To improve the effectiveness of video surveillance systems, this research suggests a new method of motion detection. This paper has proposed three modules: background modelling, alarm trigger, and object extraction. When compared to other approaches, the PRO method for motion detection is noticeably more effective. The sgn function is used to predict the background intensity in the first calculation. It is suggested that the MSDE approach be used to construct the flexible backend model. The formula for the underlying model is as follows: Multiple moving objects can be detected with more precision using the Simple Statistical Difference (MSDE) approach than using the SDE method alone. This is because the MSDE technique uses a multi-modal backdrop model to produce a binary mask $D(x, y)$ of moving objects. To describe the adaptive background in the DCT domain, the RADCT algorithm [4] employs a modified RA technique. The RADCT approach generates the adaptive backdrop by using a DCT coefficient, as opposed to the classic RA method's focus

on pixel intensity. In this paper, we introduce a unique motion detection strategy for use in stationary camera surveillance systems. Our method incorporates three proposed components that together enable full detection of moving objects. The proposed BM module begins with the development of a one-of-a-kind two-stage background matching technique that combines speedy matching alongside precise matching. With the help of the proposed AT module. The OE module will only need to process blocks that actually contain moving objects, rather than the full surrounding region. Optimum background modelling's (OBM) primary goal is to isolate the steady signal from the next frame in the video stream. Quickly matching, employing the stable signal trainer, and figuring out the best background pixels are the three key components of OBM. The PRO technique outperforms MSDE, SDE, SSD, and RADCT in terms of precision across the field. Only the PRO approach achieves an overall accuracy of 80% or above across the field [8]. Compute time can be reduced by 14.09% using the proposed AT module. Using only a few simple computations, the suggested AT module can improve motion detection performance overall.

As a central topic of study in the field of computer science, motion detection has inspired a wide variety of methods for its study and solution. To establish whether an item is in motion between a set number of frames, say three, a three-frame difference algorithm is used. Anti-theft and anti-destruction initiatives utilize cutting-edge technologies. With their help, we can track and record an object's every movable detail in real time. However, they largely stand by and do nothing to prevent or reduce criminal activity. In 1965, closed-circuit television (CCTV) was the first step toward what is now known as video surveillance. Whenever there is activity on a live footage feed, it is picked up by a system. When motion is detected, the software will sound an alarm and save the current clip for further review. With the tapes wearing out and the VCR's storage capacity capped at eight hours, video storage was never reliable. It is crucial to be aware of the many possibilities of video surveillance in the areas of security and the monitoring of assets. Organizations need to watch how their proposed video surveillance system is built and run to make sure it doesn't invade people's personal space too much. The latest in video monitoring technology should make users feel very safe. According to Nafisiaty Mbabu, a person may only rest easy if he knows he would be alerted of any possibility of his stuff getting robbed in real-time. Creating a monitor and control setup that could identify movement in real-time footage. As soon as motion is detected, an alarm will sound and the footage will be saved for analysis. One way to deal with a potential security threat is to set off an alert. Motion can be detected in a webcam's footage by analyzing a series of frames captured at a constant rate (frames per second). The AVI format is a hybrid video and audio container. Frames from the video are saved in a sequential order. Pixels from many photos are added up to create a histogram [9]. A greater entropy value is indicative of activity in the area surrounding a given pixel. The best value of T can be determined using the entropy-based threshold method.

For example, computer vision is widely used in the sectors of security and surveillance for the purpose of detecting and monitoring abnormal activities. Detecting irregularities involves looking for things that are out of the conventional or unexpected. Whenever one observes a pattern that deviates from a baseline of expected

behaviours, one is witnessing an anomaly. The suggested system is designed to accurately detect and categorize firearms. COCO is a picture dataset that was made by its creator and features commonly used objects with their labels. The dataset was trained with a Single Shot Detector (SSD) model that was trained in SAS over the course of 2669 iterations with the COCO dataset and the SSD VGG-16 Architecture. [19]. An average of 74% MAP and 59 frames per second can be achieved with these images. Anaconda's python `-i anaconda-xlsd.py` script generates a CSV export of the data. Using an SSD in place of RCNN for weapons identification reduces data loss to 0.05 percent. 72% and 67% accuracy in identifying AK-47s, M1911s, and Smith & Wessons, respectively. As a whole, IEEE's Faster R-CNN is accurate 84.6% of the time and runs at a pace of 1.606 s/frame. The chart evaluates the precision of various firearms against a pre-labeled dataset, including the AK-47, Colt M1911, Smith & Wesson Model 10, UZI Model, Remington, and more. A trained model was created for five distinct firearms using SSSD and R-CNN. These firearms include the AK-47, the Smith & Wesson Model 10, the Colt M 1911, the UZI Model, and the Remington Model. SSD and Faster RCNN achieve better results than self-created image datasets when using a pre-labeled dataset like A K47. The improved performance of RCNN (1.606s/frame) is mediocre in comparison to SSD. However, the accuracy of the faster RCNN is much higher at 84.6%. [25]. THigh-end DSPs and FPGA kits can be trained to do this for larger datasets.

Background modeling plays a pivotal role in motion detection by reliably identifying the static elements in video scenes. A natural scene in a video usually consists of dynamic objects such as shaking trees, swaying curtains, undulating surfaces of the water, waving flags, etc. A reliable method for spotting moving items is made possible by accurately identifying the image's background. There are three stages to video processing: acquiring an image, processing a background image, and de-noising the foreground. Several algorithms are available [3] for this task. The threshold is applied for determining whether a pixel within an image is classified as being a part to the background or the foreground. In this context, N indicates the total count of pixels in an image, while D denotes the spacing between each individual pixel. Additionally, M signifies the dimensions of the image's size. The use of information geometry involves the isolation of entities from their surrounding, which is achieved through the utilization of convex contours and the analysis of the number of edges or corners present on these contours. The Sum of Absolute Differences (SAD) is used as a means of detecting the existence of motion. This method represents the second proposed approach for the identification of non-rigid objects showing a specific geometric shape. The process of motion detection can be described as follows: Acquired images must first undergo preprocessing, during which noise and other defects are fixed. Geometric shapes are then isolated in a second process (square, rectangle, circle, and ellipse). We are using Microsoft Windows 8.1 Professional Compiler C++ under Microsoft Visual Studio and OpenCV 2.4.10 (Open Source Computer Vision Library) as well as Intel(R) Core(TM) i3-380M (2.53 GHz) for our analysis. The objective is to identify moving objects that have a definite geometric outline (square, rectangle, circle, and ellipse). An object's speed determines how long it takes to spot it in motion. If we know where the object was located at two different times, we can figure it out. Pixels are used to determine the distance, and a reduced form of Newton's trigonometric equation is used to convert the

result to meters (m). To assess how well our system performed, we employed the F1-measurement family of metrics in addition to recalling r and precision p . When compared to the second method, the first can identify moving objects travelling at speeds greater than 1.19 m/s [15]. The F1 metric is a hybrid of the scene's object recall and detection accuracy. The final result for an object identification experiment is based on a weighted average of the two individual measurements (F1). We can see from the figures that F1 is superior to the original method when it comes to detecting items that have a certain geometric shape (FP, R, and P). False positives manifest when an object is obscured by another, cast into shadow, or subject to a sudden increase or decrease in brightness. One method can identify things based on their geometric shape (circular or quadrilateral) Image analysis is another method that has been used to solve similar problems.

The importance of computer technology in human detection continues to grow. Computer scientists are constantly creating new algorithms to process both simple and complex jobs, such as video and audio analysis. Many different strategies for identifying people in photographs have been developed by researchers. Intensity maps like Visual Saliency show where people tend to look based on where they see the most contrast. Image content is used as a metric for measuring visual attention. When we have many data points that share similarities, we can use clustering to group them into a single category. We were able to identify people's motion patterns in the video by employing the k-means algorithm. This holds significance in relation to another widely studied topic, namely the re-identification of persons. This study is focused solely on expanding upon the HOG [2]. For this, the Ohio State University Color-Thermal Pedestrian Dataset has been used. HOG features use a convolutional neural network to analyze RGB human photos in order to detect people in them. The fundamental algorithm for human detection takes as input a picture with a fixed-size window. Images are segmented into windows of this size using convolutional techniques, and HOG features are then calculated for each segment. Calculations showed a remarkably high recall of 0.93 [16]. Developing an effective model for identifying people in surveillance footage needs an equal measure of precision and imagination. Using a visual saliency model allowed us to make educated guesses about the possible locations of people within the frame. While other models struggle on the SILICON Dataset, the deep Multi-Layer Network presented in [14] excels on the MIT Saliency Benchmark. The approach works well enough for our purpose, which is human detection in video surveillance. As compared to Normal images, the performance of the Saliency-windowed image increased by a factor of 76.866. Saliency-windowed video frames were added to the HOG + SVM classifier, which resulted in an increase in Human Detection accuracy [2]. Optical motion tracking is something we want to implement. People in motion can be identified in surveillance footage using a combination of Flow and Visual Saliency windowing. In addition, the k-means algorithm has played a significant role in identifying motion patterns in video data, which is essential for re-identifying individuals. This is particularly relevant in the context of person re-identification, where the challenge is to match individuals across different camera views.

In conclusion, this section provides an in-depth exploration of the academic landscape surrounding video surveillance, with a particular focus on motion detection,

violence recognition, and firearms identification. Researchers have made significant strides in these areas, leveraging advanced techniques, neural networks, and multimodal data fusion to enhance security and safety measures. These findings underscore the continual evolution of surveillance technology in response to emerging security challenges.

Chapter 3

Dataset

Real-world violent incidents are a problematic issue for public and personal safety. Addressing this challenge, the application of Artificial Intelligence (AI) and Machine Learning (ML) models are crucial to be developed as an effective strategy to suppress or respond to such situations. To accurately predict and classify violent incidents, these models require to-the-point and comprehensive data. The critical issue at hand is the timely recognition and response to such instances of violence. To face the challenge, we find the urgency to develop AI models with the right datasets. To allow these models to differentiate between violent and non-violent situations, it is crucial to have access to a dataset that encompasses a diverse range of real-world violence scenarios. Such a dataset should be equipped with attributes that capture the key characteristics of each incident. The effectiveness of AI and Machine Learning models in this context is profoundly influenced by the diversity, quality and quantity of the training data they receive. A dataset that contains all required attributes for differentiating different types of violence from non-violence is vital to optimizing model training and testing. Currently, available online datasets of this specific nature remain very limited. This lack of such a complete dataset made our team undertake the step of creating a compilation of a dataset tailored to address this critical need. For this purpose, we have collected relevant videos from many online sources to create a unique dataset that we have named the ‘BBW Dataset’. These videos contain a broad spectrum of scenarios, capturing a variety of environments, lighting conditions, and situations. In addition to this, we have chosen to utilize the “Real Life Violence Situations Dataset” sourced from Kaggle [23]. This existing dataset was selected because it features a significant amount of footage from the Indian region, demonstrating resemblances with the situation in Bangladesh. This mixture helped us in creating a bigger and better dataset to train our model. This enlargement step, supported by detailed annotation, enables a rich repository capable of refining our model’s understanding of diverse events. Real-world violent incidents are multidimensional, encompassing an overload of complex factors. These factors include the number of parties involved, contextual information, location, the form of violence, and the outcomes of these incidents. While these variables contribute to a more subtle understanding of violent situations, they also introduce variability that challenges the precision of predictive models. Therefore, it is imperative for the dataset to contain both routine and exceptional instances of violence, enabling AI and Machine Learning models to navigate this complexity with greater accuracy. The dataset utilized in our research serves as the cornerstone of

our violence detection framework. Its comprehensiveness, coupled with its relevance to real-world scenarios, equips our models with the necessary foundation to distinguish violence from non-violence, thereby enhancing the accuracy and effectiveness of our model.

3.1 Dataset Description

Our violence detection framework leverages two distinct datasets to enhance the precision and comprehensiveness of its training and testing phases:

1. **Real Life Violence Situations Dataset(RLVS) (Kaggle):** This dataset forms the foundation of our research efforts, containing a total of 1000 videos depicting real-life violent incidents and an equal number of 1000 videos showcasing nonviolent scenarios [23]. This initial dataset provides a diverse range of video footage, enabling our model to discern between violent and nonviolent behaviours effectively. From the figure 3.1 we can see the visual representation.

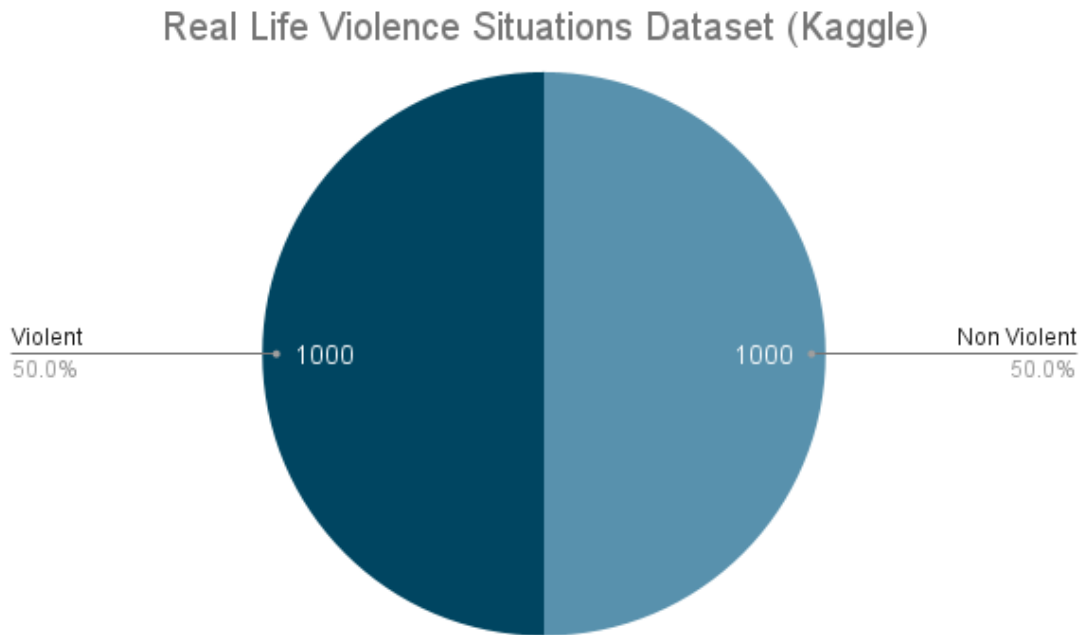


Figure 3.1: RLVS Dataset

2. **Beating-Burning-Weapon (BBW) Violence Dataset:** We have meticulously compiled a unique dataset to fulfil the demand for a more specialized and nuanced approach to violence detection, which we have named the ‘Beating-Burning-Weapon (BBW) Violence Dataset.’ This dataset encompasses 1,150 violent videos, meticulously categorized and labelled as follows: 625 videos as ‘Beating,’ 252 videos as ‘Burning,’ and 273 videos as ‘Use of Weapon.’ These videos were judiciously selected based on specific criteria, including pixel quality, frame-per-second (fps) rates, and video duration, sourced from various online platforms. Furthermore, we have incorporated 850 videos from the ‘RLVS Dataset’ [23] available on Kaggle to enrich our research. In

a nutshell, this unique dataset comprises a total of 2000 videos, distributed among three distinct types of violent acts ('Beating,' 'Burning,' and 'Use of Weapon'), with 850 nonviolent videos included for comprehensive analysis. Figure 3.2 illustrates the categorization.

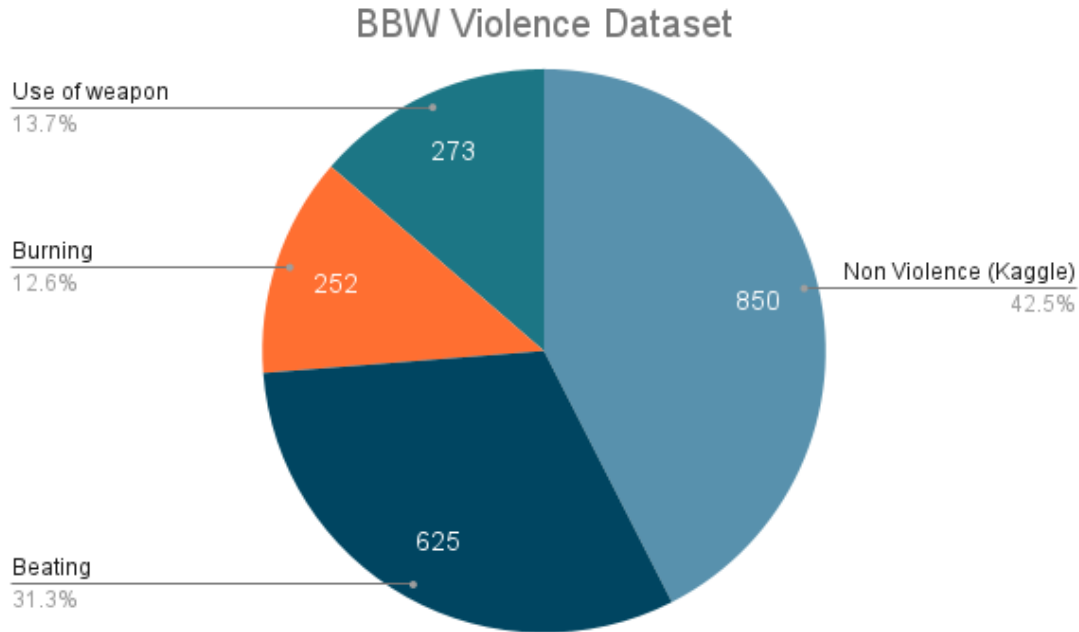


Figure 3.2: BBW Violence Dataset

The dataset's feature extraction phase encompasses a rich array of attributes drawn from diverse sources, offering a comprehensive perspective on the underlying video content. This rigorous feature extraction process is applied to two datasets comprising 1,000 violent and 1,000 non-violent videos, with the violent video subset encompassing a wide spectrum of real-life violent scenarios in various settings and conditions.

Our feature extraction methodology combines both static and dynamic approaches to comprehensively analyze and classify the behaviour and characteristics of violent incidents within the videos. These extracted features are subsequently subjected to in-depth scrutiny using a feature analyzer, enabling the derivation of a multitude of insightful metrics.

During the analysis phase, numerous key metrics of interest are computed to shed light on the nature and patterns of real-world violence situations. These metrics encompass various facets, including the severity, intensity, frequency, and contextual factors associated with each incident. By quantifying these attributes, we gain a nuanced understanding of the nuances within violent behaviours and their contextual implications.

The culminating stage involves the application of a discriminant function or classifier, which leverages the extracted features and their associated metrics to determine whether a given input should be classified as a violent or non-violent situation. This

decision-making process, informed by the rich feature set, underpins the foundation of our violence detection system, ensuring its ability to effectively discern and classify diverse forms of violent behaviours within surveillance video sequences.

Categories	Average Duration (sec)	Minimum Duration (sec)	Maximum Duration (sec)
NonViolence	5.41	2.90	375.75
Violence	5.1	1	179.91

Table 3.1: RLVS Dataset (Kaggle)

Categories	Average Duration (sec)	Minimum Duration (sec)	Maximum Duration (sec)
NonViolence(Kaggle)	5.14	2.77	179.91
Beating	4.98	2.90	11.32
Burning	4.87	3.01	7.00
Use of Weapon	6.63	2.97	12.93

Table 3.2: BBW Violence Dataset

3.2 Dataset Collection

The process of gathering and curating our datasets involved a systematic approach to ensure relevance and diversity:

1. **Real Life Violence Situations Dataset (Kaggle):** We obtained the first dataset, titled the “Real Life Violence Situations Dataset”, from Kaggle [23]. This dataset was selected for its extensive collection of 1,000 violent and 1,000 nonviolent videos, offering a broad spectrum of real-world scenarios.
2. **BBW Violence Dataset:** The creation of our second dataset was a meticulous endeavour. We initiated this process by extracting 850 nonviolent videos from the Kaggle dataset. To diversify our collection, we categorized 625 videos as ‘Beating’ and 34 videos as ‘Use of Weapon’ based on stringent criteria such as pixel quality, fps rates, and video duration. Subsequently, we meticulously sourced an additional 239 videos depicting ‘Use of Weapon’ scenarios and 252 videos portraying ‘Burning’ violence from various online platforms. This comprehensive approach ensured that our custom dataset encompassed a total of 2000 videos, distributed across the three specific types of violent behaviours, ‘Beating,’ ‘Burning,’ and ‘Use of Weapon’ alongside an efficient number of nonviolent videos.

These two datasets, each with their unique characteristics and composition, form the cornerstone of our violence detection research, empowering our model to discern a wide spectrum of violent and nonviolent behaviours within surveillance video footage.

3.3 Data Pre-Processing

As the RLVS [23] from Kaggle has two different folders containing violent videos and non-violent videos respectively, we have taken one variable for directories and one list for all the video file names. Then we created two different video paths for violent and non-violent folders. We have iterated the whole dataset and converted all the videos into a single format (.mp4). The procedure of extracting individual frames or images from a given dataset, which may comprise many types of multimedia data including images and videos, is typically denoted as frame extraction. In contrast, to the process of extracting frames from a video database, the task of frame extraction from a dataset entails the acquisition of frames from diverse sources or formats, as opposed to a coherent video sequence. The frames that were extracted possess the potential to serve as input for the purpose of training machine learning models or for executing diverse computer vision tasks. Furthermore, it is possible to assign labels or annotations to the frames for purposes such as object recognition, scene comprehension, or activity Identification.

In the first method of preprocessing, we have taken 51 frames for each video and processed it through the dense optical flow method [30]. Then we proceed to the optical flow [30] implementation stage. The conversion of each frame to grayscale is a common practice in order to streamline the computation of optical flow [30]. Grayscale images exclusively encompass luminance data, which may prove adequate for the purpose of motion detection activities. Moreover, spatial smoothing approaches, such as the application of Gaussian blurring, have the potential to effectively reduce noise and improve the resilience of optical flow [30] computing. The implementation of smoothing techniques may effectively reduce the presence of minor fluctuations and enhance the precision of flow estimation.

The computation of dense optical flow [30] is performed on a per-pixel basis across consecutive frames. The primary concept revolves around the estimation of the movement of individual pixels between consecutive frames. The typical procedure frequently involves the resolution of the optical flow equation [30], which establishes a connection between the spatial and temporal gradients of image intensity. Furthermore, flow field visualizations can be created by overlaying the computed dense optical flow vectors [30] onto the original frames, which can assist in visual analysis and solving issues. These visualizations facilitate the comprehension of motion patterns inside the video. Lastly, the dense optical flow vectors [30] may undergo further post-processing procedures, such as motion filtering or trajectory analysis, depending on the specific application. This technique has the potential to extract significant motion data or effectively eliminate extraneous noise. Afterwards, we get optical flow based 50 different frames at an equal distance as shown in 3.4.

Then, we have specified the height and width to which each video frame will be resized in our dataset. For first method, we have selected a 64×64 resolution for each video. Moreover, we specified the number of frames of a video that will be fed to the model as one sequence. which is 50. Now, in the video file we count the total number of frames in that video. Then, we get the interval after which frames will be added to the list by dividing the total number of frames in that video by the total

sequence length.

$$SkipFramesWindow = \max(TotalVideoFramesCount/Sequencelength, 1) \quad (3.1)$$

For example, if we get 80 frames in a video then we will take every 5th frame from the video in order to process our dataset. To resize the data, we will convert the interval frame into our selected resolution for this process. With the continuation of the previous example, after taking that 5th frame we will convert it into 64×64 resolution. Then normalize the resized frame and get the value between 0 and 1. Afterwards, we get the 50 pre-processed frames for each video which we kept in a list. These lists are sent into the hybrid model which categorizes the video (violent or non-violent) and violent subcategory (Violent: Beating, Burning or Use of Weapon) by identifying the frame sequences.

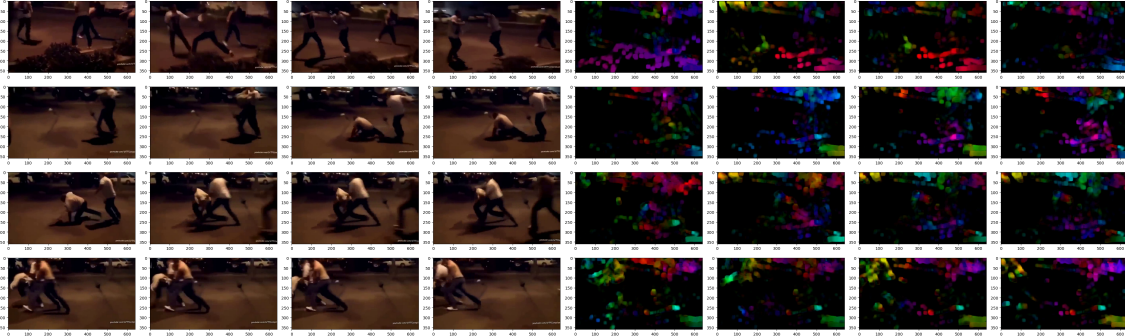


Figure 3.3: Original Frames

Figure 3.4: Processed Frames
(first method)

Then, in the second method, we are adding the initial 50 frames for each video, stacking one onto another. As a result, we get only 1 merged frame but the RGB value becomes more than 255. So, we divide each pixel of that 1 frame by 50 so that each pixel value becomes less than 255. Afterwards, similarly like the first phase, we resize, normalize and get the 1 final frame which will be feeded to our hybrid model. Finally, we will get 1 merged frame as shown in 3.5 for each video which the model will categorize if the video is violent or non-violent and violent subcategory (Beating, Burning or Use of Weapon) by identifying the frame sequences.

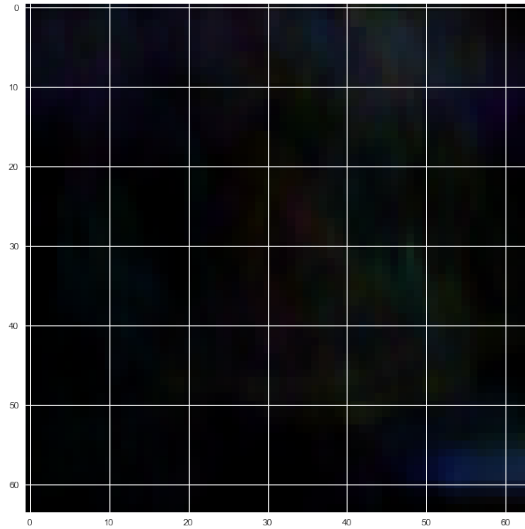


Figure 3.5: Merged Frame (second method)

Subsequently, we used one-hot encoding, which is a technique used to represent categorical data in a binary format which involves transforming each category into a binary vector, where all elements are assigned a value of zero except for the index that corresponds to the category, which is assigned a value of one. We categorized the original dataset 0, 1 for non-violence and violence respectively in the first method and 0, 1, 2, and 3 as per non-violence, beating, burning and use of weapons respectively in the second method of preprocessing.

Chapter 4

Methodology

This thesis presents a novel strategy for automating violence detection. The suggested approach utilizes machine learning and deep learning techniques, including object and motion detection, as well as pose estimation through optical flow [30] analysis. The architecture employed is a hybrid MobileNet-Bi-LSTM model. Our approach surpasses conventional techniques by capturing temporal dynamics and spatial features. The selection of models was based on their key features. The optical flow [30] analysis was selected for its efficient and accurate estimation of pixel motion, which was used in the data pre-processing. Bi-directional Long Short-Term Memory (BiLSTM) [26], [27] addresses the need to capture sequential dependencies and temporal context within video frames and the MobileNetV2 [22] was used for its accuracy in image classification. All these models have been applied in our analysis based on their theoretical reliability and practical effectiveness in tasks similar to the identification of criminal activities in video recordings.

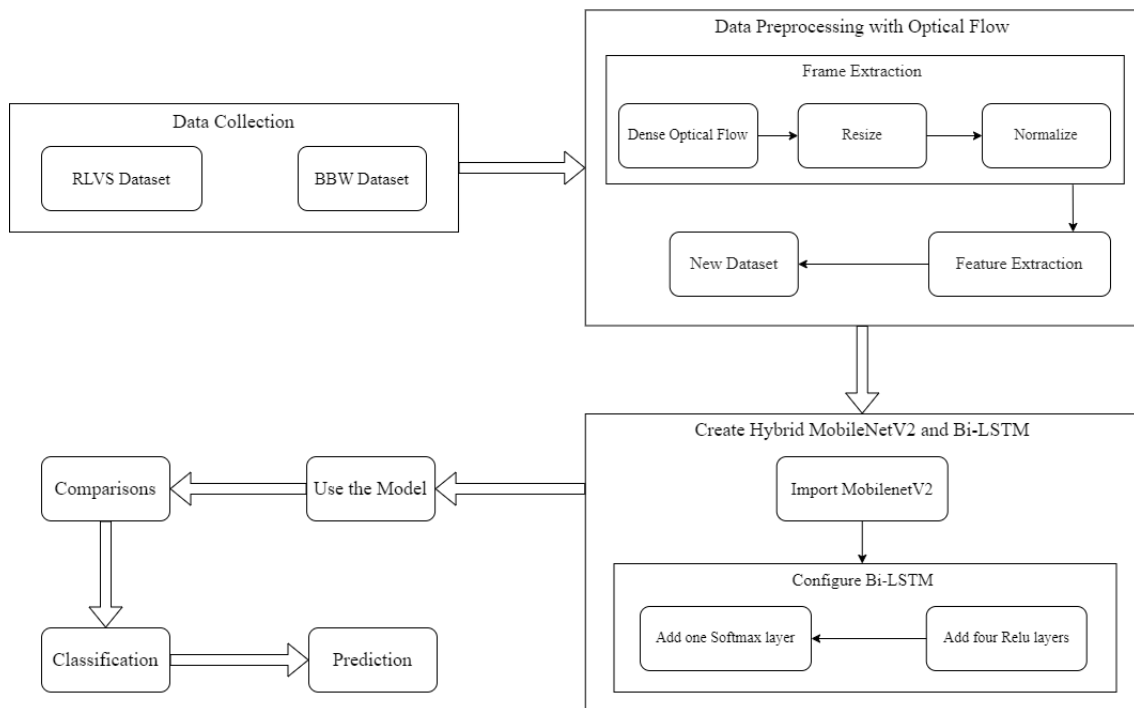


Figure 4.1: Workflow

4.1 Model Specifications

Our violence detection framework integrates Optical Flow [30], MobileNetV2 [22], and BiLSTM [26], [27] models, harnessing their unique capabilities to create a comprehensive system for identifying violent behaviours in surveillance videos. Optical flow captures motion [30], MobileNet [22] extracts spatial features, and Bi-LSTM [26], [27] explores sequential dependencies, collectively enhancing precision and accuracy. This balance between computational efficiency and performance ensures the practical and reliable deployment of the system in real-world scenarios.

4.1.1 Dense Optical Flow

Dense optical flow [30] is a powerful computer vision technique implemented within our violence detection framework to comprehensively analyze motion patterns in video sequences. This method goes beyond the basic optical flow [30] by calculating motion vectors for every pixel in each consecutive frame, creating a dense grid of motion vectors across the entire image. The underlying principle behind dense optical flow is the preservation of fine-grained motion details, enabling a granular understanding of how each pixel moves over time. This is particularly valuable in violence detection, as it allows us to capture subtle motion variations that may indicate violent behaviours, such as aggressive gestures or rapid movements. Dense optical flow [30] operates by tracking the displacement of image pixels from one frame to the next, considering variations in pixel intensity. It relies on mathematical formulations, with one of the common equations being the Lucas-Kanade method, which minimizes an energy function to estimate flow vectors. The choice to integrate dense optical flow [30] in our framework is driven by the need for precise motion analysis, which is essential to identifying violent acts accurately. By using dense optical flow [30], our model gains the ability to discern fine-grained motion patterns, enhancing its overall effectiveness in violence detection within surveillance video footage.

Optical flow [30] estimation can be represented mathematically by the Lucas-Kanade method. The Lucas-Kanade method makes the assumption that the displacement of the image content between two adjacent instants (frames) is negligibly large and roughly constant in the area surrounding the point p under consideration. Therefore, it is safe to assume that the optical flow equation holds for every pixel inside a window with the origin at p . Namely, the local image flow (velocity) vector (V_x, V_y) needs to meet the following conditions:

$$\begin{aligned} I_x(q_1)V_x + I_y(q_1)V_y &= -I_t(q_1) \\ I_x(q_2)V_x + I_y(q_2)V_y &= -I_t(q_2) \\ &\dots \\ &\dots \\ &\dots \\ I_x(q_n)V_x + I_y(q_n)V_y &= -I_t(q_n) \end{aligned} \tag{4.1}$$

Here, we consider a window containing pixels q_1, q_2, \dots, q_n . The image I is analyzed by evaluating the partial derivatives $I_x(q_i), I_y(q_i), I_t(q_i)$ with respect to position x, y and time t , at the point q_i and at the current time.

$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \vdots \\ I_x(q_n) & I_y(q_n) \end{bmatrix} \quad v = \begin{bmatrix} V_x \\ V_x \end{bmatrix} \quad b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(q_n) \end{bmatrix} \quad (4.2)$$

Utilizing deep learning methods, dense optical flow estimation can deliver precise and trustworthy motion data between frames, assisting numerous computer vision applications.

4.1.2 Bi-directional Long Short-Term Memory (BiLSTM)

BiLSTM [26], [27] is a type of recurrent neural network architecture that is capable of processing sequential data in both forward and backward directions. The BiLSTM model [26], [27], which is an advancement of the conventional LSTM, exhibits the capability to efficiently capture contextual information from both past and future sequences. In contrast to unidirectional LSTM, BiLSTM [26], [27] is capable of processing data in both forward and backward directions, which enhances its ability to detect patterns over a longer time frame. For this reason, this architecture plays a pivotal role in our framework’s temporal sequence learning. Violence detection requires an understanding of not only spatial but also temporal context. Bi-LSTM [26], [27] excels in capturing sequential dependencies within video frames, enabling the model to discern patterns of behaviour over time.

As we explore the utilization of LSTM as a chain model for time sequence processing. The distinctive feature of LSTM lies in its implementation of memory cells to substitute hidden layer nodes, thereby effectively addressing the issues of gradient vanishing and gradient explosion. The model acquires temporal information of the EEG signal by inputting continuous time sequences. Thesis: The LSTM’s weight between the hidden layer and the output layer exhibits recyclability and possesses significant memory capacity for extended information sequences. The structure of an LSTM unit consists of three gate control units, namely the forget gate, input gate, and output gate, with their respective calculation formulas defined by equations, which are:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t &= f_t \times C_{t-1} + i_t \times \tilde{C}_t \\ O_t &= \sigma(W_O \cdot [h_{t-1}, x_t] + b_O) \\ h_t &= \tanh(C_t) \times O_t \end{aligned} \quad (4.3)$$

Here, the notation used in the context of time sequences, including the representation of input time sequences as x_t , the sigmoid function represented by σ , the weight matrix denoted by W , and the bias vectors associated with the weights represented

by b terms. The continued use of the feature is determined by the forget gate f_t . The information of the previous state and the current state simultaneously inputs into the σ function. The responsibility of the input gate is to update the state of the LSTM unit. The cell state, denoted as C_t is a fundamental component in cell state. The hidden output of the backward layer is h_t . In the context of LSTM units, the output gate O_t is responsible for regulating the output values that are passed on to the subsequent LSTM unit.

The Bi-LSTM [26], [27] network adds a backward layer to learn the future emotion information, which is an extension of the past, in comparison to the aforementioned unidirectional LSTM network. The core computation in a Bi-LSTM [26], [27] unit can be expressed through the following equations:

$$\begin{aligned}
 h_t^f &= \tanh\left(W_{xh}^f x_t + W_{hh}^f h_{t-1}^f + b_h^f\right) \\
 h_t^b &= \tanh\left(W_{xh}^b x_t + W_{hh}^b h_{t+1}^b + b_h^b\right) \\
 y_t &= W_{hy}^f h_t^f + W_{hy}^b h_t^b + b_y
 \end{aligned} \tag{4.4}$$

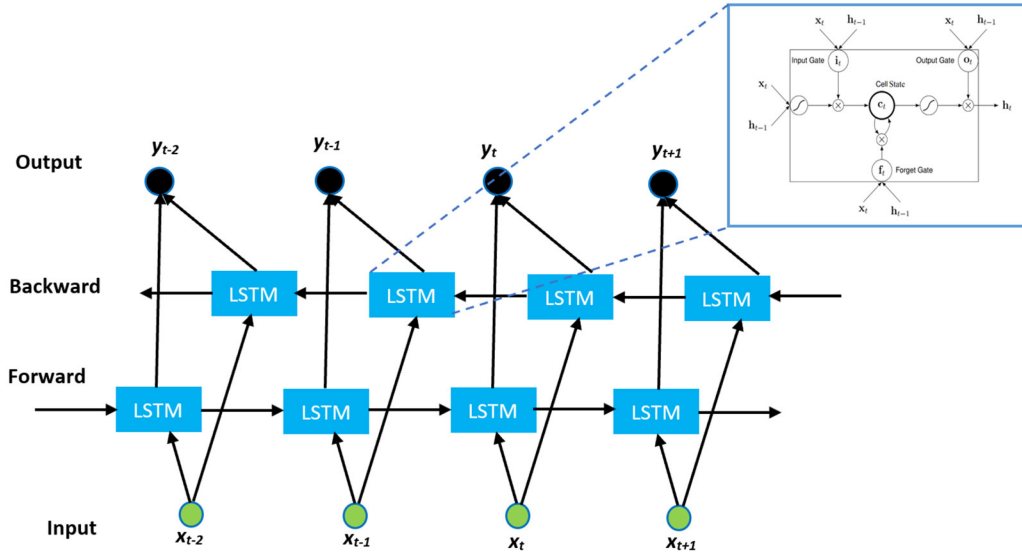


Figure 4.2: Bi-LSTM

There are two distinct hidden layers referred to as the ‘forward’ hidden layer and the ‘backward’ hidden layer. The ‘forward’ hidden layer h_t^f calculates the input in ascending order, $t = 1, 2, 3, \dots, T$. Whereas, the ‘backward’ hidden layer h_t^b considers the input in descending order, $t = T, \dots, 3, 2, 1$. Finally, they both are combined to generate output y_t .

The BiLSTM [26], [27] model effectively tackles the issue of long-term dependencies in a sequence, including the challenge of gradient vanishing, by utilizing the gating mechanisms of LSTMs. This particular capability holds great significance within our current context, as it is imperative to uphold continuity and extract relevant data from lengthy video sequences.

4.1.3 MobileNetV2

MobileNetV2 [22] is incorporated into our framework as an integral component of our feature extraction process. MobileNetV2 [22] is widely recognized for its ability to maintain a high level of accuracy while exhibiting low latency and minimal computational expense. Its efficiency and effectiveness in image classification tasks make it an ideal choice. By utilizing MobileNet [22], we can efficiently convert each video frame into feature vectors that capture spatial information. The desirable nature of this attribute is particularly evident in real-time criminal activity detection scenarios, where immediate action is often critical and computational capacity may be limited.

The MobileNetV2[22] architecture is founded on an inverted residual structure that incorporates linear bottlenecks. It achieves this efficiency through depthwise separable convolution, which reduces computational complexity while preserving the quality of extracted spatial information. This structure facilitates the extraction of abstract and generalized characteristics from video frames, which is pivotal in the identification of a wide spectrum of criminal behaviours. As MobileNetV2 [22] employs depthwise separable convolution, a process that combines depthwise convolution and pointwise convolution to generate feature representations from input tensors, which can be mathematically represented as:

$$Y = DWConv(X, K) = PointwiseConv(DepthwiseConv(X, K), K) \quad (4.5)$$

Here, X represents the input tensor, K is the convolutional kernel, and Y denotes the output tensor.

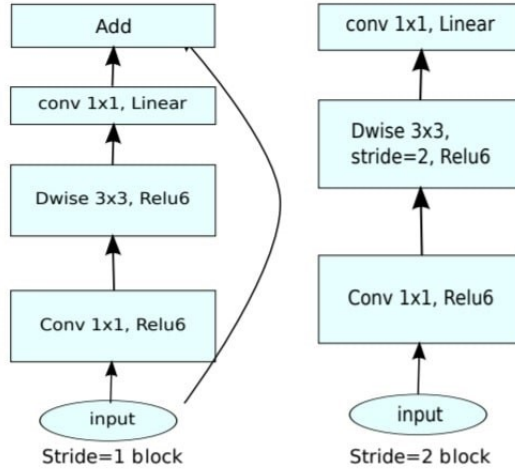


Figure 4.3: MobileNetV2

To put it in a nutshell, by integrating these models into our violence detection framework, we leverage their individual strengths to construct a holistic system that can effectively determine violent behaviors in surveillance video footage. The optical flow [30] model captures motion information, the MobileNet [22] model extracts spatial features, and the BiLSTM [26], [27] model delves into sequential dependencies, collectively enhancing the system's precision and accuracy. The equilibrium

between computational efficiency and performance accuracy guarantees the effective deployment of the system in real-world situations, making this approach both practically feasible and theoretically reliable.

4.1.4 Proposed OptiMobBi-LSTM Model

In this section, we introduce a novel approach that combines the strengths of multiple violence detection models. This Proposed OptiMobBi-LSTM Model aims to leverage the complementary characteristics of Optical Flow [30], MobileNetV2 [22], and BiLSTM [26], [27], offering a holistic solution to enhance violence detection accuracy and adaptability. By intelligently integrating these components, we envision a more robust and versatile model capable of addressing the intricacies of real-world surveillance scenarios.

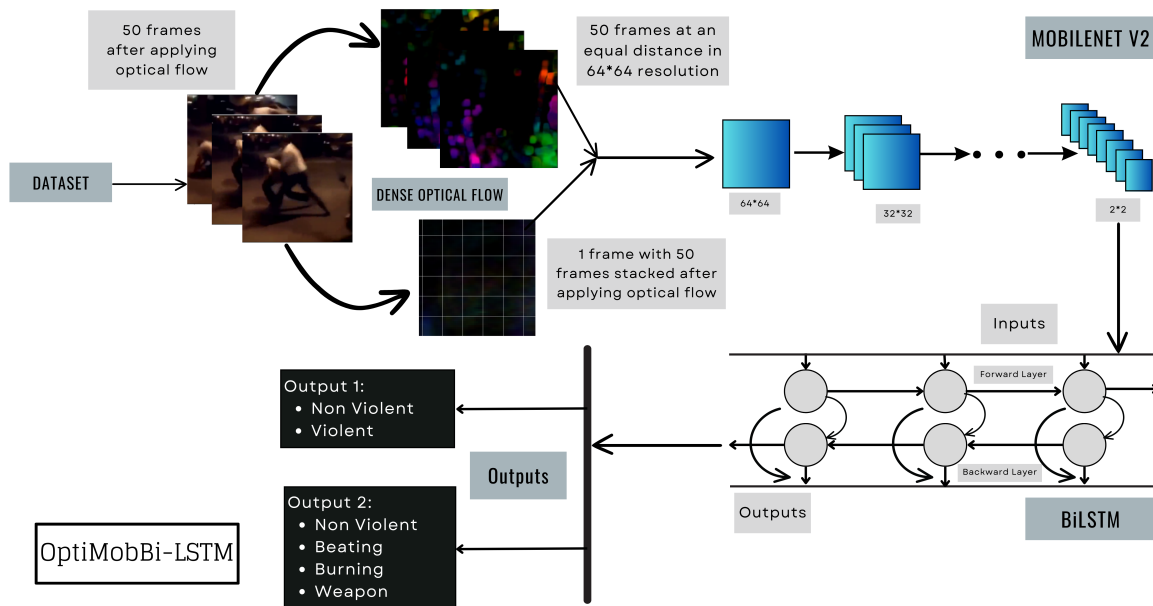


Figure 4.4: Proposed OptiMobBi-LSTM Model

Chapter 5

Result and Discussion

In this section, we comprehensively analyze the performance of our violence detection models across various datasets and methodologies. We delve into the implications of our findings, highlighting the strengths and limitations of each approach.

5.1 Performance Evaluation Measures

The evaluation of results has been conducted using established performance measures, including accuracy, precision, recall and F-1 Score. The computation of these metrics entails the consideration of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values. The variables TN and TP represent the quantities of accurately categorized negative and positive samples, respectively. The variables FN and FP represent the quantities of positive and negative samples that have been misclassified, respectively.

TN represents a negative case which also predicted negative. On the other hand, TP stands for a positive case which is also predicted positive. Similarly, FN symbolizes a positive case but predicted negative. Conversely, FP depicts a negative case but predicted positive.

Here, accuracy is a metric that quantifies the overall efficacy of a classification strategy. The calculation can be determined using the following formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (5.1)$$

Precision refers to the classifier's capacity to accurately identify instances as negative when they are indeed negative. The class-specific metric is determined by calculating the ratio of true positives to the sum of true positives and false positives.

$$Precision = \frac{TP}{TP + FP} \quad (5.2)$$

Recall refers to the capacity of a classifier to accurately identify and retrieve all instances that are classified as positive in classification. The class-specific metric is formally defined as the quotient of true positives divided by the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN} \quad (5.3)$$

Finally, The F1 score is a mathematical measure that combines precision and recall in a weighted harmonic mean. It ranges from 0.0 to 1.0, with 1.0 representing the highest possible score and 0.0 indicating the lowest possible score. In general, F1 scores tend to be lower than accuracy measures due to their use of precision and recall in their calculation. It is often recommended to utilize the weighted average of F1 scores rather than global accuracy when comparing classifier models.

$$F1 - score = \frac{2 \times (Recall * Precision)}{Recall + Precision} \quad (5.4)$$

5.2 Experimental Results

The purpose of this thesis was to develop an effective violence detection system. To achieve this, a comprehensive implementation of models was conducted, followed by a thorough analysis of the received data. The performance of two distinct methods applied to two distinct datasets was evaluated.

5.2.1 OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on Real Life Violence Situations Dataset

For Frame Selection at Equal Intervals Method in the Real Life Violence Situations Dataset from Kaggle, the results demonstrated promising performance.. The results for this model showcased impressive accuracy, standing at 90.16%. The precision, recall, and f1-score for violence and nonviolence classes were consistently high, underlining the model’s robustness in distinguishing between violent and nonviolent scenarios.

	Precision	Recall	f1-score
Nonviolence	0.90	0.91	0.90
Violence	0.91	0.89	0.90
Accuracy			0.90

Table 5.1: OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on Real Life Violence Situations Dataset

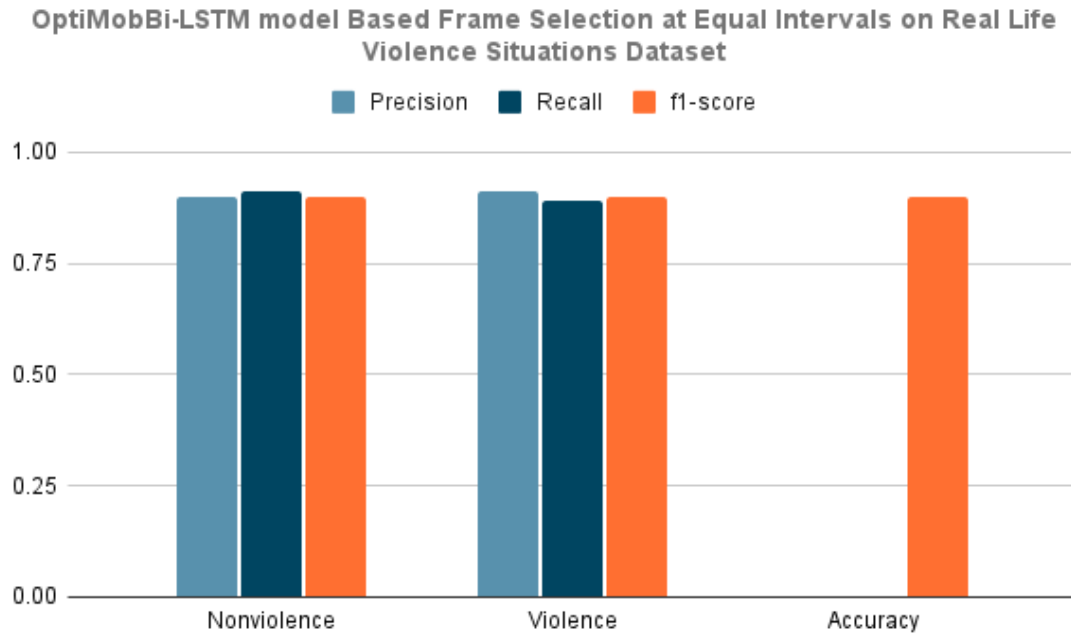


Figure 5.1: OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on Real Life Violence Situations Dataset

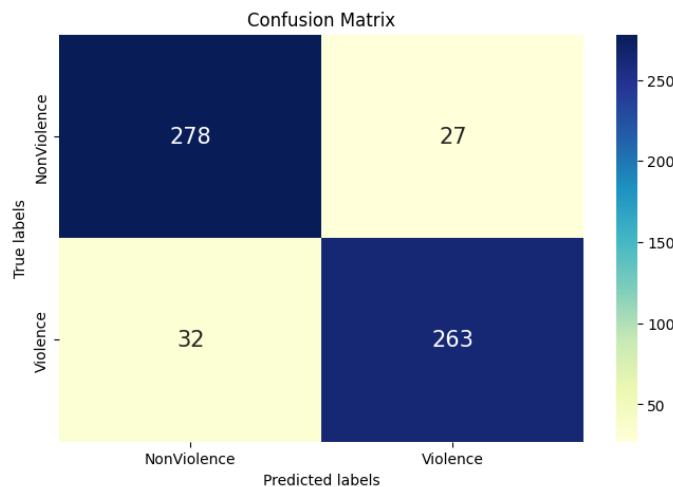


Figure 5.2: OptiMobBi-LSTM Model Based Confusion Matrix of Frame Selection at Equal Intervals on Real Life Violence Situations Dataset

5.2.2 OptiMobBi-LSTM Model Based Merged Frame Stacking on Real Life Violence Situations Dataset

As we applied our evaluation further to the Merged Frame Stacking method on the Real Life Violence Situations Dataset, where we applied the merged frame stacking technique. In this model, the accuracy reached 85%, indicating its effectiveness in discerning violence from nonviolence in the Kaggle dataset.

	Precision	Recall	f1-score
Nonviolence	0.83	0.87	0.85
Violence	0.87	0.83	0.85
Accuracy			0.85

Table 5.2: OptiMobBi-LSTM Model Based Merged Frame Stacking on Real Life Violence Situations Dataset

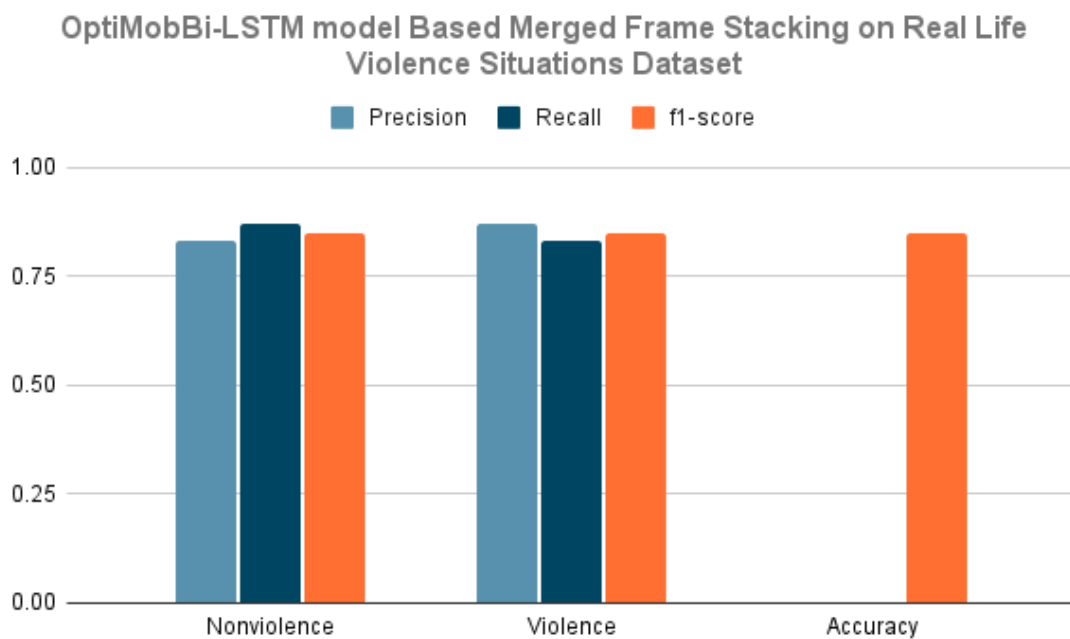


Figure 5.3: OptiMobBi-LSTM Model Based Merged Frame Stacking on Real Life Violence Situations Dataset

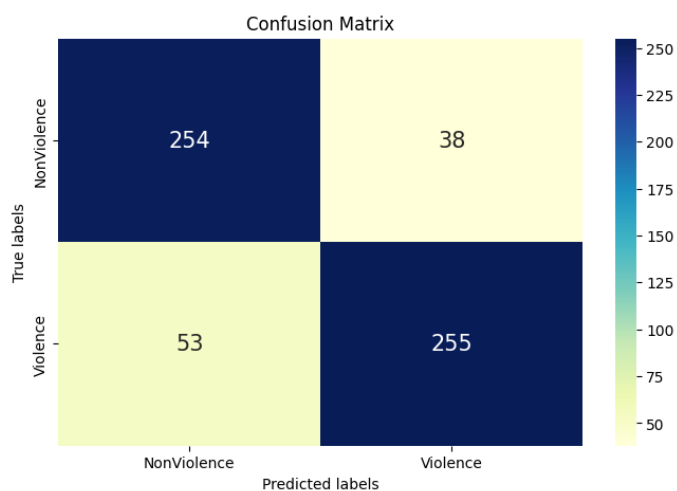


Figure 5.4: Confusion Matrix of OptiMobBi-LSTM Model Based Frame Stacking on Real Life Violence Situations Dataset

5.2.3 OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset

Our analysis extended to our meticulously created BBW Violence Dataset. For Frame Selection at Equal Intervals Method in our BBW Violence Dataset, the results demonstrated promising performance. This approach involved the selection of 50 different frames at equal intervals from each video, aiming to capture a diverse set of frames representing the video content effectively. The precision, recall, and f1-score for each violence class ('Beating,' 'Burning,' and 'Use of Weapon') along with nonviolence were notably high. Moreover, the accuracy achieved in this model was 85.32%, showcasing its effectiveness in distinguishing between different types of violent and nonviolent behaviours within the custom-compiled dataset. The tabular and visual representation of the result is shown in Table 5.3, Figure 5.5 and the confusion matrix 5.6.

	Precision	Recall	f1-score
Nonviolence	0.91	0.87	0.89
Beating	0.81	0.90	0.85
Burning	0.81	0.85	0.83
Weapon	0.81	0.72	0.76
Accuracy			0.85

Table 5.3: OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset

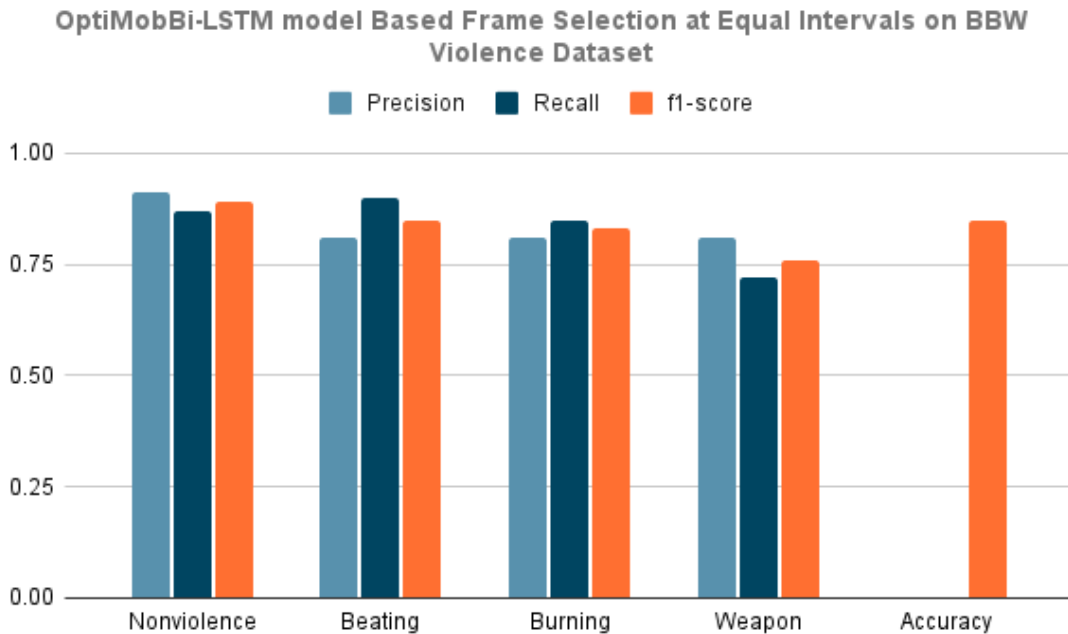


Figure 5.5: OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset

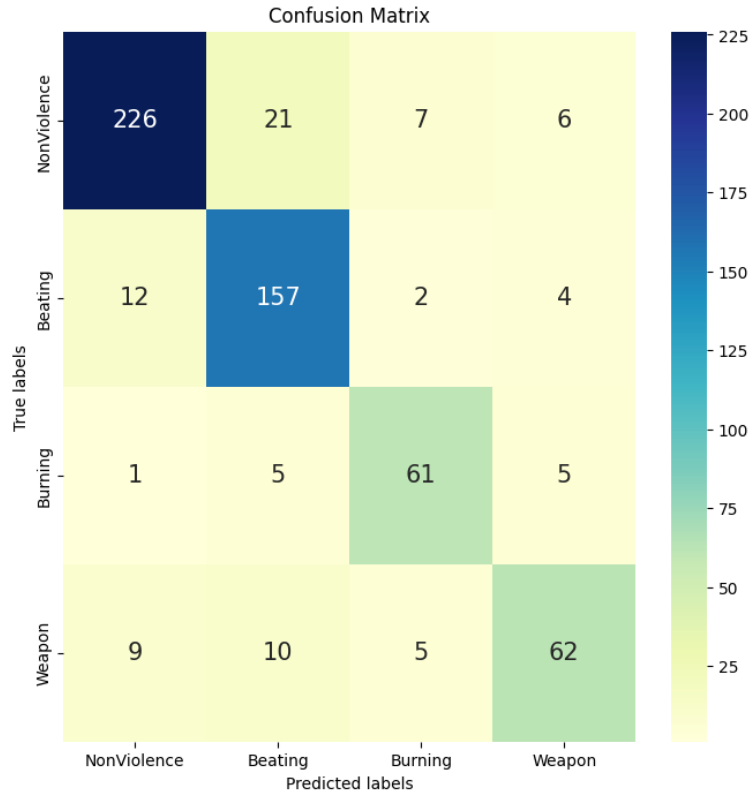


Figure 5.6: Confusion Matrix of OptiMobBi-LSTM Model Based Frame Selection at Equal Intervals on BBW Violence Dataset

5.2.4 OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset

Our second method, Merged Frame Stacking applied on our BBW Violence Dataset involved the technique of stacking the initial 50 frames from each video into a single merged frame. This approach aimed to condense video information into a single representation. The results for this model displayed a decent performance, with an overall accuracy of 74%. While precision, recall, and f1-score varied across violence classes, this method provided a reasonable baseline for violence detection within the custom-compiled dataset. The tabular and visual representation of this result is shown in Table 5.4, Figure 5.7 and the confusion matrix 5.8.

	Precision	Recall	f1-score
Nonviolence	0.80	0.77	0.79
Beating	0.79	0.72	0.75
Burning	0.54	0.72	0.62
Weapon	0.68	0.65	0.67
Accuracy			0.74

Table 5.4: OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset

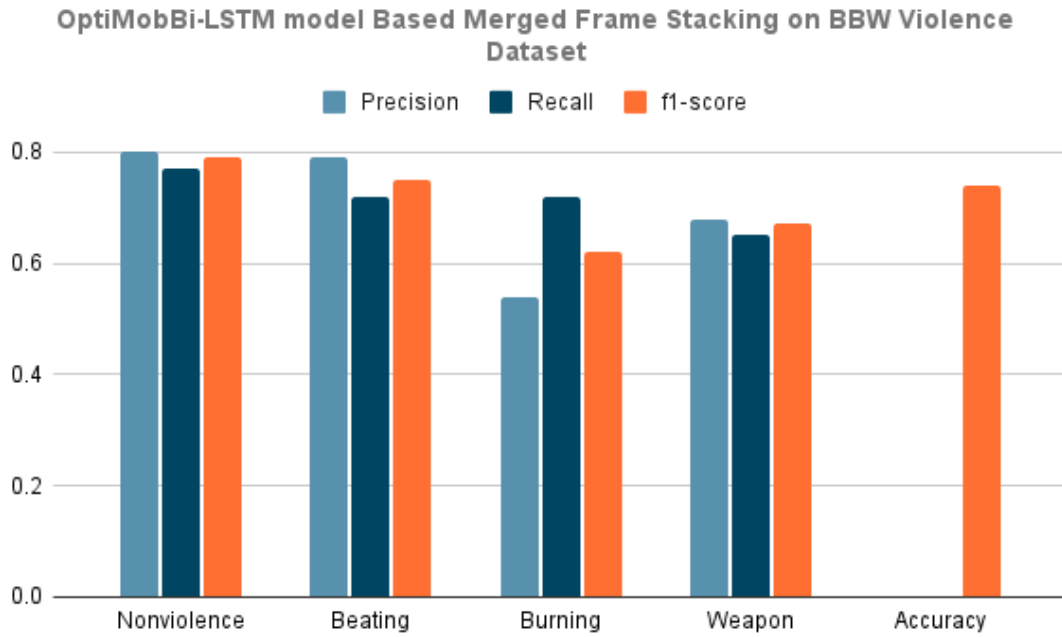


Figure 5.7: OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset

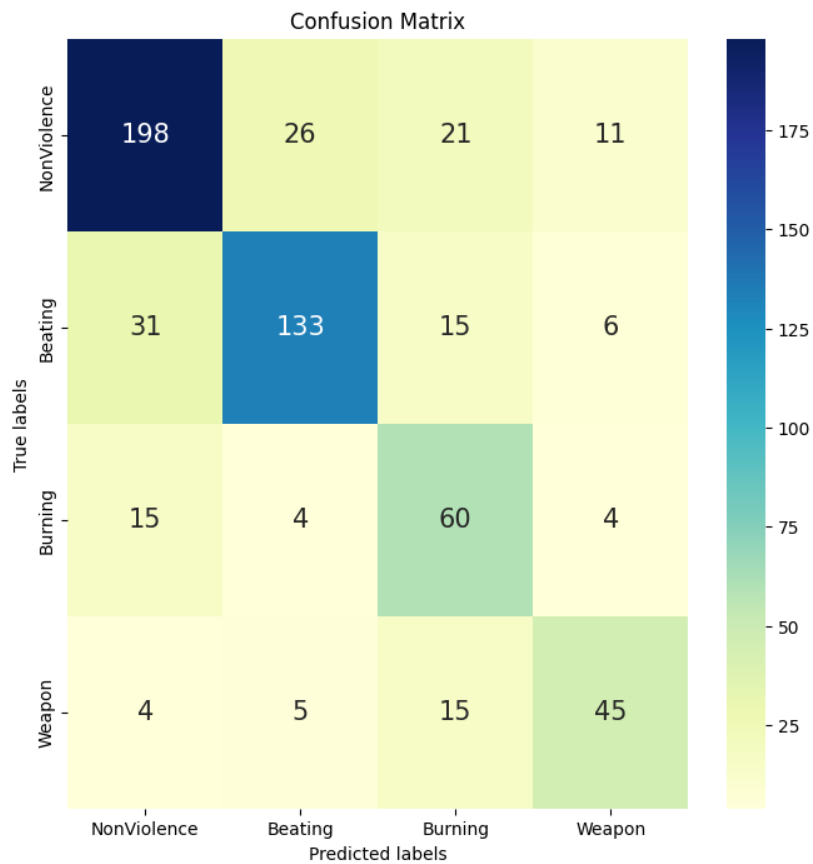


Figure 5.8: Confusion Matrix of OptiMobBi-LSTM Model Based Merged Frame Stacking on BBW Violence Dataset

Both methods employed in this preliminary analysis share the goal of effectively detecting violence within surveillance video footage. Method 1 focused on capturing 50 different frames at equal intervals from each video, providing a diverse set of frames for comprehensive representation. On the other hand, Method 2 condensed video content by stacking the initial 50 frames into a single merged frame. These methodologies serve as the initial building blocks for our violence detection system, offering alternative approaches to address diverse scenarios and computational requirements while laying the foundation for further refinement and optimization.

5.3 Discussion

The comprehensive evaluation of our violence detection methods, encompassing two distinctive datasets and two different techniques, yields valuable insights into the performance and versatility of our approach.

Frame Selection at Equal Intervals method, which involved selecting 50 frames at equal intervals, exhibited remarkable accuracy, with a noteworthy 90.16% in the RLVS Dataset. In this method, the model demonstrated the ability to distinguish between nonviolence and violence with exceptional precision and recall. In contrast, the Merged Frame Stacking method, which condensed video content into a single merged frame, displayed a lower but still respectable accuracy of 85% in the Real Life Violence Situations Dataset, indicating its robustness in identifying violent scenarios. This method offers computational efficiency advantages and serves as a practical alternative.

Within the BBW Violence Dataset, the Frame Selection at Equal Intervals Method exhibited strong performance with an accuracy of 85.32%. This approach excelled in distinguishing between different types of violent behaviours, including ‘Beating,’ ‘Burning,’ and ‘Use of Weapons.’ On the other hand, when we applied the Merged Frame Stacking on the BBW Violence Dataset, delivered a decent accuracy of 74%, providing a foundational baseline for violence detection within this dataset.

	Frame Selection at Equal Intervals Method	Merged Frame Stacking Method
Real Life Violence Situations Dataset	90.16%	85%
BBW Violence Dataset	85.32%	74%

Table 5.5: Overall Comparison

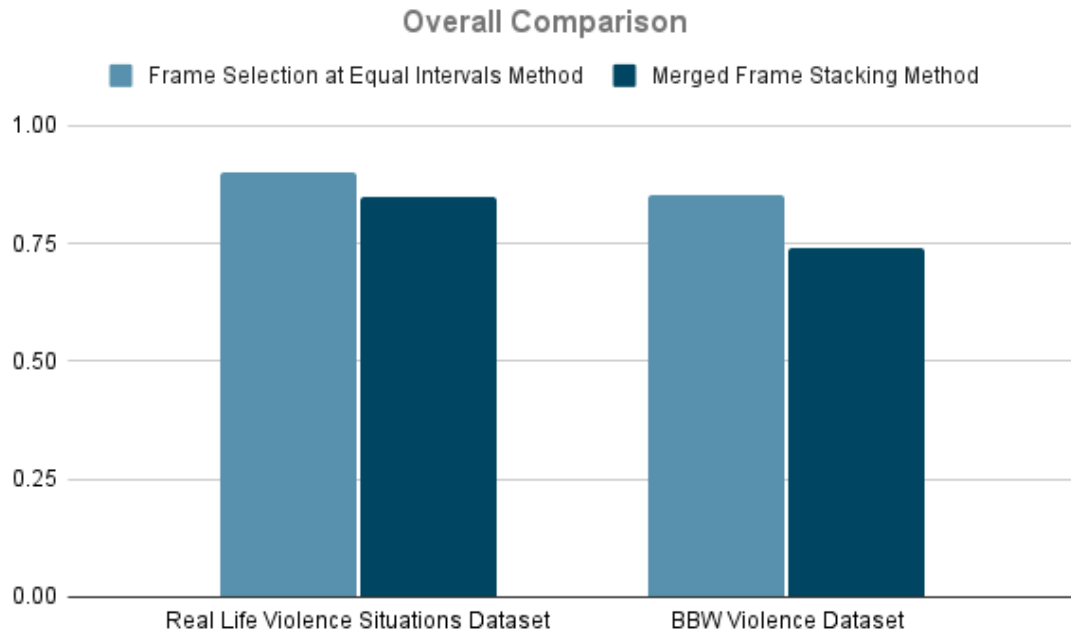


Figure 5.9: Overall Comparison

From the overall findings, we can say that the Frame Selection at Equal Intervals method consistently demonstrated higher accuracy in both datasets, making it a preferred choice for scenarios where precision is paramount. On the other hand, the Merged Frame Stacking method, while offering slightly lower accuracy, may be favoured in situations where computational efficiency is critical. These results lay the groundwork for further enhancements, aiming to achieve even greater accuracy and versatility in real-world surveillance scenarios.

Chapter 6

Conclusion

This study has presented an innovative methodology for automatic detection of violence in surveillance footage, effectively addressing the urgent requirement for improved security measures in both public and private settings. The research presented here demonstrates the utilization of a combination of machine learning and deep learning methodologies, such as object detection, motion analysis, and optical flow-based [30] pose estimation. Additionally, a hybrid MobileNet-Bi-LSTM architecture is employed to effectively capture both temporal dynamics and spatial features, surpassing conventional approaches. A thorough framework has been developed by the careful augmentation of datasets and the rigorous annotation of numerous scenarios, lighting conditions, and acts including ‘Beating,’ ‘Use of Weapons,’ and ‘Burning.’ In this research, we tested two violence detection methods on the two previously mentioned datasets. The ‘Frame Selection at Equal Intervals’ method achieved higher accuracy, 90.16% in the Real Life Violence Situations Dataset and 85.32% in the BBW Violence Dataset, making it a precise choice. On the other hand, the ‘Merged Frame Stacking’ method, offering computational efficiency, achieved respectable accuracies of 85% and 74% in the RLVS and BBW Violence Datasets respectively. This paradigm not only differentiates between acts of violence and non-violence, but also enables a more nuanced understanding of the characteristics of violent incidents. As a result, it has the potential to bring about significant changes in proactive security measures and the timely identification of threats. The effectiveness of the research is substantiated through empirical validation using graphical representation and comparative analysis. This holds the potential to revolutionize the field of surveillance systems by facilitating the efficient detection and prevention of violence.

6.1 Future Works

The discipline of violence detection presents a multitude of possible avenues for further research. In order to enhance the accuracy and adaptability of our models, it is important to implement certain measures by investigating advanced deep learning architectures and ensemble methodologies. The utilization of transfer learning from extensive datasets and the ongoing process of fine-tuning on specific violence detection tasks offers a potential avenue towards improving the overall generalization of models. The process of converting our models into real-time implementations for surveillance systems is a significant stage, accompanied by the application of incre-

mental learning techniques to adjust to the changing patterns of violence. Moreover, the incorporation of multimodal methodologies that incorporate audio and textual data in conjunction with visual clues has the potential to provide an extensive understanding of violent events. The prioritization of ethical factors, such as privacy concerns and bias prevention, remains of utmost importance. Finally, the addition of dataset size and diversity, potentially achieved through collaborative efforts with relevant agencies, has the potential to strengthen the resilience and practicality of models in real-world scenarios. The forthcoming initiatives have the objective of progressing the domain of violence detection, with a focus on guaranteeing precision and ethical accountability in practical security implementations.

Bibliography

- [1] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: <https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf>. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.
- [3] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, “Review and evaluation of commonly-implemented background subtraction algorithms,” in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4760998.
- [4] W. Wang, J. Yang, and W. Gao, “Modeling background and segmenting moving objects from compressed video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 670–681, 2008. DOI: 10.1109/TCSVT.2008.918800.
- [5] W. Liu, P. Miller, J. Ma, and W. Yan, “Challenges of distributed intelligent surveillance system with heterogenous information,” *Procs. of QRASA, Pasadena, California*, Jan. 2009.
- [6] F. D. M. d. Souza, G. C. Chávez, E. A. d. Valle Jr., and A. d. A. Araujo, “Violence detection in video using spatio-temporal features,” in *2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, 2010, pp. 224–230. DOI: 10.1109/SIBGRAPI.2010.38.
- [7] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, “Violence detection in video using computer vision techniques,” in *Computer Analysis of Images and Patterns*, P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 332–339, ISBN: 978-3-642-23678-5.
- [8] S.-C. Huang, “An advanced motion detection algorithm with video quality analysis for video surveillance systems,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 1–14, 2011. DOI: 10.1109/TCSVT.2010.2087812.
- [9] A. Rajpurohit, A. Agarwal, M. Gaikwad, K. Garg, and V. Inamdar, “Securing public places using intelligent motion detection,” in *2012 IEEE International Conference on Engineering Education: Innovative Practices and Future Trends (AICERA)*, 2012, pp. 1–4. DOI: 10.1109/AICERA.2012.6306742.

- [10] S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013. DOI: 10.1109/TPAMI.2012.59.
- [11] M. Paul, S. M. E. Haque, and S. Chakraborty, *Human detection in surveillance videos and its applications - a review - eurasp journal on advances in signal processing*, Nov. 2013. [Online]. Available: <https://doi.org/10.1186/1687-6180-2013-176>.
- [12] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, “Vsd, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation,” *Multimedia Tools and Applications*, vol. 74, May 2014. DOI: 10.1007/s11042-014-1984-4.
- [13] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *CoRR*, vol. abs/1406.2199, 2014. arXiv: 1406.2199. [Online]. Available: <http://arxiv.org/abs/1406.2199>.
- [14] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multi-level network for saliency prediction,” in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 3488–3493. DOI: 10.1109/ICPR.2016.7900174.
- [15] L. Guezouli and H. Belhani, “Automatic detection of moving objects in video surveillance,” in *2016 Global Summit on Computer & Information Technology (GSCIT)*, 2016, pp. 70–75. DOI: 10.1109/GSCIT.2016.14.
- [16] V. Gajjar, Y. Khandhediya, and A. Gurnani, “Human detection and tracking for video surveillance: A cognitive science approach,” in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2805–2809. DOI: 10.1109/ICCVW.2017.330.
- [17] L. Guezouli, H. Boukhetache, and I. Kebi, “Human detection by surveillance camera,” *International Journal of Robotics Applications and Technologies*, vol. 6, no. 1, pp. 21–33, 2018. DOI: 10.4018/ijrat.2018010102.
- [18] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.
- [19] N. Jain, S. Yerragolla, T. Guha, and Mohana, “Performance analysis of object detection and tracking algorithms for traffic surveillance applications using neural networks,” in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019, pp. 690–696. DOI: 10.1109/I-SMAC47947.2019.9032502.
- [20] M. Ramzan, A. Abid, H. U. Khan, *et al.*, “A review on state-of-the-art violence detection techniques,” *IEEE Access*, vol. 7, pp. 107 560–107 575, 2019. DOI: 10.1109/ACCESS.2019.2932114.
- [21] D. J. Samuel R., F. E. G. Manogaran, *et al.*, “Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm,” *Computer Networks*, vol. 151, pp. 191–200, 2019, ISSN: 1389-1286. DOI: <https://doi.org/10.1016/j.comnet.2019.01.028>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1389128618308521>.

- [22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, 2019. arXiv: 1801.04381 [cs.CV].
- [23] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, “Violence recognition from videos using deep learning techniques,” in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2019, pp. 80–85. DOI: 10.1109/ICICIS46948.2019.9014714.
- [24] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, “A novel violent video detection scheme based on modified 3d convolutional neural networks,” *IEEE Access*, vol. 7, pp. 39 172–39 179, 2019. DOI: 10.1109/ACCESS.2019.2906275.
- [25] H. Jain, A. Vikram, Mohana, A. Kashyap, and A. Jain, “Weapon detection using artificial intelligence and deep learning for security applications,” in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 193–198. DOI: 10.1109/ICESC48915.2020.9155832.
- [26] Y. Verma, *Complete guide to bidirectional lstm (with python codes)*, Nov. 2021. [Online]. Available: <https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/>.
- [27] S. Aljumah and L. Berriche, “Bi-lstm-based neural source code summarization,” *Applied Sciences*, vol. 12, p. 12 587, Dec. 2022. DOI: 10.3390/app122412587.
- [28] H. Gupta and S. T. Ali, “Violence detection using deep learning techniques,” in *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, 2022, pp. 121–124. DOI: 10.1109/ICETCI55171.2022.9921388.
- [29] R. Vijeikis, V. Raudonis, and G. Dervinis, “Efficient violence detection in surveillance,” *Sensors*, vol. 22, no. 6, p. 2216, Mar. 2022, ISSN: 1424-8220. DOI: 10.3390/s22062216. [Online]. Available: <http://dx.doi.org/10.3390/s22062216>.
- [30] C.-e. Lin, *Introduction to motion estimation with optical flow*, Apr. 2023. [Online]. Available: <https://nanonets.com/blog/optical-flow/>.