VIOLENT SCENE DETECTION IN VIDEOS

BY

YEW KYNN MAN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology
(Perak Campus)

JANUARY 2019

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: ____

____

____

Academic Session: _____________

I ____

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
1. The Library is allowed to make copies of this dissertation for academic purposes.

	Verified by,
_________________________	_________________________
(Author's signature)	(Supervisor's signature)
Address:	
_________________________	
_________________________	_________________________
_________________________	Supervisor's name
Date: ____________________	**Date**: ____________________

VIOLENT SCENE DETECTION IN VIDEOS

BY

YEW KYNN MAN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology
(Perak Campus)

JANUARY 2019

DECLARATION OF ORIGINALITY

I declare that this report entitled "**VIOLENT SCENE DETECTION IN VIDEOS**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : ___________________________

Name : ___________________________

Date : ___________________________

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude towards my supervisor Dr. Tan Hung Khoon who gave me the opportunity to be exposed to deep learning. Throughout this entire duration of this project he has continuously guided me and gave me very useful advices despite my shortcomings. I would like to thank him from the bottom of my heart for being so generous by providing me with the resources for this project.

Next, I would like to thank both my parents for their unconditional love and support. I truly appreciate what they have done for me because without them I would never be where I am today.

Finally, I would like to thank my friends for being with me through thick and thin throughout my degree years. You guys have helped and taught me so much, my degree life would never be the same without you guys.

ABSTRACT

Long term exposed to violent content may cause harm to an individual, especially younger children. Nowadays, video sharing and streaming websites are becoming more and more widespread which makes exposure to unwanted violent content much more frequent and inevitable. While there exist many types of violent content, encountering content with physical violence seems to be more common compared other types of violence. Unlike the other types of violent content, physical violence can usually be identified by using visual cues and often associated with certain actions. Recent Convolutional Neural Networks (CNN) has shown great success in visual tasks such as image recognition and classification tasks. Furthering the success, Convolutional Neural Networks extended to video data and has shown that CNN can effectively extract and learn important features for complex tasks such as human action recognition. In this project, a desktop application for violent scene detection and localisation is built using multimodality deep learning architecture for violent scene detection. Besides that, this application also provides users with the interface to view and filter detected violent scenes. To examine the generalisation capability of this application on different video types, this application is tested on Hollywood movies and web-based videos. Based on testing results, despite less optimal training process, this application is able to detected violent scene in Hollywood movies quite well, MAP2104 in Hollywood movies is 0.56. The performance on web-based videos on the other hand is poorer with MAP2014 of 0.43.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS v

ABSTRACT.. vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES x

LIST OF ABBRIVATIONS xi

Chapter 1: Introduction 1

1.1 Problem Statement and Motivation 1

1.2 Project Scope and Objective 2

1.3 Impact, Significance and Contribution 2

Chapter 2: Literature Review 4

2.1 Organisation of Literature Review 4

2.2 Violent Dataset 4

2.3 Conventional Methods 8

2.4 Deep Learning Methods 10

Chapter 3: System Design 14

3.1 Block Diagram 14

3.2 Use Case Diagram 15

3.3 Application Flowchart 17

3.4 Methodology 19

3.4.1 CNN Architecture 19

3.4.2 Two Stream CNN Input 21

3.4.3 Spatial Stream CNN Configuration 22

3.4.4 Temporal Stream CNN Configuration 22

3.4.5 Late Fusion 23

3.4.6 Training and Testing Dataset 23

Chapter 4: Implementation Details 24

4.1 System Specification 24

4.1.1 Hardware Specification 24

4.1.2 Software Specification 24

4.2 Sampling Non-Violent Segments 24

4.3 Spatial Stream Training 25

4.4 Temporal Stream Training 25

4.5 Testing 26

4.6 Graphical User Interface 27

Chapter 5: Evaluation 33

5.1 Dataset Evaluation 33

5.2 Test Results 34

5.2.1 Main Task Results 34

5.2.2 Generalisation Task Results 39

Chapter 6: Conclusion 42

REFERENCES 43

APPENDIX 46

Samples of Train Set Data 46

POSTER 49

PLAGARISM CHECK RESULT 50

LIST OF FIGURES

Figure 2.1: Statistics of Movies and Videos in VSD2014 5

Figure 2.2: Participant Performance for Main Task 6

Figure 2.3: Participant Performance for Generalisation Task 6

Figure 2.4: System Overview of Violent Scene Detection using Deep Neural Networks by (Dai et al., 2014) 8

Figure 2.5: System Overview Violent Scene Detection using Concept-based Fusion Approach by (Sjöberg et al., 2014) 9

Figure 2.6: Two-stream Architecture for Action Recognition (Simonyan and Zisserman, 2014) 11

Figure 2.7: Violent Scene Detection System overview of (Dai et al., 2015) 12

Figure 3.1: Block Diagram 14

Figure 3.2: Use Case Diagram for Violent Scene Detection Application 15

Figure 3.3: Flowchart for Violent Scene Detection Application 17

Figure 3.4: Overview of Two Stream CNN Architecture for Violent Scene Detection ... 19

Figure 3.5: ResNet50 Architecture 20

Figure 3.6: Spatial and Temporal Stream CNN Input 21

Figure 3.7: Optical Flow between Two Consecutive Frames 21

Figure 4.1: Application Main Screen 27

Figure 4.2: Application File Browser 28

Figure 4.3: Video File Loaded from Video Folder 29

Figure 4.4: Violent Scene Detection Process 30

Figure 4.5: Filter by Top N Violent Scenes 31

Figure 4.6: Filtering by time 32

LIST OF TABLES

Table 5.1: Main Task Results 34

Table 5.2: Categories of Violent Scenes Detected by Spatial Stream CNN..................... 36

Table 5.3: Violent Scenes Predicted by Temporal Stream CNN...................................... 37

Table 5.4: False Positive Scenes Detected by Temporal Stream CNN 38

Table 5.5: Main Task Performance Figure for Each Participating Team 39

Table 5.6: Generalisation Task Results 39

Table 5.7: Comparison of Least Violent Images with False Positive Images 41

Table 5.8: Generalisation Task Performance figure for Each Participating Team 41

LIST OF ABBRIVATIONS

ConvNet/ CNN	Convolutional Neural Network
Cd	Compact Disc
RGB	Red, Green and Blue
MAP	Mean Average Precision
LSTM	Long Short-Term Memory
VSD2014	Violent Scene Detection Dataset 2014
STIP	Space-Time Interest points
MFCC	Mel-Frequency Cepstral Coefficient
HOG	Histogram of oriented gradient
HOF	Histogram of optical flow
MBH	Motion boundary histogram
TrajShape	Trajectory shape
DNN	Deep neural network
SVM	Support Vector Machine

Chapter 1: Introduction

1.1 Problem Statement and Motivation

In the recent decade internet penetration rate has drastically increased, Malaysia's Internet penetration in 2018 has risen to 85.7 per cent from just 70.0 per cent in 2015 according to an article from (Straits Times 2018). Because of the conveniences that the internet brings to us, many of our daily activities has been gradually shifted onto the internet. Consuming media content online in particular has become a big part of our daily activities, statistics shows that there is more than more than 500 million hours of videos watched each day (27 Video Stats for 2017) which goes to show that there is a big market for online media, no wonder more and more video sharing and streaming websites are mushrooming all over the internet. Although these platform does provide us with the benefit of having access to a larger variety of content and not to mention better content quality, these platforms are not without their downsides. As children are having their own personal devices at a younger and younger age these days, many adults are concerned that their children might be exposed to unwanted harmful content online. Acknowledging these concerns many video sharing or video streaming platforms use methods such as community flagging or employing human assessors for manually filtering harmful content. Although these methods can reduce the amount of harmful content from being spread, they do not scale well when presented with very large amount of videos. For reference, according to a (Wordstreamcom. 2019) for every hour 72 hours of video is uploaded to YouTube.

One of the types of harmful content that can often be found on the internet is violent content. Studies have shown that long term exposure to violent content can cause unwanted side effects on mental health Long term exposure to violent content can cause one to be desensitized towards violence, which will result in gradual decrease of empathy of an individual towards people around them. Some individual, especially younger children might tend to imitate violent actions and act more aggressively towards others. Others tend to become very fearful about the world around them (Huesmann, L. R. 2007). Therefore, to solve this problem a violent scene detection application can be built to detect violent scenes so that further actions can be taken for the violent video.

Violent content can be divided to different categories such as physical, sexual, psychological and emotional. Unlike other categories which require more semantic information, physical violence can often be associated with certain actions. Recent deep learning architecture has proven to be able to achieve good performance in action recognition task. Since physical violence are often associated with particular actions such as fighting, falling, shooting, explosion etc., deep learning architecture for action recognition could be adopted for detecting physical violence in videos.

1.2 Project Scope and Objective

1. **To develop a violent detection system that is able to detect and localise violent scenes in videos.**

   In order to detect violent scenes in videos, a multimodality deep learning CNN will be trained. The multimodality deep learning CNN architecture will be trained and tested on video data in Violent Scene Detection Dataset 2014. In addition to detection, the violent scene detection system will be able to determine the start and end frame for each scene and determine the degree of violence that is present.

2. **To build a desktop application that allows users to view and filter violence scenes.**

   To allow users to have a detailed visual view of the detected violent scenes, an application will be built. Beside viewing, the application will allow users to filter through violent scenes so that they will have more control as to view more violent or less violent scenes.

1.3 Impact, Significance and Contribution

The main beneficiary of this application would be the users of video sharing or streaming websites. With this violent detection application violent videos can be blocked from being uploaded and there will be lesser circulation of violent content. Parents can be less worried of their children being constantly exposed to unwanted violent content.

For individuals that are hired by web sharing or streaming website companies to assess harmful content, filtering violent content on long videos will become less tedious. Since all the scenes are segmented and scenes that are likely to contain violence are automatically detected, human assessors will only have to review the detected violent scenes, this will help them to significantly narrow down the amount of footage human assessors have to go through.

Besides benefiting the human assessors, violent scene detection application will also be able to benefit the web sharing and streaming companies itself, this is because by using machines for violent scene detection, significant amount of detection task to be automated. This will allow violent content filtering to be scalable on large amount of videos.

Chapter 2: Literature Review

2.1 Organisation of Literature Review

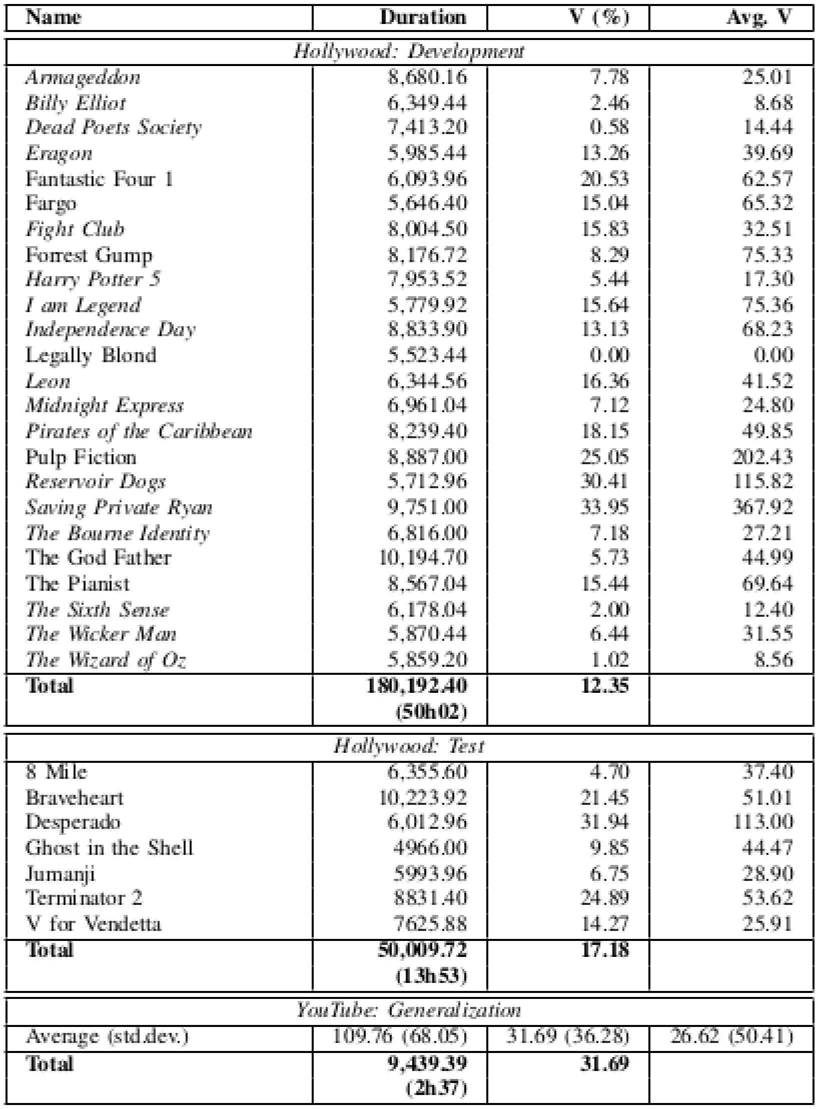
This literature review is organized into 3 sections. Section 2.2 will be used to introduce the dataset that will be used for training and testing. Section 2.3 will be used to review two existing systems that uses conventional features. Section 2.4 will be discussing deep learning approach for video data and one existing system that uses deep learning approach for violent scene detection.

2.2 Violent Dataset

A common problem when building a violence detection system is that there are wide variety of violence, many people will have different interpretations towards what a violent detection system should be able to detect. Many researchers have attempted to create or gather data for their methods of violence detection. However, those data are often closed or only adapted to specific context of a certain method. The lacking of common definition of violence has caused it to be difficult for different violent detection methods to be compared. Acknowledging this problem, (Schedl et al., 2015) prepared an annotated dataset for violent scenes which can be easily referenced and obtained for benchmarking emerging research task for violent scene detection.

Annotation of violent segment in VSD2014 is based on a subjective definition closer to targeted real-world scenario of violence which is physical violence that one would not let an 8-year-old child watch. Annotation were created by human assessors in a bottom-up manner where master annotators would cross check with the violence annotation created by regular annotators.

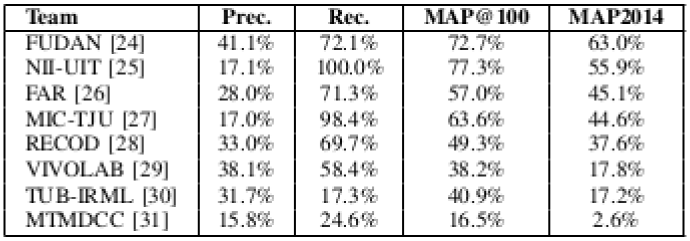
VSD2014 dataset consist of binary segment annotation for violent in 31 Hollywood movies and 86 web video clips. The release year of annotated movies are between 1990 and 2000. The range of violence in these movies are from movies that are very violent to movies with almost no violence. Segments are annotated at frame level and multiple consecutive violent action are merged into one segment. In addition to violent binary segment annotations, annotation for high level concepts that are mostly related to violent content are also provided. These additional annotations can be broken down into visual and audio concepts where visual concepts include presence of blood, fights, fire, guns, cold arms, car chase and gory scenes while audio concepts include gunshots, explosion and scream. The statistics of individual movies are shown in Figure 2.1.



Name	Duration	V (%)	Avg. V
Hollywood: Development			
Armageddon	8,680.16	7.78	25.01
Billy Elliot	6,349.44	2.46	8.68
Dead Poets Society	7,413.20	0.58	14.44
Eragon	5,985.44	13.26	39.69
Fantastic Four 1	6,093.96	20.53	62.57
Fargo	5,646.40	15.04	65.32
Fight Club	8,004.50	15.83	32.51
Forrest Gump	8,176.72	8.29	75.33
Harry Potter 5	7,953.52	5.44	17.30
I am Legend	5,779.92	15.64	75.36
Independence Day	8,833.90	13.13	68.23
Legally Blond	5,523.44	0.00	0.00
Leon	6,344.56	16.36	41.52
Midnight Express	6,961.04	7.12	24.80
Pirates of the Caribbean	8,239.40	18.15	49.85
Pulp Fiction	8,887.00	25.05	202.43
Reservoir Dogs	5,712.96	30.41	115.82
Saving Private Ryan	9,751.00	33.95	367.92
The Bourne Identity	6,816.00	7.18	27.21
The God Father	10,194.70	5.73	44.99
The Pianist	8,567.04	15.44	69.64
The Sixth Sense	6,178.04	2.00	12.40
The Wicker Man	5,870.44	6.44	31.55
The Wizard of Oz	5,859.20	1.02	8.56
Total	**180,192.40** **(50h02)**	**12.35**	
Hollywood: Test			
8 Mile	6,355.60	4.70	37.40
Braveheart	10,223.92	21.45	51.01
Desperado	6,012.96	31.94	113.00
Ghost in the Shell	4966.00	9.85	44.47
Jumanji	5993.96	6.75	28.90
Terminator 2	8831.40	24.89	53.62
V for Vendetta	7625.88	14.27	25.91
Total	**50,009.72** **(13h53)**	**17.18**	
YouTube: Generalization			
Average (std.dev.)	109.76 (68.05)	31.69 (36.28)	26.62 (50.41)
Total	**9,439.39** **(2h37)**	**31.69**	

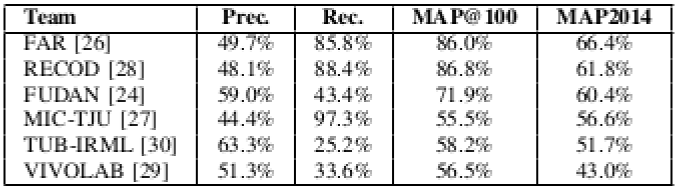
Figure 2.1: Statistics of Movies and Videos in VSD2014

In the 2014 edition of violent scene detection task, 2 violence detection tasks were carried out. For both tasks, participants were asked to identify start and end frames of violent scenes. The main task required participants to test their performance of violent detection algorithms on 7 movies in the Hollywood test set while for generalisation task violent detection algorithms were tested on 87 short web videos. During training, any features that are present in the 24 Hollywood development set can be used for violent scene detection which most teams resorted to using multimodality models. Most common features that were used were audio features such as MFCC and video features such as dense trajectories implemented using histogram of oriented gradients (HOG), histogram of optical flow (HOF), or motion boundary histogram (MBH). Some teams also use static image features such as SIFT, colourfulness, saturation, brightness and hue.



Team	Prec.	Rec.	MAP@100	MAP2014
FUDAN [24]	41.1%	72.1%	72.7%	63.0%
NII-UIT [25]	17.1%	100.0%	77.3%	55.9%
FAR [26]	28.0%	71.3%	57.0%	45.1%
MIC-TJU [27]	17.0%	98.4%	63.6%	44.6%
RECOD [28]	33.0%	69.7%	49.3%	37.6%
VIVOLAB [29]	38.1%	58.4%	38.2%	17.8%
TUB-IRML [30]	31.7%	17.3%	40.9%	17.2%
MTMDCC [31]	15.8%	24.6%	16.5%	2.6%

Figure 2.2: Participant Performance for Main Task



Team	Prec.	Rec.	MAP@100	MAP2014
FAR [26]	49.7%	85.8%	86.0%	66.4%
RECOD [28]	48.1%	88.4%	86.8%	61.8%
FUDAN [24]	59.0%	43.4%	71.9%	60.4%
MIC-TJU [27]	44.4%	97.3%	55.5%	56.6%
TUB-IRML [30]	63.3%	25.2%	58.2%	51.7%
VIVOLAB [29]	51.3%	33.6%	56.5%	43.0%

Figure 2.3: Participant Performance for Generalisation Task

As shown in Figure 2.2 and 2.3 The metric for evaluating performance were using MAP100 and MAP2014, other relevant metrics used include precision and recall. The purpose of using MAP2014 in comparison to MAP100 is to have a more suitable measure for violent scene detection system where the better violent detection algorithm can detect a wider variety of violence. For MAP100, algorithms could predict multiple small segments within the same violent segment while not predicting on other violent segment and still obtain a high score. MAP2014 on the other hand, penalises these algorithms and rewards algorithms that are able to predict from more violent segment instead. MAP2014 considers a hit if a predicted segment overlaps with corresponding ground truth segment by more than 50% or vice versa and most importantly several hits on the same ground truth segments is considered as one true positive and once a violent segment is considered detected subsequent hits on the same segments are ignored.

2.3 Conventional Methods

Most participants in MediaEval 2014 violent detection task made use of a combination of multiple different features that are present in video-based data, majority of these work use conventional features for their algorithms.

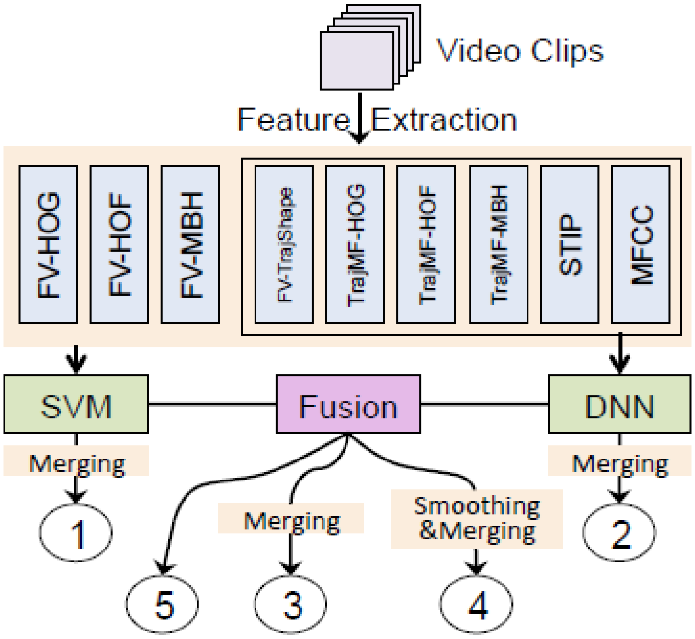


Figure 2.4: System Overview of Violent Scene Detection using Deep Neural Networks by (Dai et al., 2014)

The violent scene detection system by (Dai et al., 2014) shown in figure 2.4 which had the best MAP2014 performance for main task combined visual, audio and trajectory-based features for detecting violent scenes. The visual feature extracted was STIP, audio feature that were extracted was MFCC and trajectory-based motion features were HOG, HOF, MBH and TrajShape. In order to capture the relationship of distinct features for each clip, all their distinct features were used as input for their regularised DNN based video classifier. The first layer of their DNN performs feature abstraction for each input. The second layer which has special regularised weights performs fusion of features from different input. Their last layer classifies the fused features into violent and non-violent prediction. During test time when they were required to detect violent scenes, they applied several techniques to obtain prediction for violent segments. First, they partition the video into multiple 3-second-long clips and detect each clip for violent content. Then, they smoothen the prediction scores by averaging the scores in 3 clip windows to eliminate some incorrect predictions. In addition to that, all individual neighboring violent or non-violent prediction were merged by averaging their prediction value to obtain segment level prediction score. Their score smoothing and clip merging technique has been noted to significantly improve their violent scene detection performance their main task.

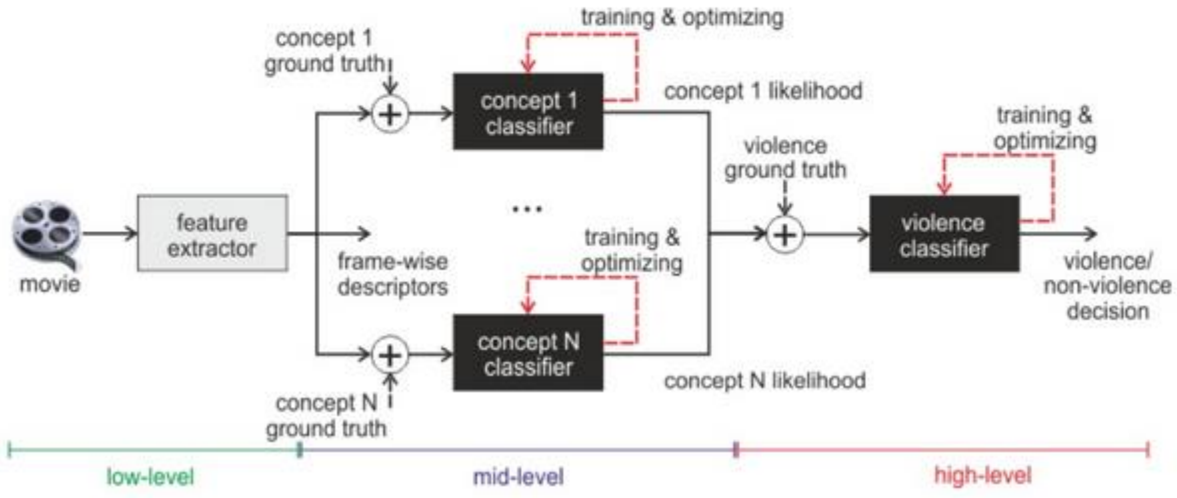


Figure 2.5: System Overview Violent Scene Detection using Concept-based Fusion Approach by (Sjöberg et al., 2014)

Another participating group (Sjöberg et al., 2014) also using conventional features has shown that for more mixed styled video data such as web-based YouTube videos, using violent related concept fusion as shown in Figure 2.5 for detecting violent scenes can be quite effective as their algorithm that uses mid-level concept prediction score as input for their violence classifier has performed consistently well on web-based. Their training process is broken down into two levels. In the first level of training, they extract the low-level features from segments that are annotated with violent related concepts, then they use the low-level features to train their mid-level concept classifiers. For visual features they extract Colour Naming Histogram, Colour Moments, Local Binary Patterns, Colour Structure Descriptor, Grey Level Run Length Matrix and HOG for visual. For audio features they extract amplitude envelop, root-mean-square energy, zero-crossing rate, band energy ration, spectral centroid, spectral flux, bandwidth and MFCC. In the second level of training, they train their high-level violence classifier with previous concept prediction score. For all their classifiers, they use multi-layer perceptron with a single hidden layer of 512 neuron and one or multiple output neurons. During test time, to obtain the final violent prediction score they apply smoothing by using sliding median filter and thresholding prediction score using the threshold value that maximises their MAP2014 score on train set.

2.4 Deep Learning Methods

Both methods that were mentioned previously have been using conventional machine learning methods for violent scene detection. Participants had to carefully identify and select features that will be useful for prediction of a particular task. Deep learning on the other hand allows end-to-end learning which enables learning of high-level features from data in an incremental manner. This has reduced the amount of expertise needed for applying machine learning in a particular domain which is especially useful for complex tasks where useful features may be hard to determine.

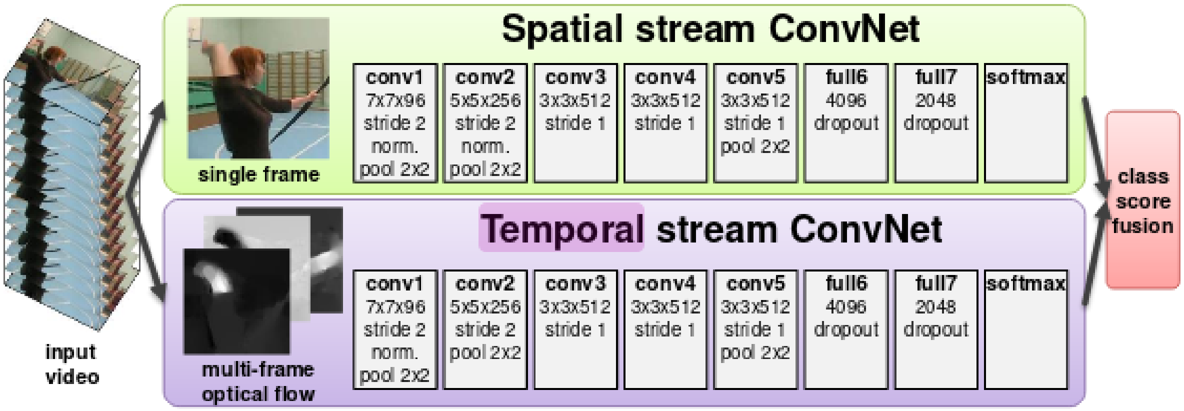


Figure 2.6: Two-stream Architecture for Action Recognition (Simonyan and Zisserman, 2014)

Recent work by (Simonyan and Zisserman, 2014) has shown that deep learning method was able to perform well compared to conventional machine learning on complex task such as action recognition. The proposed architecture was a two stream CNN architecture extended from image recognition task to action recognition task by adding an additional CNN to the architecture for extracting the additional temporal information in videos. In the proposed architecture, as shown in Figure 2.6 each stream is used to extract different components in video. The spatial stream learns the information of objects that are associated with certain action from still RGB images. The temporal stream on the other hand, learns the motion of object that is associated with certain actions from optical flow stack.

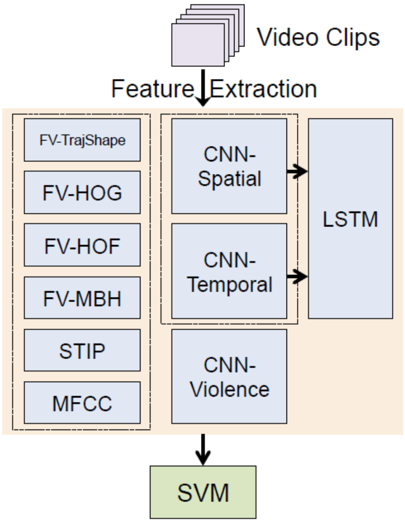


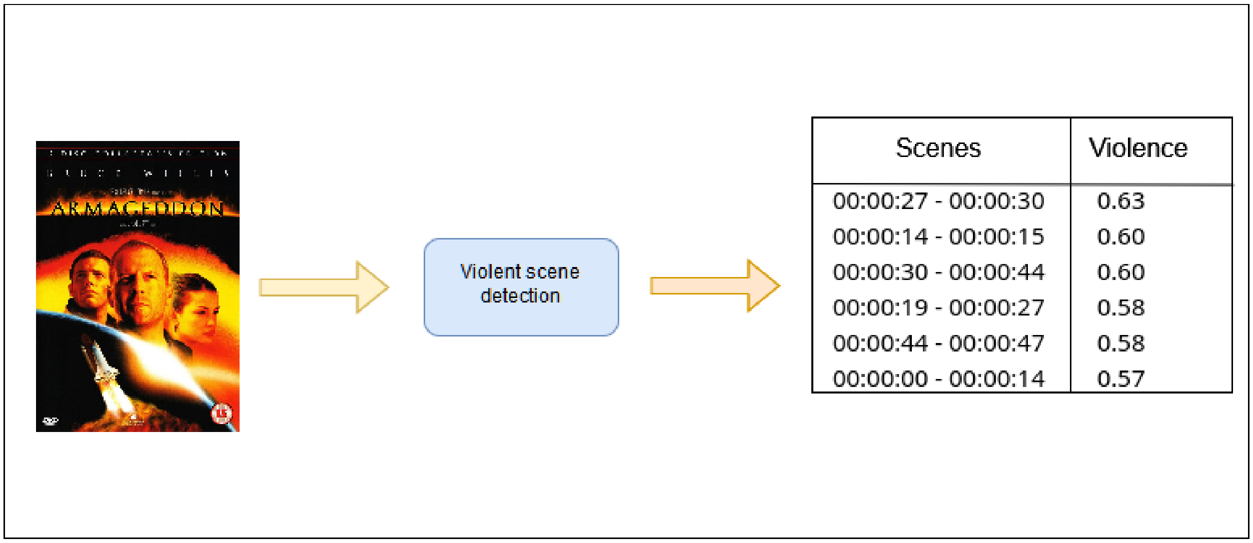
Figure 2.7: Violent Scene Detection System overview of (Dai et al., 2015)

(Dai et al., 2015) built a system to explore the effectiveness of deep learning methods for violent scene detection. Their system as shown in Figure 2.7 uses conventional features and deep learning features for violent scene detection. Similar to their previous violent scene detection system (Dai et al., 2014), the conventional features that were used included visual, auditory and trajectory-based features. Their deep learning features on the other hand, are extracted from outputs of their deep learning architectures. For extracting deep learning features from violent images, they used CNN-violence which is an AlexNet model trained on a subset of 2614 ImageNet classes related to violent. For extracting spatial and temporal information they used 2 stream CNN architecture that was proposed by (Simonyan and Zisserman, 2014) and trained the spatial steam on ImageNet data and the temporal stream on optical flow data. To allow retention of long-term information from their two stream CNN model they stack LSTM architecture on top of each CNN stream. From the results of their violent scene detection system, they shows that deep learning architectures were able to outperform conventional methods in extracting useful features for complex tasks such as violent scene detection.

One common obstacle faced when training deep learning architectures is that video-based data are usually limited in both size and diversity. This might become a problem when training deep ConvNets as they require a large amount of training samples to achieve optimal performance. Deep ConvNets trained on small video-based datasets are often confronted with high risk of over-fitting. (Wang et al., 2016) has proposed several techniques to training deep ConvNet with small video datasets. First method is to provide good weight initialisation for ConvNet using transfer learning. For spatial stream CNN, weights can be initialised using pretrained weight of models that are trained on huge dataset. Similarly, temporal steam CNN can also use pretrained weight from spatial stream CNN by using cross modality pretraining. Second technique for training deep ConvNet on small datasets is to add regularisation such as dropout after global average layers to prevent overfitting during training. Another method to reduce overfitting effect is to use enhanced data augmentation to provide the training model with more variation of training data. The enhanced data augmentation methods used by (Wang et al., 2016) was random horizontal flipping and cropping from 4 corners and centre with height and width randomly selected between {256, 224, 192, 168} pixels.

Chapter 3: System Design

3.1 Block Diagram



Scenes	Violence
00:00:27 - 00:00:30	0.63
00:00:14 - 00:00:15	0.60
00:00:30 - 00:00:44	0.60
00:00:19 - 00:00:27	0.58
00:00:44 - 00:00:47	0.58
00:00:00 - 00:00:14	0.57

Figure 3.1: Block Diagram

Figure 3.1 shows the block diagram of this violent scene detection application. Users can select videos that they desire for detection. This application will detect and localise each violent scene contained within a video and then allow user to playback or filter detected violent scenes.

3.2 Use Case Diagram

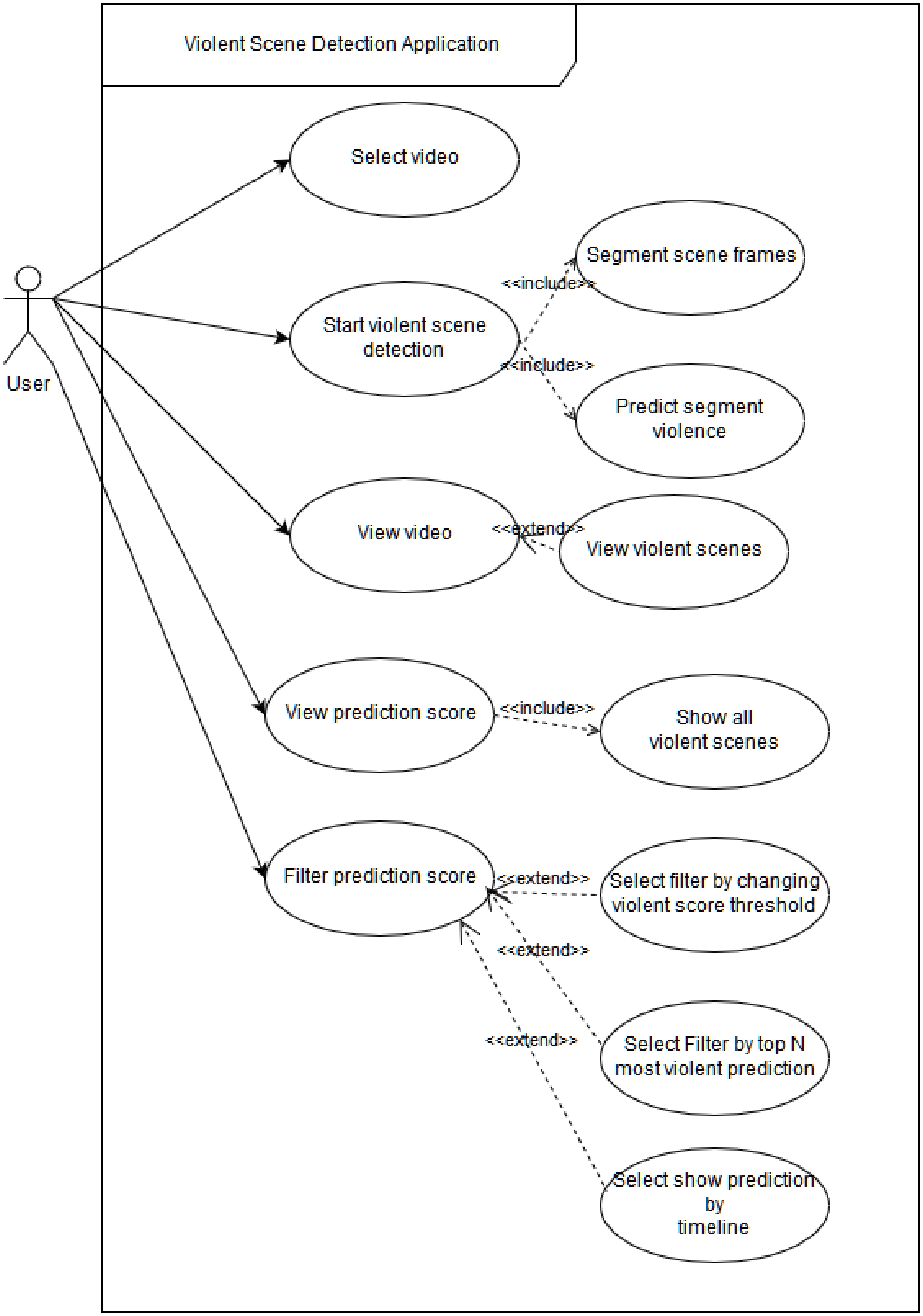


Figure 3.2: Use Case Diagram for Violent Scene Detection Application

As shown in Figure 3.2, user will be able to perform these actions in this violent scene detection application.

1. Selecting video
   - User can select the directory for video or file that contain videos for detection.
2. Detect violent Scenes
   - User can start the detection process.
   - During detection, the application will start by identifying the start and end frame of each scene. Then the application will start detecting violence for each scene.
3. Play selected video
   - Users can choose to either play the entire video or individual violent scene.
4. View predicted score
   - After detection process is complete, users will be shown the scenes that are detect to contain violence. Users can also click on the segment to view the video.
5. Filter prediction score
   - User can filter out unwanted results to show segments which are most important to violence such as top N most violent scenes, according to video timeline or top most violent scenes by thresholding violence score.

3.3 Application Flowchart

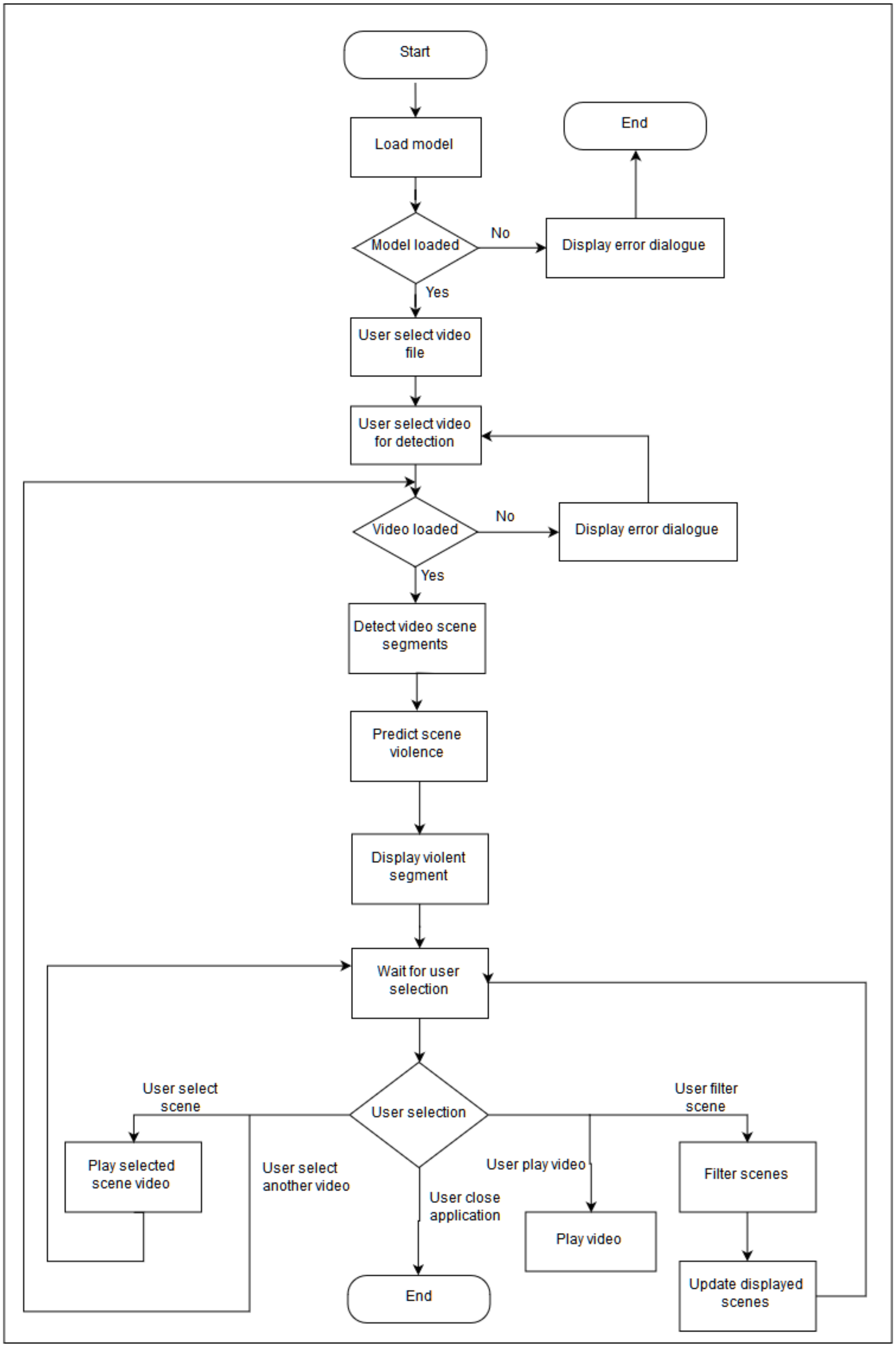


Figure 3.3: Flowchart for Violent Scene Detection Application

Figure 3.3 shows the flowchart for this violent scene detection application. When users start up the application, both spatial and temporal model will first be loaded into memory. Then users will be allowed to choose their video or file of their choice. After selecting their desired video of video folder, users can start the prediction process by double clicking on the video name. During the detection process, the application first has to segment the video into individual scenes. After all scenes has been segmented, the application starts to detect for violence in each scene using the deep learning models that were loaded beforehand. After detection is complete, the application will display the scenes that has violence score of over 0.5. Each scene segment shown to the user is labelled with the start and end time of the scene as well as the prediction score for the entire scene. Users can then choose to play the entire video or individual scenes. In addition to that, users can also filter violent segments by changing the detection threshold or by selecting the top N most violent scenes.

3.4 Methodology

Video data is made up of spatial and temporal components, the spatial component is the information of objects in each frame while the temporal component is the motion of object across consecutive frames. Usually violent scenes have very distinct visual cues and these visual cues could be present in either components of video data. For example, scenes that contain explosions, blood, gore or firearm have more distinct visual cues in the spatial component while scenes that contain fights or falls have more distinct visual cues in the temporal component. In order to learn the features of violent scenes from both components, a two stream CNN architecture is used violent scene detection.

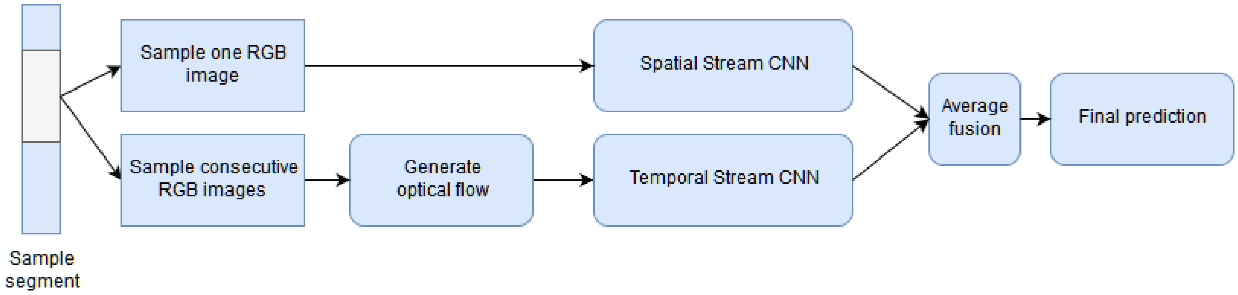


Figure 3.4: Overview of Two Stream CNN Architecture for Violent Scene Detection

As seen in the Figure 3.4 the two stream CNN architecture uses 2 different modalities in video data for violent scene detection. The spatial stream CNN is modeled for the spatial component of video data, it extracts and learns useful features from individual RGB frames. Temporal stream CNN on the other hand is modeled for the temporal components of video data and it extracts and learns useful motion features from optical flow stack generated from consecutive RGB frames. Finally, prediction scores for both modalities are merged by using late fusion.

3.4.1 CNN Architecture

For this two stream CNN architecture, both spatial and temporal stream CNN uses a variant of residual network shown in Figure 3.5. The reason for choosing ResNet50 is because it has a good balance between performance and number of parameters. Besides that, residual networks are also easier to train because they have additional identity shortcut connections that reduces the effect of vanishing or exploding gradient during training.

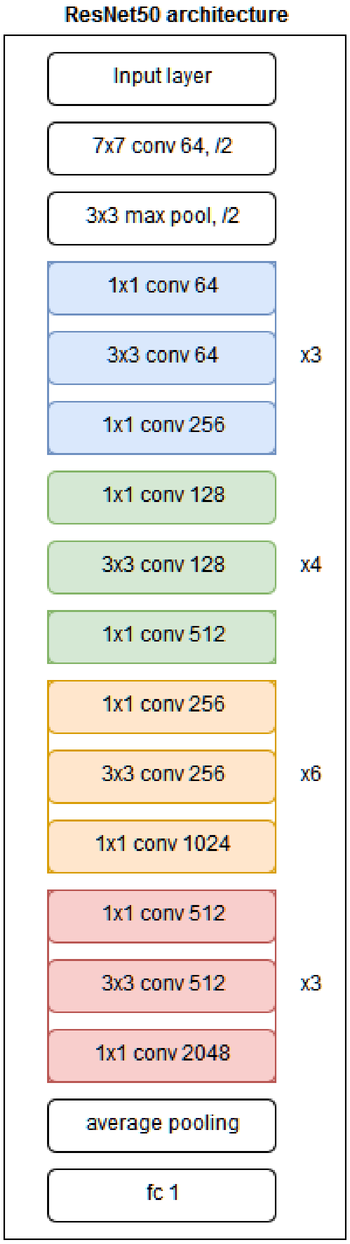


Figure 3.5: ResNet50 Architecture

3.4.2 Two Stream CNN Input

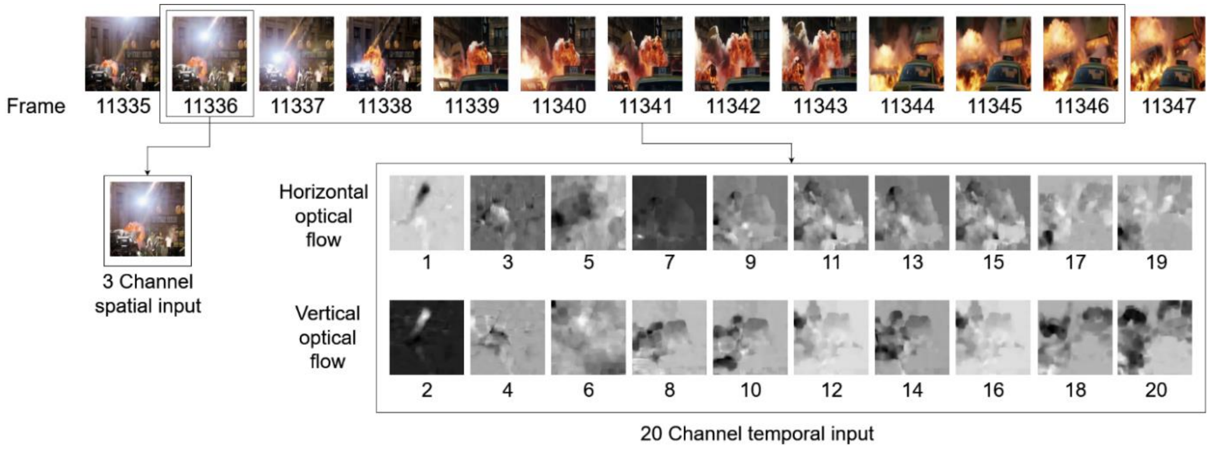


Figure 3.6: Spatial and Temporal Stream CNN Input

Spatial stream input: As shown in Figure 3.6, for each sample the input for spatial stream CNN will be a 3-channel image with shape of 224x224x3.

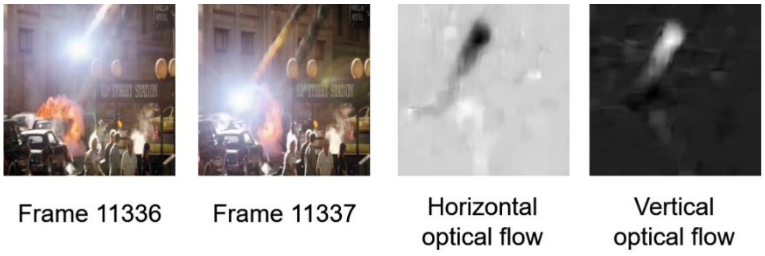


Figure 3.7: Optical Flow between Two Consecutive Frames

Temporal stream input: As shown in Figure 3.6, for each sample the input for temporal stream CNN will be a stack for optical flow frames with a shape of 224x224x20. The 20-channel optical flow frame stack is produced by stacking 10 dense optical flow frame pairs generated from 11 consecutive RGB frames. As shown in Figure 3.7 calculating optical flow between 2 consecutive RGB frames produces a pair of horizontal and vertical optical flow frames.

3.4.3 Spatial Stream CNN Configuration

Since spatial stream CNN will be taking RGB images at input, the input shape of ResNet50 is unchanged as 224x224x3. In order to allow transfer learning from image classification task to spatial stream CNN, the weights of ResNet50 trained on ImageNet dataset is loaded as pretraining for spatial stream CNN. Then, the last fully connected layer of the original ResNet50 is replaced with a new fully connected layer of one neuron with sigmoid as its activation function. This is because unlike image recognition task that has to classify 1000 classes, spatial stream CNN only has to classify violence and non-violence. To avoid overfitting during training, all weights except for the final fully connected layer are set to be non-trainable.

3.4.4 Temporal Stream CNN Configuration

For motion information dense optical flow between 11 consecutive RGB frames will be generated. The resulting temporal input is a stack of 10 consecutive dense optical flow frames pair with shape 224x224x20. Since the temporal input has 20 channels instead of 3 from the spatial stream, the shape of the first convolutional layer has to be altered to have 20 channels. Then, in order to allow transfer learning from image classification task to temporal stream CNN, ImageNet weights for ResNet50 trained on ImageNet dataset is loaded across modality to fit the shape of filter weights in the temporal stream CNN. This is done by modifying weights of the first convolutional layer of shape 7x7x3 to fit the temporal stream CNN’s first convolutional layer of shape 7x7x20 by averaging the weights across the RGB channels and replicating this average by the number of channels of temporal input. Similar to the spatial stream CNN the last fully connected layer of the original ResNet50 is replaced with a new fully connected layer with one neuron and sigmoid as its activation function. For temporal stream CNN, none of the weights in the temporal stream CNN are set to non-trainable so that temporal CNN will be able to learn new temporal features.

3.4.5 Late Fusion

Merging of both streams is done by late fusion, for each sample prediction score from both fully connected layer of spatial stream CNN and temporal stream CNN are averaged to obtain the joint score.

3.4.6 Training and Testing Dataset

The performance of two stream CNN for violent scene detection is trained and tested similarly to the main task and generalisation task of VSD2014 violent scene detection. During training for both spatial and temporal steam, 24 Hollywood movies in Hollywood development set are used for training and validation. As for testing, two stream CNN for violent scene detection will be tested on all movies except Terminator 2 and Ghost in the Shell in the Hollywood test set for main task and all 87 web videos in the generalisation set for generalisation task.

Chapter 4: Implementation Details

4.1 System Specification

4.1.1 Hardware Specification

- Intel Core i7-6700 @ 3.40GHz x 8
- 16GB RAM
- Nvidia GTX1080 GPU 8GB GDDR5X

4.1.2 Software Specification

- Python 3.6
- OpenCV library is used for all image related processing such as sampling, generating optical flow frames and data augmentations.
- Keras which is a high-level neural network API with TensorFlow as backend is used for constructing and training the CNN models for violent scene detection.
- PySceneDetect Library is used to detect start and end frames from each scene in videos.
- PyQt5 is used to create the graphical user interface for violent scene detection application.

4.2 Sampling Non-Violent Segments

When gathering non-violent samples, only those segments that meet certain criteria are selected. First to ensure that non-violent scene do not contain any violent or violence related frames, consecutive frames that are not labelled with violence or other violent related visual concepts such as presence of blood, fights, presence of fire, presence of guns, presence of cold arms, car chases and gory scenes are selected. Then to avoid sampling scenes that contain less useful information or less scene diversity, non-violent scenes that have short duration are filtered out.

4.3 Spatial Stream Training

During each epoch of spatial stream training, one RGB frame from between each annotated sample start and end frame is randomly sampled. Then the sampled image is resized to 340 pixels in width and 256 pixels in height. The resized RGB image then undergoes data augmentation such as:

- random horizontal flips
- random brightness range shift
- RGB jittering
- random width and height cropping from a list of 256, 224, 192, 168 pixels

After augmentation, each individual RGB image is then resized back to 224x224 so that it matches the input shape of the spatial CNN. During training, weights are learnt using mini batch stochastic gradient descent with Adam optimizer with each batch consisting of 32 images. Learning rate is set to 1e-4 so that the best weights could be saved before overfitting. Training of spatial stream is stopped when there were signs of overfitting.

4.4 Temporal Stream Training

Unlike RGB frames that can be sampled on the fly, optical flow frames between 2 frames require considerable larger amount of time to be generated. Therefore, all training and validation optical flow frames are generated to disk prior to temporal stream training from violent and non-violent segment in movies from train set. To generate optical flow frame, 2 RGB frames are first converted to grayscale then an algorithm based on Gunner Farneback’s in OpenCV is used to generate the dense optical between frames. After each pair of dense optical flow frames is generated, the optical flow frames are then linearly rescaled from floating point values to a range of 0-255 and saved to disk.

During each epoch of temporal stream training, 10 consecutive pairs of optical flow frames from between each annotated sample start and end frames are randomly sampled. Consecutive sampled frames are ensured not to exceed segment end annotation. Each consecutive pair of horizontal and vertical optical flow frames are resized to 340 pixels in width and 256 pixels and stacked to generate a 20-channel input. The resized 20 channel input then undergoes data augmentation such as:

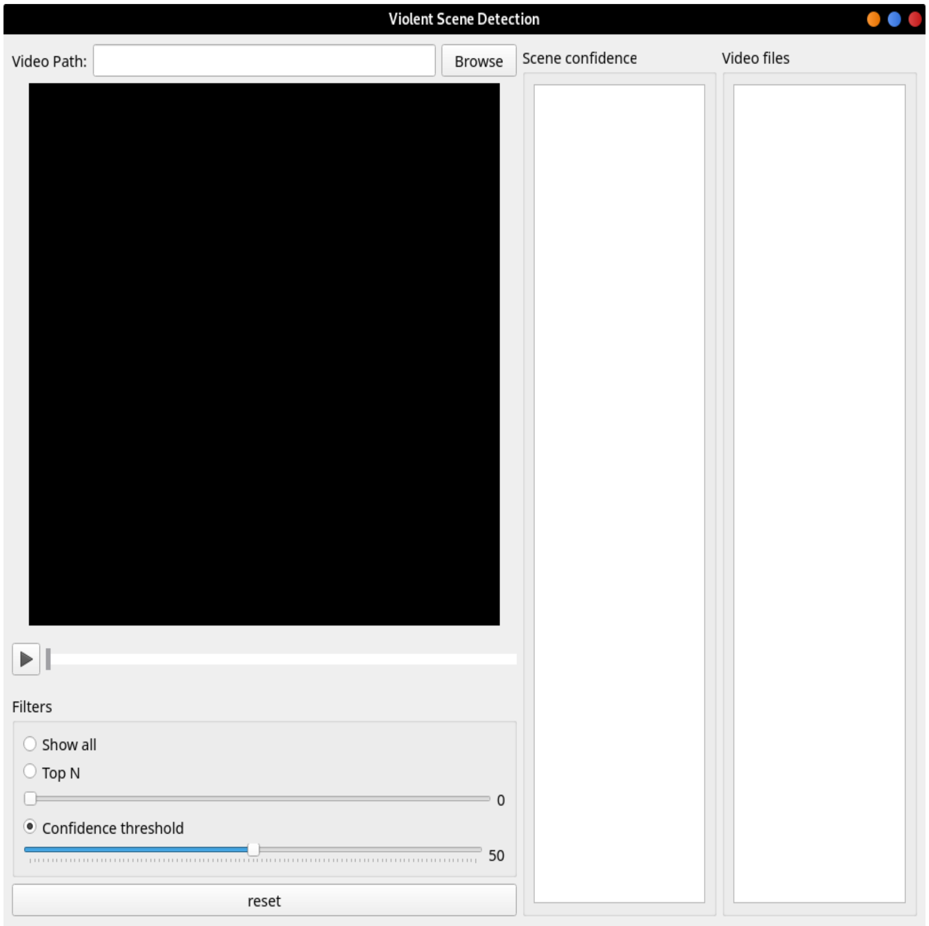
- random horizontal flips
- zero centering of optical flow pixels
- random width and height cropping from a list of 256, 224, 192, 168 pixels

After augmentation, each stack of optical flow frames is then resized back to 224x224x20 input so that it matches input shape of the temporal CNN. Similar to spatial stream, weights are learnt using mini batch stochastic gradient descent with Adam optimizer with each batch consists of 32 images. Learning rate is set to 1e-3 and decreased to 1e-4 after certain amount of epoch where validation accuracy does not increase.

4.5 Testing

At test time, start and end frames for each scene is first determined by using PySceneDetect library. For each scene, a final violent confidence score is obtained by averaging the prediction score of multiple samples with equal temporal spacing. The first sample is taken from the first frame of a scene while consecutive samples will be sampled 3 seconds apart from the previous sample until the end of scene. For each sample, one RGB frame will be sampled for spatial stream CNN while 11 RGB frames will be used to generate 10 optical flow frame pairs for temporal stream CNN. Using the previously sampled input, 10 ConvNet inputs for each stream will be obtained by random cropping and flipping of four corners and centre of each RGB frame and optical flow frame stack. Scores for each sample is obtained by averaging the score for both spatial and temporal stream CNN.

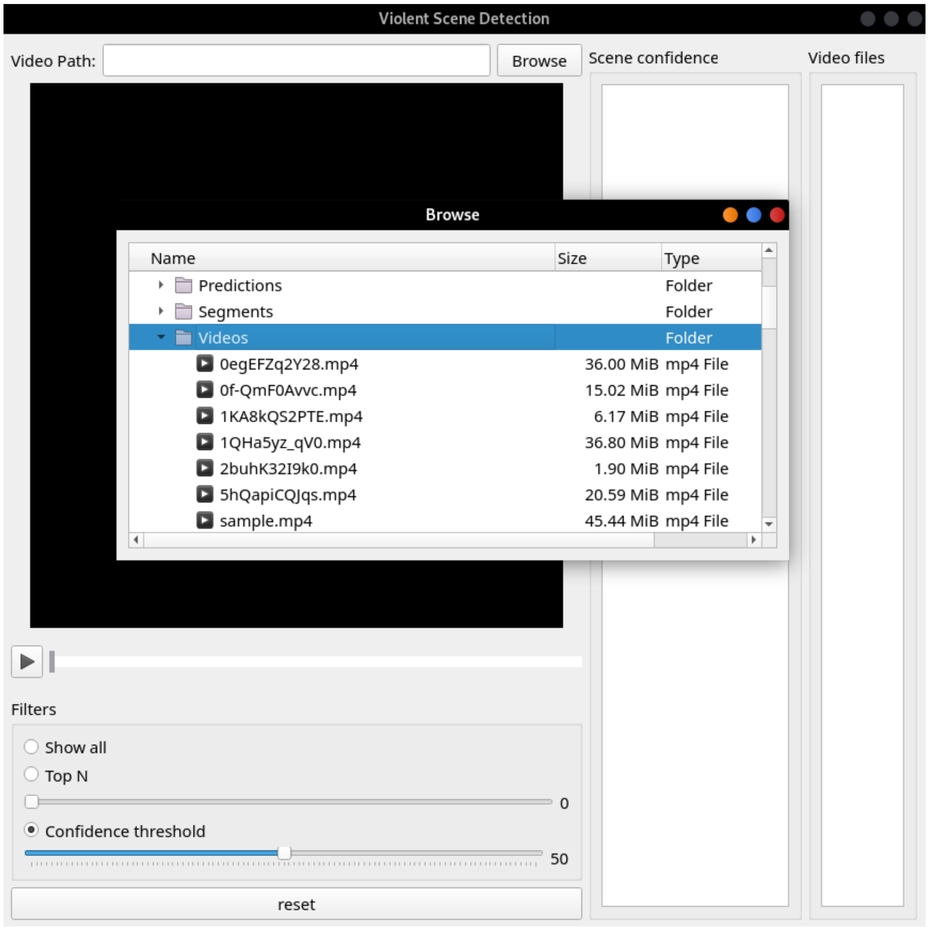
4.6 Graphical User Interface



○ Show all
○ Top N 0
● Confidence threshold 50

Figure 4.1: Application Main Screen

When the application is launched user will be directed to main page as shown in Figure 4.1. User will then have to browse and select their video file or video folder for detection.

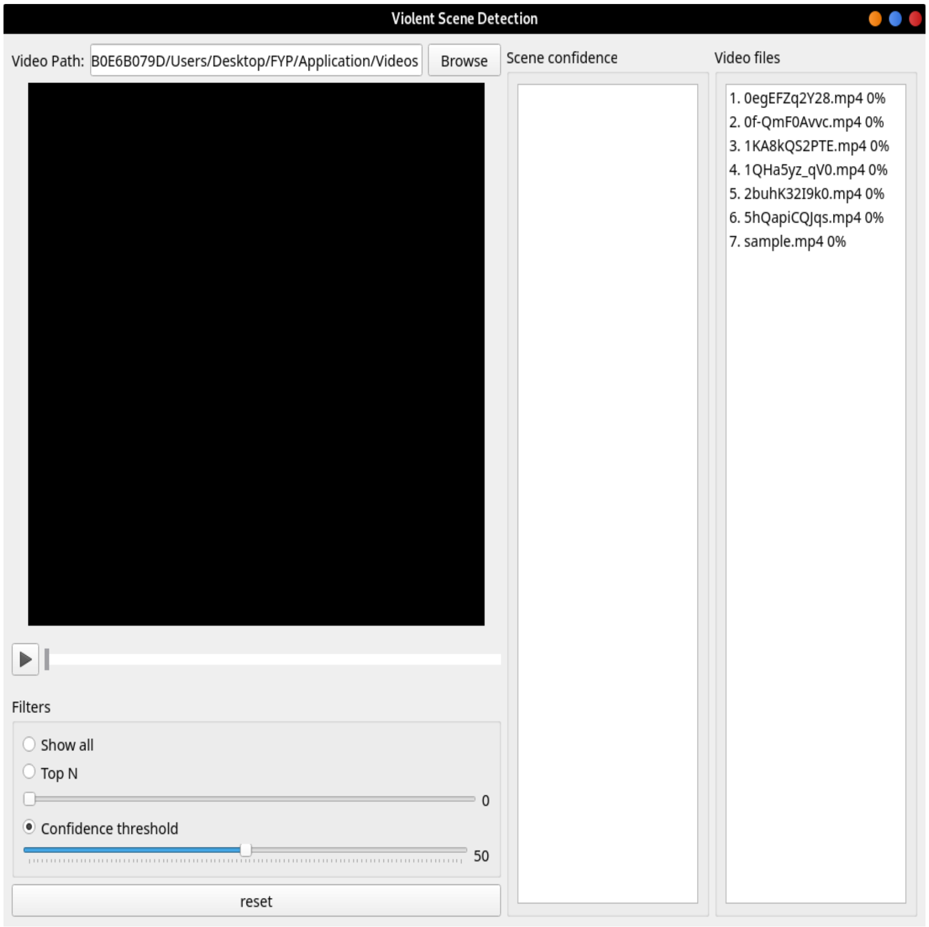


Name	Size	Type
Predictions		Folder
Segments		Folder
Videos		Folder
0egEFZq2Y28.mp4	36.00 MiB	mp4 File
0f-QmF0Avvc.mp4	15.02 MiB	mp4 File
1KA8kQS2PTE.mp4	6.17 MiB	mp4 File
1QHa5yz_qV0.mp4	36.80 MiB	mp4 File
2buhK32I9k0.mp4	1.90 MiB	mp4 File
5hQapiCQJqs.mp4	20.59 MiB	mp4 File
sample.mp4	45.44 MiB	mp4 File

○ Show all
○ Top N 0
● Confidence threshold 50

Figure 4.2: Application File Browser

User will then have to select their video file or video folder for detection using the file browser as shown in Figure 4.2.



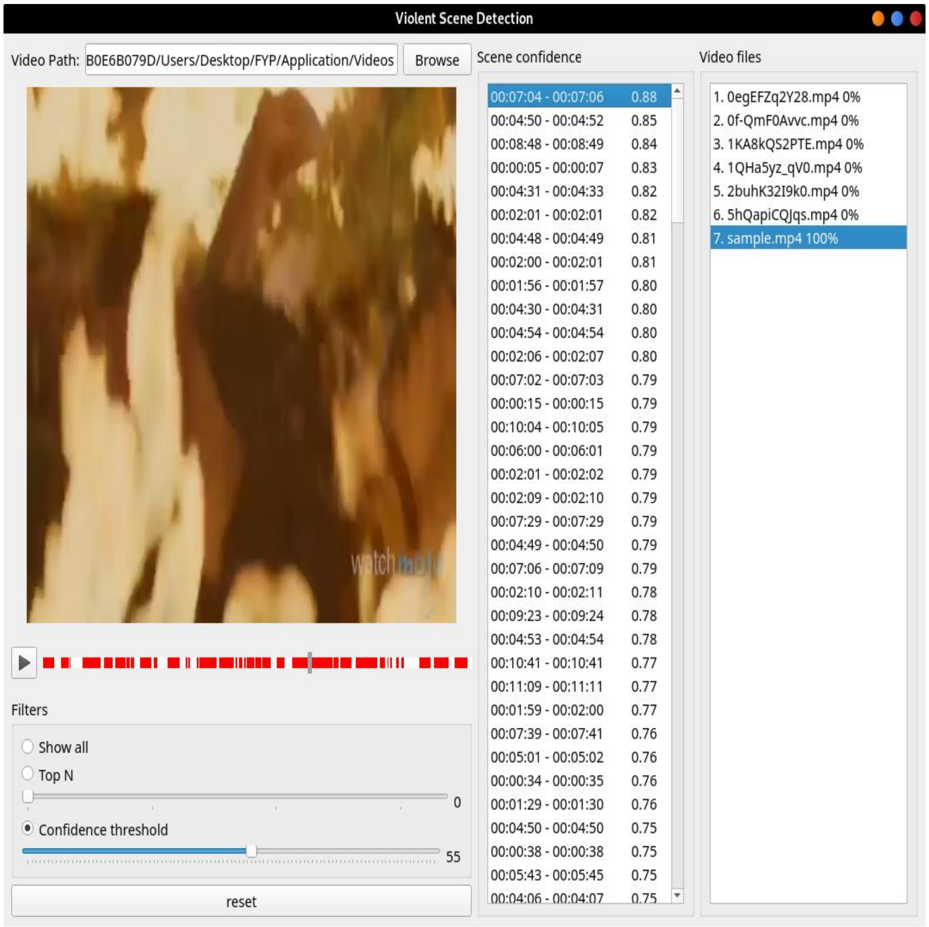
Video Path: B0E6B079D/Users/Desktop/FYP/Application/Videos

1. 0egEFZq2Y28.mp4 0%
2. 0f-QmF0Avvc.mp4 0%
3. 1KA8kQS2PTE.mp4 0%
4. 1QHa5yz_qV0.mp4 0%
5. 2buhK32I9k0.mp4 0%
6. 5hQapiCQJqs.mp4 0%
7. sample.mp4 0%

○ Show all
○ Top N 0
● Confidence threshold 50

Figure 4.3: Video File Loaded from Video Folder

After user has selected a video or folder, all video files will be shown in the video files tab as shown in Figure 4.3. they can start the violent scene detection process by double clicking on the video in the video files tab.



Violent Scene Detection

Video Path: B0E6B079D/Users/Desktop/FYP/Application/Videos

Scene confidence

00:07:04 - 00:07:06	0.88
00:04:50 - 00:04:52	0.85
00:08:48 - 00:08:49	0.84
00:00:05 - 00:00:07	0.83
00:04:31 - 00:04:33	0.82
00:02:01 - 00:02:01	0.82
00:04:48 - 00:04:49	0.81
00:02:00 - 00:02:01	0.81
00:01:56 - 00:01:57	0.80
00:04:30 - 00:04:31	0.80
00:04:54 - 00:04:54	0.80
00:02:06 - 00:02:07	0.80
00:07:02 - 00:07:03	0.79
00:00:15 - 00:00:15	0.79
00:10:04 - 00:10:05	0.79
00:06:00 - 00:06:01	0.79
00:02:01 - 00:02:02	0.79
00:02:09 - 00:02:10	0.79
00:07:29 - 00:07:29	0.79
00:04:49 - 00:04:50	0.79
00:07:06 - 00:07:09	0.79
00:02:10 - 00:02:11	0.78
00:09:23 - 00:09:24	0.78
00:04:53 - 00:04:54	0.78
00:10:41 - 00:10:41	0.77
00:11:09 - 00:11:11	0.77
00:01:59 - 00:02:00	0.77
00:07:39 - 00:07:41	0.76
00:05:01 - 00:05:02	0.76
00:00:34 - 00:00:35	0.76
00:01:29 - 00:01:30	0.76
00:04:50 - 00:04:50	0.75
00:00:38 - 00:00:38	0.75
00:05:43 - 00:05:45	0.75
00:04:06 - 00:04:07	0.75

Video files

1. 0egEFZq2Y28.mp4 0%
2. 0f-QmF0Avvc.mp4 0%
3. 1KA8kQS2PTE.mp4 0%
4. 1QHa5yz_qV0.mp4 0%
5. 2buhK32I9k0.mp4 0%
6. 5hQapiCQJqs.mp4 0%
7. sample.mp4 100%

Filters

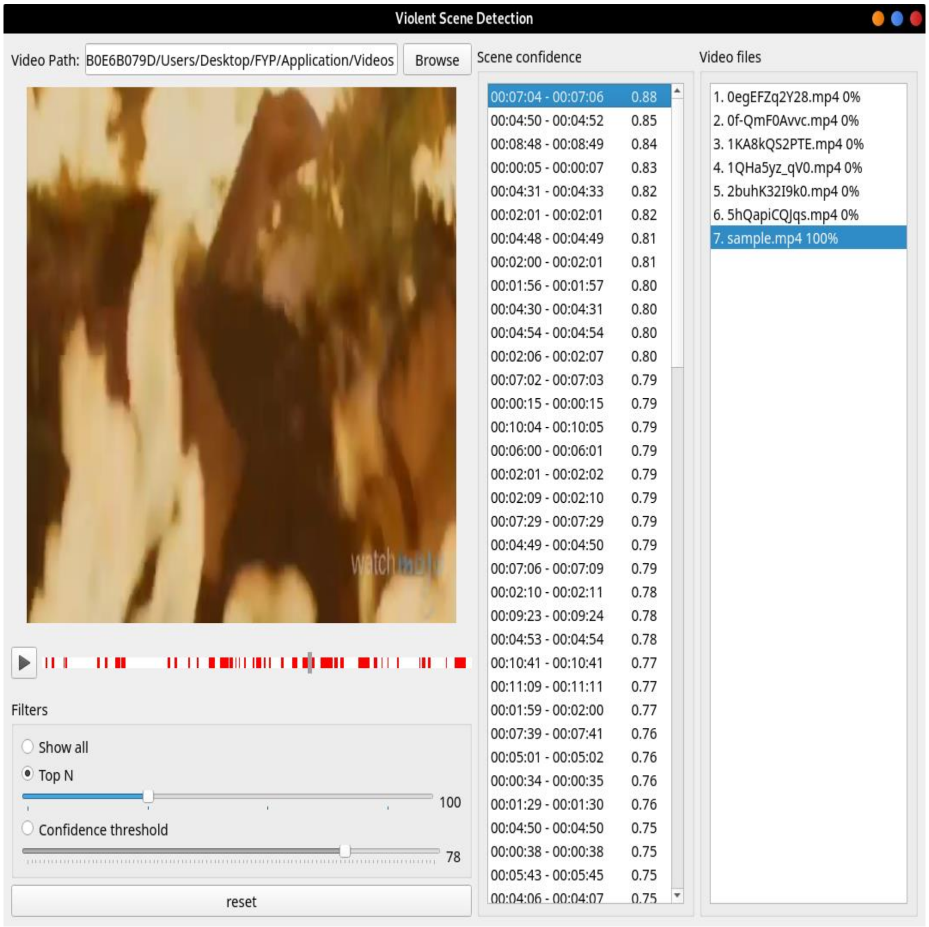
○ Show all

○ Top N 0

● Confidence threshold 55

Figure 4.4: Violent Scene Detection Process

During detection as shown in Figure 4.4, detection progress will be shown beside the movie name. After detection is complete, the application all will automatically present the user with all the detected violent scenes that have violent score of 0.5 and above. Users can change the threshold of violent score by selecting the value from the slider. In addition, users can first view the entire video or control the video by selecting the button under the video player. Optionally users can choose the play the video of each individual violent scene by clicking on the scene start and end time.



Violent Scene Detection

Video Path: B0E6B079D/Users/Desktop/FYP/Application/Videos

Scene confidence

00:07:04 - 00:07:06	0.88
00:04:50 - 00:04:52	0.85
00:08:48 - 00:08:49	0.84
00:00:05 - 00:00:07	0.83
00:04:31 - 00:04:33	0.82
00:02:01 - 00:02:01	0.82
00:04:48 - 00:04:49	0.81
00:02:00 - 00:02:01	0.81
00:01:56 - 00:01:57	0.80
00:04:30 - 00:04:31	0.80
00:04:54 - 00:04:54	0.80
00:02:06 - 00:02:07	0.80
00:07:02 - 00:07:03	0.79
00:00:15 - 00:00:15	0.79
00:10:04 - 00:10:05	0.79
00:06:00 - 00:06:01	0.79
00:02:01 - 00:02:02	0.79
00:02:09 - 00:02:10	0.79
00:07:29 - 00:07:29	0.79
00:04:49 - 00:04:50	0.79
00:07:06 - 00:07:09	0.79
00:02:10 - 00:02:11	0.78
00:09:23 - 00:09:24	0.78
00:04:53 - 00:04:54	0.78
00:10:41 - 00:10:41	0.77
00:11:09 - 00:11:11	0.77
00:01:59 - 00:02:00	0.77
00:07:39 - 00:07:41	0.76
00:05:01 - 00:05:02	0.76
00:00:34 - 00:00:35	0.76
00:01:29 - 00:01:30	0.76
00:04:50 - 00:04:50	0.75
00:00:38 - 00:00:38	0.75
00:05:43 - 00:05:45	0.75
00:04:06 - 00:04:07	0.75

Video files

1. 0egEFZq2Y28.mp4 0%
2. 0f-QmF0Avvc.mp4 0%
3. 1KA8kQS2PTE.mp4 0%
4. 1QHa5yz_qV0.mp4 0%
5. 2buhK32I9k0.mp4 0%
6. 5hQapiCQJqs.mp4 0%
7. sample.mp4 100%

Filters

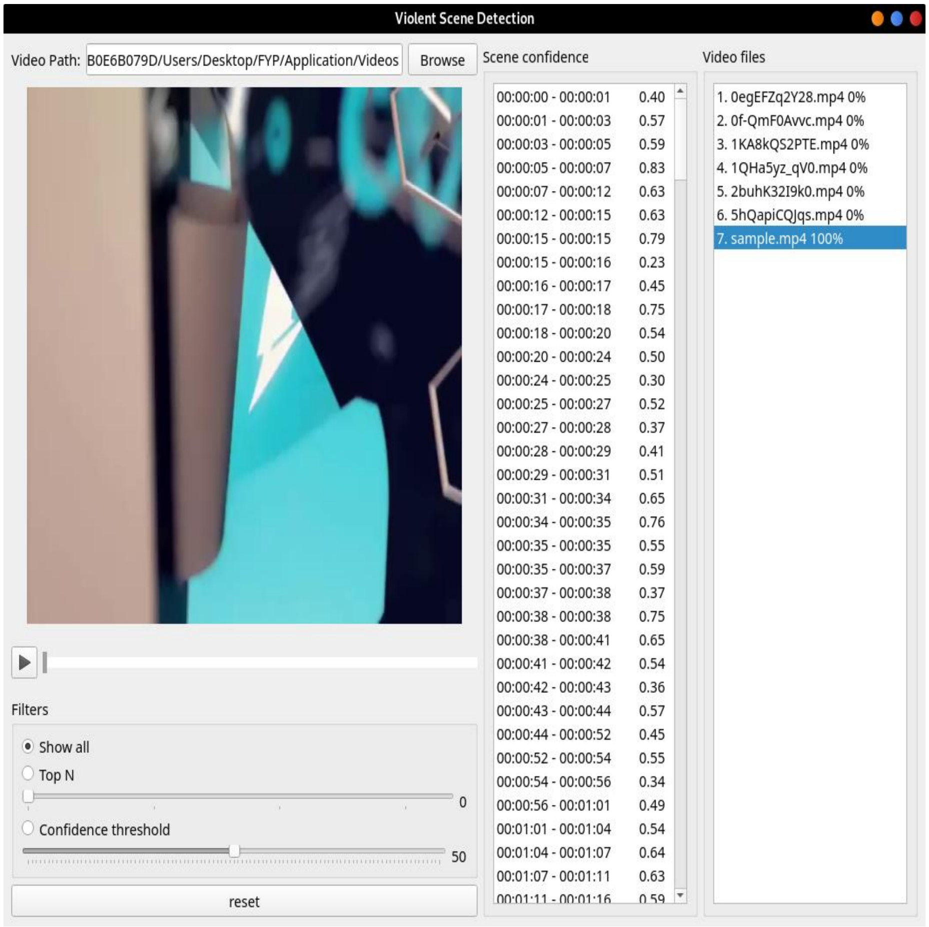
○ Show all

● Top N 100

○ Confidence threshold 78

Figure 4.5: Filter by Top N Violent Scenes

As shown in Figure 4.5, if user wants to view the top N most violent scenes, user can select top N radio button and select the value using the slider below the radio button.



Video Path: B0E6B079D/Users/Desktop/FYP/Application/Videos

Scene confidence

00:00:00 - 00:00:01	0.40
00:00:01 - 00:00:03	0.57
00:00:03 - 00:00:05	0.59
00:00:05 - 00:00:07	0.83
00:00:07 - 00:00:12	0.63
00:00:12 - 00:00:15	0.63
00:00:15 - 00:00:15	0.79
00:00:15 - 00:00:16	0.23
00:00:16 - 00:00:17	0.45
00:00:17 - 00:00:18	0.75
00:00:18 - 00:00:20	0.54
00:00:20 - 00:00:24	0.50
00:00:24 - 00:00:25	0.30
00:00:25 - 00:00:27	0.52
00:00:27 - 00:00:28	0.37
00:00:28 - 00:00:29	0.41
00:00:29 - 00:00:31	0.51
00:00:31 - 00:00:34	0.65
00:00:34 - 00:00:35	0.76
00:00:35 - 00:00:35	0.55
00:00:35 - 00:00:37	0.59
00:00:37 - 00:00:38	0.37
00:00:38 - 00:00:38	0.75
00:00:38 - 00:00:41	0.65
00:00:41 - 00:00:42	0.54
00:00:42 - 00:00:43	0.36
00:00:43 - 00:00:44	0.57
00:00:44 - 00:00:52	0.45
00:00:52 - 00:00:54	0.55
00:00:54 - 00:00:56	0.34
00:00:56 - 00:01:01	0.49
00:01:01 - 00:01:04	0.54
00:01:04 - 00:01:07	0.64
00:01:07 - 00:01:11	0.63
00:01:11 - 00:01:16	0.59

Video files

1. 0egEFZq2Y28.mp4 0%
2. 0f-QmF0Avvc.mp4 0%
3. 1KA8kQS2PTE.mp4 0%
4. 1QHa5yz_qV0.mp4 0%
5. 2buhK32I9k0.mp4 0%
6. 5hQapiCQJqs.mp4 0%
7. sample.mp4 100%

Filters

● Show all

○ Top N 0

○ Confidence threshold 50

Figure 4.6: Filtering by time

As shown in Figure 4.6, if user wishes to view all the prediction scores according to time, users can select the show all option to list all the scenes by ascending order with their prediction.

Chapter 5: Evaluation

5.1 Dataset Evaluation

The train set of VSD2014 dataset consist of 24 Hollywood movies. These movies have varying amount of violence ranging from very violent to almost no violent. In total there are 392 annotated violent segments. Each annotated violent segment has an average of 1470 frames which is around 59 seconds of violent actions. Most of the violence are physical violence but some violence is shown indirectly through facial expression instead of the violent action itself. Although there is a wide variety of violence, generally they can be grouped into these categories, images are included in appendix for reference:

- Rioting
- Falling
- Fighting with or without weapons
- Shooting with guns
- Burning with fire
- Shooting with magical projectiles
- Explosion
- Falling of rubbles
- Gory scenes with dead people or large amount of blood
- Combination on multiple violent elements
- Violence that has to be guessed from actor's facial expression

For testing, there are 2 tasks used to evaluate the performance of violent scene detection. First task will be the main task where the two stream CNN will be tested on 5 Hollywood movies. The generalisation task will be testing on 87 web videos to determine the generalisation capability on different types of videos. In order to compare the performance of two stream CNN for violent scene detection, the standard evaluation metrics in VSD2014 such as MAP100, MAP2014, precision and recall are calculated.

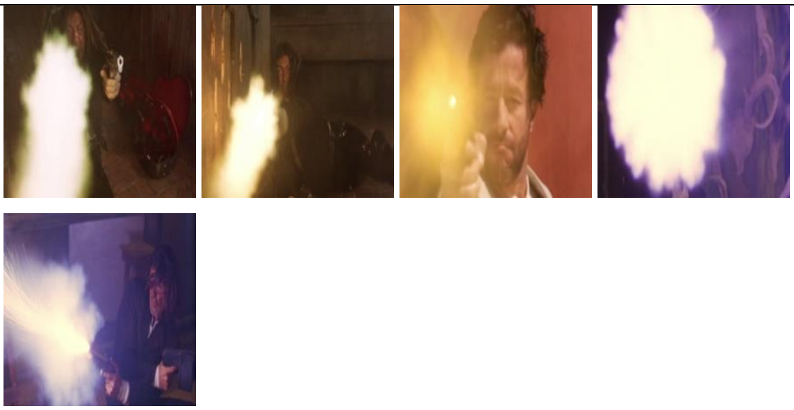
5.2 Test Results

5.2.1 Main Task Results

	Precision	Recall	MAP100	MAP2014
Spatial stream CNN	0.40	0.58	0.74	0.41
Temporal stream CNN	0.47	0.53	0.71	0.29
Two stream CNN	0.46	0.57	0.79	0.56

Table 5.1: Main Task Results

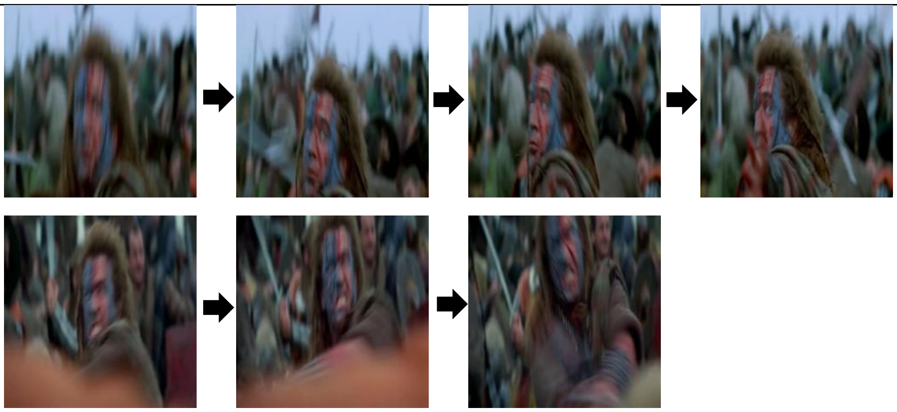
Spatial stream CNN: Based on samples from top 100 most violent prediction, spatial stream CNN was able to detect violent scenes as seen in Table 5.2 such as presence of fire, gunshot, explosion and presence of blood. Spatial stream CNN was also able to detect violence that are more motion-based such as fighting and falling. One observable feature that is common among detected motion-based violent scenes is that they generally contain significant amount of blur regions. For still images, motion-based violence such as fighting or falling often results in blur regions, spatial stream CNN could have associated presence of large amount of blur region with violence.

Violence category	Sample violent images
Presence of fire	
Gunshot	
Explosion	

Presence of blood	
Motion based violence	

Table 5.2: Categories of Violent Scenes Detected by Spatial Stream CNN

Temporal stream CNN: Based on samples from top 100 most violent temporal stream CNN prediction, most violent scene detected by temporal stream CNN generally have large changes in motion across multiple frames either from objects or camera movement. Table 5.3 shows samples from top 100 detected violent scene.

Scene description	Sample violent scenes
Sword fighting	

Man strangling women	
Fist fighting	
Man harassing women	

Table 5.3: Violent Scenes Predicted by Temporal Stream CNN

As seen in Table 5.1, the MAP2014 for temporal stream CNN is significantly lower compared to spatial stream CNN despite both having quite similar MAP100. This is because for MAP100 in temporal stream CNN the top 100 most violent scenes were dominated by fight scenes from the movie Brave Heart. However, after MAP2014 removed repeating detection that were from same violent scenes, detected scenes from Brave Heart reduced from 66 to 36 which is why the MAP2014 is lower. As for why MAP2014 for temporal stream CNN is low, from observing the top 100 detected scenes in MAP2014, it seems that temporal stream CNN has difficulty differentiating between violent scene with large movement from non-violent scenes with large movement as shown in Table 5.4, this could be because unlike in still images that contain multiple visual cues such as colour, depth, contrast and possible motion from blurry region to detect violent scenes, optical flow frames only contain the motion information which could be insufficient in some cases for differentiating between violent and non-violent scenes that has large motions.

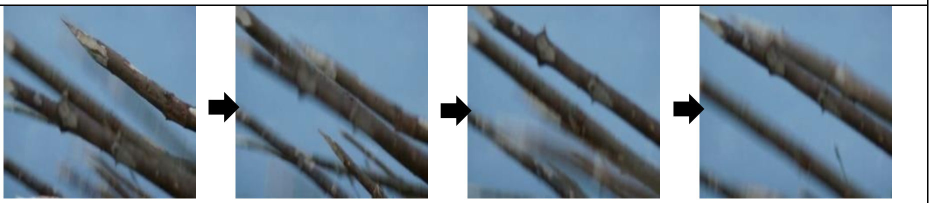
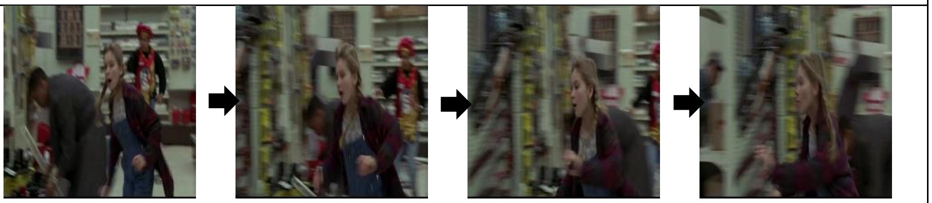
Scene description	False positive scenes
Waving wooden spears	
Girl running	
Boy running	
Man jumping into his car	
Man running	

Table 5.4: False Positive Scenes Detected by Temporal Stream CNN

Two stream CNN: For main task violent scene detection, merging both spatial stream CNN and temporal stream CNN increases the performance of MAP2014 significantly. This shows that both spatial and temporal modalities are complementary and effective for detection of violent scenes. Table 5.5 shows that two stream CNN was able to perform quite well for violent scene detection in movies compared to other participants.

Participant	Precision	Recall	MAP100	MAP2014
FUDAN (Dai et al., 2014)	0.41	0.72	0.72	0.63
Two stream CNN	**0.46**	**0.57**	**0.79**	**0.56**
NII-UIT (Lam et al., 2014)	0.17	1.00	0.77	0.55
FAR (Sjöberg et al., 2014)	0.28	0.71	0.57	0.45
MIC-TJU (Zhang et al., 2014)	0.17	0.98	0.63	0.44
RECOD (Avila et al., 2014)	0.33	0.69	0.49	0.37
VIVOLAB (Castán et al., 2014)	0.38	0.58	0.38	0.17
TUB-IRML (Acar et al., 2014)	0.31	0.17	0.40	0.17
MTMDCC (Bruno et al., 2014)	0.15	0.24	0.16	0.02

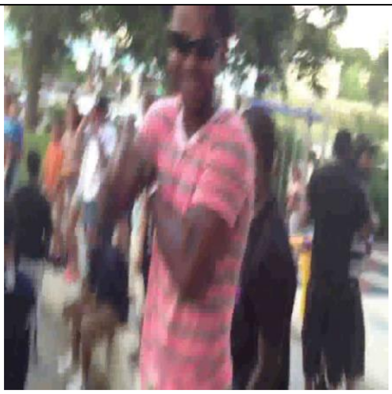
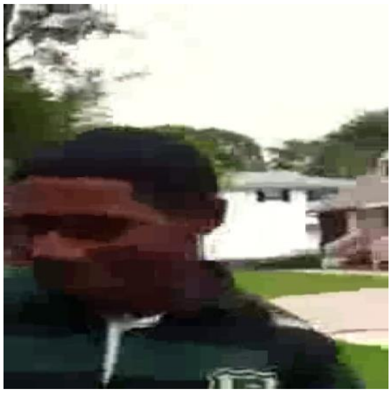
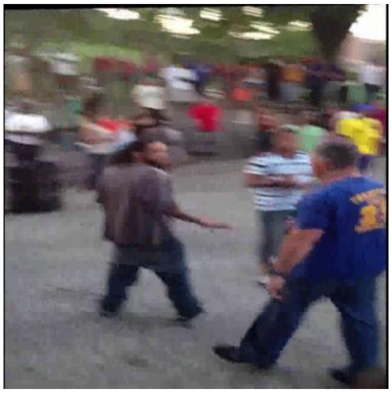
Table 5.5: Main Task Performance Figure for Each Participating Team

5.2.2 Generalisation Task Results

	Precision	Recall	MAP100	MAP2014
Spatial stream CNN	0.32	0.93	0.67	0.19
Temporal stream CNN	0.47	0.70	0.68	0.42
Two stream CNN	0.37	0.91	0.75	0.43

Table 5.6: Generalisation Task Results

Spatial stream CNN: For generalisation task, the performance of MAP2014 for spatial stream CNN performed significantly worst compared to in main task. As seen from Table 5.6 spatial stream CNN has high recall and low precision, this shows that spatial stream CCN suffer from large amount of false positive detection for web videos. The reason for this could be because features learned from Hollywood scenes are unable to be generalised to web videos. One difference between Hollywood movies and web videos is that most web videos in the generalisation dataset have poor video quality. Web videos especially user recorded videos often have large camera movements or low camera resolution which results in the recorded footage to have large amount of blur. Since spatial stream CNN associates blur regions with violence, non-violent scene that contain large amount of blur region would often be incorrectly detected as violent scenes. This can be seen in Table 5.7 that shows samples from top 100 detected scenes where non-violent scenes with close to no blur regions are correctly classified while non-violent scenes with large amount of blur regions would be incorrectly classified as violent scenes.

Least violent images	False positive images
     	  

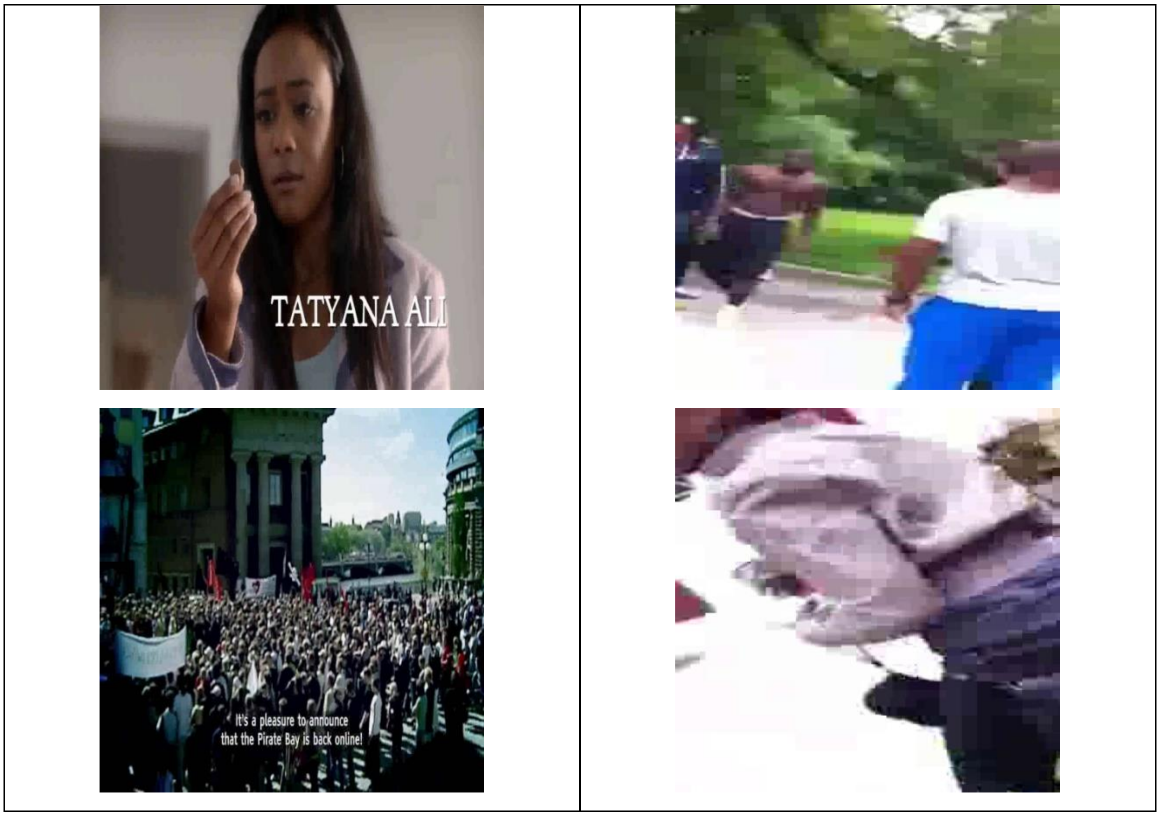

TATYANA ALI
It's a pleasure to announce that the Pirate Bay is back online!

Table 5.7: Comparison of Least Violent Images with False Positive Images

Two stream CNN: For generalisation task violent scene detection, the performance of MAP2014 for two stream CNN is poorer compared to main task. This is because the spatial stream CNN performed poorly on web videos which caused the improvement from merging both modalities to become less significant. As shown in Table 5.8 compared to other participants two stream CNN performed the worst on generalisation task.

Participant	Precision	Recall	MAP100	MAP2014
FAR (Sjöberg et al., 2014)	0.49	0.85	0.86	0.66
RECOD (Avila et al., 2014)	0.48	0.88	0.86	0.61
FUDAN (Dai et al., 2014)	0.59	0.43	0.71	0.60
MIC-TJU (Zhang et al., 2014)	0.44	0.97	0.55	0.56
TUB-IRML (Acar et al., 2014)	0.63	0.25	0.58	0.51
VIVOLAB (Castán et al., 2014)	0.51	0.33	0.56	0.43
Two stream CNN	**0.37**	**0.91**	**0.75**	**0.43**

Table 5.8: Generalisation Task Performance figure for Each Participating Team

Chapter 6: Conclusion

In this project, an application is built to detect and localise violent scene in videos. This application uses two stream CNN to extract spatial and temporal features from videos. Two stream CNN is trained on VSD2014 dataset and evaluated on Hollywood movies and web videos. According to the results, spatial and temporal modality has proven to be complementary and useful for violent scene detection. As for the performance on different type of videos, two stream CNN seems to perform considerably well on Hollywood movies. However, violent scene detection on web videos especially user generated videos seem to perform much poorer because the features learnt in spatial stream CNN from violent movie scenes are not able to be generalised to web videos. Currently, because the number of samples in VSD2014 is considerably small for training deep learning architectures, the spatial stream CNN is trained as a feature extractor to prevent overfitting during training, However, in future when there is more training data, spatial stream CNN can be retrained to learn more violent related features which could improve the performance for detection in web videos. Other modification to current two stream architecture can also be considered such as using more powerful model for temporal stream CNN and different fusion methods.

REFERENCES

1. Straits Times 2018, Malaysia's Internet penetration is now 85.7 per cent. Available from: <https://www.nst.com.my/business/2018/03/346978/malaysias-internet-penetration-now-857-cent>. [19 March 2018]

2. Insiviacom. 2017. Insivia Marketing + Web Design. [Online]. [27 March 2019]. Available from: https://www.insivia.com/27-video-stats-2017/

3. Wordstreamcom. 2019. Wordstreamcom. [Online]. [27 March 2019]. Available from: https://www.wordstream.com/blog/ws/2017/03/08/video-marketing-statistics. [12 March 2019]

4. Schedl, Markus & Sjöberg, M & Mironică, Ionuţ & Ionescu, Bogdan & Lam, Vu & Jiang, Y.-G & Demarty, Claire-Hélène. (2015). VSD2014: A dataset for violent scenes detection in Hollywood movies and web videos. Proceedings - International Workshop on Content-Based Multimedia Indexing. 2015. 10.1109/CBMI.2015.7153604.

5. Dai, Q., Zhao, R.W., Wu, Z., Wang, X., Gu, Z., Wu, W. and Jiang, Y.G., 2015, September. Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In MediaEval.

6. Simonyan, K. and Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).

7. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L., 2016, October. Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision (pp. 20-36). Springer, Cham.

8. Dai, Q., Wu, Z., Jiang, Y.G., Xue, X. and Tang, J., 2014, October. Fudan-NJUST at MediaEval 2014: Violent Scenes Detection Using Deep Neural Networks. In MediaEval.

9. Sjöberg, M., Mironica, I., Schedl, M. and Ionescu, B., 2014, October. FAR at MediaEval 2014 Violent Scenes Detection: A Concept-based Fusion Approach. In MediaEval.

10. Lam, V., Le, D.D., Phan, S., Shin'ichi Satoh and Duong, D.A., 2014, October. NII-UIT at MediaEval 2014 Violent Scenes Detection Affect Task. In *MediaEval*.

11. Zhang, B., Yi, Y., Wang, H. and Yu, J., 2014, October. MIC-TJU at MediaEval Violent Scenes Detection (VSD) 2014. In *MediaEval*.

12. Avila, S., Moreira, D., Perez, M., Moraes, D., Cota, I., Testoni, V., Valle, E., Goldenstein, S. and Rocha, A., 2014. RECOD at MediaEval 2014: Violent scenes detection task. In *CEUR Workshop Proceedings*. CEUR-WS.

13. Acar, E. and Albayrak, S., 2014, October. TUB-IRML at MediaEval 2014 Violent Scenes Detection Task: Violence Modeling through Feature Space Partitioning. In *MediaEval*.

14. Castán, D., Rodríguez, M., Ortega, A., Orrite, C. and Lleida, E., 2014, October. ViVoLab and CVLab-MediaEval 2014: Violent Scenes Detection Affect Task. In *MediaEval*.

15. do Nascimento Teixeira, B., 2014, October. Mtm at mediaeval 2014 violence detection task. In Working Notes Proceedings of the MediaEval Workshop.

16. Castán, D., Rodríguez, M., Ortega, A., Orrite, C. and Lleida, E., 2014, October. ViVoLab and CVLab-MediaEval 2014: Violent Scenes Detection Affect Task. In *MediaEval*.

APPENDIX

Samples of Train Set Data

Violence category	Violent image		
Explosion	70878 ori: 70723 – 81233 offset: 70723 – 81233 gap: 5 offset: 0 ext: 0		
Shooting with guns	41748 ori: 41599 – 44698 offset: 41599 – 44698 gap: 1 offset: 0 ext: 0	BRAUN E224 E225 23575 ori: 23422 – 24265 offset: 23422 – 24265 gap: 1 offset: 0 ext: 0	121935 ori: 121921 – 124556 offset: 121921 – 124556 gap: 14 offset: 0 ext: 0
Rioting	35272 ori: 35272 – 35308 offset: 35272 – 35308 gap: 5 offset: 0 ext: 0	108489 ori: 108435 – 108787 offset: 108435 – 108787 gap: 1 offset: 0 ext: 0	79011 ori: 78901 – 79050 offset: 78901 – 79050 gap: 5 offset: 0 ext: 0
Falling of rubble	10140 ori: 9968 – 10235 offset: 9968 – 10235 gap: 1 offset: 0 ext: 500	82111 ori: 82012 – 82409 offset: 82012 – 82409 gap: 1 offset: 0 ext: 0	31210 ori: 31157 – 31498 offset: 31157 – 31498 gap: 1 offset: 0 ext: 500

APPENDIX

Fighting	
Burning with fire	
Falling	
Gory scenes	
Shooting with magic projectiles	

Violence guessed from facial expression	88241 ori: 88214 – 88540 offset: 88214 – 88540 gop: 1 offset: 0 ext: 0	111235 ori: 111220 – 111786 offset: 111220 – 111786 gop: 5 offset: 0 ext: 0	7887 ori: 7873 – 7902 offset: 7873 – 7902 gop: 1 offset: 0 ext: 0
Combination of multiple violent element	85879 ori: 85859 – 91709 offset: 85859 – 91709 gop: 1 offset: 0 ext: 0	ori: 2070 – 2388 offset: 2070 – 2388 gop: 1 offset: 0 ext: 0	

POSTER

Violent Scene Detection in Videos

By: Yew Kynn Man
Supervisor: Dr. Tan Hung Khoon

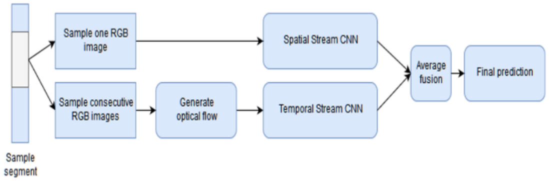
Introduction

Consuming media content online in particular has become a big part of our daily activities which has brought rise to video sharing and streaming websites. Because of the sheer amount of video footage uploaded each day, manually filtering violent content could be ineffective. This project aims to build a violent scene detection application that can detect and localise violent scenes in videos using deep learning so that significant amount of detection task can be handled by machine.

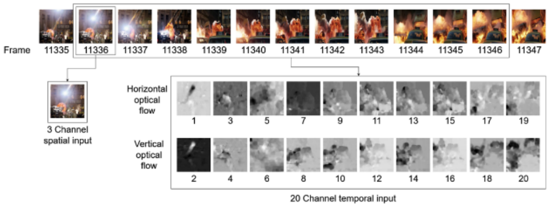
Objective

- To develop a violent detection system that is able to detect and localise violent scenes in videos.
- To build a desktop application that allows users to view and filter violence scenes .

Methodology



Modality



Evaluation Results

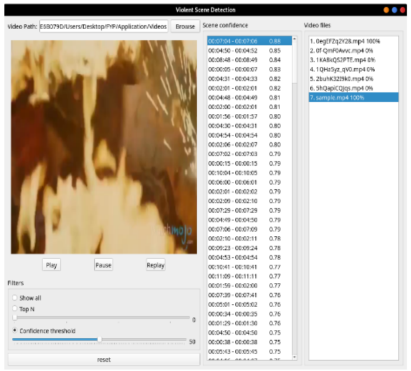
	Precision	Recall	MAP100	MAP2014
Spatial stream CNN	0.40	0.58	0.74	0.41
Temporal stream CNN	0.47	0.53	0.71	0.29
Two stream CNN	0.46	0.57	0.79	0.56

Table 5.1: Main task results

	Precision	Recall	MAP100	MAP2014
Spatial stream CNN	0.32	0.93	0.67	0.19
Temporal stream CNN	0.47	0.70	0.68	0.42
Two stream CNN	0.37	0.91	0.75	0.43

Table 5.5 Generalisation task results

User Interface



Video Path: E6B079D/Users/Desktop/FYP/Application/Videos

Scene confidence

00:07:04 - 00:07:06	0.88
00:04:50 - 00:04:52	0.85
00:08:48 - 00:08:49	0.84
00:00:05 - 00:00:07	0.83
00:04:31 - 00:04:33	0.82
00:02:01 - 00:02:01	0.82
00:04:48 - 00:04:49	0.81
00:02:00 - 00:02:01	0.81
00:01:56 - 00:01:57	0.80
00:04:30 - 00:04:31	0.80
00:04:54 - 00:04:54	0.80
00:02:06 - 00:02:07	0.80
00:07:02 - 00:07:03	0.79
00:00:15 - 00:00:15	0.79
00:10:04 - 00:10:05	0.79
00:06:00 - 00:06:01	0.79
00:02:01 - 00:02:02	0.79
00:02:09 - 00:02:10	0.79
00:07:29 - 00:07:29	0.79
00:04:49 - 00:04:50	0.79
00:07:06 - 00:07:09	0.79
00:02:10 - 00:02:11	0.78
00:09:23 - 00:09:24	0.78
00:04:53 - 00:04:54	0.78
00:10:41 - 00:10:41	0.77
00:11:09 - 00:11:11	0.77
00:01:59 - 00:02:00	0.77
00:07:39 - 00:07:41	0.76
00:05:01 - 00:05:02	0.76
00:00:34 - 00:00:35	0.76
00:01:29 - 00:01:30	0.76
00:04:50 - 00:04:50	0.75
00:00:38 - 00:00:38	0.75
00:05:43 - 00:05:45	0.75

Video files

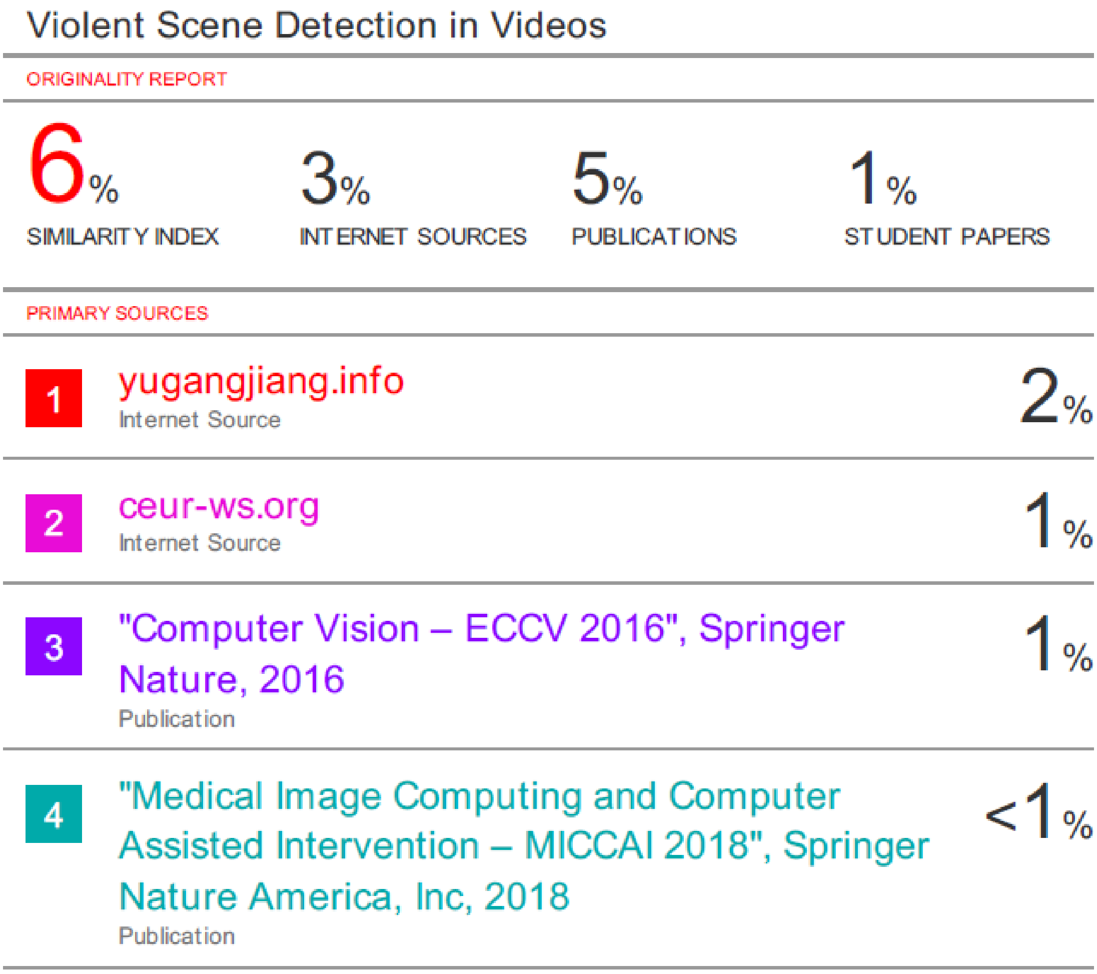
1. 0egEFZq2Y28.mp4 100%
2. 0f-QmF0Awc.mp4 0%
3. 1KABkQS2PTE.mp4 0%
4. 1QHa5yz_qV0.mp4 0%
5. 2buhK32I9k0.mp4 0%
6. 5hQapiCQJqs.mp4 0%
7. sample.mp4 100%

Conclusion

- Spatial and temporal modality are complimentary for violent scene detection.
- Two stream CNN is able to detect violent scenes in Hollywood movies but features learnt in spatial stream CNN is unable to be generalised to web videos

PLAGARISM CHECK

RESULT



Violent Scene Detection in Videos

ORIGINALITY REPORT

6%	3%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	yugangjiang.info Internet Source	2%
2	ceur-ws.org Internet Source	1%
3	"Computer Vision – ECCV 2016", Springer Nature, 2016 Publication	1%
4	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2018", Springer Nature America, Inc, 2018 Publication	<1%