T.C.
MARMARA UNIVERSITY
FACULTY of ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

CSE 4197 Engineering Project I Analysis and Design Document

Title of the Project

Real-Time Violence Detection

Group Members

150115031-Alperen BAYAR
150115052-Nuri YILDIZ
150115057-Büşra YAĞCI

Supervised by
Prof. Çiğdem EROĞLU ERDEM

Table of Contents

1 INTRODUCTION

This section describes the problem our project aims to solve, the scope and aims of our project, the benefits of our project, and explanations of some of the terms and abbreviations we use throughout this article.

1.1 Problem Description and Motivation

This part, includes the problem we have chosen to solve, the explanations of this problem and our motivation.

1.1.1 Problem Statement

Violence is a behavior that disrupts the peace of society and endangers its security, and states and private institutions take various measures against it. The term "violence" referred to in this article refers to people fighting with each other and damaging their environment. Even today the most commonly used method to prevent and respond to such violence is the presence of security forces. Nevertheless, with the development of technology, security cameras also play a major role in the fight against violence and have become an indispensable part of our lives. They work 24/7 for security.

There are thousands or millions of CCTV and some of them are always monitored also by people. Because threats like violence or abnormal human activity must be detected immediately in places like military zones, energy plants, etc. At present human surveillance in a few certain places like military zones or energy plants costs very much in terms of manpower and financial reasons, this is why it seems impossible to generalize surveillance to all CCTVs in the world. So, most of the CCTVs are only recording. These records are used to identify criminals and solve cases, in other words, to help security forces. Therefore, immediate action cannot be taken in cases of violence in systems without human surveillance. So, real-time surveillance either cost very much or it is nearly impossible. In Figure 1 and Figure 2, we see the security camera footage that contains images of violence.

Figure 1: Two women fighting in the street.

Figure 2: A man is kicking a parked car.

1.1.2 Problem Description and Motivation

Due to increasing violence, citizens and government are trying different security methods. Surveillance cameras or CCTVs are one of the methods in addition to night watchmen and police.

CCTVs can be used to deter crime or violence besides helping security forces. Because monitoring the cameras allows immediate action against violence. It has been found that CCTV surveillance contributes 16% in crime prevention [1]. This ratio can be reached 51% in the car parks. As can be seen, CCTV surveillance may have a deterrent effect as much as security guards against crime.

The number of cameras must be sufficient to ensure adequate prevention for the crime. London is currently one of the cities with the most CCTVs. It is estimated 500.000 CCTV cameras in London and if an observer would be responsible for 8-10 cameras, only the monthly observer salary cost would be $ 1.25 billion [2]. The estimated number for the world is 25 million [1]. This number means that, as much as the number of soldiers in Russia, CCTV observers are needed to ensure minimum quality surveillance on a global scale. In Figure 3 we see an observer watching the security cameras.

.

Figure 3: An example of a CCTV observer

Given these high costs and the failure of human surveillance, it is possible to avoid economic losses and loss of lives by using autonomous systems. Advanced computer vision technologies can be used for the dissemination of these autonomous systems. Considering the CCTV records collected to date, it is seen that there is a sufficient dataset for the deep learning to be done for the detection of violence.

1.2 Scope of the Project

The system will be installed on computers where cameras are monitored in areas where violence is desired. And the data from the computers where the system is installed (no matter the day or night view) will be processed. The transferred data will be processed with intervals. And it will be determined if there is an anomaly in that period. If there is an anomaly, the type of anomaly is detected. Two types of anomalies are detected. These are vandalism and human violence. Contact information will be taken from authorized persons during the installation of the system and if any violence is detected, they will be warned by an SMS with the identity of the related camera.

- The notification will be made in writing only.
- The whole event will be in the field of view of the camera (for example, one of the two people is completely visible, the other side can't be detected, such as entering and leaving the camera angle).
- The system will only detect the following conditions:
    - Is there a fight between people?
    - Is there any vandalism?
- It will be treated in the same way as the force used by the security forces (violence by the security forces will not be in a separate segment).
- The person to be notified must have access to a device from which they can receive SMS.
- The computer on which the system is installed must be connected to the Internet to use the SMS notification feature.
- Authorities will decide whether or not to intervene.
- Non-contact anomaly events are not included. For example:
    - Gun wounding.
    - Throw something.
    - Verbal fight.

1.3 Aims of the Project

Aims of the project are listed below.

Real-time violence detection:

Automatic monitoring of security cameras, detecting violence with 90% accuracy (close to state of the art) and informing security forces with a maximum delay of 1 minute.

Sub-goals of the project:

As mentioned in section 1.1.2, non-autonomous classical surveillance systems are entirely dependent on manpower. In these old systems, one observer is expected to be responsible for 8-10 cameras [2]. With the autonomous surveillance system, the number of cameras that the observer is responsible for can be increased. In this way, the required manpower can be reduced.

With the above mentioned target, we anticipate that the cost of surveillance will decrease as the number of observers needed is reduced.

The following aims are planned for the components of the whole system.

Violence detection

90% (close to state of the art) detection of the presence or absence of violence in the monitored camera (UCF Anomaly Detection Dataset.).

Violence Recognition

90%(close to state of the art) successful recognition of violence (fighting) among people or vandalism (environmental damage) (UCF Anomaly Detection Dataset.).

Real-time notification

As soon as the violence is detected and recognized, to send information to the authorized person as a notification (SMS).

1.4 Success Factors and Benefits

In the following section success factors and benefits of the project are listed.

1.4.1 Success Factors

- Violence detection will determine whether there is any violence in the data coming from the camera. The accuracy score of determining violence is 90% (close to state of the art) on our dataset which is based on the UCF Crime dataset.
- Violence recognition is used to classify action is fighting or vandalism. The accuracy score of the recognition event is 90% (close to state of the art) on our dataset, which is based on the UCF Crime dataset.
- The real-time violence notification system will send the authorized person the type of violence and the identity of the camera via SMS to the authorized person with a delay of up to 1 minutes.

1.4.2 Benefits

The benefits of this project can be listed as below:

- State and citizen spending on security will be reduced.
- The manpower spent on security will be reduced. Thus, the manpower saved can be used for other purposes.
- Human errors will be minimized.
- People will feel safer.
- Because of late intervention, loss of life will be prevented.
- We can distinguish between vandalism and fighting, so the authorities will be able to work with different priorities for different events. (For example, intervening in a fight will be more important than vandalism.)

1.5 Definitions, Acronyms, and Abbreviations

Definitions are listed in Table 1.

Accuracy Score	The ratio of the number of correct predictions to the total number of test samples.
Google Colab	A tool and environment created by Google to help machine learning education and research.
F1 Score	A statistical measure to rate performance.
Feature Extraction	A process of dimensionality reduction, reducing the amount of data that must be processed.
Fighting	All kinds of fights in which the body is used.
Precision Score	How often prediction is correct when the model predicts positive.
Recall	The number of correct positive results divided by the number of all relevant samples.
Vandalism	The crime of intentionally damaging things in public places.
Violence Detection	To determine if there is violence using a trained model.
Violence Recognition	If there is violence, to determine whether the type of violence is fighting or vandalism.

Table 1: List of definitions.

Abbreviations are listed in Table 2.

ACM DL	Association for Computing Machinery Digital Library
C3D	A convolutional 3D model for feature extraction from videos
CNN	Convolutional Neural Network
CVPR	Conference on Computer Vision and Pattern Recognition
CCTV	Closed Circuit Television: security cameras, surveillance cameras
NTU	Nanyang Technological University
ICCV	International Conference on Computer Vision
IEEE	Institute of Electrical and Electronics Engineers
LSTM	Long short-term memory
UCF	University of Central Florida

Table 2: List of abbreviations.

2 RELATED WORK

Violence detection with surveillance cameras is a field of computer vision. Conference on Computer Vision and Pattern Recognition (CVPR) and International Conference on Computer Vision (ICCV) are among the most important conferences to reach state-of-the-art violence detection methods. There are a lot of approaches from traditional methods to very new techniques. Machine learning is considered a traditional method and most of the new articles contain methods that use deep learning.

- A Review on State-of-the-Art Violence Detection Techniques

This article is a good choice for the beginning of this research. It mentions a research methodology to access all articles on the subject that need to be reviewed using the "regular expression" with all relevant terms. In this article, using this methodology, a total of 2853 articles were reached from 5 different sources: IEEE Explore, Science Direct, ACM DL, Springer, and Google Scholar. Based on the years of publication and the technologies it uses, it has been eliminated up to 29 articles. These 29 articles have been examined under three main headings: "Violence Detection Using Machine Learning Techniques", "Violence Detection Using SVM" and "Violence Detection Using Deep Learning". All research begins by collecting data sets. UCF-101 and Hockey data sets are the most popular used for the detection of violence. After the data is cleared, feature extraction is performed using different methods. C3D, ViF, Motion Blobs methods are some of the methods used for feature extraction. Once these features are normalized, they are used as input to the model trained with the above-mentioned data sets to determine whether there is violence [3]. Figure 4 illustrates these steps.

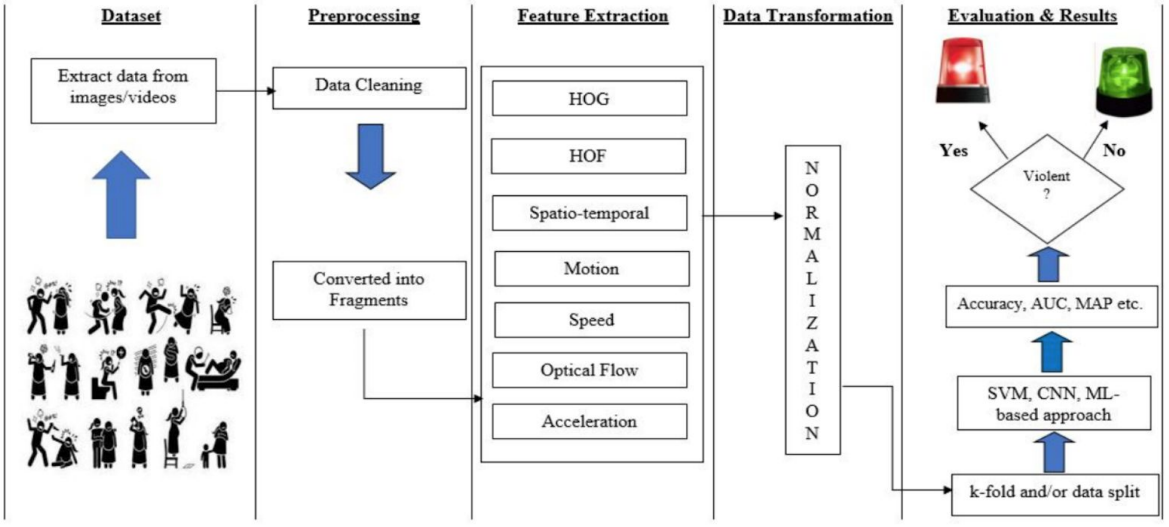

Figure 4: Main steps for violence detection techniques [3].

These articles provide the most up-to-date information about the properties of the data set, the properties of the images in the data set, the types of feature extraction methods according to the field of use and about to determine the number of frames [3].

- Violent Activity Recognition Without Decoding

In this work, violent activity recognition was performed. Motion vectors method was used. The Motion vectors were obtained from compressed videos. And their features were extracted for both each frame and between frames. Then these are given to the Region Motion Vector (RMV) descriptor.

SVM is used to classify the RMV and detection violence. VVAR10 dataset has 296 positive compressed videos and 277 negative compressed videos. The sources of the videos are YouTube, UCF50, UCF sports, and HMDB51. VVAR10 dataset is used for testing. And the system achieves 96.1% success in detecting violence. The outstanding feature of this work compared to other approaches in the literature is that it is quite fast because it uses compressed videos [4].

- Real-Time Violence Detection

This method used both CNN and SVM for violence detection. The model has three-part: feature extraction, SVM model, and label fusion.

CNN extracts two features. First one, single frames to extract the features of appearance and the second one difference of ordered frames to separate the features of motion. Then, SVM used these features like a classifier.

The label fusion method is used in this method. It compounds the motion information and the appearance information is used for attaining the result of detecting violence. Hockey Fight and Violent Crowd datasets are used in the experiments. The results of this method are better than the existing methods in many realistic scenes in terms of accuracy [5].

- Violence Detection Using Spatiotemporal Features With 3D CNN [6]

This method used a triple-staged structure for violence detection. A light-weight CNN model detected persons to reduce the cost of processing unusable frames.

3D CNN extracted spatiotemporal features of these remaining 16 frames. Softmax classifier received them as input. After the violence detection, related units are informed. Used datasets are the Violent Crowd, Hockey and Violence in Movies. Figure 5 shows some sample screenshots from videos of these datasets. The success rate can reach up to 98%. The results of the model are better than state-of-the-art methods like Hough, Forest, VIF, SVM, and 2D CNN, SHOT and others due to classification success metrics which are accuracy, precision[6].

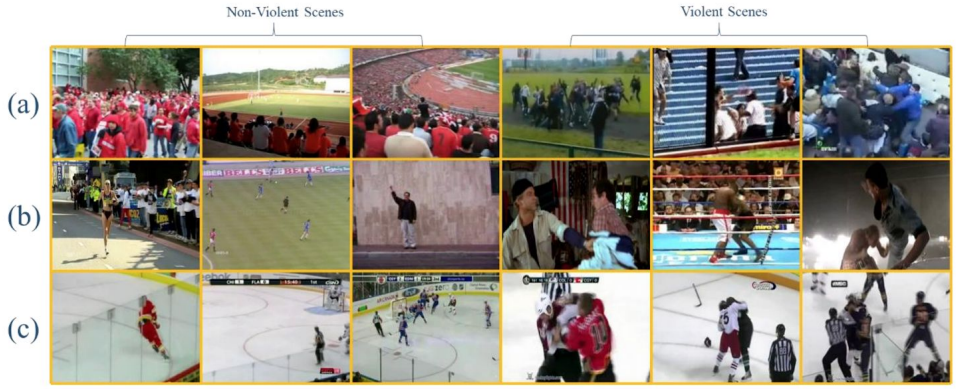

Figure 5: Sample video frames randomly selected from: (a) violent crowd, (b) violence in movies, (c) hockey fight [6].

- Real-world Anomaly Detection in Surveillance Videos

This method detected anomalies by used both violent and non-violent videos. Instead of calculating all segments containing violent, labeled videos are used because it is very costly to calculate all segments.

A dataset created which is consisting of anomalies to verify this approach. Dataset consists of 1900 videos that have 128 hours total was created by gathering different datasets, 950 normal and 950 violent. This article uses a pre-trained 3D ConvNet for C3D feature extraction. Secondly, to determine is there a violence or not, Deep Multiple Instance Learning (MIL) model is used [7]. Figure 6 shows the flow diagram of the anomaly detection approach for this work.

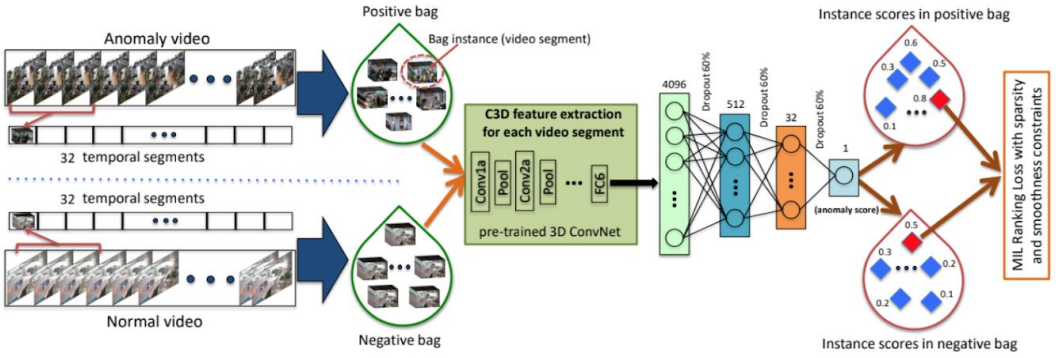

Figure 6: The flow diagram of the proposed anomaly detection approach. Given the positive (containing anomaly somewhere) and negative (containing no anomaly) videos [7].

In this work, each of the videos is separated into multiple segments. Each segment corresponds an instance. After features are obtained, A previously trained neural network model is trained which contains a loss function. The loss function evaluates the ranking loss among samples in bags.

Even though this two- steps approach can be slow for real-time violence detection, published their open- source project is very helpful. When anomaly detection is performed using this data set, it has a better success rate than other approaches in the literature.

- Learning Spatiotemporal Features with 3D Convolutional Networks

A simple approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks trained on a large scale supervised video dataset is proposed in this work. A homogeneous architecture with small 3 × 3 × 3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets [8]. Visualization of 3D and 2D convolution operations as shown as Figure 7.

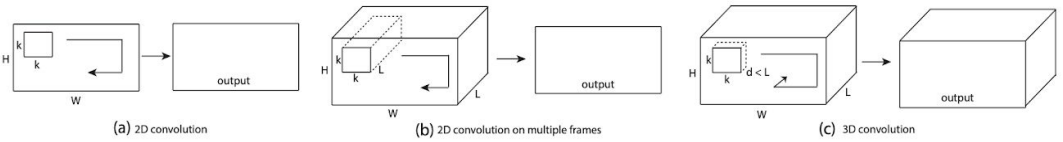

Figure 7: Visualization of 3D and 2D convolution operations [8].

A convolutional network is trained on a labeled video dataset to obtain spatiotemporal features. "A homogeneous architecture with small 3 × 3 × 3 convolution kernels in all layers is among the best-performing architectures for 3D ConvNets " [8]. The C3D features are quite convenient for classification. Using these properties, a classifier method achieved higher success rates than other works in the literature. Results are achieved 52.8% success rate on UCF101 dataset. Action recognition results on UCF 101 are shown in Table 3.

Method	Accuracy (%)
Imaganet + linear SVM	68.8
iDT w/ BoW + linear SVM	76.2
Deep networks [15]	65.4
Spatial stream network [16]	72.6
LRCN [17]	71.1
LSTM composite model [18]	75.8
C3D (1 net) + linear SVM	82.3
C3D (3 nets) + linear SVM	85.2
iDT w/ Fisher vector [19]	87.9
Temporal stream network [16]	83.7
Two-stream networks [16]	88.0
LRCN [17]	82.9
LSTM composite model [18]	84.3
Conv. pooling on long clips [20]	88.2
LSTM on long clips [20]	88.6
Multi-skip feature stacking [21]	98.1
C3D (3 nets) + iDT + linear SVM	90.4

Table 3. C3D compared with state-of-the-art methods. First part: features with SVM; Second Part: works using RGB frames; Third Part: works using multiple feature combinations [8].

Table 3 contains different feature extraction methods and their accuracy scores. According to this table above, the best accuracy is achieved when C3D and iDT are used together. iDT is a feature rest on optical flow gradients.

- Compressed Video Action Recognition

Many of the features used in video processing methods are used in video compression. Considering that videos are generally used in compressed format, it will be more efficient to use compressed format when processing video. The information in MPEG compression format is located in I-frames. There is no required to use extra methods to achieve them. Instead, a more efficient processing method is applied by focusing on P-frames with updates.

When experiments perform on UCF-101 dataset this method has better results than other works in the literature. Video compression is free of cost method and using this method has made the work very efficient.

3 SYSTEM DESIGN

In this section, technical details of the system, metrics to be used, dataset and professional considerations will be mentioned.

3.1 System Model:

Our project consists of 4 stages and these stages are explained below.

- Taking Videos from CCTV

We will take videos from the surveillance camera in pieces and run our system with these pieces.This system is summarized in figure 8.

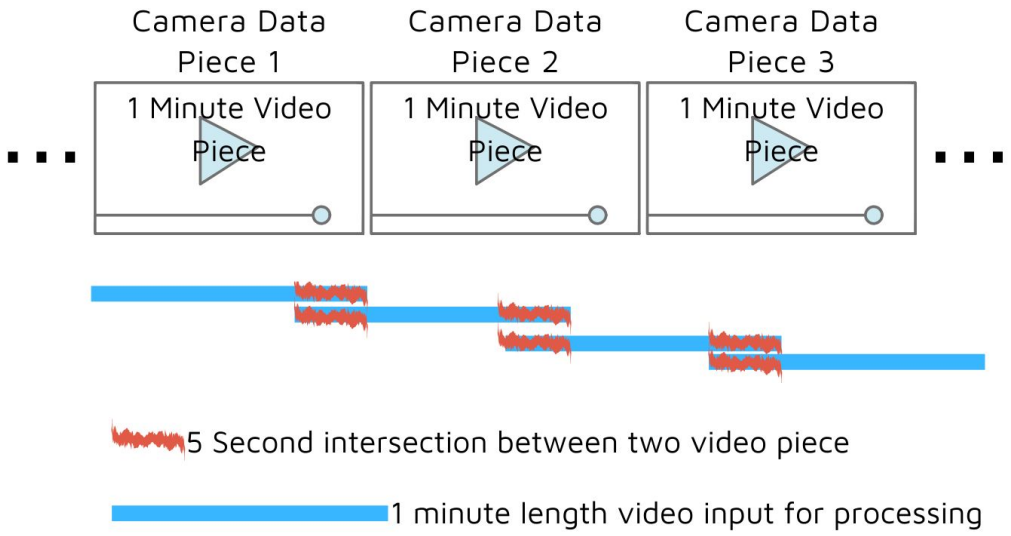

Figure 8: Process of taking videos from CCTV.

Since the average length of videos in the dataset to be used is 1 minute, we will take videos from the camera in 1-minute lengths. 5 seconds of each video track will intersect with each other. The reason we do this is to prevent problems in detecting events that will occur at the beginning or at the end of the video tracks.

In the following steps, Convolutional Neural Network models will be used. Convolutional Neural Networks are useful for image recognition and classifications. Input images taking as an array of pixels.

Each input image will pass through a series of convolution layers with filters, pooling, fully connected layers. Then Softmax classifier is used to obtain a probabilistic value between 0 and 1. Figure 9 shows the flow of CNN to process an input image.

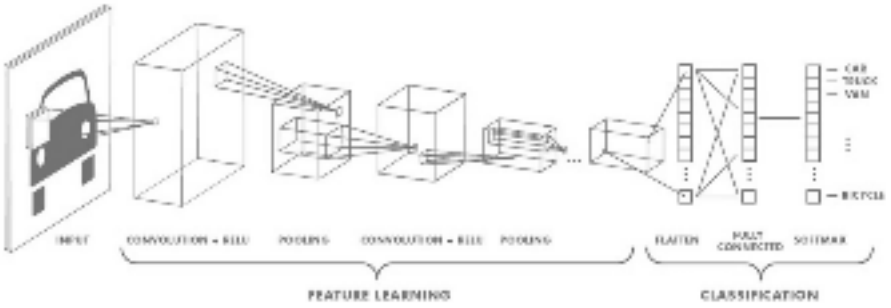

Figure 9: Neural network with many convolutional layers [14].

- Feature Extraction

In the first stage of video processing projects, it is necessary to represent the video with numerical values. The process of numerically representing the content of a video is called "Feature Extraction". Images from the camera will be feature extracted first, making them available for later models. At this stage, we will take the feature extraction layer of C3D Video Classifier model developed by Facebook [13]. 3D convolutions extract both spatial and temporal components related to the movement of objects, human actions, human scene or human-object interaction and the appearance of these objects, people and scenes. “This makes it a very generic video feature representation for various video-related tasks such as action localization, and event detection without the need for fine-tuning for each task” [10]. Layers of the C3D model are shown as Figure 10.

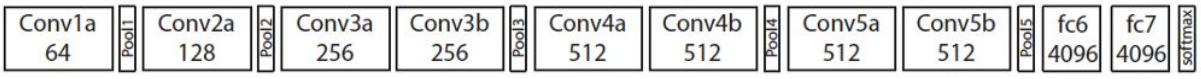

Figure 10: Layers of the C3D model [14].

- Violence Detection

Next, the violence detection model will work to determine if there is violence in the video that has completed feature extraction. At this stage, the model will be developed using CNN architecture. While the model is being trained, only videos of fighting and vandalism will be used as violent videos, and other events(shooting, road accidents, throwing something, walking, running, etc.) will be considered as non-violent. UCF Crime and NTU CCTV Fights dataset will be used in the training of this model. If there is no fighting or vandalism in the incoming video, this video will not be considered violent. In this way, we will be able to determine the existence of fighting or vandalism events more accurately and faster. The training of the violence detection model is shown in Figure 11.

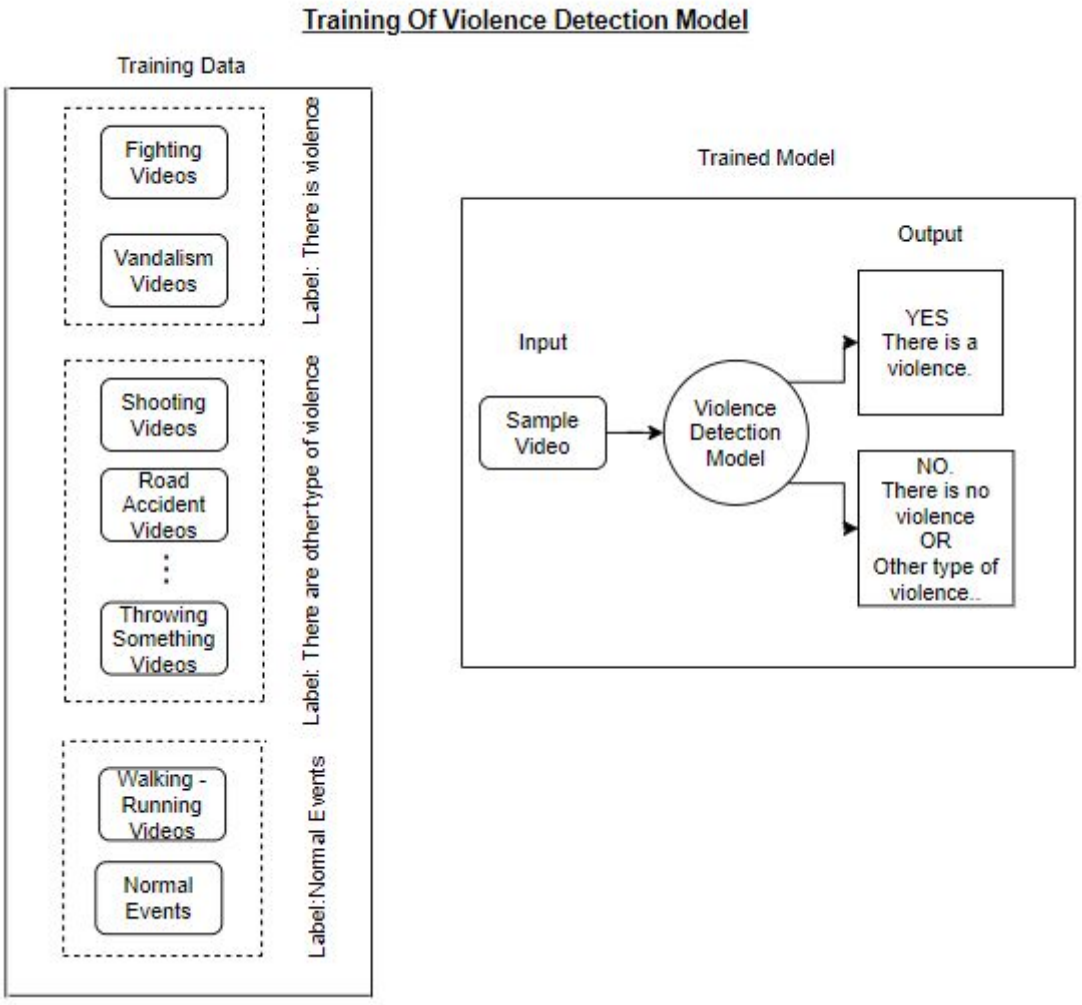

Figure 11: Training of violence detection model.

- Violence Recognition

Following the model that detects violence, a model will work to recognize whether the violence detected is fighting or vandalism.

CNN architecture will be used when training the recognition model. Only the fight and vandalism videos will be used while training the model. The output of the model will indicate what kind of violence the incoming video is. The training of the violence detection model is shown in Figure 12.

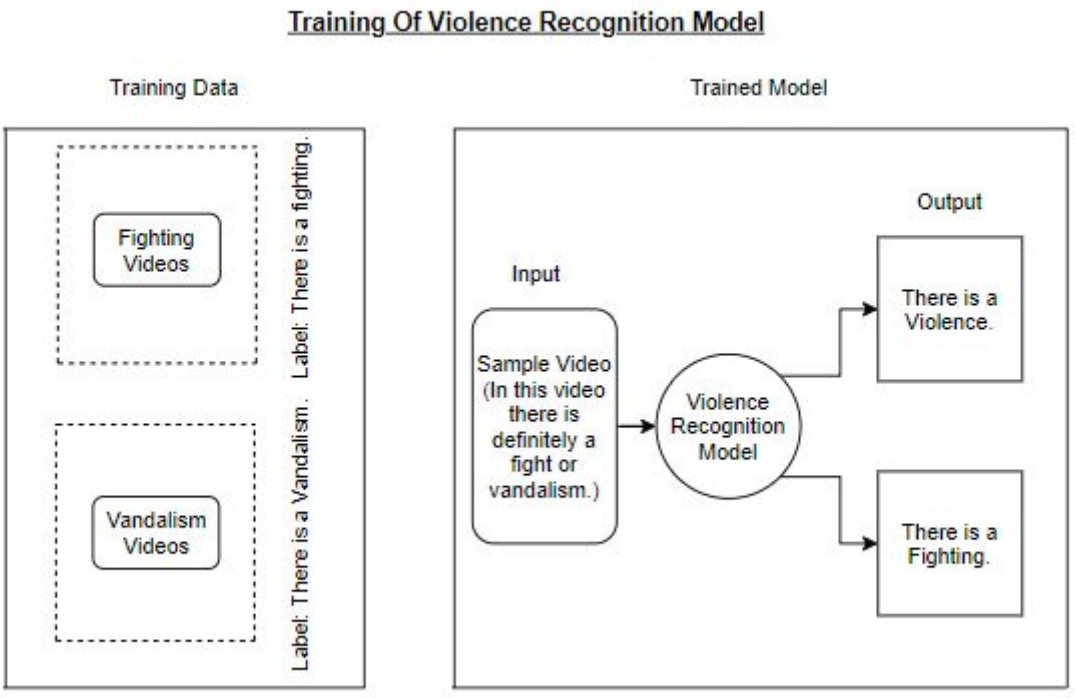

Figure 12: Training of violence recognition model.

The recognition model and detection model are made separately to ensure that the project runs at a faster rate in real-time. Thanks to this dual system, if there is no violence (fight or vandalism) in the video, we will not need to recognize this violence. This will prevent loss of recognition time. Given the low incidence of violence on the security camera, trying to recognize a non-violent video will be a waste of time. In addition, the system's fragmented operation will improve its maintainability.

3.2 Flowchart

There are 3 main models in the system. The first one is the feature extraction model, which will convert the videos into a text file of numbers. The second model is the violence detection model, which will tell if there is violence in the video. The third model is the violence recognition model, will determine the type of violence. If there is violence, the violence recognition model will work and send the type of violence detected together with the camera information to the authorized person via SMS. If there is no violence, a new piece of video will be processed before the 3rd model works. This flow is shown in Figure 13.

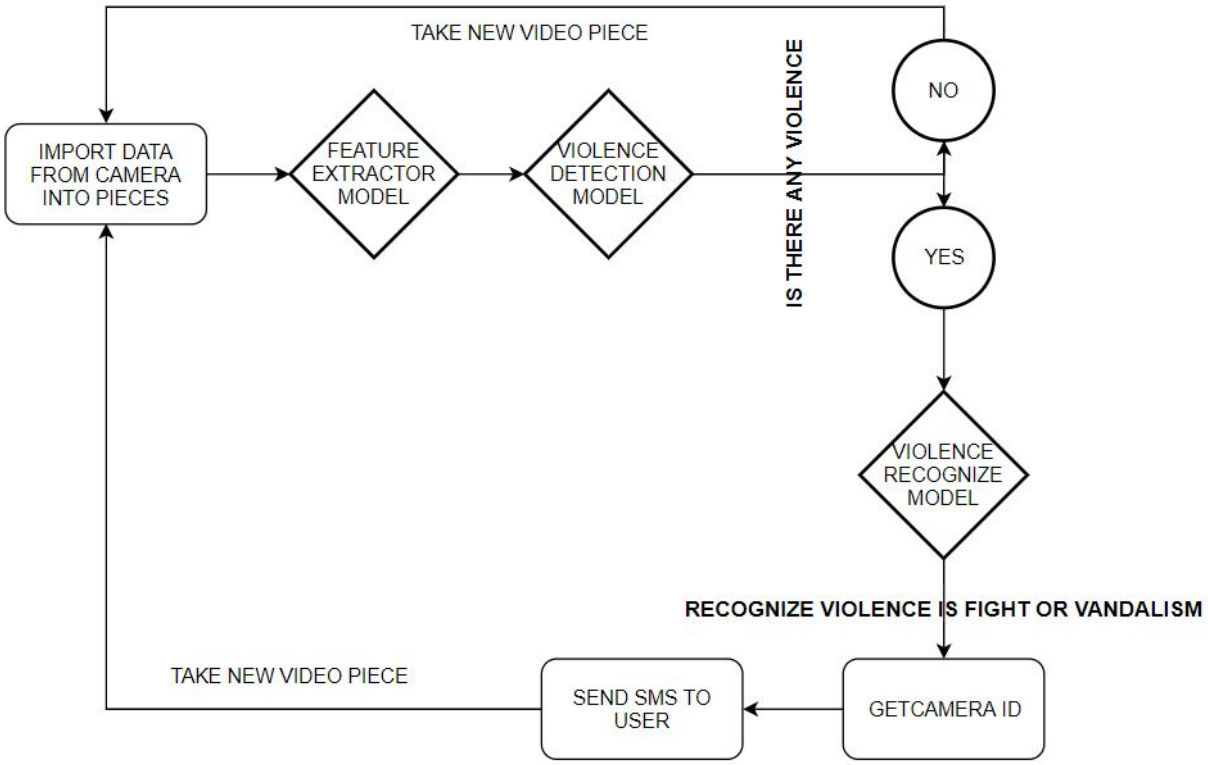

Figure 13:Flowchart of our system.

3.3 Comparison Metrics

The system we developed has a cascaded architecture, the overall performance of which depends on the performance of each unit. The process starts with a feature extraction model. Naturally, the performance of all steps, feature extraction model, detection model, and classification model steps depend on the quality of the video, video size, and position of the camera. These used models are the Convolutional Neural Network. And they have same performance metrics. Precision, recall, accuracy and F1 score are the performance metrics for these models.

The concepts used in calculating these accuracy scores are given in Table 4.

	Negative (predicted)	**Positive (predicted)**
Negative (actual)	True Negative	False Positive
Positive(actual)	False Negative	True Positive

Table 4:Definition table of some metrics.

- Accuracy

Accuracy is the ratio of the number of correct predictions to the total number of test samples. It tells us whether a model is being trained correctly and how it may perform generally. But it works well if only each class has an equal number of elements.

- Precision

Precision tells how often prediction is correct when the model predicts positive. So precision measures the portion of positive identifications in a prediction set that were actually correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall

Recall is the number of correct positive results divided by the number of all relevant samples so recall represents the proportion of actual positives that were identified correctly.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- F1 Score

F1 Score is a measure of the test's accuracy It is the Harmonic Mean between precision and recall. The value of F1 Score can be between 0 and 1. When the F1 score is equal to 1, the model is considered to work perfectly.

$$\text{F1} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.4 Dataset or Benchmarks

We decided to use well-known datasets and we wanted to test our system with the specified datasets. So, we used UCF-Crime and NTU CCTV datasets.

- UCF-Crime Dataset

This dataset contains 13 different types of violence and 1900 non-violent videos. The most common types of violence were taken into consideration when choosing the types of violence. Realistic anomalies are Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. Figure 14 shows examples of anomalies in the UCF 101 dataset. The distribution of video lengths is shown in Figure 15 [11].

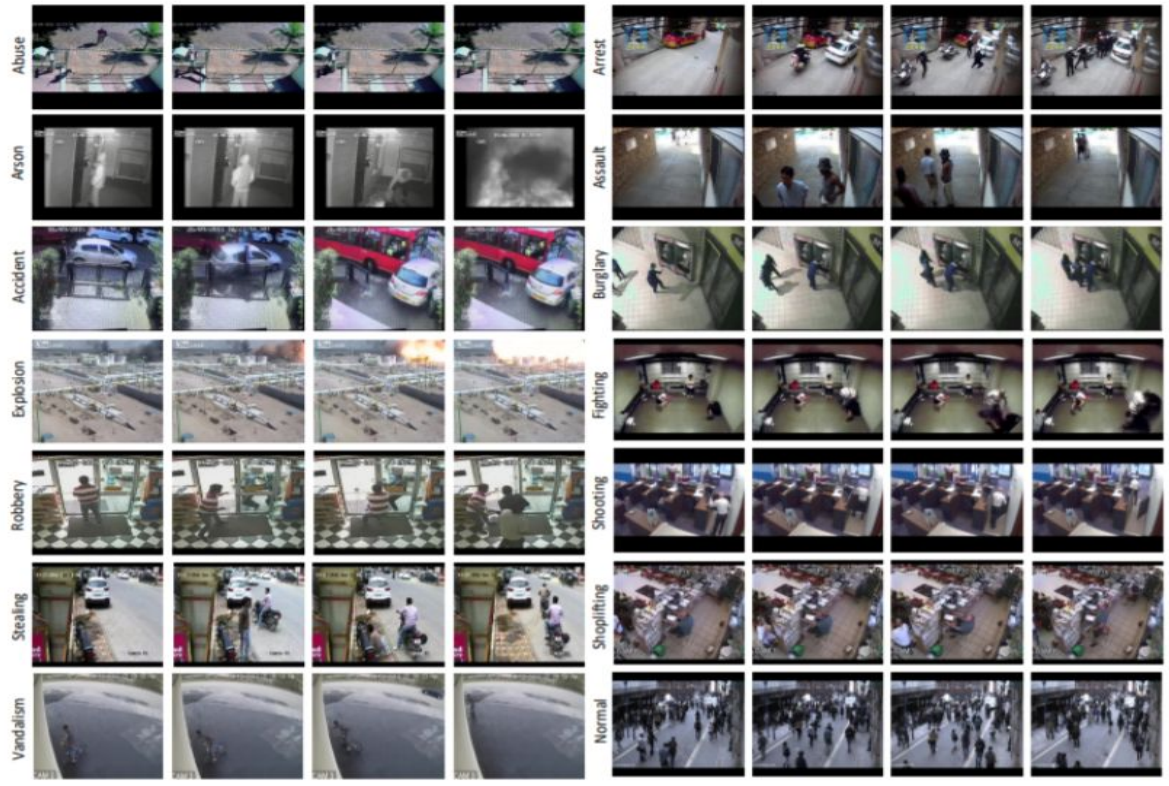

Figure 14: Examples of anomalies in UCF 101 Dataset [11].

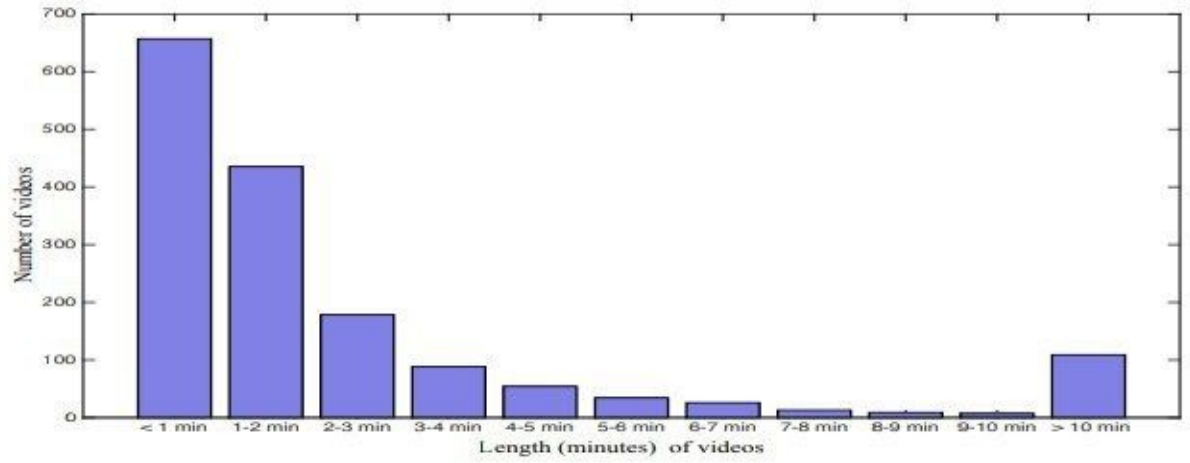

Figure 15: Distribution of videos according to length (minutes) in the training set [8].

- NTU CCTV Fight Dataset

CCTV-Fights Dataset [12] contains 1,000 violent videos, recorded from CCTVs or mobile cameras. these videos were selected and collected from Youtube. Fights can contain various range of actions, for example, punching, kicking, pushing, with two persons or more, etc. It was excluded videos that did not came directly from a CCTV recording, as well as videos with heavy special effects. The size of the dataset is 7.2 GB. Example of frames in NTU CCTV dataset as shown in Figure 17.

Fig 16 shows a summary of the NTU CCTV Dataset statistics.

	Videos	Duration (hours)	Fight Instances	Instances Average per video
All	1,000	17.68	2,414	2.41
CCTV	280	8.54	747	2.67
Non-CCTV	720	9.13	1,667	2.32

Figure 16: NTU CCTV Dataset Video Statistics. [12]

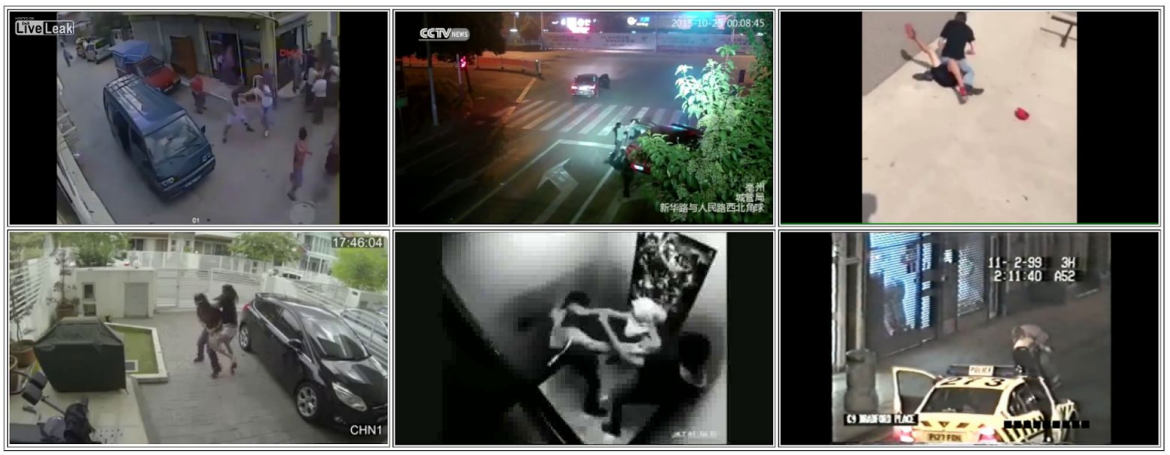

Figure 17: Sample frames of NTU CCTV Dataset. [12].

3.5 Professional Considerations

In the following section, professional considerations will be explained in 6 chapters.

3.5.1 Methodological considerations/engineering standards

- Git is a version control system that provides strong coordination skills between team members. That's why this common VCS is chosen.
- Python is the most popular language for the implementation of deep learning applications. So, many source codes to help can be found easily. Therefore, Python the programming language will be used in this project.
- For visualization of the progress of the project, flowchart diagrams will be used.
- Google Drive was used to communicate between team members and the supervisor and to keep documentations.

3.5.2 Economical

There are security cameras in many places with people, especially crowded. The number of these cameras is very much, and footage needs to be constantly monitored in order to be able to intervene early in a fight or similar situations. However, this is very costly because it requires too many people. Thanks to the project, there is no need for any person to watch the footage, thus providing substantial financial gain to the institutions where the system is used.

There is also a lot of economic damage in the environment in which these incidents take place. The project provides early intervention for such incidents and minimizes environmental damage.

3.5.3 Health and Safety

The most important impact of the project is on health and safety. Thanks to the project, violence can be intervened early. In this way, security forces can control events much faster without

growing. This raises the security level considerably. At the same time, early intervention minimizes the number of people injured and prevents mortality.

3.5.4 Social

The project may even prevent many events from starting because it is known that the event will be intervened early and there is no possibility of overlooking. Decreasing violence in society enables people to trust each other and live without fear. As a result, the project creates a more peaceful society.

3.5.5 Ethical

As we started the project, we read many articles and received ideas from them, but we created our own method. The data sets that we use in the project are published with the articles we review and are available to everyone. Therefore, people in these datasets have permission to use their images of this kind of project.

3.6 Legal Considerations

There is no legal issue for the projects. Security camera footage will be processed on the owner's device and no data will be shared with the outside. No connection is needed when detecting violence. The connection will only be used to notify the authorized person. The researches which we used to base the level of our project are free to use licenses. The databases we will use are publicly available for research purposes.

3.7 Risk Management

- The system cannot make the right decision;

  -Detecting violence in non-violent situations,
  -Inability to detect violence in cases of violence.

In case of such a problem, the level of violence can be determined by a function to be added to the system. With this function, the magnitude of the violence in the video can be determined.

- The system cannot work near real-time;

Real-time intervention in fighting events is more vital than vandalism. The vandalism detection feature will be removed from the system to prevent loss of time due to vandalism detection.

4 SYSTEM ARCHITECTURE

Our system works in 4 steps as shown in the Figure 18. The camera and the authorized person only influence our system as input-output. The input of the system is the video that comes in parts from the camera. The output is the message that sends the information of the camera where the violence is detected. Information to the authorized person will be made via SMS.

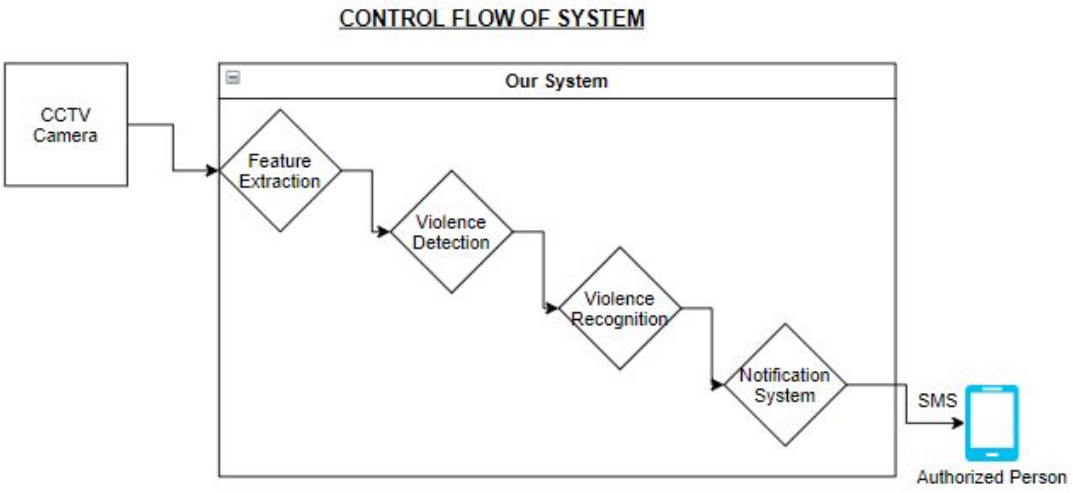

Figure 18: Control Flow of System

5 EXPERIMENTAL STUDY

Since we will use the feature extraction model as pre-trained, we will basically need to test the following situations:

- Feature extraction model run time for a 1-minute video.
- Precision-recall, accuracy and f1 scores of the violence detection model.
- Precision-recall, accuracy and f1 scores of the violence recognition model.
- 1 minute video processing time while each model is running independently.
- 1 minute of video processing time with models working together.
- According to the authorized person in real-time, the system is running in real-time (via SMS).

Tests of the above test conditions were not completed. However, we observed the intensity of each frame of a pre-trained model. Graph of violence detection of a pre-trained model as shown as Figure 19 and 20. Y values indicate the percentage of violence in the frame being processed.

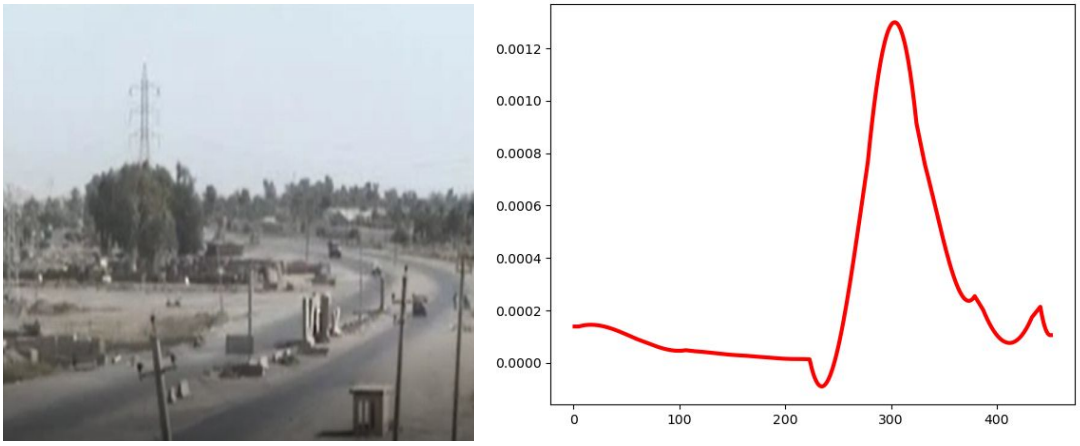

Figure 19: Violence intensity of non-violence frame. Note the y values, smaller than 0.1.

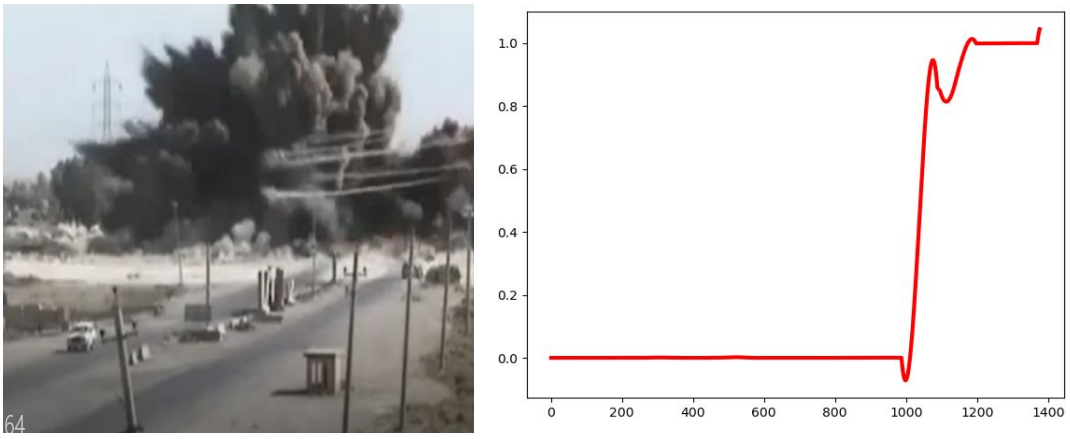

Figure 20: Violence intensity of violence frame. Note the y values greater than 0.9.

6 TASK ACCOMPLISHED

The current status of the project, the meeting logs and the division of tasks will be discussed below.

6.1 Current State of the Project

The completed tasks are shown below:

- Literature Survey for the components is done.
- Related works are examined.
- We decided to use the dataset.
- Completed the necessary procedures to obtain dataset.
- We have decided on the model we will use for feature extraction(C3D).
- All videos in our dataset have features extracted.
- We started to implement a violence detection model.
- We have made our computer's camera usable to simulate the security camera.
- We started to create test videos.

6.2 Task Log

The logs of the meetings until the current stage are as follows.

Meeting #1	Decisions and Notes
Objectives: Research About Topic	● The articles to be searched about the project were determined. ● The scope of the project was clarified. ● The project was broken into pieces and everyone's responsibility was determined.
Period: 1 Week	
Date: 09.10.2019	

Meeting #2	Decisions and Notes
Objectives: Research About Topic	● PSD preparation should start. ● Presentations about different methods were discussed. ● An online course has been determined to follow up on deep learning. ● The appropriate dataset was determined (UCF Crime-NTU CCTV Fights).
Period: 1 Week	
Date: 16.10.2019	

Meeting #3	Decisions and Notes
Objectives: Research About Feature Extraction	● The procedures for obtaining the dataset have been completed. ● The model for feature extraction was researched. ● Alternative feature extraction models identified.
Period: 1 Week	
Date: 23.10.2019	

Meeting #4	Decisions and Notes
Objectives: Research About Violence Detection **Period:** 1 Week **Date:** 30.10.2019	● The methods for detection of violence were determined (Optical flow-based features Spatio-temporal CNN). ● PSD is complete ● We will check if we can achieve the same scores by trying out the models in the researched articles.

Meeting #5	Decisions and Notes
Objectives: Feature Extraction Model **Period:** 1 Week **Date:** 6.10.2019	● It was decided to use C3D for feature extraction. ● It was decided to use COLAB to run the C3D model. .

Meeting #6	Decisions and Notes
Objectives: Feature Extraction Model **Period:** 2 Week **Date:** 20.11.2019	● Feature extraction was applied to the videos in the dataset and turned into text. ● We started running a pre-trained model for violence detection.

Meeting #7	Decisions and Notes
Objectives: Simulating CCTV Violence Detection Model **Period:** 2 Week **Date:** 04.12.2019	● We ran a model that was made before with our own data. ● We decided to use a webcam to simulate a security camera. We have implemented the necessary system for this. ● We started to prepare the fall presentation.

Meeting #8	Decisions and Notes
Objectives: Fall Presentation ADD	● Deep learning courses continue. ● Fixed bugs related to the presentation. ● ADD started to be written.
Period: 2 Week	
Date: 18.12.2019	

6.3 Task Plan with Milestones

6.3.1 Description of Task Phases

Phase 1: Literature survey of anomaly detection, violence detection, and recognition.
Phase 2: Examining existing trained models and testing their performance.
Phase 3: Selecting data set and implementing the violence detection model.
Phase 4: Fine-tuning of detection model.
Phase 5: Implementing and fine-tuning of violence recognition model.
Phase 6: Testing on real data.
Phase 7: Implementing the real-time notification system.

6.3.2 Division of responsibilities and duties among team members

Implementing of feature extraction model: Alperen Bayar-Büşra Yağcı.
Testing of feature extraction model: Büşra Yağcı.
Implementing of violence detection model: All together.
Testing of violence detection system: Nuri Yıldız.
Implementing of violence recognition model: Nuri Yıldız-Büşra Yağcı.
Fine-tuning and testing all system: Alperen Bayar.
Implementing of real-time system: Alperen Bayar-Nuri Yıldız.

6.3.3 Timeline

Time planning as explained in figure 21.

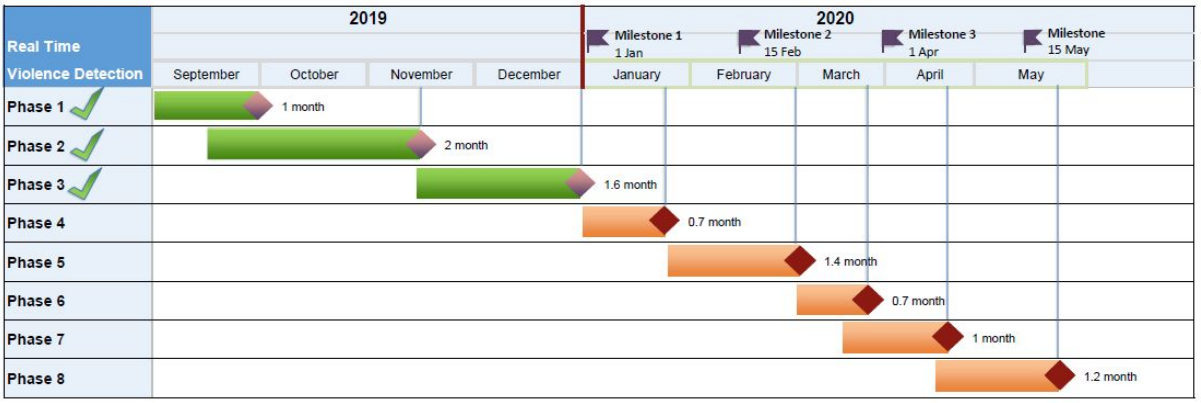

Figure 21: Gantt chart of time plan.

7 REFERENCES

[1] “CCTV surveillance for crime prevention”, Eric L. Piza Brandon C. Welsh David P. Farrington Amanda L. Thomas, First published: 24 March 2019, https://doi.org/10.1111/1745-9133.12419(Date of access: 28/10/2019)

[2] https://www.glassdoor.com/Hourly-Pay/Los-Angeles-Metro-CCTV-Observer-Hourly-Pay-E16590_D_KO18,31.htm, CCTV Observer Salaries(Date of access: 28/10/2019)

[3] “A Review on State-of-the-Art Violence Detection Techniques”, Muhammad Ramzan, Adnan Abid, Hıkmatullah Khan, Shahmid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8782115&isnumber=8600701(Date of access: 28/10/2019) In IEEE, 2019.

[4] J. Xie, W.Yan, C. Mu, T. Liu, P. Li, and S.Yan, ``Recognizing violent activity without decoding video streams," Optik, vol. 127, no. 2, pp. 795801, Jan. 2016.

[5] Q. Xia, P. Zhang, J. Wang, M. Tian, and C. Fei, ‘‘Real-time violence detection based on deep Spatio-temporal features,’’ in Proc. Chin. Conf. Biometric Recognit., 2018, pp. 157–165.

[6] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, ‘‘Violence detection using spatiotemporal features with 3D convolutional neural network,’’ Sensors, vol. 19, no. 11, p. 2472, May 2019.

[7] “Real-world Anomaly Detection in Surveillance Videos”, Waqas Sultani, Chen Chen, Mubarak Shah, https://arxiv.org/pdf/1801.04264.pdf, 04 Jan 2020

[8] “Learning Spatiotemporal Features with 3D Convolutional Networks”, Du Tran, Lubomir Bourdev, RobFergus, Lorenzo Torresani, Manuhar Paluri, http://vlg.cs.dartmouth.edu/c3d/c3d_video.pdf (Date of access:04/01/2020)

[9] “Compressed Video Action Recognition” Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, Philipp Krähenbühl https://arxiv.org/abs/1712.00636 (Date of access:04/01/2020)

[10]https://mc.ai/quick-overview-of-convolutional-3d-features-for-action-and-activity-recognition-c3d/ (Date of access:04/01/2020)

[11] https://webpages.uncc.edu/cchen62/dataset.html (Date of access:04/01/2020)

[12] http://rose1.ntu.edu.sg/Datasets/cctvFights.asp (Date of access:04/01/2020)

[13] C3D Feature Extraction https://github.com/facebookarchive/C3D (Date of access:04/01/2020)

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

[16] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

[17] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014.

[18] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In ICML, 2015.

[19] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. CoRR, abs/1405.4506, 2014.

[20] J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, 2015.

[21] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond the gaussian pyramid: Multi-skip feature stacking for action recognition. CoRR, abs/1411.6660, 2014.