A Computer Vision based Approach for Stalking Detection using CNN-LSTM Hybrid Model

by

Shahriar Iqbal
18101643
Murad Hasan
18301253
Md Billal Hossain Faisal
18301066
Md.Musnad Hossin Neloy
22141032
Md. Tonmoy Kabir
18301245

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2022

Declaration

Hereby it is proclaimed that:

1. The submitted thesis is our own unique work created while pursuing undergraduate degree at BRAC University.
2. The thesis does not include previously published or written content by a third party, unless properly referenced with complete and correct referencing.
3. There is no content in the thesis that has been approved or submitted for any other degree or certificate at a university or other institution.
4. We have recognized every major source of assistance.

Student's Full Name & Signature:

Shahriar Iqbal
18101643

Murad Hasan
18301253

Md Billal Hossain Faisal
18301066

Md.Musnad Hossin Neloy
22141032

Md. Tonmoy Kabir
18301245

Approval

The thesis titled "A Computer Vision based Approach for Stalking Detection using CNN-LSTM Hybrid Model" submitted by

1. Shahriar Iqbal (18101643)
2. Murad Hasan (18301253)
3. Md Billal Hossain Faisal (18301066)
4. Md.Musnad Hossin Neloy (22141032)
5. Md. Tonmoy Kabir (18301245)

Of Spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 28, 2022.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Md. Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Associate Professor and Charperson(CSE)
Department of Computer Science and Engineering
Brac University

Abstract

The next level of revolution toward a better world could involve combining human security with machine intelligence. In recent years, stalking in public areas has become a pervasive issue, and women are disproportionately affected. In order to solve the problem, we want to design a model that can identify public-space stalking. There have been several study papers and publications written on the topic of stalking. However, most of them relied on spatial co-occurrence for the detection of suspicious actions and face recognition, which does not adequately address the problem. Using a hybrid mix of CNN and LSTM, we explain in our study a model for determining the presence of a stalker situation utilizing a dataset of video footage. The proposed model was evaluated using two approaches: one using manual feature extraction and the other using dynamic feature extraction. The manual feature extraction approach was evaluated with three distinct machine learning classifiers (SVM,KNN, and Random Forest), whereas the dynamic feature extraction method was examined with two different CNN models (VGG16 and ResNet50) and a CNN-LSTM hybrid model. The CNN-LSTM hybrid model has the highest accuracy of any of these models, at 89%. Experiment results indicate that the CNN-LSTM hybrid model detects a stalking scenario with a spatio-temporal advantage and provides a better classification result than other models.

Keywords: Machine Learning; Stalking;Non-Stalking; Prediction; LSTM; CNN; Neural Networks; Classification;

Acknowledgement

All thanks is due to the Almighty Allah, because of whom our thesis was finished without substantial interruption. Second, we would like to thank our co-advisor, Md. Tanzim Reza sir, for his assistance and advise with our work. He assisted us anytime we needed assistance. Thirdly, we would like to express our appreciation to our supervisor, Dr. Md. Golam Rabiul Alam, and co-supervisor, Mr. Tanzim Reza, for their advice and invaluable assistance during the whole thesis process. Finally, without the unwavering support of our parents, this may not be feasible. With their kind assistance and prayers, we are now on the brink of graduating.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The planet is constantly expanding and changing because of the tremendous strides that have been made in science and technology in the modern era. The rapid advancement of digital technologies has created a plethora of new options for us to develop the planet. By lowering human effort and solving human difficulties, we may apply our understanding of fundamental development in technological innovation to promote human wellbeing. To illustrate, a variety of problems and crimes in our country are being resolved via the use of various technologies. These include artificial intelligence, machine learning, deep learning, computer vision, and camera networking. Violence against women is one of the most disturbing aspects of crime in our country today. Physical assault and cyberbullying are both forms of violence against women. It is contingent upon the circumstances and the nature of the perpetrator's relationship with the victim. Several incidents of violence against women garnered widespread attention and were addressed via the use of information and communication technologies. However, a great number of different kinds of violence have gone unnoticed. In criminology, stalking analysis has developed into a significant social issue. Stalking can be classified into two types. One is cyberstalking, and the other is street stalking, or in certain ways, physical stalking. Among these two forms, cyberstalking has already captured the majority of the public's attention and has sparked enormous public outrage in recent years. Additionally, many people are working diligently to put an end to cyberstalking. However, if we look closely, physical stalking has not been the one to work on it in a systematic manner. While information and communication technologies are advancing and improving across all spheres of social life, they have not kept pace with this cultural shift in instances of physical stalking. Physical stalking has a devastating effect on women. Knowing the deadly implications of physical stalking on women might shock some. To begin, stalking is a distinct type of criminal behavior in which one individual harasses or monitors another by observing, following, spying, or looking at them. Other academics have defined stalking in a variety of ways in numerous philosophical research papers and publications. The researchers narrated the term stalking as the malicious, purposeful and iterative pestering of another person which generally accompanied by a genuine intimidation of harm against the victim [1]. The definition of stalking in terms of criminology includes a wide variety of acts and are established by the precinct, local and laws, and other types of systematic discretion. As per medical experts, it's defined by intrusive actions (e.g. threatening, following, spying, getting unsolicited phone and email calls) that provide the victim a false sense of

security and make her believe she's in a dangerous situation that she has no control over [33]. Additionally, statistics from a variety of countries illustrate the severity of physical stalking. Bianca Fileborn examined 292 people across Australia in 2016 and discovered that 65.1% of them were gazing at women on the street [18]. As a result, the victim's physical, mental, and social lives are severely impacted, as is her overall quality of life. Meanwhile, camera footage has been important in identifying a potential stalking situation. We can use computer vision, machine learning and deep learning to solve the crime of physical stalking.

1.1 Problem Statement

Nowadays, stalking women in public places is a widespread problem across the world. However, we are not sufficiently aware of the dangers associated with this widespread problem. Every woman and girl in the world has encountered street harassment, particularly stalking, at least once throughout their lives. As a result, many women and girls are forced to take a step back or are unable to walk freely on the street.

Despite the pervasiveness of information communication technology in many spheres of social life, physical stalking has not been a factor in studies on how to operate compactly. Numerous academics, writers of computer science, and technologies have focused on detecting various human behaviors, but physical stalking detection has yet to gain the prominence it deserves among these behaviors. Though there is some work on it collaborating with computer vision techniques or other algorithms. However, it should be increased since it is one of the first stages of every major crime. As a consequence of the psychological and mental issues that might develop from stalking. Furthermore, according to data from all around the world, stalking should be given top attention, with urgent steps taken to reduce the incidence of stalking.

According to one study, the existence of stalker patterns in surveillance videos can be identified through the use of a noble parameter called "Spatio-Temporal Co-appearance." However, in this case, two people cannot be detected with a single camera; several cameras are required to identify individuals in different locations and times. This parameter can offer an effective method for decreasing the range and speeding the investigation process, but does not provide ideal answers for a proper stalking situations and moreover it is much costly. [32]

In our research, we will seek to identify a possible case of stalking. To be more specific, we will concentrate on whether or not a stalking scenario is happening at a particular moment. We may examine the precise spatial connection between two individuals appearing from any angle. We will use CNN and LSTM together to get the right pattern of steps to identify a stalking scenario.

1.2 Objective

Our ultimate objective is to design a hybrid model that can be installed in a camera that will work as a detector of suspicious physical stalking behavior and identify whether it falls on stalking scenario or non-stalking scenario. This application can further help security enforcement authorities to identify and get evidence against any potential stalker. And it will also ensure everyone safety on the streets by giving a proper evidence of a potential stalking situation to detect the stalkers. Our thoughts on building the model is to identify the stalking scenario in very few frames.

Chapter 2

Literature Review

Stalking is viewed as a dangerous offense in the modern world due to the direct and imperceptible effects it has on our society. In our very neighboring country, India, a survey on street harassment of women was conducted. This report's outcome was heartbreaking. Almost 67 percent of those polled said they had been stalked on the street. However, the most tragic aspect is that the majority of them no longer feel safe leaving their homes. Though discussion of this subject began in 1944, no ground-breaking initiative was taken to avert the crux. However, the gravity of the trespass has recently compelled the world to take action to avert the labyrinth. Numerous preparatory measures have been taken in recent years to address stalking issues. These preventative measures may include mobile applications, CCTV camera surveillance, and approaches based on neural networks. A drone- based surveillance approach is described in papers [35] and[22].

Aadesh Guru Bhakt Dandamudi along with his other authors [35] primarily discusses a CNN-based model for drone image technology. The author proposes a model in which each person is detected within a frame and pose estimation is used to identify humans using key points. Additionally, this model must recognize human positions on the same scale. The significance of this model is that it enables more precise classification of human actions.

Bhattacherjee, S. and Somashekhar [22] developed a flying device that resembles a bird and is capable of continuously monitoring any suspicious activity that occurs in its vicinity. It is capable of locating suspects or any human being with the bare minimum of human effort. This intelligent device is also responsible for transmitting a signal to the nearest police station containing the suspect's location and for tracking the suspect's movement. This device is schematically similar to a bird that is following an aerodynamic model and is equipped with a 360-degree rotational camera. Additionally, it makes use of an Arduino Uno processor and a GPS tracker worn on the body to navigate to the suspected object's location. This model makes use of a mat lab image comparison algorithm [SIFT ALGORITHM] to reconcile the latest image with the reference image.

Furthermore, another method of stalking darn is through mobile applications. This approach discusses the feasibility of a mobile-based application for rescuing stalking victims. The Stalking Dran model includes an emergency button that can alert

rescue personnel (police, other people who use this app, guardians).

Sidhu, R. S., and Sharad, M. [20] considered two distinct application areas in order to ensure confidentiality in a variety of locations. One is in a secure location where CCTV cameras will begin recording when the likelihood of a "critical situation" occurring is high (triggered ON). Second, in public locations where they conduct a preliminary examination of critical situations. Researchers can do this by reducing the amount of memory used, ensuring privacy, and avoiding legal repercussions.

Wei Niu and Jiao Long along with their other co-authors [4] proposes a software framework based on real-time activity detection and recognition. This software detects and recognizes human activity in real time. The researchers introduced intelligent control and fail-over mechanisms to detect activity and track a moving person. The detection process is herculean due to frame differencing and feature correlation techniques. Additionally, the author developed a scheme that employs the relative positions of people, velocity, and individual trajectories as parameters for identifying targeted stalker groups. This straightforward scheme can be used in place of more complicated models. However, in this paperless world, camera calibration errors and a lack of pixel processing make the paperless efficient.

According Jianquan Liu; Duncan Yung and their co-authors,[32] stalker patterns were identified through the introduction of a noble parameter dubbed "spatiotemporal co-appearance." This means that the same pair of people have been accused of stalking in different time periods. Neo-face detection is successfully used to identify the same person in multiple frames. Despite the fact that this paper accomplished a great deal, this system will remain focused on a single point. For instance, stalking cases will be detected when friends, family members, and couples walk side by side. Too many false-positive cases could erode the system's usefulness in practice.

False-positive cases are extremely manageable in surveillance-based systems. As a result, identifying appropriate stalker patterns is highly emergent. [9], [11] papers have been shown to be effective in locating stalker specimens. Bhaumik, G., Mallick and their other two cowriters attempted to categorize unnatural events using six fundamental human emotions (Joy, Sadness, Anger, Surprise, Fear, and Disgust). Numerous facial postures are used as indicators of whether or not the approach is natural. He, L., Wang, D., and Wang, H. [9] differentiated between distinct behavioral patterns in their paper. Thirteen distinct physical movements have been pre-programmed as parameters. Pang, J. M., Yap, V. V., and Soh, C. S. [11] , used the Cartesian coordinate formula to calculate the position of body joints detected by the Microsoft Kinect sensor. The pilot results indicate that this proposed method has an average accuracy of 95.83 percent for punching poses and nearly flawless detection of kicking and normal behavior.

[14] Tiwari, C., Hanmandlu, M., and Vasikarla, S. thought distinguishing stalker patterns and eye movement can also be a game changer. Eye movement, according to cognitive visuomotor theory, is a very rigid phenomenon that characterizes a stalking scenario. What a person is thinking at the moment can be deduced from his or her glancing, as proposed by the "Eye-Mind hypothesis." The non-linear entropy

of a criminal's eyes is significantly greater than that of a normal person. Additionally, rapid eye movement is another indicator of stalkers' patterns. Nominating eye movement is an effective technique for setting the ambit stalkers group apart.

[8] Wiliem, A. and his co-authors, determined that suspicious conduct based on ordinary occurrences is another technique to detect stalker activities. However, the term "abnormal activity" is context-dependent. For instance, while running and whistling are uncommon occurrences in general, they can be common and natural in certain circumstances. To achieve the best results and to strengthen the approach, locating those exceptional cases based on the context is a critical task.

Shakya, S., Sharma, S.,& Basnet, [19] proposed another fantastic technique for detecting stalking using facial expression analysis. Video data is converted into an image sequence. The faces, noses, eyes, mouths, and upper bodies of people are identified using a color-filtered image algorithm. There are a number of telltale signs, such as the suspect's wrathful lips, the corners of his mouth showing anguish, and his crimson complexion, which shows either intoxication or anger. Due to the fact that greater entropy regions of the human body have an easier time picking up on feelings. Additionally, a PCI algorithm is used to extract features, which calculates the coefficient, latent, and score. Additionally, the researchers used Euclidian distance calculation to predict facial expressions. All algorithms are implemented using the Viola-Jones AdaBoost technique. The HMM, Bayesian, or Kalman methods are used to track a moving person.

[27] Bisagno, N. along with his other two cowriters, thought calculating crowd flux can be an extremely effective tool for stalking detection. A prudent use of resolution can enhance surveillance's dynamic nature. Providing the same resolution throughout the surveillance area can occasionally result in a loss of resolution. Thus, using a constant population to determine whether or not a space is crowded is a very noble idea. Additional resolution in the dense zone can help ensure the model's accuracy. Dividing the entire area into local and global coverage zones is an excellent idea. This strategy ensures a minimum resolution in each segment of the area, with an emphasis on the crowded zone, resulting in a highly productive, cost-effective, and practical system. Another method of determining if a location is crowded is to analyze a data set. Additional surveillance of those alarming areas may make the approach more effective.

Chapter 3

Background Study

3.1 Statistics

The culture of Bangladesh is male-dominating. Consequently, in the event of an embarrassing occurrence, society blames women for it. Besides, all the law enforcement agencies like police, sheriff departments can not take any action simply because of proper evidence. As a result, girls are not getting proper justice and are deprived of moving freely in the streets. Besides, it is also hampering the economic growth in our country indirectly. As women cannot ensure their safety in the street, they cannot go to their workplaces. For this reason, our country is not getting any contribution from girls and women. This is one case of stalking where the victim is a woman.

There is a wide spectrum of trauma that may be inflicted on the victim of stalking, many of which are often misinterpreted by society as a whole. Starting from mental health to physical health and social life, stalking has been a big issue that affects negatively on this three segments in a person's life. In the case of mental health, stalking has a great number of negative effects. And this involves things like perplexity regarding numerous topics as well as denial, self- doubt, questioning whether what is happening is unreasonable, and wondering if they are over- reacting. Furthermore, they experience feelings of frustration, guilt, humiliation, and self-blame.

Stalking can lead the victim to feel helpless, which may result in depression, anxiety, panic attacks, and agoraphobia (fear of leaving the house, never feeling safe). For agoraphobia, a person tends to get themselves isolated and make themselves introverted. Fear of stalking can change their personality to become more suspicious, aggressive, and unable to trust others, which sometimes leads them to suicidal thoughts. Besides, it can result in them being unable to sleep (generally occurring because of overthinking), which results in difficulty concentrating and remembering things [10]. It should be noted that a survey held in India reported short-term mental and physical health issues such as stress and anxiety (63%), sleeping and eating disorders (29%), and self- blaming, which can be turned into long-term issues [15]. In addition, the victim may be pushed to question her beliefs about her invulnerability and control over her life as a consequence of the menacing behavior of the stalker, which may lead to post-traumatic stress disorder in the victim (PTSD) [2].

These mental and psychological problems ultimately lead the victims to many physical health problems. Effects of chronic stress include headaches, hypertension, dizziness, shortness of breath, gastrointestinal problems, heart palpitations, and sweating can occur due to excessive mental pressure [10]. In Australia, Pathe and Mullen surveyed 100 stalking victims. By this survey, they got to know that these women were suffering from mental and physical difficulties. This survey shows that women endured hypervigilance (83%), sleep disturbances (74%), intrusive recollections or flashbacks (55%), suicidal thoughts (24%), and various somatic complaints or a worsening of serious medical conditions such as asthma [3].

These psychological issues can have a negative influence on a victim's social life. It may lead victims to stay home most of the time, leave jobs, change careers, or drop out of school. Several studies show the consequences for the victim's employment. According to Pathe and Mullen's study of 100 victims, it shows that stalking had a mischievous impact on their occupational functioning. Among these, 6 reporting major lifestyle changes [3]. Victims tend to avoid places where the stalker might be more frequent. About 83% of the victims had modified their colloquial activities because of the fear of being stalked or followed [3]. Among these surveyed victims, 53% reported mitigation of work or educational institute attendance [3]. 54% reported that they have left their job because of the fear of being stalked[3].

If we look hermetically at the statistics, we can understand how repeatedly people are suffering from being stalked in public places. In the countries of Asia, stalking incidents happen very often, and the victim cannot file a case because there's no strong evidence to catch the main culprit. In Bangladesh, from 2011-2020, almost 3000 girls have faced stalking in public places directly or indirectly [41]. Because of that, many women have committed suicide. More than 13% of college women are stalked during the very first year of their college life [23]. In India, stalking cases are increasing very alarmingly. There were 6,266 cases reported among 100,000 women in 2015, and the percentage has more than doubled since 2017 [40]. Another survey conducted in India shows that women are experiencing 97% visual, 67% stalking while accessing public places in their everyday lives [15]. Not only in Asia, but all the developed countries are also facing the same problem. In Germany, 25 percent of women experienced harassment in the form of persistent staring in the year 2020 [34]. On the other hand, sexual insult percentages are quite high in the UK in the same year, which is 10% [42]. About half of all women who are murdered in the United States and Canada are killed by a person with whom they have or had an intimate relationship. Ninety percent of these deaths are preceded by some kind of stalking. At least 200 thousand stalkers are on the loose, says forensic psychiatrist Dr. Park Dietz, and one in twenty American women will be stalked at some point in her lifetime [2]. Surveys from Nepal also show a significant percentage of harassment and stalking carries a big portion of this harassment, which is almost 63% [39]. Also and perhaps most critically, all of the murders and rapes we have seen so far began with a street stalking.

3.2 CNN

In the subject of computer vision, the study of feature extraction and classification algorithms has long been a significant research focus. Conventional image processing methods based on many algorithms are incapable of adequately reproducing the original visual data. Consider the hardware-dependent Artificial Neural Network (ANN) scenario. It often demonstrates unexpected behavior, which causes us concern. There is no defined process for building the architecture of artificial neural networks; rather, the correct network topology is built through experimentation and error. Another classification method is the RNN, which has issues with falling and increasing gradients. In addition, RNNs are challenging to train. CNN offers a gradient-parameterized end-to-end learning model that can be trained [25] . A CNN that has been properly taught may get moreknowledge of the image's properties. CNN broadens the concepts of receptive field and shared weights, hence reducing the number of training parameters and simplifying the network model [44] . As a deep, feed-forward artificial neural network, convolutional neural networks (CNNs) have proved to perform well in computer vision tasks including image classification and detection [28] . Millions of photographs from a wide range of topics are available to CNN. An image's spatial information may be used by CNNs to automatically learn complicated properties like texture and color. The aggressiveness of CNN increases the variance of an input.

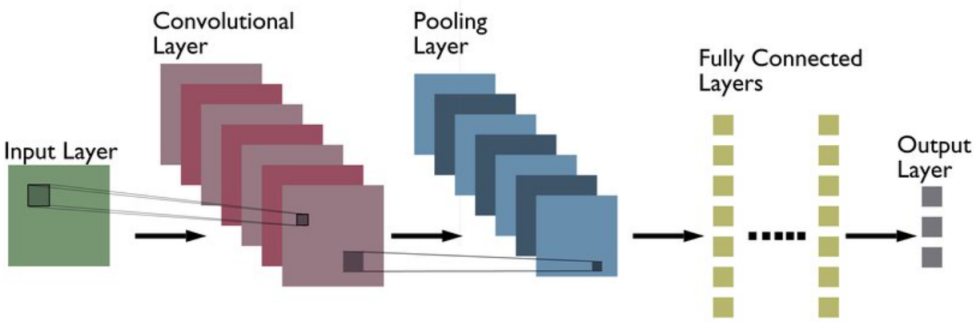

Figure 3.1: CNN Architecture

There are four primary levels in the CNN framework. Examples of these layers include an input layer, a convolutional layer, a pooling layer, and a fully connected layer [44] . The forward propagation and BP (Back Propagation) techniques are extensively used in CNN training to learn the layer-connection weights, bias and other parameters. Images and their labels are used in a supervised learning approach to adjust network parameters during training. This results in a model that is light in weight.

The input to a convolutional neural network is generally the original picture. [25] The pooling layer, also known as the down-sampling layer, is often placed after the convolutional layer and down samples the preceding feature map according to a predetermined pattern. [25] The rules include max-pooling, average-pooling, stochastic pooling, overlapping pooling, etc. The pooling layer function focuses mostly on two factors: The first is to lower the number of dimensions in the feature map. The second is to maintain scale invariance.

As input to a fully-connected network, the image feature maps are concatenated to form a one-dimensional feature vector. The output of a fully connected layer may be determined by doing a weighted summation on the layer's input and then using the activation function to react. The BP (back propagation) technique is used to alter the weight parameters of neural networks [25]. The most critical parameters for CNN optimization are the convolution kernel parameters, the weights of the pooling layer, the weights of the full-connected layer, and the bias parameters. At its most fundamental level, it is possible to calculate the partial derivative for each layer parameter, learn some correlation between residual and network weight, then use this correlation rule to fine-tune network parameters such that their output is more in line with training data's expectations.

CNN's training objective is to minimize the network's loss function [25]. Gradient descent is used to modify the trainable parameters of each layer in the CNN architecture after the training process. The strength of back propagation may be controlled by varying the pace at which new information is learned. As the procedure advances, the weight and bias are updated.

3.3 LSTM

The LSTM is a time-series data processing extension of the recurrent neural network [43]. In cases when there is a minimal time lag between events, recursive neural networks are an excellent choice. LSTM was created to address the two primary issues with RNN [31]. As time goes on, it fades into insignificance. A normal recurrent layer does not include a structure that controls the individual memory flow, therefore the "future" time-steps will have essentially little memory of initial inputs until the number of time-steps grows to a substantial extent. When the number of time-steps is increased, this will be the case.

However, RNN does not deliver high learning accuracy when the gradient value is so tiny [43]. As a result, it is unable to store information for extended sequences. LSTM has been developed to solve the vanishing gradient issue and provide more stability in long-term processes. LSTM uses the idea of an internal loop to keep only the important data and get rid of the rest [31].

All information in an LSTM cell is sent via three gates. These are the input gate, the output gate, and the forget gate [31]. The working flow of these gates in LSTM is shown below using a diagram:

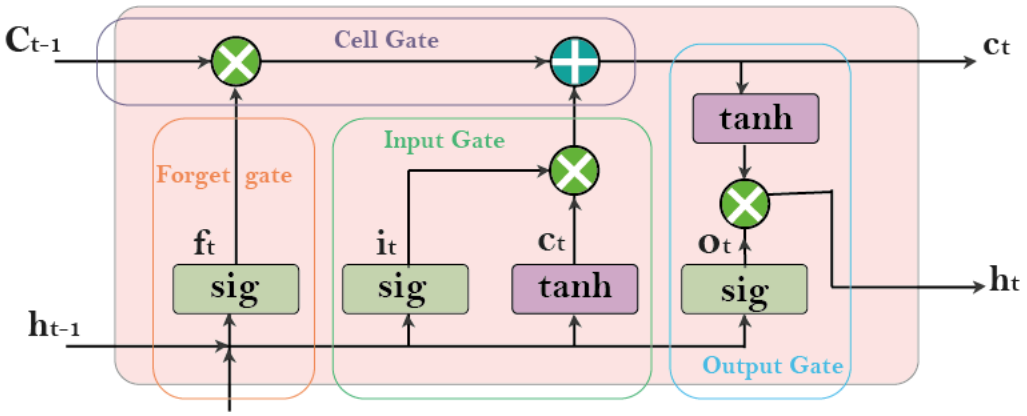

Figure 3.2: LSTM Architecture

The forget gate determines what knowledge must be recalled and what information may be forgotten. There is also an activation function named sigmoid function which processes data coming from the current input that is x(t) and the hidden state that is h(t-1). The output of Sigmoid is a number between zero and one, depending on the parameters. It decides whether or not a chunk of the previous output is necessary. Cells utilize this value of f(t) to point-by-point multiplication.To keep track of the current state of the cell, the input gate runs the following programs. To begin with, the current state X(t) and the previously concealed state h(t-1) are fed into the next sigmoid function. Afterwards between zero and one, which is very significant (not-important). The tanh function will then get the identical data from the hidden and current states. The tanh operator creates a vector (C(t)) with all conceivable values between -1 and 1 in order to govern the network. Point-by-point multiplication may be performed on the values obtained by the activation functions.

The next step is to determine how the data from the new state should be stored in the cell state. As a consequence, the preceding cell state, C(t-1) and forget vector f(t) gets multiplied. In the event that the result is 0, the values in the cell state will be cleared. The next step the model does is take the value produced by the input vector i(t) and conducts point-by-point addition on it. Later it updates the cell state. This gives the network a new cell state that it may refer to as C(t).

The output gate is the one that decides what the value of the next hidden state will be. This state is responsible for storing all of the previous inputs. To begin, the third sigmoid function is fed the values of the current and prior concealed states. After that, the tanh function is applied on the new cell state that was generated from the existing cell state. A point-by-point multiplication of these two outputs is performed. The network chooses what information the concealed state should convey based on the final result. Predictions are based on this secret state.

At last, the newly established cell state as well as the newly established hidden state are carried on to the subsequent time step. In conclusion, the forget gate identifies whatever knowledge from the preceding stages is required for the current step's outcome. Gates in the input and output sections are responsible for determining what information may be added to the next concealed state.

Additionally, CNN has the capacity to learn spatial characteristics, and it was able to provide excellent results in image processing in our scenario. The ability to extract spatial and temporal characteristics of data from CNN and LSTM collaboration will lead to an enhancement in the detection system's overall accuracy.

3.4 Facial Landmark

In our system, face component analysis will be essential in recognizing legitimate stalking occurrences. As a result, facial landmarks can be a very effective technique for analyzing facial points.
Face landmarks are clusters of distinctive facial features that can be used to identify a person. Only a few facial landmarks must be identified one by one on the face. Such as the eyes, nose, mouth, and eyebrows. Face recognition, face tracking, gesture understanding, and face registration all necessitate the use of accurate facial landmarks and analysis of facial characteristics. Face-related tasks such as gaze detection, expression understanding, face recognition, face tracking and face registration are all impacted by these key functions. It's easier to think of primary and secondary landmarks as two separate categories of facial landmarks. With these devices, there is a wide range of accuracy that is dependent on the application. Facial recognition and face tracking operations benefit from primary landmarks because they are more precise. On the other hand, the secondary group of landmarks has a greater impact on facial expressions. While the distinction between these two responsibilities is not always crystal clear.

Figure 3.3: Output of Facial Landmark Detection

To extract facial landmarks, a variety of methods are employed. These methods include Convolutional Neural Networks (CNN), Robust Discriminative Regression (RDR), Semantic Segmentation, and so forth. Despite the fact that all of these extraction methods are not completely optimized, all have their own set of drawbacks that need to be considered. Face recognition algorithms are used in conjunction with neural networks and fuzzy logic to improve the recognition rate of faces. Object detection, tracking, and alignment are all made easier with landmark annotation. Pre-information can be used in two ways prior to doing image annotation.One is semantic information, such as a person's ID for facial recognition or the name of an object for content-based image retrieval. Other than that, there is geometry/landmark data. The aim of the algorithm is to detect facial components such as eyes, eyebrows, nose and nose holes, as well as mouth and face contours from a captured face, which is accomplished through the placement of landmarks on each component. As part of our approach, 68 points are assigned to six specific components of the face. Face shapes range from 1 to 17, left eye brows range from 18 to

22, right eye brows range from 23 to 27, left and right eyes range from 37 to 48, and so on. The nose can be found in the range of 28-36, and the mouth shape can be found in the range of 49-68 [26]. Following the successful detection of the landmarks that correspond to our recruitment, we will proceed to the next step, which will be the determination of the facial angle, alignment, and progression, among other things.

3.5 Head Pose Estimation

We were able to distinguish important face features such as the nose, mouth, eye, eye brows, and facial contour using our facial landmark implementation. We take the typical 3D coordinates of those facial landmarks and attempt to estimate the nose tip's rotational and translation vectors.

Figure 3.4: Output of Head Pose Estimation

Once we have the required vector, we may project the three-dimensional points onto a two dimensional surface that matches to the image. The Euler angle can be used to indicate the rotation vector (roll, pitch, yaw). We will use the word coordinate approach to convert the 2D location of a feature point to a 3D location. Below are the coordinates.

1. Tip of the nose (0.0, 0,0,0.0)
2. Chin (0.0, -330.0, -65.0)
3. Left corner of the left eye (-225.0f. 170.0f, -135.0)
4. Right corner of the right eye (-225.0, 170.0, -135,0)
5. Left corner of the mouth (-150.0, -150.0, -125.0)
6. Right corner of the mouth (150.0, -150.0, -125.0).

3.6 Random Forest

Use of Random Forest classifier, which is generated from decision trees that have been constructed using diverse subsets of the provided information, may enhance predicted accuracy. The random forest model is able to make more accurate predictions of outcomes than it could if it relied solely on a single decision tree because it makes use of multiple decision trees. The greater the number of individual trees in the forest, the higher the accuracy will be and the lower the likelihood that it will be overfit. When Segal, Mark R [6] utilizing random forest tree predictors, the x represents the observed input (covariate) vector of length p and is associated with a random vector X. The k are independent random vectors that are identically distributed. Our primary concentration is on numerical regression, but we do occasionally discuss classification problems (categorical outcomes). The observed data, which will be used for training, are presumed to be independent of one another in this model. As a result of the random forest combining the predictions of multiple trees in order to determine the classification of the dataset, it is possible that some decision trees will predict the correct output while others will not. When all of the trees are taken together, however, they are able to correctly forecast the outcome. As a consequence of this, the two assumptions that will make a better Random forest classifier are as follows:

1. It is necessary for there to be some actual values in the feature variable of the dataset in order for the classifier to be able to make accurate predictions rather than guesses.

2. The predictions from each tree must have very low correlations with one another.

The first stage is to construct a random forest by combining N decision trees, and the second step is to generate predictions for each tree that was constructed in the previous phase. Follow these steps to get things done at work: Make a random selection of K data points from the training set Use the data you've selected to create decision trees (Subsets). If you want to create N decision trees, enter N in the appropriate field. Continue the process a second time. Assign the new data points to the category with the most votes from each decision tree's predictions.Random Forest is an algorithm that can be applied to both classification and regression analysis. It is capable of handling large and highly dimensional datasets. As a result, the model's accuracy improves, and the problem of overfitting is avoided. Less reliant on training data. However, despite its versatility, random forest is not better suited for regression than classification.

3.7 KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms and is based on the Supervised Learning technique. A new case/data is compared to the available cases using a KNN algorithm and placed in the most similar category. When new data is entered, the KNN algorithm compares it to all previously saved data to determine classification. This means that the KNN algorithm can easily categorize new data. When dealing with Classification problems, the KNN algorithm

is most commonly used. Because this algorithm is non-parametric, that is why the underlying data is left completely open, hence no assumptions can be made about its accuracy or completeness. Because it does not instantly learn from the training set, it is referred to as a "lazy learner" algorithm. This dataset is stored in the KNN algorithm until fresh data is classified into a category that is strikingly similar to the newly discovered data. In the work of Zhou Yong, Li Youwen and Xia Shixiong [37], the classification of the test sample is established by comparing it to the K training samples, which are samples that are geographically near to the test sample. After that, it places the sample under the classification that has the best chance of being accurate.

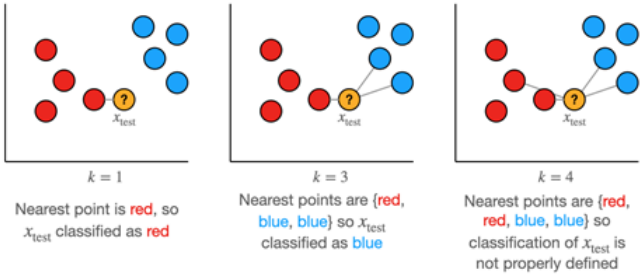

Figure 3.5: KNN classification algorithm [45]

It is not difficult to put it into practice. It can function well even with noisy training data. It's possible that the effectiveness of the training will increase with more data. It is imperative that the value of K be established at all times, despite the fact that doing so is not always easy. The distance between each pair of data points in every training sample must be computed, which contributes to the high cost of the computation.

3.8 SVM

A state-of-the-art big margin classifier is known as a Support Vector Machine (SVM). This classifier can categorize data points even if they are not otherwise linearly differentiable because of the high-dimensional feature space [5]. Through the use of support vector machines (SVMs), it is able to fully capitalize on the inherent geometric intuition. Once a solution has been found, the SVM is a good generalization property of the solution that ensures it is the unique (global) solution. In addition to this, there is also a common ground or formulation for the class separable and the class non-separable problems and it is accomplished by the integration of appropriate penalty factors of arbitrary degree into the optimization cost function [7]. Overall, SVM is capable of handling classification as well as regression on linear and non-linear datasets.

Chapter 4

Proposed Model

4.1 System Model

A typical algorithmic workflow for our research has been demonstrated below. Our data required preliminary processing before being incorporated into the model. The work flow illustrates how we compared two approaches to video classification and selected the optimal algorithm for detecting stalking situations.

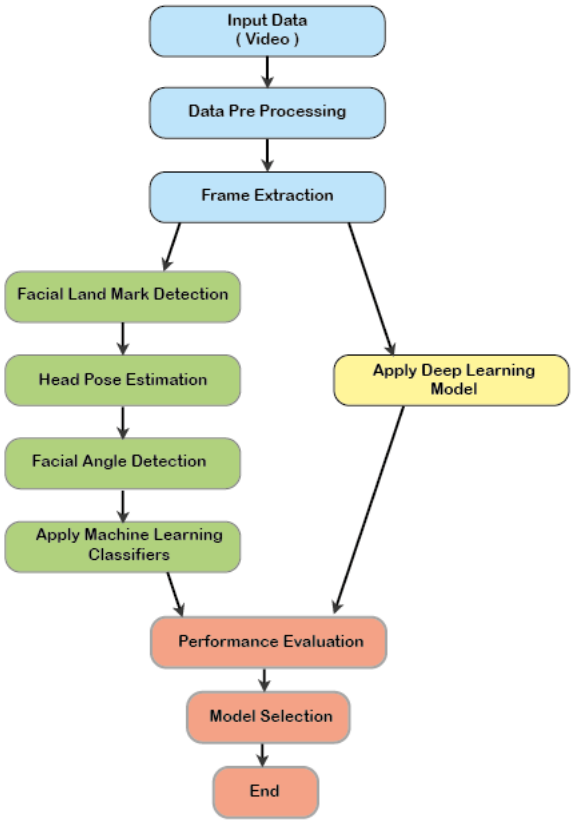

Figure 4.1: Top-level view of the proposed model

4.2 Data Collection

A study's correct outcome is contingent upon the collecting of data. Without a proper dataset, it is hard to conduct an impactful study. Despite the fact that we used multimedia data (video) in our study, the majority of researchers used pre-prepared data sets. We required video evidence of one person stalking another for our study. This type of film is extremely difficult to get. Not every location has access to this type of data. We began collecting information from a range of various sources on an individual basis. YouTube is a significant source of data points. However, the stalker's face was obscured in most of the video footage. As a result, we were unable to retrieve a sufficient amount of data from YouTube. Additionally, we acquired footage from a number of feature films and television series. We clipped very short footage from the films according to our needs. The video recordings are approximately 9-10 seconds in length. The quality of the video 640×480p. Despite searching approximately 150-200 videos, we did not obtain a greater amount of data. We excluded data in which the face of a suspected person is obscured. In order to strengthen our reliance in the data, we asked feedback from everyone, as well as from our supervisor.

4.3 Data Pre-processing

Our dataset required some form of pre-processing, which included applying Mask-R-CNN on the video footage then extracting frames from them. Mask-R-CNN has been used to each video to eradicate the background. After that, per video we retrieved 20 frames. Below is a brief summary of Mask-R-CNN.

4.3.1 Mask-R-CNN

As deep learning technology has developed rapidly in recent years, computer vision researchers have achieved great progress in object identification. There are two types of deep learning models now used in object detecting techniques. One phase in the detection process begins with the creation of a region proposal, followed by a categorization of all possible regions inside that area. In most cases, this will need to be tweaked to fit the specific area. R-CNN, Fast R-CNN, and Mask-RCNN [38] are all members of the CNN family. As for the second, there are single-stage detection methods like YOLO or SSD that do not need the area proposal step but instead produce the object's class probability and location coordinates immediately.

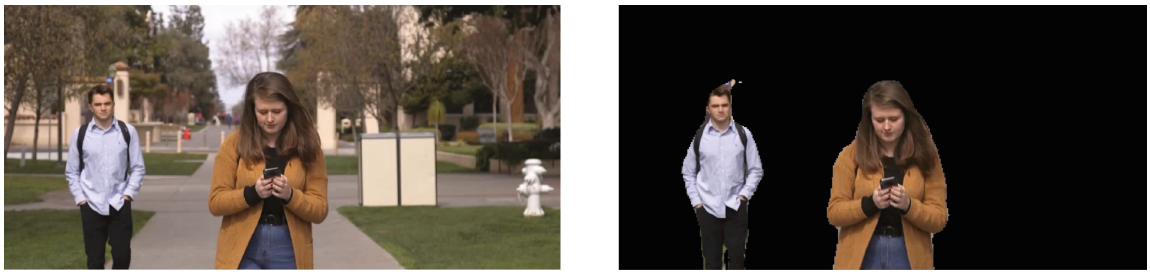

Figure 4.2: Background subtraction using Mask-R-CNN

Yolo's method is efficient, but it has a few drawbacks, such as the fact that it can produce false positives in the background area. It also has a low average position calculation compared to the CNN family, making it a poor performer overall. In addition to this, Yolo has a lesser recall and a higher position error rate[38]. When compared to other CNN family members, R-CNN is more complex and takes up a significant amount of physical memory. As a result of the fact that each region proposal in R-CNN needs to extract features from the CNN network, the running speed is rather slow [29]. Because Fast R-CNN selective search is used to extract regional proposals, it is also very time consuming [12]. Yolo is faster than Mask R-

CNN. In an experiment on basis of time consumption Yolo need 5.48544 ms average while Mask R-CNN used 67.632 ms [16]. On the other hand, Mask R-CNN has a very high recognition rate, which is significantly higher than YOLO's. In paper [29], by using 2049 images Yolo given the recognition rate 78% while in Paper [36] Mask R-CNN provided recognition rate 95.70% which is very rich.

Mask R-CNN (regional convolutional neural network) is an extension of Faster R-CNN. The backbone (ResNet, VGG, Inception, etc.) and the region proposal network make up the first stage. These networks run once per picture to provide a collection of region recommendations. Region recommendations are parts of the map that include the item.

The second step predicts the bounding boxes and object classes for each of the indicated regions created in stage one. Each suggested area may be of any size, but fully linked layers in networks need a fixed-size vector to make predictions. Based on either RoI pooling or RoIAlign, potential areas are sized.

Using Quicker R-CNN for object class and bounding box prediction is much faster.Mask R-CNN predicts segmentation masks for each area of interest by adding an extra branch. Based on the Faster R-CNN algorithm, this is a development of the original (RoI). The output of the RoIAlign layer is then sent to the Mask head, which has two convolutional layers. It creates masks for each region of interest (RoI), making it possible to segment a photo pixel by pixel.

In terms of picture segmentation, the Mask R-CNN is a convolutional neural network (CNN). This variation of a deep neural network recognizes items in a picture and provides a high-quality segmentation mask for every occurrence. Mask-R-CNN is very powerful algorithm for extracting feature from any image. Mask R-CNN algorithm is able to handle object detection and image segmentation operation together. So, segregating multiple objects in a single image is not a big deal by using Mask-R-CNN algorithm.

4.4 Methodology

4.4.1 VGG16

Visual Geometry Group(VGG) is a conventional deep Convolutional Neural Network architecture which consist of several layers and a stack of compact convolutional filters. It serves as the basis for ground-breaking object identification model which in turn helps attain high accuracy on large-scale images. In most cases, it is built using 13 layers of convolutional layers and three layers of completely linked layers [30]. The VGG network is just a deep convolutional neural network that was designed with great attention given to the optimal configuration of the network's layers depth in order to avoid increasing the network's complexity while simultaneously training more accurate and efficient models. This model is determined by increasing the number of feature maps or convolutions in the network as its depth grows. There are two distinct models available in VGG which are VGG16 model and VGG19 model.

The VGG16 network is a variant of the VGG network that consists of 16 convolution layers and has a small receptive field of 3 x 3. The VGG16 network is very comprehensive, as it has around 138 million individual parameters in total. The architecture of VGG16 is relatively straightforward and highly consistent. It has a total of 5 Max pooling levels with a dimension of 2x2 and 3 completely linked layers that come after the layer with the max pooling layer. After this, there are three layers that follow, all of which are linked to one another and a softmax classifier is used to complete the classification process. In addition, the procedure of activating ReLu is performed on each and every hidden layer.

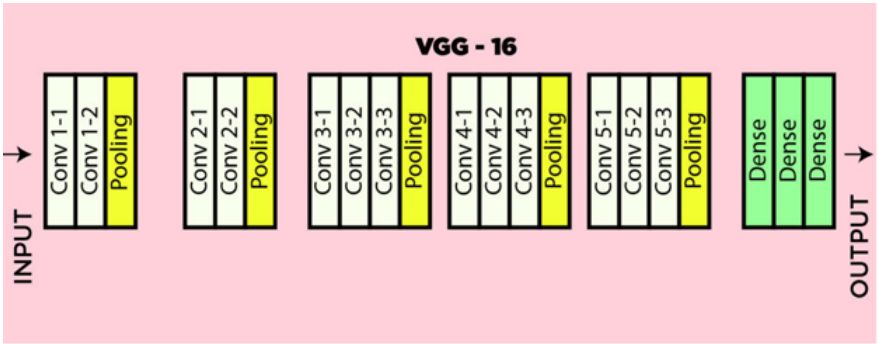

Figure 4.3: VGG-16 architecture

4.4.2 ResNet

In order to solve challenging problems such as image recognition and image classification, it is essential to construct deeper neural networks with additional layers to improve classification and recognition accuracy. The purpose of adding additional layers is to allow those layers to learn ever more convoluted features as they are implemented. There is a maximum depth threshold that can be reached while using the standard convolutional neural network model. But, there is a problem that arises

when we add more layers to the standard convolutional neural network model. As soon as deeper networks start to converge, a degradation problem is identified. This problem occurs when the network depth increases and as a consequence accuracy ultimately reaches a saturation point and subsequently rapidly declines [17]. There is evidence to indicate that a deeper network will have larger levels of training error and test error [17]. As a result of this, residual neural network (ResNet) is used in order to improve accuracy while simultaneously reducing the amount of errors encountered during training and testing. The ResNet is a type of feed-forward neural network wherein every layer is linked to other layer in the network [24]. With the help of ResNet, the challenge of training very deep networks has been significantly simplified. Resnets are often formed using Residual Blocks as its building blocks which is giving below.

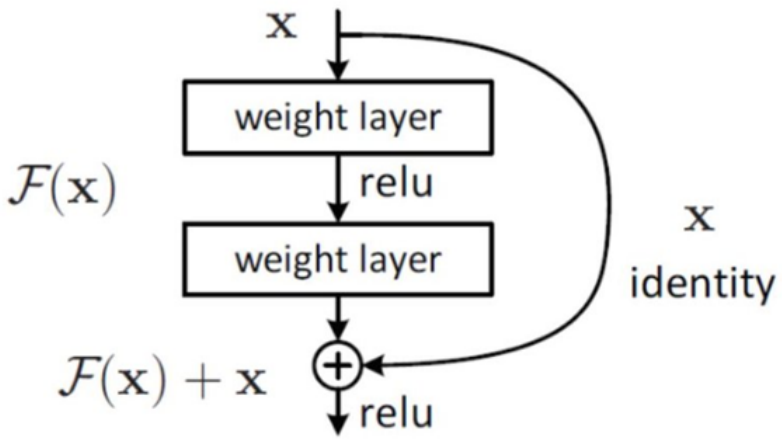

Figure 4.4: ResNet Residual Block

Theoretically, the stacked nonlinear layers fit another mapping of by identifying the required underlying mapping as H(x) in order to fit another mapping. After that, the initial mapping is transformed into the form F(x)+x. The F(x) behaves similarly to a residual, from where the term "residual block" comes. The formulation of F(x) +x can be realized in Residual Blocks with the use of feedforward neural networks with skip connections. The skip connection is the fundamental component of residual blocks which enables the skipping of one or more layers. The execution of identity mapping is the primary objective of the skip connections, and these outputs are being added to the stacked output layers without adding any additional parameters or complexity to the computational process. The issue of vanishing gradients in deep neural networks can be overcome with ResNet's skip connections, which give an alternate shorter channel for the gradient to flow through [21]. In addition to this, it is helpful for the model to get an understanding of the identity functions, since this assures that the top layer will operate effectively in the same manner as the lower layer.

As a whole, we can claim that ResNet makes it easier to link any two layers with the same feature-map size through direct connections. In addition to that, it can automatically scale up to a large number of layers without demonstrating any op-

timization challenges. As a result of this, we can achieve more accuracy with fewer errors during training and testing.

4.4.3 ConvLSTM

Convolutional LSTM basically extension of FC-LSTM [13]. FC-LSTM shows too redundancy for spatial data. To resolve these issues Convolutional LSTM has been initiated. Both the input-to-state and the state-to-state transitions in a convolutional LSTM have a structure that is similar to that of a convolution. By layering numerous Convolutional LSTM layers and building an encoding-forecasting structure, we can construct a network model not just for precipitation nowcasting problem but for other general spatiotemporal sequence forecasting issues. When dealing with spatiotemporal data, FC-LSTM suffers from the use of complete connections in transitions between inputs and states when no spatial information is recorded. As a solution to this problem, a disctinct aspect of our design is that the Convolutional LSTM's inputs X1,..., Xt, cell outputs C1,..., Ct, hidden states H1,..., Ht, and gates it, ft, ot are all three-dimensional tensors whose final two dimensions are spatial (rows and columns) [13]. In Convolutional LSTM structure, firstly, it transforming 2D image into 3D tensor. Before performing the convolution, It is necessary to provide padding in order to guarantee that the outputs have the same number of rows and columns as the inputs. By analyzing the inputs and past states of its neighbors, the Convolutional LSTM predicts a cell's future state. The below figure demonstrates the design of our proposed convLSTM architecture:

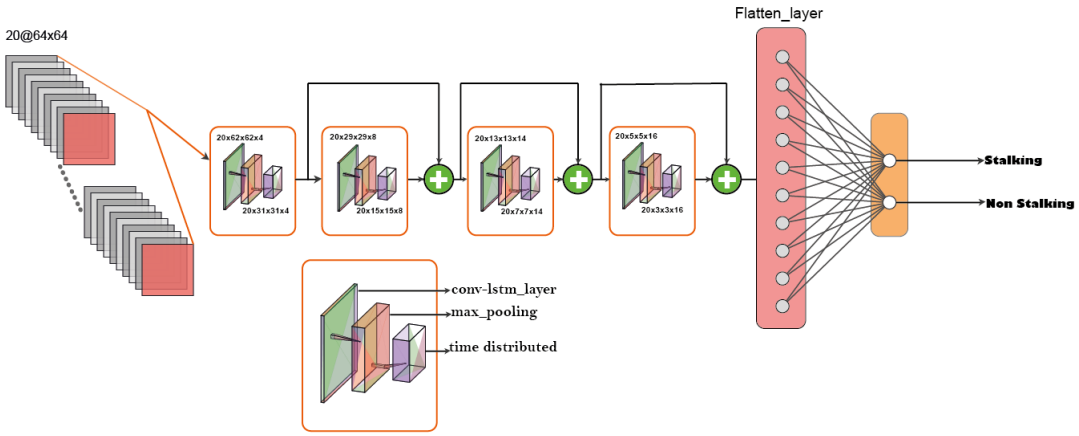

Figure 4.5: ConvLstm architecture

This convLSTM architecture is divided into some layers which are given below:

1. In the convLSTM layer, the matrix multiplication step that is normally performed at each gate in the LSTM cell is replaced with a convolution operation.

2. MaxPooling layers are being used to decrease data size in the images being processed. Basically, max pooling is a pooling technique for determining the maximum value for a feature map's patches. Additionally, it leads in a faster convergence rate by selecting better invariant features consequently improving generalization performance.

3. The TimeDistributed layer provides support for the manipulation of data in the form of time series or individual frames extracted from a frame. Because of this, it is feasible to make use of a layer for each of the inputs.

4. The matrix can be easily converted into an output vector with the help of a flatten layer. The process of flattening a layer usually includes converting each of the dimensional arrays that are created as a consequence of pooling feature mappings into a single extended continuous linear vector.

5. Neurons that are connected to each other are found in the dense layer. It can be found at the very end of the convLSTM architecture. In order to modify the dimension of the vectors, a dense layer must be employed and each neuron must be involved in this process.

The suggested convLSTM that will be implemented in this research will have an architecture that comprises of 4 blocks, followed by a flattening layer and then a dense layer. Each block is made up of three layers: the convLSTM layer, the MaxPooling3D layer and the TimeDistributed layer. Even though there are the same number of frames in each block, their individual sizes are becoming smaller as additional layers are added and there are more channels as a result. Images with dimensions of 20 x 64 x 64 x 3 are fed into the model as input (number of frames, width, height, number of channels). Following that, the first convLSTM layer decreased the amount of pixels in the input to 62 x 62 (height x width) while simultaneously increasing the number of channels to 4. A MaxPooling3D layer can be found after the first convLSTM layer. This layer halves the size of the frames, making them 31 x 31(height x width) without affecting the total number of channels. The number of frames in the second convLSTM layer of the second block remains the same, but the size of the pixels has been decreased to 29 x 29 and the number of channels has been increased from 4 to 8. Following the completion of MaxPooling3D, it will drop the pixel count to 15x15, much like the initial MaxPooling3D layer. In a way quite similar to the previous blocks, the third and fourth convLSTM layers each have a size that is 13x13 and 5x5 correspondingly. Following the third and fourth convLSTM layers, there are MaxPooling3D levels that have pixel sizes of 7x7 and 3x3 correspondingly, the same as the preceding MaxPooling3D layers. After the first four blocks, there will be a layer that has flatten layer, followed by a dense layer and these are located at the very end of the network. The process of flattening involves transforming the data into a one-dimensional array that consists of 2880 neurons so that it can be input into the dense layer. Finally, in the dense layer, there are just two neurons, and the output indicates whether or not there is a stalking scenario.

Chapter 5

Result And Analysis

We have gone through mainly in 2 types of classifications. One is through the Dynamic feature extraction and another is through the Manual feature extraction. Among these two approaches, ResNet, VGG16, and ConvLSTM falls under Dynamic feature extraction and SVM (Support Vector Machine), KNN (k-nearest neighbors), and Random Forest goes to Manual feature extraction. To measure the robustness and efficiency of any model, some parameters are needed which includes accuracy, precision, recall, loss, F1 score, and support. These parameters are defined as follows:
Precision: The ratio of the number of true positives to the total number of true and false positives is the definition of precision.
Recall: The term "recall" refers to the proportion of true positives relative to the total number of true positives as well as false negatives.
F1 Score: The F1 is a normalized harmonic mean that takes into account both precision and recall. The better the projected accuracy of the algorithm is going to be, the nearest the value of the F1 score is to being equal to 1.0.
Support: The number of times the class really appears in the dataset is referred to as its support. It is not dependent on the particular model being analyzed; rather, it analyzes the performance evaluation process.

5.1 Manual Feature Extraction

We utilized the face landmark and head position estimation approach to manually extract features. We received 68 points for face landmarks but only picked six of them. These are the nose tip, the chin, the left eye left corner, the right eye right corner, the left mouth corner, and the right mouth corner. We obtained these six points from the people in each frame. These values were obtained in terms of x and y coordinates. As a result, each individual has 12 coordinate values. We also obtained two essential pieces of information from head pose estimation: Yaw and Roll, where Yaw represents the left and right pose angle and Roll represents the in-plane rotation angle. Each individual received one roll and one yaw value. So we collected a total of 28 points (12*12 co-ordinate value and 4 roll and yaw values for two people) from each frame with a victim and a stalker. We classified our frame as stalking or non-stalking based on these 28 characteristics. These labelled frames were used to train the human feature extraction algorithms.

Used Model	Accuracy	Precision	Recall	F-1 Score
SVM	68%	68%	67%	67.5%
KNN	77%	68%	55%	46%
Random Forest	81%	81%	82%	81%

Table 5.1: Performance of Manual Feature Extraction Algorithms

SVM provided us with the lowest accuracy of 68 percent. And the precision, recall, and F1 score are 68, 67, and 67.5 percent, respectively. KNN produced somewhat better results than SVM. We obtained an accuracy of 77 percent, precision of 68 percent, recall of 55 percent, and F1 score of 46 percent via KNN. However, when compared to these two algorithms, Random Forest produced the best results. We received an accuracy of 81%, precision of 81%, recall of 82%, and F1 score of 81%. However, these outcomes fall short of our expectations. As a result, we've moved on to Dynamic feature extraction.

5.2 Dynamic Feature Extraction

To develop a dynamic feature extraction architecture, batch size and epoch must be specified. We have used several batch sizes and epochs to train our dataset with various architectures. In VGG16, for instance, the batch size and epoch are 32 and 5, respectively. Same epoch and batch size should also apply to ResNet. We choose an epoch size of 25 and a batch size of 8 to provide the highest degree of precision in Conv-LSTM.

5.2.1 ResNet50

We have tried ResNet up to epoch 5 level. Taking further epochs is a waste of time and resources since the rate of accuracy improvement has paused. By this point in time, we have achieved 67.35 % accuracy. Other parameters' values are 66% for precision, 61% for recall, and 60% for F1-score.

We can easily observe that the accuracy decreases with each increasing epoch. This is an obvious evidence of an overfitting issue, as shown in the graph above.

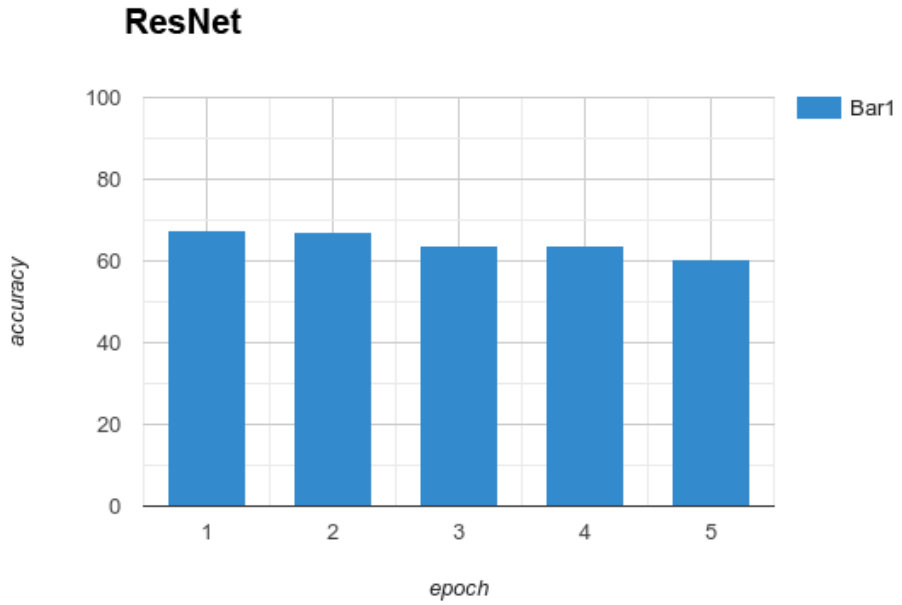

Figure 5.1: Bar Graph Of Accuracy For ResNet

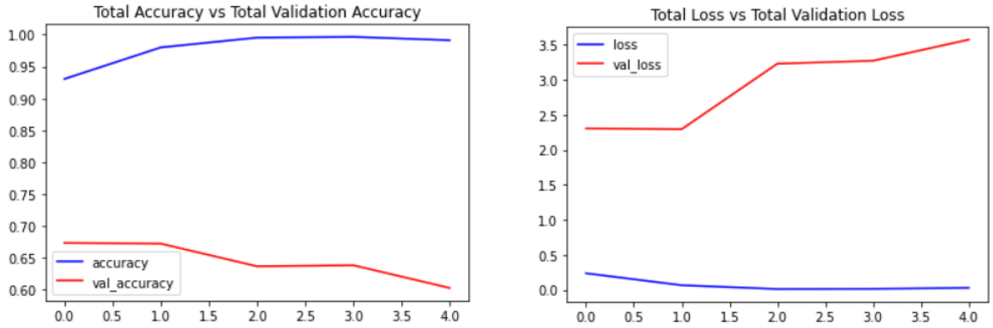

Figure 5.2: Training Accuracy vs Testing Accuracy And Training Loss vs Testing Loss Of ResNet

5.2.2 VGG16

We also used VGG16, which produced somewhat better results than ResNet but was still insufficient. Gain accuracy is 77.38 %, while precision, recall, and F1-score are 89%, 86%, and 87% respectively.

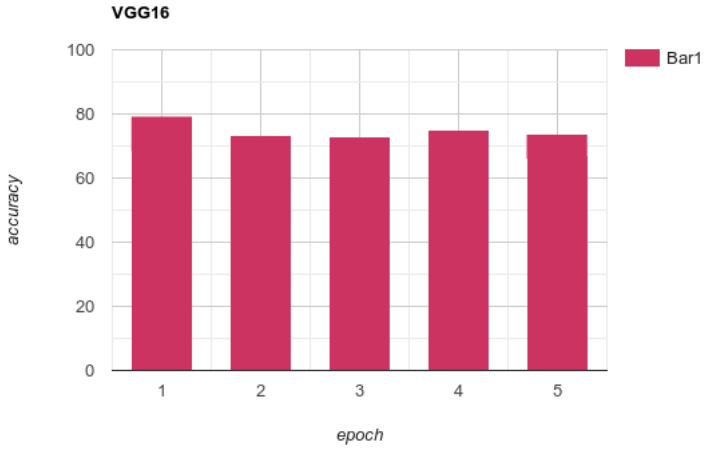

Figure 5.3: Bar Graph Of Accuracy For VGG16

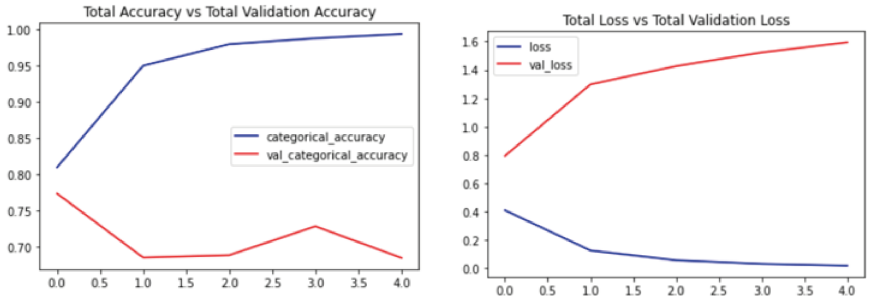

Figure 5.4: Training Accuracy vs Testing Accuracy And Training Loss vs Testing Loss Of VGG16

We couldn't acquire any acceptable precision in VGG16 since the validation loss was too high. VGG16 is solely concerned with spatial features. It is not intended to extract temporal functions. As a result, we picked the ConvLSTM model.

5.2.3 ConvLSTM

Training loss is concerned with the fitting of training data, while validation loss is concerned with the fitting of fresh data. We discovered an accuracy of 88.57 percent in our Conv-LSTM model, which is higher than any other model.Furthermore, to demonstrate the model's overall performance, we determined the accuracy, recall, and F1-score values. The numbers are 68%, 67%, and 65% respectively.

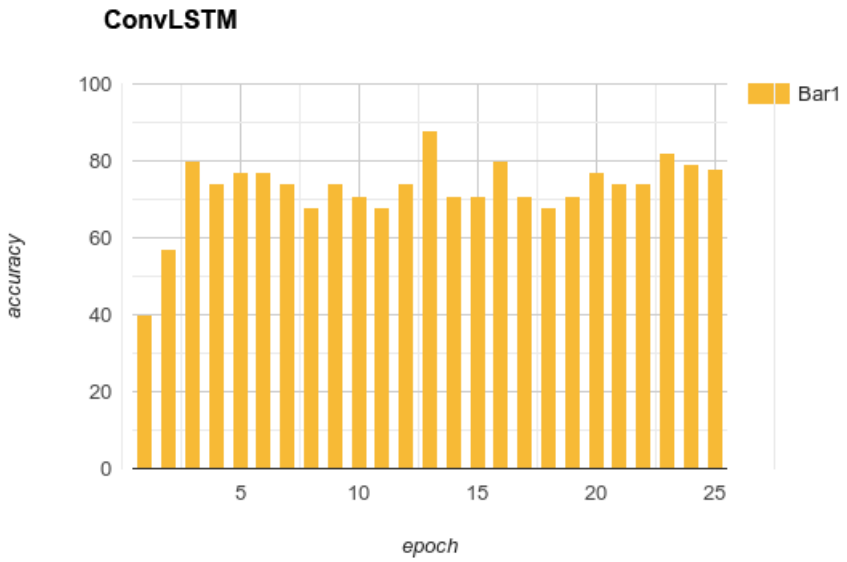

Figure 5.5: Bar Graph Of Accuracy For Convlstm

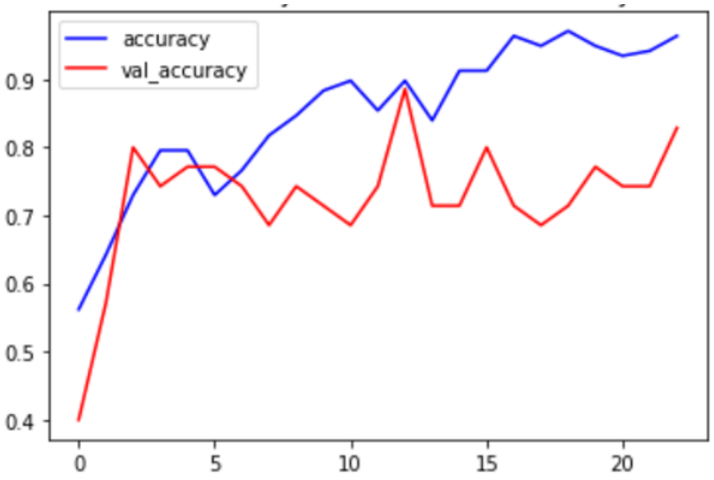

Figure 5.6: Training Accuracy vs Testing Accuracy Of ConvLSTM

After assessing all of the manual feature extraction algorithms, we found that Random Forest approaches had the best accuracy (81%). Furthermore, we excluded ResNet because to its overfitting issue and VGG16 due of its lesser accuracy. And, among dynamic feature extraction algorithms, ConvLSTM achieved the greatest accuracy of 88.57%.

Used Model	Accuracy	Precision	Recall	F-1 Score
ResNet	67.35%	66%	61%	60%
VGG-16	77.38%	89%	86%	87%
ConvLSTM	88.57%	68%	67%	65%

Table 5.2: Performance of Dynamic Feature Extraction Algorithms

Even though we trained extracted features in several machine learning classifier models after manual feature extraction, these models have certain drawbacks. Because we classified our features frame by frame, there is a possibility that the same video frame may appear in both the training and testing datasets, causing machine learning bias. We chose ConvLSTM over Random Forest based on this reasoning. As previously stated, frames in Random Forest may mix across the defined classes, but there is no such possibility in ConvLSTM since we trained the ConvLSTM model in a 3D architecture where the input takes all of the frames of each video. Furthermore, the spatio-temporal characteristic is another reason to choose ConvLSTM over other algorithms.

Chapter 6

Conclusion

Wrapping up the whole paper, if we analyze minutely, we can observe that before occurring any kind of crime there will a plan by the criminal, and in that plan, stalking is a common and must thing to happen first. As, this is typically one of the primary levels of a crime if this can be prevented a huge sum of a misdemeanor can be diminished like ladies badgering, on-road burglary, etc. There has been a lot of research done on stalking and other types of suspicious behavior, but this study use a hybrid model that combines CNN and LSTM neural networks in an effort to get more accurate results than previous research has managed to achieve. Mask-R-CNN made a significant role in order to pre-process the video footages. However, more research and making this model more hybrid is suggested to increase the accuracy of the objective.

Bibliography

[1] K. M. Abrams and G. E. Robinson, "Stalking part i: An overview of the problem," *The Canadian Journal of Psychiatry*, vol. 43, no. 5, pp. 473–476, 1998.

[2] ——, "Stalking part i: An overview of the problem," *The Canadian Journal of Psychiatry*, vol. 43, no. 5, pp. 473–476, 1998.

[3] ——, "Occupational effects of stalking," *The Canadian Journal of Psychiatry*, vol. 47, no. 5, pp. 468–472, 2002.

[4] W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in *2004 IEEE international conference on multimedia and expo (ICME)(IEEE Cat. No. 04TH8763)*, IEEE, vol. 1, 2004, pp. 719–722.

[5] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, IEEE, vol. 3, 2004, pp. 32–36.

[6] M. R. Segal, "Machine learning benchmarks and random forest regression," 2004.

[7] M. E. Mavroforakis and S. Theodoridis, "A geometric approach to support vector machine (svm) classification," *IEEE transactions on neural networks*, vol. 17, no. 3, pp. 671–682, 2006.

[8] A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda, "Detecting uncommon trajectories," in *2008 Digital Image Computing: Techniques and Applications*, IEEE, 2008, pp. 398–404.

[9] L. He, D. Wang, and H. Wang, "Human abnormal action identification method in different scenarios," in *2011 Second International Conference on Digital Manufacturing & Automation*, IEEE, 2011, pp. 594–597.

[10] R. D. MacKenzie, T. E. McEwan, M. T. Pathé, D. V. James, J. R. Ogloff, and P. E. Mullen, *Stalking: Ein Leitfaden zur Risikobewertung von Stalkern-das" Stalking Risk Profile"*. Kohlhammer Verlag, 2014.

[11] J. M. Pang, V. V. Yap, and C. S. Soh, "Human behavioral analytics system for video surveillance," in *2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014)*, IEEE, 2014, pp. 23–28.

[12] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5325–5334.

[13] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[14] C. Tiwari, M. Hanmandlu, and S. Vasikarla, "Suspicious face detection based on eye and other facial features movement monitoring," in *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, IEEE, 2015, pp. 1–8.

[15] R. Bhattacharyya, "Street violence against women in india: Mapping prevention strategies," *Asian Social Work and Policy Review*, vol. 10, no. 3, pp. 311–325, 2016.

[16] C. A. Cuevas and C. M. Rennison, *The Wiley handbook on the psychology of violence*. John Wiley & Sons, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] A. Schetzer, *Street harassment of lgbti people rife, la trobe university study finds*, Apr. 2016. [Online]. Available: https://www.theage.com.au/national/victoria/street-harassment-of-lgbti-%20people-rife-la-trobe-university-study-finds-20160406-gnzq0z.html.

[19] S. Shakya, S. Sharma, and A. Basnet, "Human behavior prediction using facial expression analysis," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, 2016, pp. 399–404.

[20] R. S. Sidhu and M. Sharad, "Smart surveillance system for detecting interpersonal crime," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, 2016, pp. 2003–2007.

[21] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[22] S. Bhattacherjee and G. Somashekhar, "Artificial intelligence to impart surveillance, tracking, & actuation on suspicious activities," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, IEEE, 2017, pp. 1–5.

[23] N. Fouad, R. Bubar, L. Jennings, *et al.*, "Report to the standing committee on the status of women faculty at colorado state university," *Colorado State University, Fort Collins*, 2017.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[25] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," in *2017 36th Chinese control conference (CCC)*, IEEE, 2017, pp. 11 104–11 109.

[26] G. Amato, F. Falchi, C. Gennaro, and C. Vairo, "A comparison of face verification with facial landmarks and deep features," in *10th International Conference on Advances in Multimedia (MMEDIA)*, 2018, pp. 1–6.

[27] N. Bisagno, N. Conci, and B. Rinner, "Dynamic camera network reconfiguration for crowd surveillance," in *Proceedings of the 12th International Conference on Distributed Smart Cameras*, 2018, pp. 1–6.

[28] R. L. Galvez, A. A. Bandala, E. P. Dadios, R. R. P. Vicerra, and J. M. Z. Maningo, "Object detection using convolutional neural networks," in *TENCON 2018-2018 IEEE Region 10 Conference*, IEEE, 2018, pp. 2023–2027.

[29] R.-C. Chen *et al.*, "Automatic license plate recognition via sliding-window darknet-yolo deep learning," *Image and Vision Computing*, vol. 87, pp. 47–56, 2019.

[30] M. F. Haque, H.-Y. Lim, and D.-S. Kang, "Object detection based on vgg with resnet network," in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, IEEE, 2019, pp. 1–3.

[31] Z. Kong, Y. Cui, Z. Xia, and H. Lv, "Convolution and long short-term memory hybrid deep neural networks for remaining useful life prognostics," *Applied Sciences*, vol. 9, no. 19, p. 4156, 2019.

[32] J. Liu, D. Yung, S. Nishimura, and T. Araki, "Stalker retrieval on surveillance videos using spatio-temporal coappearance," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, 2019, pp. 127–134.

[33] D. A. Maran, B. Loera, and A. D'Argenio, "Health care professionals' knowledge of stalking perpetrators, victims, behaviors, and coping strategies: A preliminary study among italian hospitals," *The Scientific World Journal*, vol. 2019, 2019.

[34] P. by D. Clark and F. 13, *Sexual harassment in germany 2018*, Feb. 2020. [Online]. Available: https://www.statista.com/statistics/1096411/street-harassment-germany/.

[35] A. G. B. Dandamudi, G. Vasumithra, G. Praveen, and C. Giriraja, "Cnn based aerial image processing model for women security and smart surveillance," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2020, pp. 1009–1017.

[36] W. Jia, Y. Tian, R. Luo, Z. Zhang, J. Lian, and Y. Zheng, "Detection and segmentation of overlapped fruits based on optimized mask r-cnn application in apple harvesting robot," *Computers and Electronics in Agriculture*, vol. 172, p. 105 380, 2020.

[37] Y. Liu, A. Huang, Y. Luo, *et al.*, "Fedvision: An online visual object detection platform powered by federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 172–13 179.

[38] J. Messing, M. Bagwell-Gray, M. L. Brown, A. Kappas, and A. Durfee, "Intersections of stalking and technology-based abuse: Emerging definitions, conceptualization, and measurement," *Journal of family violence*, vol. 35, no. 7, pp. 693–704, 2020.

[39] M. S. Rosenbaum, K. L. Edwards, B. Malla, J. R. Adhikary, and G. C. Ramırez, "Street harassment is marketplace discrimination: The impact of street harassment on young female consumers' marketplace experiences," *Journal of Retailing and Consumer Services*, vol. 57, p. 102 220, 2020.

[40] I. Shreya Raman, *In 2018, india reported a stalking case every 55 minutes. the actual number may be even higher*, Feb. 2020. [Online]. Available: https://scroll.in/article/952903/in-2018-india-%20reported-a-stalking-case-every-55-minutes-the-actual-number-may-be-even-higher.

[41] R. I. Sifat, "Sexual violence against women in bangladesh during the covid-19 pandemic," *Asian journal of psychiatry*, vol. 54, p. 102 455, 2020.

[42] T. R. Statista, "28.03. 2020 tarihinde https://www. statista. com/statistics/1108088/products-and-services-people-spend-more-ondue-to-the-covid-19-pandemic/. adresinden erişilmiştir. su, s., wong, g. ve shi, w.(2016). epidemiology, combination and pathogenesis of coronavirus," *Journal of Trends Microbiol*, vol. 24, pp. 490–502, 2020.

[43] M. Abdallah, N. An Le Khac, H. Jahromi, and A. Delia Jurcut, "A hybrid cnn-lstm based approach for anomaly detection systems in sdns," in *The 16th International Conference on Availability, Reliability and Security*, 2021, pp. 1–7.

[44] Y. Luo and B. Yang, "Video motions classification based on cnn," in *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, IEEE, 2021, pp. 335–338.

[45] Yousefnami, *Why does increasing k decrease variance in knn?* Nov. 2021. [Online]. Available: https://towardsdatascience.com/why-does-increasing-k-decrease-variance-in-knn-9ed6de2f5061.