T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

DETECTION OF HUMANS IN VIDEO STREAMS USING CONVOLUTIONAL NEURAL NETWORKS

Ameen Mudher Abbas ALDULAIMI

Master of Science

Supervisor

Asst. Prof. Dr. Sefer KURNAZ

Istanbul, 2021

DETECTION OF HUMANS IN VIDEO STREAMS USING CONVOLUTIONAL NEURAL NETWORKS

by

Ameen Mudher Abbas ALDULAIMI

Electrical and Computer Engineering

Submitted to the Graduate School of Science and Engineering

in partial fulfilment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2021

The thesis titled **"**DETECTION OF HUMANS IN VIDEO STREAMS USING CONVOLUTIONAL NEURAL NETWORKS**"** prepared and presented by **"**Ameen Mudher Abbas ALDULAIMI**"** was accepted as a Master of Science Thesis in Electrical and Computer Engineering.

_____

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Thesis Defense Jury Members:

Asst. Prof. Dr. Sefer KURNAZ	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Dr. Oguz KARAN	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Dr. Zeynep ALTAN	Faculty of Engineering and Architecture, Beykent University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Ameen Mudher Abbas ALDULAIMI

signature

DEDICATION

I devote and pledge this research work to my supervisor who is salient for guiding me through the whole research work as well as my family for always assisting me in my hard time.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Asst. Prof. Dr. Sefer KURNAZ for guiding and helping me along the way in writing this dissertation. Discussing my progress, problems, and ideas with my supervisor Asst. Prof. Dr. Sefer KURNAZ a couple of times every week helped me tremendously in understanding the logic behind the research. It made me better realize the technical need for this research work. Without his time and help none of this research would have been possible, and I'm grateful to him for giving me the opportunity to work on this interesting subject. I am also very thankful to my parents as well as my friends who always supported me through whole of this journey.

ABSTRACT

DETECTION OF HUMANS IN VIDEO STREAMS USING CONVOLUTIONAL NEURAL NETWORKS

ALDULAIMI, Ameen Mudher Abbas,

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: 06/2021

Pages: 61

In this thesis, I have focused on deep learning approaches like CNN for the detection of humans in video streams. This thesis has mainly focused on training the neural networks or other models with enough amounts of data so that it achieves desirable results. Starting with the basics of neural network where a neuron, the smallest unit in deep learning field, is defined and explained, I have elevated the research to the topic where I could detect the humans in video streams. First the neural networks have been introduced in this research and then training of the neural networks has been attained with both the CPU and the GPU. The main goal of this thesis is to grow a method for the detection of Humans in Video Streams Using Convolutional Neural Networks. This research intent to train the proposed system with Convolutional neural network on any open-source dataset of video objects. And for the methodology of proposed system, CNN would be utilized. CNN method is one of the most advance and efficient method of getting better results by combination of different agents to perform any task on set of rules. The Convolutional neural network Libraries will be used in python. Dataset for video-based human detection will be procured from any open-source repository like Kaggle like UCL or Open Sets and the name of the datasets are Pascal VOC 2007 and shopping dataset. For the implementation of this thesis, Python would be forwarded in use as it is the most efficient tool for processing large scale data with more accuracy. In an algorithm

called matrix form of back propagation, the use of multiple GPUs has been made along with CUDA kernel and cuBLAS library (Python Library). Application of human face recognition has been implemented using the pre trained model Facenet (face recognition system) and Deep Convolutional Neural Networks. After analyzing neural networks and its facial recognition application, other approaches of deep learning have been given force to. I have used the library OpenCV along with deep learning approaches to implement face recognition, Image registration and YOLO Object Detection and Recognition with the tensor flow and keras environment supported by anaconda for python. We have achieved the accuracy of 91.6%.

Keywords: CNN, YOLO, Deep Learning, Video Streams.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

AI	:	Artificial Intelligence
ANN	:	Artificial Neural Network
CNN	:	Convolutional Neural Network
DT	:	Decision Tree
FPS	:	Frame per Second
GPU	:	Graphic Processing Unit
IDS	:	Intrusion Detection System
ICS	:	Intrusion Classification System
LR	:	Logistic Regression
IoT	:	Internet of Things
ML	:	Machine Learning
RF	:	Random Forest
SVM	:	Support Vector Machine

1. INTRODUCTION

1.1 INTRODUCTION

With the ubiquitous deployment of surveillance cameras in public and private places, the use of such lenses for diverse uses such as crowd monitoring and abnormal activity detection is still in high demand. Seen cameras are usually installed in places where surveillance is necessary, such as malls, railroad and bus stops, office buildings, and academic facilities, due to its low cost. In addition, video monitoring methods are commonly used in informal places such as homes and child care centers to track different activities carried out by elderly babies and kids in order to improve safety. [1]. These cameras mounted in and around different places and locations help in detecting suspicious behavior and avoiding forthcoming dangers and threats to the common public. In addition to traffic identification, recognize faces, and pose estimation prediction, noticeable devices are used in exterior monitoring systems. They are also commonly used for object recognition, leading to the development of many databases for various region detection. Experiments were carried out on the database throughout that study for 2 purposes: first, the photos in this database are of an ideal size of educational purposes, and it is widely used in improving intelligence capabilities within academic institutions. The images captured from a camera in the dataset provides information about the pedestrians walking towards and away from the camera, while a second camera placed orthogonal to the first camera also collects the side view information. Therefore, the detection of abnormal events captured using the two cameras separately poses a challenge for both frame-based and patch-based detections. Several handcrafted features have been utilized in the past to detect abnormalities [2]. The recent evolution of machine learning techniques in various computer vision applications and the advancement in big data analysis and GPU computing provided an opportunity to implement a deep learning algorithm for human detection [3]. Deep convolutional neural networks were not intensively explored in the past for surveillance purposes as they require abundant image data sets in order to achieve the desired outputs. Therefore, the aim of this research is to utilize visible camera images for obtaining spatiotemporal features using the three-dimensional convolutional neural network (C3D) technique. There are two main advantages of the C3D network [4]. First, feature representation is automatically learned from the training data and second, it contains both spatial and temporal information. The motivation behind performing this research is to compare the abnormality detected from a deep

learning feature with that of the abnormal event detected from the baseline handcrafted features [5].This research will entail training the Pascal VOC 2007 and shopping dataset and Kaggle dataset and extracting features using C3D.

1.2 PROBLEM STATEMENT

Detecting humans in crowded environments have been always a challenging task for public safety and security. For monitoring any suspicious activities in the crowd, surveillance cameras play an important role for detection, recognition and tracking of the individuals. Surveillance video is a fundamental technology that provides higher-level applications for observing and analyzing different activities of the crowd. There are various events that occur within the crowd for which a video surveillance system is being set up. Among all, one of the events that have been monitored extensively from past several years is the abnormal event detection. Person identification by humans has long been regarded as an effective feature of a listening device for a range of security measures. Many scientists have used a variety of handmade tools to identify such anomalies in the scene. The main purpose of this research project was to evaluate the efficiency of a machine learning system and its implementation to social beings. For validating such performances on a dataset, two separate modes of human detections are considered, one is frame-based detection and another is patch-based detection. These two detection approaches have been widely used in literature for visualizing the outputs [6]. The abnormality in the video frames is visualized through frame-based detection whereas the locations of the abnormal patches in the frame can be detected through patch-based detection [7].

These days, in advance research for Human detection, the patch-based approach has been popularly used because of its higher accuracy rate than frame-based detection. For extracting features in patch-based detection, spatiotemporal non- overlapping patches are generally used for the purpose. However, there has not been enough investigation on utilizing features from spatiotemporal non-overlapping as well as overlapping patches for human detection using a deep neural network. Based on the study of literature of deep learning applications for human detection, the following research questions are required to be addressed [8].

i. How does the application of a deep learning technique for frame-based or patch-based human detection differ from the existing human detection obtained from handcrafted features?

ii. How does the extraction of deep learning features using non-overlapping spatiotemporal patches and overlapping spatiotemporal patches affect the final outcome?

1.3 AIMS AND OBJECTIVES

The main objective of this thesis is to develop a technique for the detection of Humans in Video Streams Using Convolutional Neural Networks. This research intent to train the proposed system with Convolutional neural network on any open-source dataset of video objects. And for the methodology of proposed system, CNN would be utilized. CNN method is one of the most advance and efficient method of getting better results by combination of different agents to perform any task on set of rules. The Convolutional neural network Libraries will be used in python. Dataset for video-based human detection will be procured from any open-source repository like Kaggle like UCL or Open Sets and the name of the datasets are Pascal VOC 2007 and shopping dataset. For the implementation of this thesis, Python would be forwarded in use as it is the most efficient tool for processing large scale data with more accuracy. Many existing deep learning methods like convolutional neural network (CNN) or deep CNN (CNN) are capable to cover only one part that is appearance based feature extractions. For motion information, other techniques such as optical flow, histogram of gradient (HOG), are applied. Both of these appearance and motion information gathered from different methods are then to be combined at the later stage to obtain human detection. This is a lengthy and time consuming procedure and thus it is required to choose three-dimensional convolutional neural network (C3D) which is capable enough to extract the features for both appearance and motion in the same context.

1.4 CONTRIBUTION

The contribution in this thesis that we used commonly used datasets for detection, such as ImageNet, Pascal VOC and Microsoft COCO are mainly based on static photographs. Some pedestrian detection benchmarks, such as Caltech and ETH, are recorded from vehicle-mounted

camera, which has a fixed point of view. The platform we used is Anaconda Navigator Also, some object tracking benchmarks, such as MOT and YouTube-Bounding Boxes Dataset, are not practicable for us due to different ways of annotation: the former annotates all kinds of objects, including non-human objects, while the latter is designed for single-object tracking, so that only one object is annotated for each video. However, detection in videos capturing natural viewing behavior, is made more challenging due to unpredictable movement of head, including rotation and fast motion, and motion blur becomes an unavoidable and most severe case in our dataset. The characters of data pattern are different from all the datasets mentioned above. Primarily, the main innovation of this thesis for the domain of machine learning and computer visions will be the integration and combination of new and existing video-based Human detection. The Convolutional neural network (CNN) will be used for the training and testing.

1.5 RESEARCH QUESTIONS

This thesis adopts single-frame detection for humans in video streams from head-mounted cameras, i.e. the detection is only based on the content of video at each frame and no temporal information is included. In this way, it can provide a prior study for the human tracking problem. On the other hand, owing to the promising development in CNNs, we strive to research on how robust different structures of CNNs (You look only once (YOLO) [6] and Single Shot Multibox Detector (SSD) [7]) perform on challenging videos with fast motion and strong blur, and how the structures may influence the results. The trend of solutions to detection problems with CNN technology can be summarized from Chapter 2: first of all, the accuracy of CNNs on detection problems is improving with carefully designed network structures. Most of the detection models are modified based on classification models. However, different from the classification problem, as more factors are supposed to be predicted, prior knowledge can be important when designing networks. For example, scale, region proposal and aspect ratio of objects. Secondly, rather than a domain specific model, a model with generalization ability is more popular in recent development. [9] adopts joint training strategy from two datasets so that it even extends the network with the ability to detect object with unseen classes. Thirdly, current technologies are already capable to detect objects at high speed as well as relatively high accuracy. This factor is of high importance in practical applications [10].

1.6 THESIS STRUCTURE

Chapter 1 will include an outline of the study, as well as a brief review of each chapter. The context information and related work presented in Chapter 2 will be important to the thesis's topics. This study uses Python to detect humans in video streams using convolutional neural networks, as well as different methods to monitor video artefacts.

Following this, several topics related to object tracking will be describe Theoretical outline of the approach proposed in this study will be presented in Chapter 3. It describes the technique's design as well as the new framework for application and contrast, and several sub-methods for application. The approach will be trained in Python and evaluated using a database of these Individual identification with deep neural networks architectures. This information is displayed as well. This chapter will present the findings as well as lab practice, which will be contrasted. The fifth chapter will discuss the thesis is summarized and concluded in the final chapter. This chapter includes a discussion of the findings, as well as the thesis's concerns and consequences, as well as a summary of possible studies that may be undertaken further to investigate the topics discussed, as well as a bibliography.

2. RELATED WORK

In this chapter, we organize the literature survey in a few sections. In the first section, we review the detection of humans in video streams and related concepts. The method of convolutional neural networks will be discussed. Work on video processing has been performed in the image processing domain for years and has addressed multiple issues such as pedestrian detection [11], unusual activity recognition [12], and correlation value [13]. [13]. There are different solutions that have been provided to tackle these problems, still, there is growing demand for developing a video descriptor which is generic, compact, efficient and simple to solve large-scale video tasks. The rapid progress of deep learning methods in the imaging technology has motivated various pre-trained convolutional neural network (ConvNet) models [14] for extracting image features. However, two-dimensional convolutional neural network (2D CNN) has some limitations when both appearance and motion are considered together for any human detections. The 2D filters used in CNN covers only the appearance-based detection and for motion features, other methods like optical flow and histogram of the gradient are to be applied. Due to the lack of motion, modelling in such 2D image based deep features, three-dimensional convolutional neural network (C3D) [15] has been chosen for the experiment. Spatiotemporal features are learnt with deep 3D ConvNet (convolutional neural network) where both appearance and motion are considered. The other reason for choosing the C3D is that it provides more local information as compared to any other video descriptors.

For continuous improvement on various forms of motion capture tasks, 3D Convolution is thought to be effective in the sense of huge supervised learning databases and deep hierarchical algorithms. It contains all the four qualities which a video descriptor should have as mentioned in the previous paragraph. Some of the 3D ConvNet properties which have been analyzed in the paper [16] are mentioned here:

i. 3D convolutional neural network can model both appearance and motion simultaneously.

ii. 3D filter of the size 3x3x3 covers the temporal length as well as spatial.

2.1 CONVOLUTIONAL NEURAL NETWORK

An Artificial Neural Network is a type of neural network that extracts helpful info by sorting stimuli using convolutional [17]. A feature vector map was produced by merging a region proposal with such a convolutional layer. To obtain the most valuable data from a given task, the macros are selected depending on filtered. CNN's detect the key expectations of a mission automatically. Computer vision, eventually happen, text interpretation, speaker identification, language processing, and other fields have used CNN's. One activation function, one or even more hidden nodes, or one output units make up a Learning Algorithm. Porous material, convolution layers, pooling layer, normalization surface, corrected legacy planning layer, fully associated layer, and so on are examples of these sheets. Inside - framework of a Deep Neural Networks, there must be at least one fully connected layer. Figure 2.1 depicts the entire architecture of a Convolutional Neural Network.

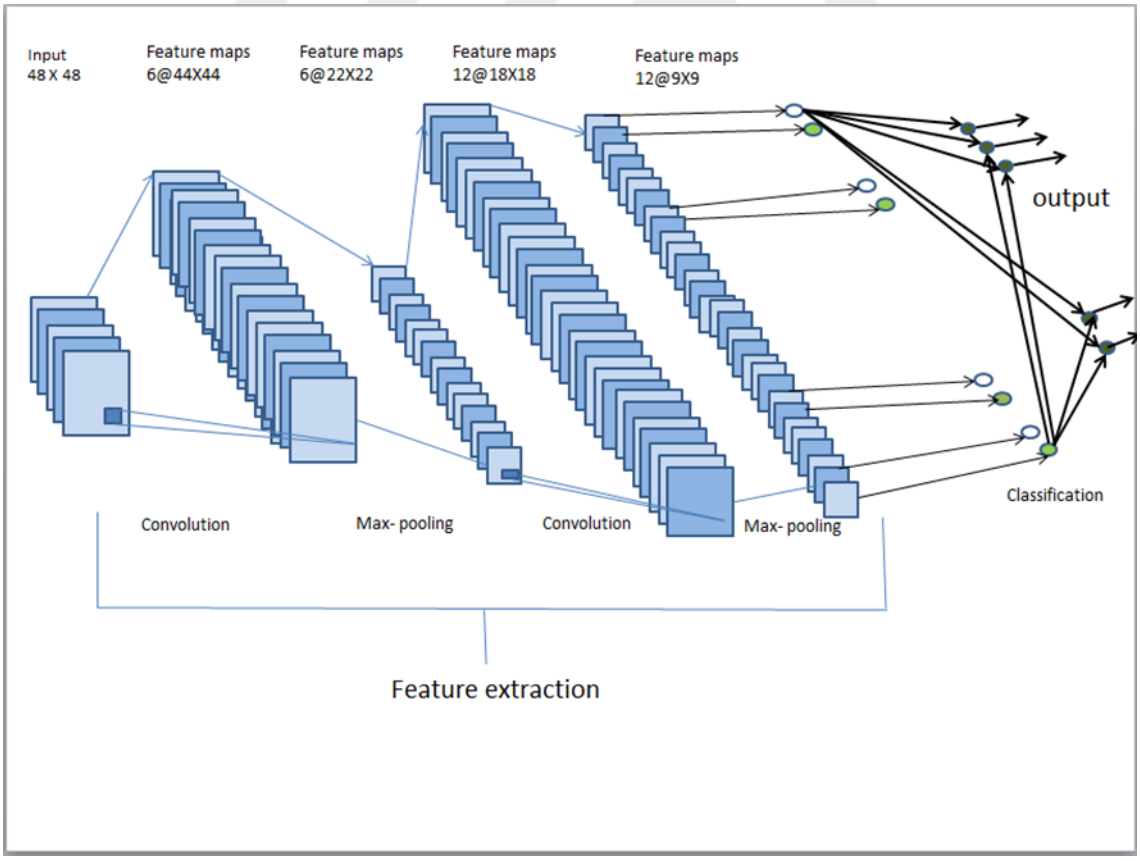

Figure 2.1: Architectures of Convolutional Neural Network

2.1.1 Input Layer

Only digital images are used in CNN's input nodes. [width x height x depth], which is a vector of image pixels, are the three aspects of these pictures. For instance, if the input signal is (width=32, height=32, depth=3), or [32x32x3], the density stood for Red, Green, and Blue CNN, or Hue CNN. In addition, the input image must be divided by two several time. To work with input layer, at first the three-dimensional matrix has to be reshaped into a single column. For example, if there is an image of dimension 32x32=1024, it needs to be converted into 1024x1 before passing it into the input. For 'x' training examples the dimension of input will be (1024, x) Figure2.2 shows the input layer of a CNN.

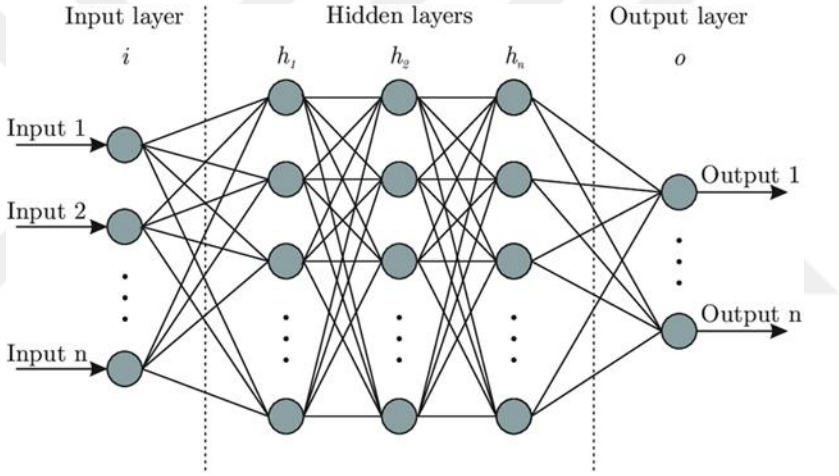

Figure 2.2: Input Layer

2.1.2 Dense Layer

The dense layer is a non-linear surface in a computer programmer. The equations for these structures are the same as for normal surfaces, with the exception that dense layers pass the final result through a non-linear input layer. One of the deep network's distinguishing features is its ability to design any linear formula. Although there are some limitations for example for any given input vector, the output will always be the same. Dense layer neither can produce different answers on the same input nor can detect the repetition in time. Figure2.3shows the dense layer of a CNN.

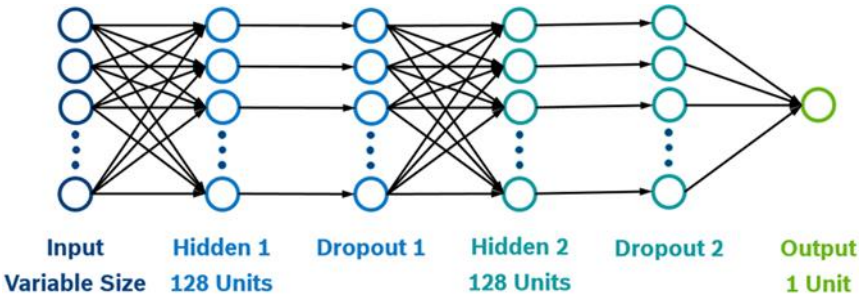

Figure 2.3: Dense Layer

2.1.3 Convolution Layer

A Convolutional Network's central major component is the Convolutional layers, which handles the majority of the computing time. The convolutional layers use filters to create variables and model parameters where it needs to be extracted and teaches. In a sentence, this is a feature extractor layer. Input images are compared with segments to find out the changes. These segments are known as features. This layer selects one or more features from the input images and creates one or multiple matrix and dot product with the image matrix. It also calculates the output neurons connected to the local regions.

Furthermore the whole computation gives out a result known as the convolution layer. For a given I x I image size and F x F filter size, the convolution result will be:

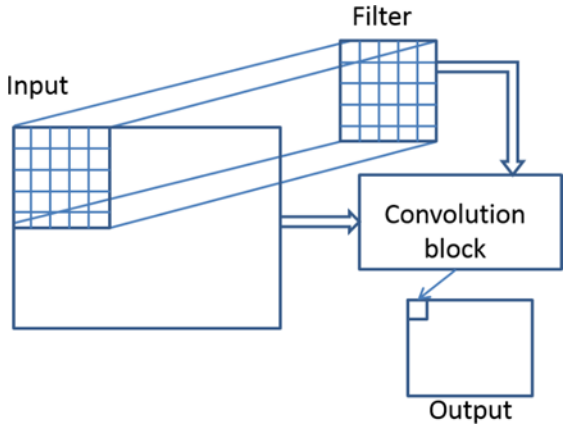

Figure 2.4: Convolution Layer

2.1.4 Rectified Linear Unit Layer

The neural network of the Rectified unit (ReLU) layer is indeed a nonlinear straight line that outputs the meaningful feedback explicitly if the feedback is good; else, it outputs nil. This layer receives the result of the convolution layer's operation., Both vertical lines will be adjusted to zero, while the higher result will remain constant. Many forms of machine learning now use ReLU as their standard activation. [18]. Model initiated with ReLU (Rectified linear unit) can be trained easily and also it achieves better performances. ReLU layer does not have any parameters or hyperparameters. Figure2.5 shows the rectified linear unit layer of CNN model.

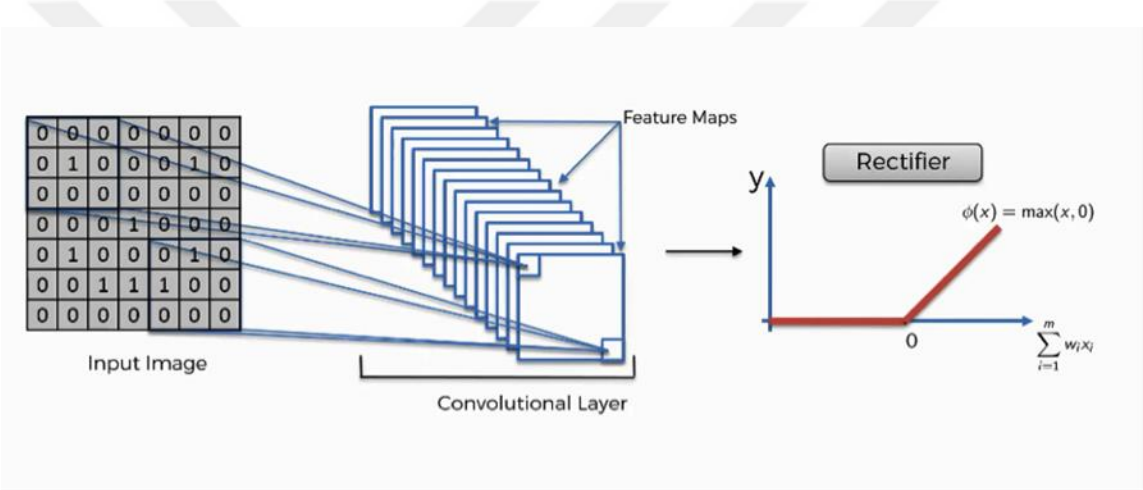

Figure 2.5: Rectified Linear Unit Layer

2.1.5 Pooling Layer

The pooling layer is a component of a Convolutional Neural Network that operates to minimize the representation's model complexity in order to minimize the amount of inputs and computations in the system. Since the max pooling is not reliant on the usability or quantities of applications in the node, it can operate separately in each feature vector. A most famous aggregation design is known as max pooling. The pooling layer has no variables, but it does have two algorithms: Filter(F) and Stride(S) (S). If the input variable is X1 x Y1 x Z1, the output variable is X1 x Y1 x Z1.

Figure2.6 shows the pooling layer of a CNN model and the max pooling mechanism.

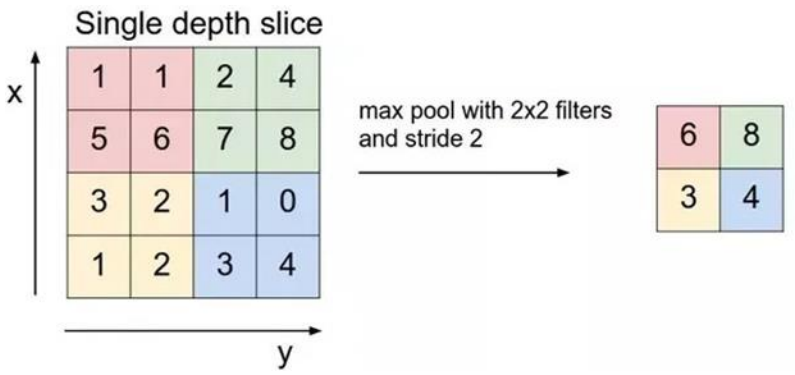

Figure 2.6: Pooling Layer

2.1.6 Normalization Layer

Normalization layer basically normalizes the activation outputs of previous layer which means it puts in a transformation that upholds the mean activation close to zero and the activation standard derivation close to 1. Batch normalization is mostly used here although there are many other alternatives like weight normalization, layer normalization, instance normalization, group normalization, spectral normalization etc. Figure 2.7 shows the normalization layer and various normalization methods.

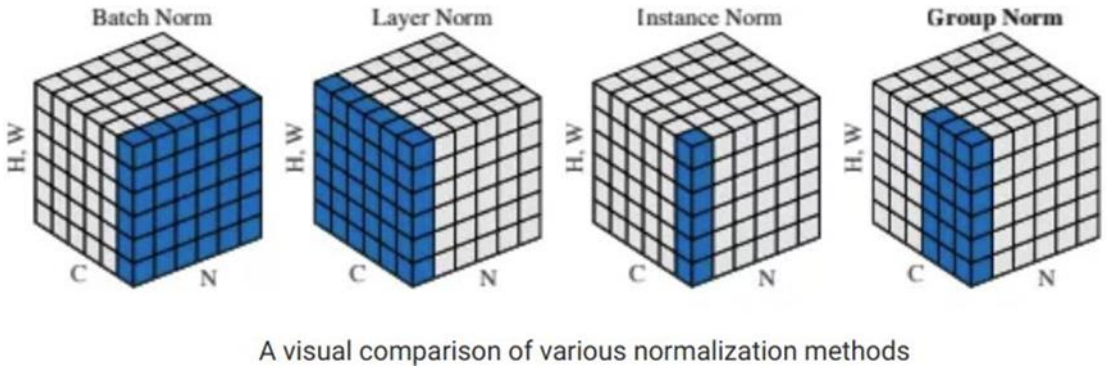

Figure 2.7: Normalization Layer

2.1.7 Fully Connected Layer

Fully connected layer normally takes an input volume and outputs an N dimensional vector where N is the number of classes that have been chosen from the program. The input volume comes from the output of convolution or relu or pool layer. After analyzing the outputs of the previous layer fully connected layer distinguishes the mostly matched features of a particular class. Fully

connected layer works on features which are high level and have particular weights. In this way fully connected layer calculates the odds for the different classes as it calculates the product outputs of the weights and the previous layer.Figure2.8 shows the fully connected layer of a CNN model.

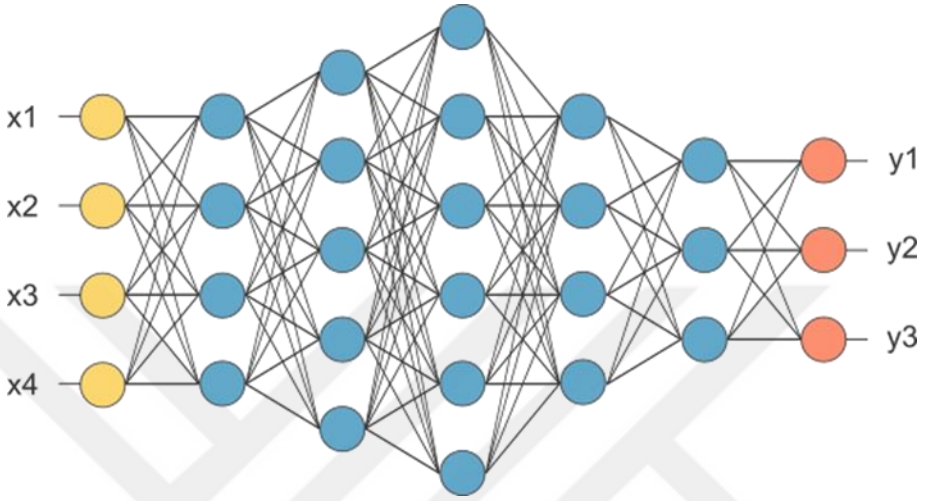

Figure 2.8: Fully Connected Layer

2.1.8 Output Layer

The convolutional layer in CNN was its output, which generates the system's end product. The output vector neurotransmitter specification is often changed in order to comprehend the outcome as well as the entire programmer. Once the information has been thoroughly educated as well as the measurements have indeed been preserved, it will compare the information to the training dataset and decide if it fits and not for a given configuration. In a paragraph, the output unit coheres and generates the end outcome appropriately. Figure2.9 shows the output or the final layer of a CNN model. In this way, CNN takes, process, trains and thus predicts for the given inputs.

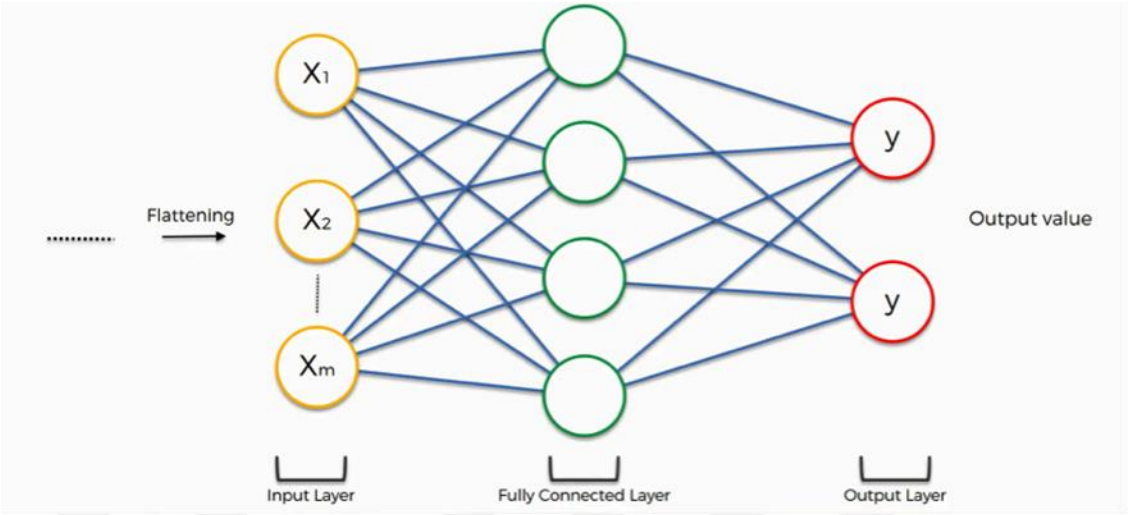

Figure 2.9: Output Layer

2.2 HUMAN DETECTION

Human identification is the job of locating all artifacts in fixed footage streams that are thought to be people. In particular, the approach is simple in the feature of social identification: extraction of potential subregions from a photo, description of the environments with adjectives, classification of the areas as person or semi, and comment techniques [19]. The previous type of identification was the Deformable Part-based Model (DPM) [20]. The method can be compared to the expansion of directed contour graphs (HOG). Both a rough regional model of the input frame and six sections of an image in a greater resolution are used to rate the expected items. HOG is used to define each data. In this manner, the HOG multi-model will address the issue of shifting perspectives. While preparation, a passive SVM (latent SVM) is being used to reduce the classification problems to a class label. Component locations are viewed as an observed variable. Because of its sturdiness, the approach had a huge impact. Individually identified characteristics, on the other hand, may be unable to explain increasingly intricate issues about items. Blur, lesion, changes in lighting, context, and other factors, in particular, presents a problem to them. As a consequence, deep learning techniques have lately been regarded as more useful techniques for object recognition, as they can learn increasingly complicated key points. [21], [22], and [23] are examples of similar research. As these early word embeddings have shown some benefits over traditional prototypes, many technologies still are carefully built, and the key concepts are extensions of past versions.

Neural Networks have also been used as a function extractor in some studies, such as [24]. Following the extraction, complexity conscious gradient practice for object tracking is applied.

Latest deep learning methods approach the issue of identification in a lot of formats. Artificial Neural Networks are amongst the most popular frameworks (CNNs). Deep CNNs have had the potential to understand details information independently, reducing their reliance on objects. In contrast to learning a database design, since big data needs a massive volume of information, learning school methods implies the opportunity to comprehend with far more information. Just a few articles have used the CNN approach in the domain expertise. [25] learn person categorization features using a CNN (e.g. backpack, hat, etc.), However, rather than making predictions explicitly, the network component only helps to increase accuracy rate by re-classifying expected artifacts as right or wrong. [26th] Build a system focused on Fast R-CNN, with a sub-network to manage comparatively tiny artefacts. [27] directly examine how the output of Faster R-CNN, a cross-class CNN capacitance, on hold machine vision is good. But for [28], which does not explicitly include detector, [29] and [30] both run studies using bridge predictive model.

2.3 OVERVIEW OF 3D CNN (C3D)

The two-dimensional convolutional neural network is a widely used deep learning architecture that exhibits state-of-the-art performance in 2D images for various tasks in detection and classification. It is such kind of deep models which can automate the feature extraction process by acting directly on the raw inputs. Various pre-trained convolutional network (ConvNet) model [31] are proposed for extracting different image features. However, 2D Convolutional Neural Network (2D CNN) is best suitable to preserve the spatial information of the images, but when it is required to conserve the temporal information of the same image, it collapses completely. It is because 2D CNN applied to multiple images in different CNN [32] are unable to compute those feature which is encoded into multiple adjacent frames in a video sequence. A similar situation happened when fusion models in [33] used 2D convolutions and the network loses their temporal signals of the input right after the first convolutional layer. Therefore, in this thesis, deep 3D ConvNets has been used which is more suitable for spatiotemporal feature learning as compared to the 2D ConvNets.

Although deep 3D ConvNets had proposed before [34] for human pose recognition, now the C3D (3D CNN) will be used for feature extraction of the abnormalities in the crowd by preserving both

spatial and temporal information of the images. The 3D convolution has been applied on an image by convolving a 3D kernel across a cube formed by piling multiple adjacent frames together [35]. From this structure, the feature maps in the convolutional layers get connected with the multiple adjacent frames in the previous layers and hence capture the motion information. Thus, the C3D is well suited for spatiotemporal feature learning. Compared to 2D ConvNets, the C3D has better ability to model temporal information and therefore contribute to various 3D convolution and 3D pooling operations.

2.3.1 C3D Training

The C3D originally trained on 'sports-1m' dataset. The network consists of 8 convolutional layers, 5 max-pooling and 3 fully connected layers followed by SoftMax layer. These layers need to be trained according to the dataset. In starting, a decision has to make that among 8 convolutional layers of the C3D, which layers are

beneficial to solve our purpose for human detection. Spatiotemporal patches obtained from the dataset are of the size 20x20 and when it passes through the original C3D layers, it is observed that most of the spatial information is lost because of the reduced patch size (1x1) at the output layer. Therefore, simplification of the convolutional layers is required so that reduced size does not lose spatial information from the patches. The following figure shows the proposed convolutional layers that will be used to train the dataset model and further for the detection purpose.

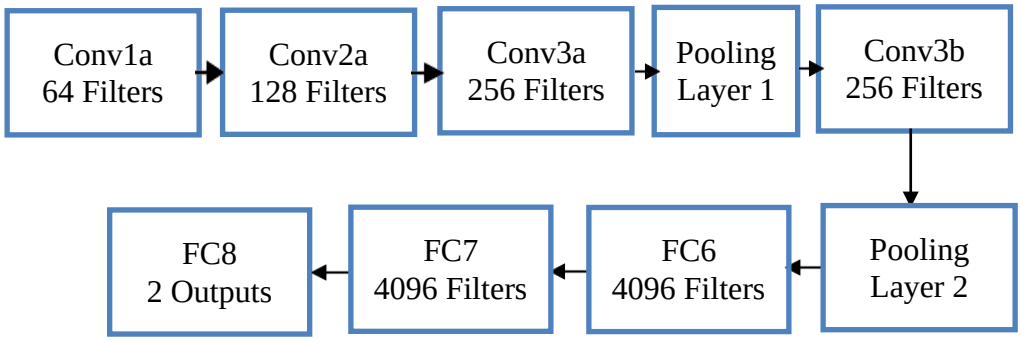

Figure 2.10: Proposed C3D architecture for training

The C3D architecture as shown in Figure 2.10 consists of 4 convolutional layers, 2 max- pooling layers and 3 fully connected layers. The convolutional layers have been specifically shortened down to 4 layers with only 2 max-pooling and 3 fully connected layers in order to analyses the feature maps of the smaller patches and obtain maximum information out of it. With the original C3D method having 8 convolutional layers, most of the information were lost as the size of the patches gets reduced to 1x1 at the end. Therefore, maximum reduced size passing through the final convolutional layer i.e. conv 3b will be 3x3, so the spatial information, in this case, will not be completely lost.

2.4 SPATIOTEMPORAL FEATURE EXTRACTION USING C3D

Abnormal event detection or detecting the unusual behavior of people in crowded scenes is a very challenging topic and research in this field is mainly focused on extracting the image features.

[36] considers the problem of human detection from localized video of crowd detection by using spatiotemporal patches of smaller dimensions. They consider the sequence of images of moving scenes represented as dynamic textures which exhibit spatiotemporal stationary properties [37]. By experiments, they showed that their approach is more reliable than the previous works [38] which rely on optical flow and the social force model. This section presents abnormal event detection using the C3D method. It is explained in different parts like network architecture, dataset, model training, detection, and baseline.

Recognizing unusual event detections in a real-world environment is applicable in many domains including video surveillance, object tracking and human behavior analysis. Current state-of-the-

art algorithms [39] provide novel unusual event detection by extracting local motion features. It has been observed that 3D CNN is well-suited for spatiotemporal feature learning because as compared to 2D CNN, 3D ConvNet has better ability to model temporal information of the images as well as to perform 3D convolution and 3D pooling operations.

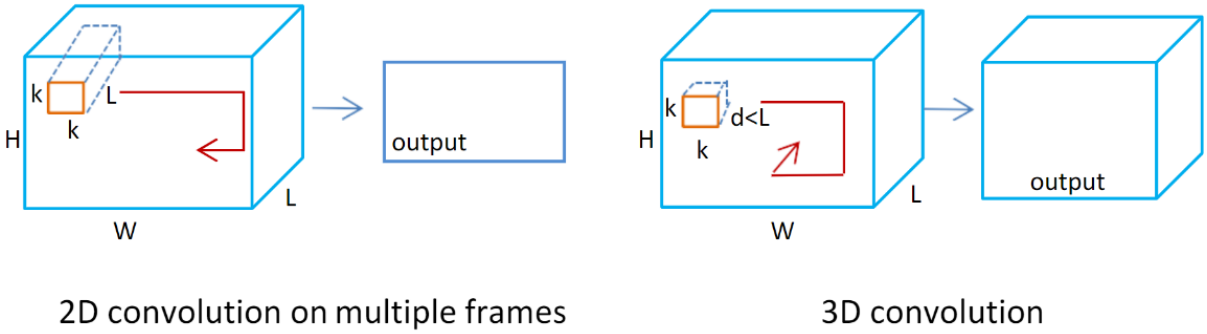

Figure 2.11: 2D and 3D convolution operations

Figure 2.11 signifies the difference between the 2D and 3D CNN methods, it is observed that whether 2D convolution is applied to a single image or on multiple images, the output only shows the spatial information as it loses the temporal information of the input signal after every convolutional operation. In comparison to this, 3D convolution is the one which preserves the temporal information resulting in an output volume. Now, to find a suitable architecture for 3D ConvNet, the experiments are conducted with a smaller dataset like the dataset. Training of the large-scale video datasets will be very time-consuming. According to the 2D ConvNet in [40], 3x3 convolution kernels belonging to smaller receptive fields yield good results. Therefore, the experiments are conducted with two different architectures, first, the kernel of 3x3 size is traversed spatially along the input frames keeping the temporal depth constant and second, traversing the kernel of size 3x3x3 along the spatial and temporal depth of the frames. Both these experiments are performed on the datasets and comparison between both results will be analyzed further.

2.5 DETECTIONS OF HUMANS IN VIDEO STREAMS

Taking the advantage of rapidly developing methods, we here focus on a specific detection task: object detection in video streams. We are also interested in how the CNN models understand human patterns in natural viewing behavior. In human behavior analysis, one of the key elements

that affect human attention is daily interactions with other people. Therefore, it is of high relevance to be able to identify and track the position of every person that is visible in a video stream recorded aligned with human vision. In our work, we take a pure detection method for human localization in video streams, so that this work can be a prior study for human tracking and human behavior analysis. For research on human behavior, the naturality is one of the most important factors to data quality. Fortunately, we get access to one of the most advanced tools, Tobii Pro Glasses 2, which is a pair of 45-gram weighed glasses, designed to capture natural viewing behavior. The videos are recorded with a head-mounted camera assembled in the glasses. As the agile glasses are able to provide the most flexible movement of vision, the videos streams usually contain fast motion, unpredictable rotation, and strong motion blur, making the problem more challenging. it can be seen that CNNs are superior to other models for computer vision problems due to their wide and deep structures. Therefore, it is also believed to be possible to gain reasonable accuracy and speed via deep CNNs in object detection in challenging video streams.

2.5.1 Template Matching

The complexity of object tracking is influenced by many factors. As has been stated in Chapter 2 of this thesis, US video streams comprise a large number of artifacts that lower the image quality. Whereas video surveillance applications may deal with a static background and moving target objects, the US video application includes at least patient movement such as respiration which conducts a non-linear transform on the frames, thus leading to a separate movement of both, background and target objects. Additionally, the transducer may be moving as well, leading to background movement, partially visible objects or objects moving completely outside the capture region. Since a B-scan of the US modality is a planar scan, a movement of the transducer may position the sectional plane such that it does not intersect with the target object anymore whereas the background remains mostly constant. Many of the presented applications also involve creating a model of the desired object. Although, this can be a simple task for applications such as player tracking in sports casts using static templates, deformable objects in US videos require dynamic models, thus complicating the modeling task. Other factors that exacerbate the tracking process comprise occlusion effects when tracking multiple objects. Player objects in sports video streams may be subject to occlusion as described by [40]. However, these occlusion effects may also occur in US videos, depending on the application. Despite the differences between the listed applications,

common characteristics can be extracted. All of them start with a set of input frames Finput in a finite or infinite video stream.

2.6 VIDEO BASED OBJECTS

Video based objects can be evaluated in the context of autonomous vehicles in order to reduce the risk of collision. To pose timing requirements of the object detection pipeline, simulations of stopping distances for a vehicle were done. By calculating a maximum allowed latency it could be used as a reference when reasoning about the speed of the algorithms. Hardware requirements was defined based on the timing requirements posed from the stopping distance simulations. YOLOv3 was initially implemented in order to evaluate its baseline performance. Following the implementation of YOLOv3, a first implementation of the object detection methods was CNN. After this, videos could be rendered out to run through the YOLOv3 algorithm, and these gradient based learning allowing testing and evaluation. Activation functions sit between layers in a network to introduce a non- linearity. Otherwise, the operations could be regenerated to a simple linear transformation and remove the benefits of building a deep model. The rectified linear unit or ReLU which is the recommendation in the activation function for neural network. It's defined as the maximum of 0 and the input value and can be described as a nonlinear function made up of two linear pieces. Because of this, it preserves many of the properties that make models easy to optimize and generalize well on new data. Since the introduction of ReLU other variants have built on its success like PReLU (Parametric Rectified Linear Unit), LReLU (Leaky Rectified Linear Activation) and ELU. A lot of research does not only concentrate on one single characteristic, it heads for combinations of biometric tokens. In this work we are concerned with the object detection and object analysis token and we focus on eyeglasses detection and segmentation as a pre-processing step for object detection and analysis applications. For robust object detection it is essential to have normalized pictures, which fulfill the ICAO requirements for Normalized object images should be interference-free and should contain a frontal pose, neutral expression and 'normal' eyeglasses with a typical frame, clear glasses and no reflections. Therefore it is necessary to detect eyeglasses and - if present - to delineate and evaluate the eyeglasses. This leads to solving problems in the areas of classification and segmentation. Normalized images which come up with all the necessary qualifications increase the reliability of object detection and detection software. The task in this work is to automatically detect the presence of eyeglasses and to locate the

eyeglasses in the object portrait image [41]. With the located eyeglasses several different parameters of the eyeglasses appearance can be calculated. All these parameters lead to decisions, whether an object image containing eyeglasses is suitable for a 'Machine Readable Travel Document' or not. Magnetic Resonance Spectroscopy (MRS) measures the spectral peaks of different compounds containing protons in the tissue of interest. Instead of producing an image, MRS data are usually reported as spectra of different metabolites. Nonetheless, it is possible to integrate under the metabolite peak of interest and generate a low resolution metabolite image as well. Studies have shown that an increased choline peak is indicative of cancerous lesions due to increased cellularity, and MRS (Magnetic Resonance Spectroscopy) can be useful for diagnosing lesions greater than 1 cm in diameter. high field magnets offer two advantages for mrs: the frequency separation between peaks is greater at higher fields making it easier to differentiate between compounds and the increased snr at high fields enables greater signal for metabolite peaks which exist at much lower concentrations than water in the body.

2.7 GRAPHICS PROCESSING UNITS

A Graphics Processing Unit (GPU) can be seen as an array of hundreds of processing units that are designed for quick performance on graphics rendering, mainly three-dimensional graphics [27]. However, as soon as tools for programming were made available by vendors, the scientific community started to use GPUs for general purpose and research. There are several Application Programming Interfaces (APIs) such as CUDA (Compute Unified Device Architecture), which is only for NVIDIA products; and the Open Computing Language abbreviated as OpenCL. These APIs allow the implementation of programs for GPUs, but the decision of which to use is most influenced by the manufacturer of the specific product. The interest in using this technology for research has had an effect on the evolution of the GPU; today's products are built with different architectures that allow general purpose applications. As mentioned earlier, GPUs can either be considered SIMD devices or be made of SIMD components, depending on the specific model. The instructions can be performed by different SIMD components, sometimes called vector units. How data is handled in GPUs is another characteristic, commonly there is a global memory that can be read or written by any processing units. There is also local memory that is only available to certain processing units; this is added because using global memory can become a bottle neck when many different processing units require access to the same data at the same time. The main coordination

mechanisms that are observed in GPUs are per- formed by the common CPU. Since the main purpose of GPUs is to help in image rendering, these are connected to a common CPU so that normal operations of the computer are performed by it. The GPU is not usually required to have an internal synchronization or coordination mechanism because the CPU is the ‘leader’ that takes these responsibilities. This can be problematic if the program that is implemented in a GPU has data dependencies between different phases [21].

3. METHODOLOGY

Our language model and accompanying scripts are written using the Python programming language, a popular language for ML tasks. There are several frameworks available for these tasks available, including Keras (neural network library), PyTorch, and TensorFlow. Though any of these frameworks would serve us well in our research, we chose Google's TensorFlow, for its wide usage and the excellent availability of tutorials and quality documentation. We specifically use the GPU variant of TensorFlow to enable GPU acceleration and train the models faster. Our code is structured as a number of scripts responsible for different tasks, from downloading and preprocessing datasets to scripts detailing the sentiment analysis based natural language processing and scripts for training and running it. We used the code as a starting point. Table 3.1 shows the specifications of the computer used to train the model. This machine was the primary on which we trained the model.

Table 3.1: Specifications of the computer used to train the model.

OS	Ubuntu 17.10
RAM	32GB
Processor	Intel i9 8700K
GPU	NVIDIA Tesla C2050 GPU

The computer ran Ubuntu 17.10 with the NVIDIA Tesla C2050 GPU and cuDNN CUDA Deep Neural Network libraries installed to take advantage of the GPU. As the primary programming language Python was chosen due to its simple syntax and popularity in the deep learning and data mining field. The code was briefly tested in pyhton and seemed to work very well. Keras and scrappy was used as the framework for training models because its high-level syntax allows fast prototyping and testing. The development environment was a Jupyter Notebook, which allowed a flexible execution of code snippets instead of running the entire program for every single change. The processing of medical image data needed a library that could handle these formats. SimpleITK is a Python binding to the ITK library written in C++. It includes many tools for image processing and is especially popular in the medical field. Other libraries were also used for smaller tasks.

The final exam of our work has already been addressed in this chapter, beginning by how they resolved the issue to construct the system and how we transformed our datasets into a format can be used with our newly developed system. To begin, a picture of our current plan was shown to give a general picture of what we'll be doing. The data summary follows, which describes our database and what it contains. Then, the database before the methods, as well as the compilation methods, were demonstrated, including a description of all the variables which were used to compilation our template.

3.1 Dataset

In this section, the datasets used in our work are introduced: Pascal VOC and video datasets extracted from video streams. It is clear that the video datasets are the target in our work, both for training and evaluation. However, we also choose Pascal VOC due to the following reasons:

i. The model we picked are YOLO and SSD, and are pre-prepared on Pascal VOC. Since our goal is to learn about humans, training on a similar but smaller database (limited to living organism data) will aid in understanding the systems' ability to learn on their own, as opposed to pre-prepared models.

ii. Pascal VOC is larger with more complete annotation. Deep learning methods benefit from large amount of data. In the experiments show that CNNs have ability to learn the information from a large dataset even the dataset is randomly made up. Therefore, training on large datasets helps models to learn better variance of training patterns, which also provides a better generalization ability and reduces the risk of overfitting.

Therefore, we also use Pascal VOC as a benchmark in our work. The differences between the two datasets are discussed in this section. We made use of Actual data gathering is a difficult task. Working with original data is often difficult because it includes several repeated lines and abnormal symbols that do not represent reality. As a result, gathered information must be processed and transformed into a database that can be used for the scientist's purposes. Correspondingly, our image is processed to exclude identical lines. Synthesized attributes have been added to better represent the equations' performance. Until executing and code, some tasks are normal, such as eliminating different lines and breaking the database into check and learn. Each optimization is tested in a variety of environmental conditions. The database for

documentary identification was collected from Kaggle, a fully accessible database. Python will be used to execute this study since it is the most effective method for manipulating large data sets with greater precision. . The justification is that in communication organizations, preparing, installing, and retaining variants for each unit, particularly for trackers, can easily be cumbersome. Python will be used to apply the Neural network for this database.

As we tend to study the human detection problem in videos capturing natural vision of human, the video dataset we use is recorded with Tobii Pro Glasses 2, which is a wearable device whose head unit is a pair of 45-gram-weighed eye tracking glasses, providing flexible movement during the natural view recording. The head-mounted camera is available for videos at quality of 1, 920 × 1, 080 pixels at 25 fps. The annotations of the video streams are only restricted to the specific human domain. The training dataset is developed based on several videos on scenario tasks, occurred in two different supermarkets/convenient stores. In these videos the participants are required to walk around in the stores and do some shopping, with Tobii Pro Glasses 2 recording the whole process. Limited by time for annotating, and since temporal information is not assumed in our methods, we decided to only extract about 100 - 300 frames (with/without human) from each video as training data. As a result, it counts to 2,914 frames from 12 videos. Within the 2,914 frames, 500 are left out for quantitative evaluation. In this paper, we will refer to the training dataset as Shopping dataset. Figure 3.1 shows some examples of training data. We also have a test dataset for qualitative research. The test dataset is videos recorded with our team in an office environment, which ensures the significant differences of scenes and participants between training and testing data. Similarly, in this paper we will refer to the test data as Office dataset.

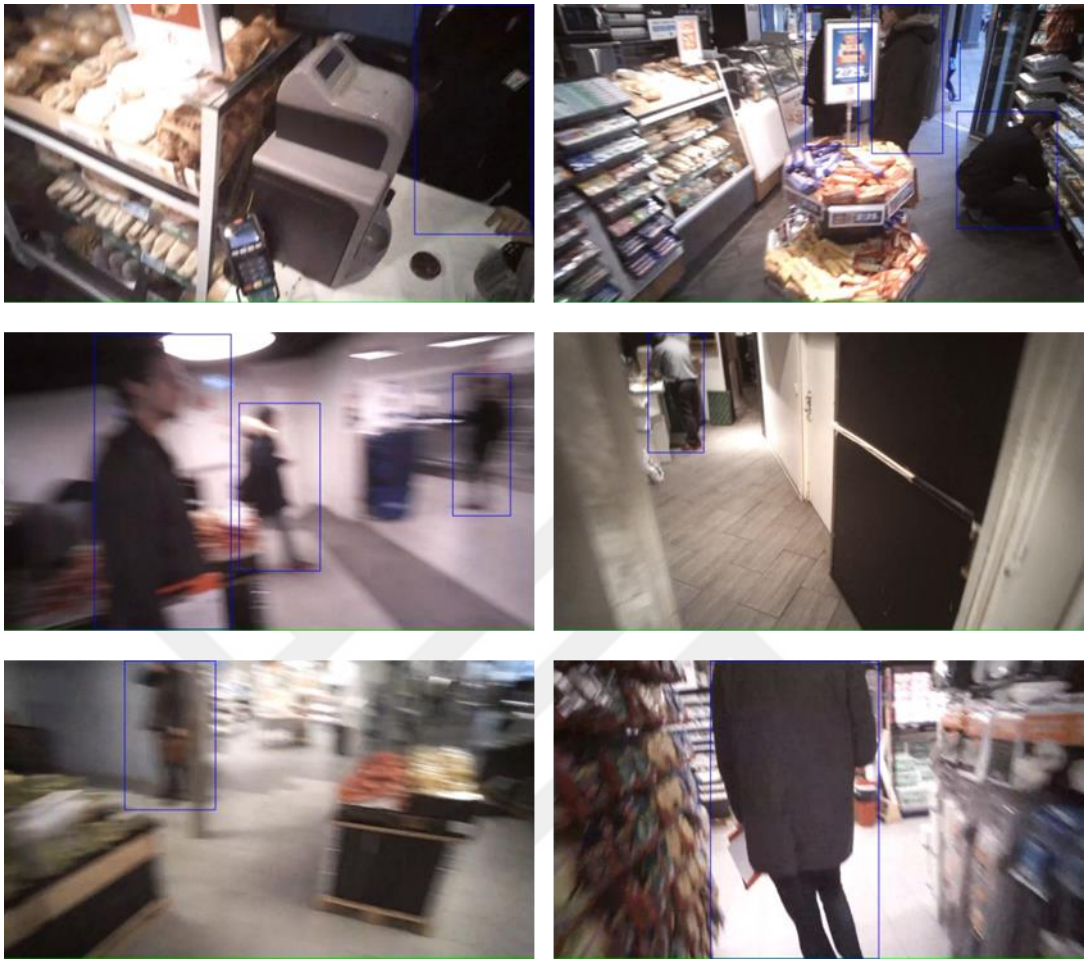

Figure 3.1: Human Detection dataset

As can be seen, compared to the Pascal VOC dataset, the images are of high resolution but includes difficult view-points and strong motion blur during fast head movement. Also, illumination is usually different from Pascal VOC, since the videos are all taken indoor.

3.1.1 Splitting the Dataset

The database is overwhelmingly weighted in favor of monitoring results. The information will be collected and during tests to establish more or less equal groups. Repeated loading will be split into 3 sets for Person sensing: train, optimize, and evaluate. Footage will be introduced during the test phase and will be the same width as the standardized test information gathering. Since the information is overwhelmingly underrepresented with groups, we have to include the usual information as well, so our training dataset will be divided into 80:20 training and testing

subsections, as is customary in the deep learning sector. For text categorization, 100000 marks will be obtained from each of the 5 groups for a total of 500000 marks, which will then be divided 80:20 for training and testing sets.

3.1.2 Feature Scaling

Scaling the information while programming machine learning is thought to be the best practice. Or else, issues such as deteriorated efficiency, irrational behavior of the optimizer, and bursting regression can arise. There are many methods for scaling information; several of them depend on understanding a directory extreme values; however, since the information will not be in a stable range, such as image pixel information, the standardization approach has been chosen, which depends on calculating the track set's variable.

3.1.3 Feature Selection

For the purposes of feature extraction and attacker detection, some data features may be more relevant than others. The research into feature extraction for object recognition has mirrored this. We will be able to correlate it research to many other researches by choosing a limited number of characteristics, and we will also be prepared to illustrate how artificial neural different shifts when adapted to various datasets.

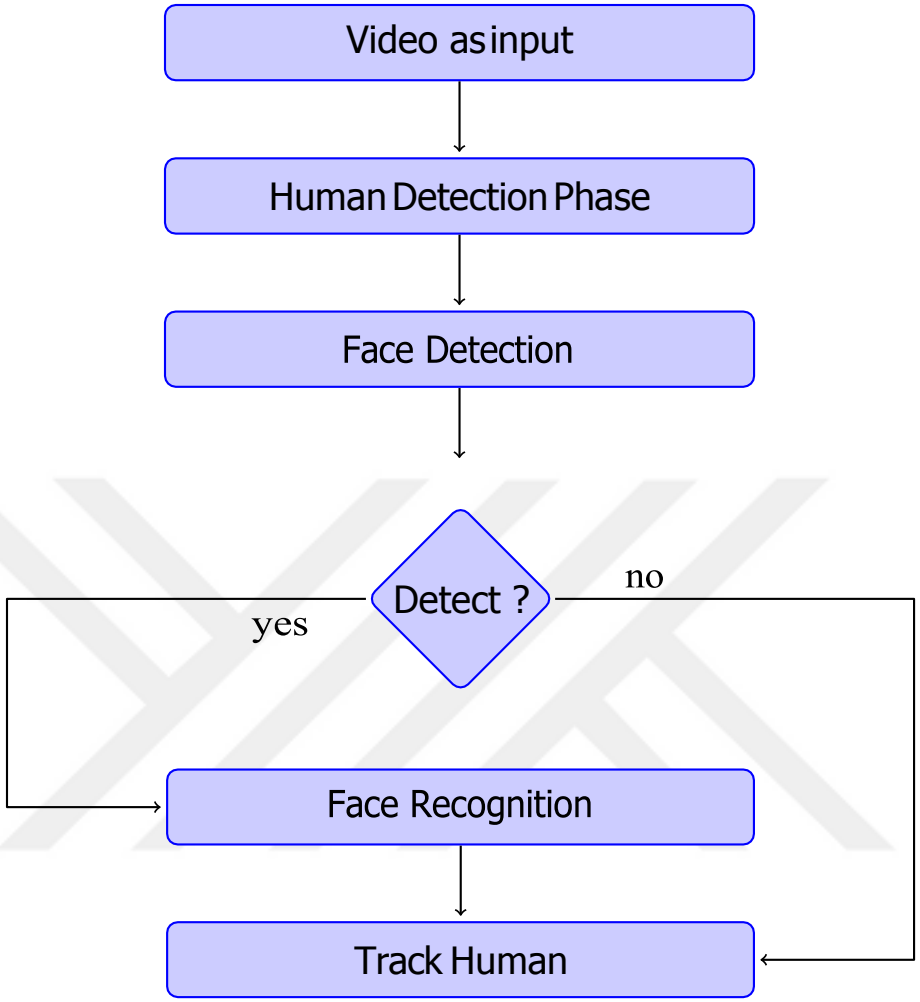

Figure 3.2: Flow chart of human detection

3.2 CONVOLUTIONAL NEURAL NETWORK APPROACH

CNN stands for Convolutional Neural Network technique, which is a classifier. The Csi connector split the entire into classes, maximizing the analysis usefulness for just any investigator. We may distinguish land linked to one category and information related to some other class depending on the various classes. Series data estimation, biometrics, and biochemical database management are only a few of the areas where CNN was used. A super plane is drawn by the Neural networks, and is said to separate the various groups.

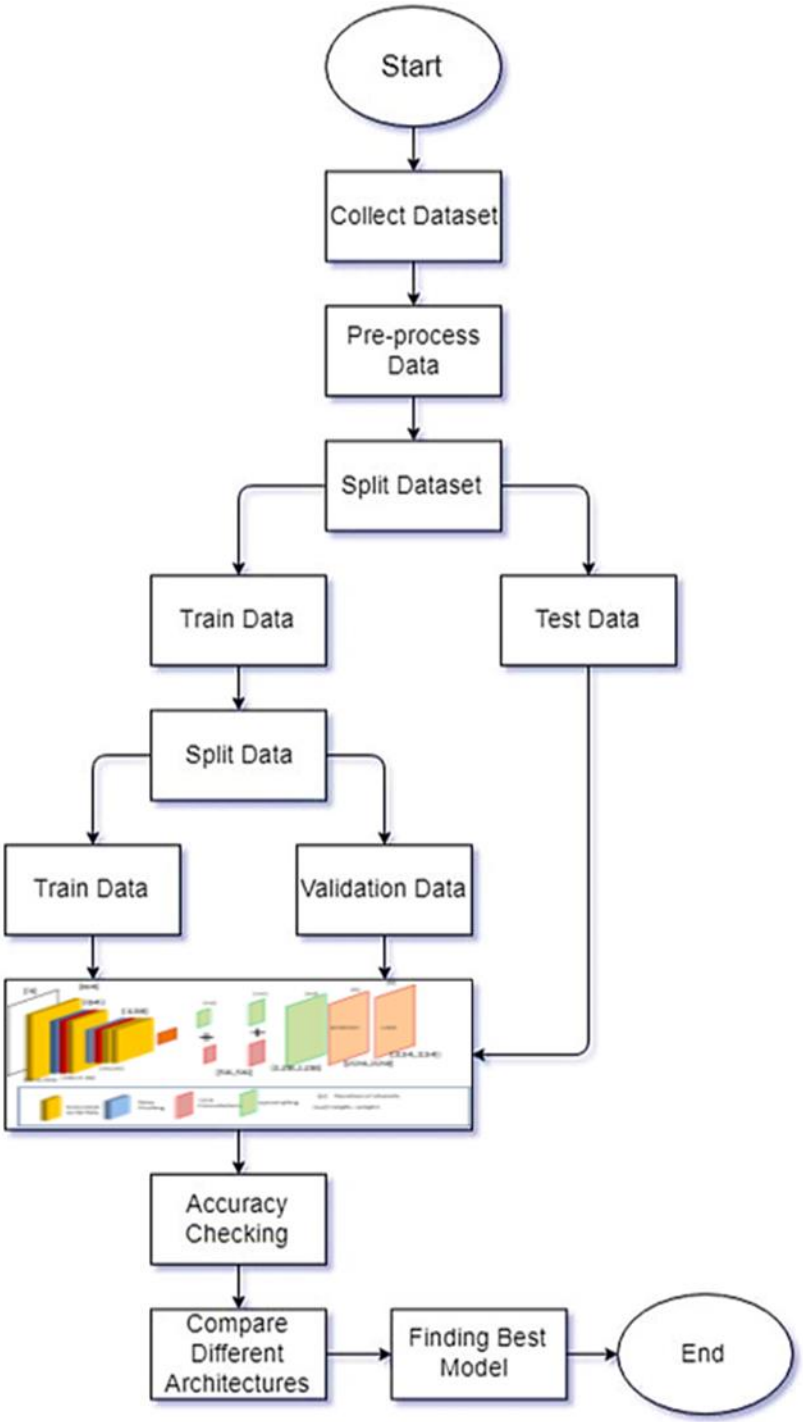

Figure 3.3: Flow chart of the proposed model using CNN model architecture.

The kernel function it's not just any line; it's a line that separates the 2 groups in such a manner that the gap between the accordance and the nearest points of both groups remains as strong as possible. In other terms, it maximizes the gap between the positive class and each of the class'

closest pieces of data. The method of drawing this dimensional space is defined in the following coefficients, and 2 distinct groups are defined.

$$w * x - b = 0 \ (1)$$

The formula is a hyper - plane formula that divides the two groups fairly. Here, w denotes the usual variable to the hyper - plane, x denotes the range between both the closest station to the hyper - plane and the higher dimensional space it, and b denotes the range here between closest station to the higher dimensional space and the higher dimensional space on its own.

$$w * x - b >= 1 \ (2)$$

If the outcome of this formula is optimistic and equivalent to one, we have the hyper - plane of the classification model, as well as any meaning 1.0 bring good class pieces of information.

$$w * x - b <= -1 \ (3)$$

LIKEWISE, IF THE VALUE IS -1, IT REDUCES THE DOMINANT CLASS'S DECISION BOUNDARY, AND ANY AMOUNT OR LESS -1 REPRESENTS DATASETS FROM THE NEGATIVE EXAMPLES.

3.3 TRAINING WITH CONVOLUTIONAL NEURAL NETWORK

The accuracy of the classifier is heavily influenced by the k value the number of Nodes defines how well the information can be separated in order to generalize the CNN application's performance. A larger n value has the benefit of reducing the variability caused by information outliers. The disadvantage or side effects is that the student develops a bias because he or she continues to overlook small trends that might contain valuable information. For our database, we then determined the ‘k' meaning that better represented our needs. CNN was our first option for classification. Even so, there was no clear sense to this. As a consequence, we wanted to determine the best price that would have the best and most accurate reliability. It operates by testing our template on a range of constraints and then comparing the results to discover the perfect outcome. In our situation, we based on a variety of k neighbor value ranging from 1 to 25 and tested the quality of k neighbor produced a most valid and precise results. It was discovered that this was the perfect fit for our collection of data. We process the data collection after deciding the price. To

prevent learning the very same information over and over, repeated rows with same meanings for all four characteristics are removed. It also avoids training with the same database, preventing unreasonable detection accuracy. The data was then divided into two groups: train or check. The train set accounts for 80% of the results, while the test set accounts for 20%.The CNN classification model is used to train the 80 percent of the results, with k set to 2. The information from the remaining 20% is then put into practice. Trying to calculate the Kl divergence among occurrences of the groups or categories established accomplishes this. After that, the lengths are reordered. After that, the top N columns are chosen, and perhaps the most common group of these lines is retrieved. Finally, we calculated the accuracy and it was very high.

3.4 TRANSFER LEARNING AND FINE TUNING

Transfer learning is a concept opposite to training-from-scratch. When trained from scratch, the networks are usually initialized with random variables. In contrast, in trans- fer learning, it makes use of some factors from a pre-trained model. Two transfer learning methods are commonly used: off-the-shelf feature extraction and fine-tuning. In the former case, a pre-trained CNN is only used as a feature extractor, but the model itself is not re-trained. In the latter case, it restores certain weights from a pre-trained model as initialization, and trains the network with different data or for different tasks. It is believed that fine-tuning is beneficial for shortening the time of training, speeding up convergence and increasing generalization. Detection CNNs are to some extent a transfer learning from classification models. This characteristic is shown as the detection models are usually developed based on pre-trained classification models. The classification models are trained on large datasets, such as ImageNet. The weights trained are a good initialization for detection models. In this way, only the modified layers need to be initialized with random variables. A lot of detection models adopt the fine-tuning strategy, for example, over feat, Detector Box and Fast R-CNN use Alex Net as pre-trained model; YOLO uses GoogLeNet and SSD uses VGG16 as backbones. In our work, two datasets are used for training: Pascal VOC 2012 and Shopping dataset., Pascal VOC is a larger dataset with more complete annotation. It can be beneficial to our model, but our final target is detection in video streams. As a result, we choose a two-step fine-tuning strategy: First of all, we fine-tune our models on Pascal VOC on only human data, and then fine-tune on Shopping dataset. In the first step, the last layer of pre-trained YOLO and SSD models are modified to fit “person” class design and the last-layer weights are replaced

with random initialization. As for the second step, the structures are kept from the first step, only the dataset is changed to Shopping dataset.

3.4.1 Weight Selection

used the proportions null in our analysis. Some post scales or the route to the weights file to be equipped as the scales are the only other options. We started from scratch to validate the system so that it could develop on its own, and then we compared each template to discover the perfect research model.

3.4.2 Pooling Layer Selection

In our design, we used Max Pooling 2D for the max pooling. There are so many different forms of max pooling, but the most general and commonly used are mutual benefit or batch normalization. If we need to finding additional across the given dataset, max pooling is a wiser choice because it uses the device's estimated return from the data set. However, since we are identifying pathogens from a physician's person's body that is only a component of the entire picture, we are using maximum pooling. Furthermore, pooling layer is a better alternative because it collects the picture's strongest attributes. It collects the most significant features, such as template borders, while max pooling removes functionalities in a smoother manner. Keras defaults to a pool size of 2X2 and a strides price of 2 if no specifications are defined. Even so, for the sake of convenience, we set the pool size to 2X2. Max pooling would give the best data point from a simple diagram and substitute it with one tiny dot for the spatial domain.

3.4.3 Dropout Selection

Dropout is a technique of regularization in CNN to prevent the over-fitting problem of the dataset in the training process. What dropout does is it reduces the complexity of the model by the given value of dropout. Dropout helps a neural network to extract more robust features while reducing the training time. In our model we used 0.25 as the value of dropout. Means, it will randomly ignore 25% of the neurons while training. The drop out value can change model to model depends on the approaches to solve a problem. We used different values for the dropout to check the best output and then we found that a value of 0.25 works the best.

3.4.4 Activation Function Selection

In a deep neural network, the credit provided receives inputs from a prior node, analyses it, and generates a result for that same nerve cell, which can then be used as the source in the next synapse. A few authentication options are open. Even so, we used only 2 kinds of neural networks in our current plan: the Sigmoid and the Relu. Operate. The fractal dimension curve is shown in Figure 3.4 for any data.

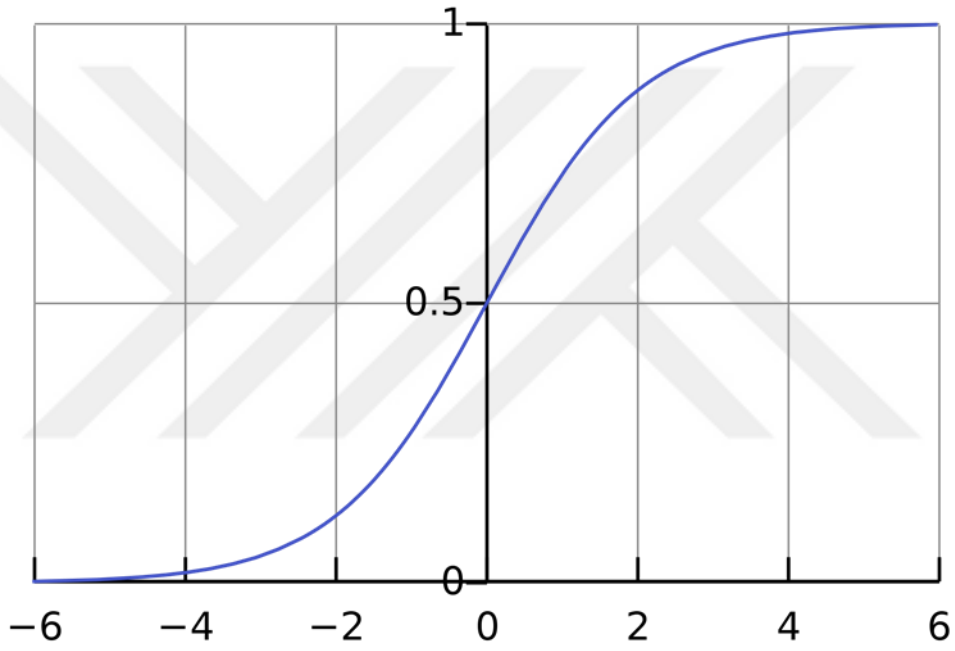

Figure 3.4: Sigmoid Function

Relu Function: A Relu component translates a value to a scale starting at 0 and ending at zero value. In other words, if the input signal is below or equivalent to 0, the cost function would be 0. In addition, if the value is larger than 0, the outcome will become the price. The Relu curves is depicted in Figure 3.3.

$$f(x) = \max(0, x)$$

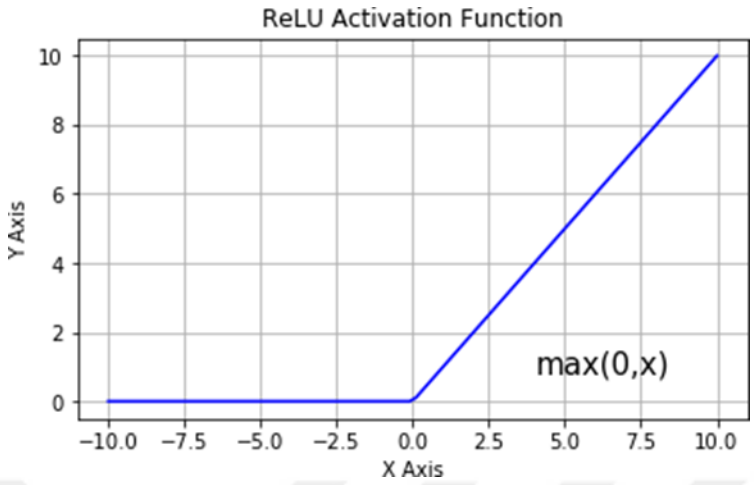

Figure 3.5: Relu Function

In this work, Convolutional Neural Networks (CNN) are applied for the purpose of feature extraction of sexually transmitted disease images. The performance of different classifier is evaluated for comparison. Augmented RGB image data i.e. multiple instances of single image are generated by using rotation and translation operations from Herlev dataset centered on nucleus are used to increase the size of cervical image dataset. A pre-trained CNN named Alex-Net trained on a natural dataset named ImageNet dataset is used to fine-tune a new CNN on cervical dataset to learn discriminative information between images containing malignant and benign cells depending on deep hierarchical features. Fine-tuned CNN is then used to classify cell patches roughly centered on nucleus.

4. RECOGNITION RESULTS

On the this database, CNN is the simple method for intruder identification and tracking. First and foremost, as compared with other methods, its precision is incredibly high. Furthermore, the implementation rate is faster. As the kernel's difficulty reduces, the process becomes quicker. In term of runtime, the other methods are almost identical.

4.1 EXPERIMENTAL SETUP

While evaluating the neural network (CNN) on the datasets in this chapter, we took into account two values. They are proportion precision and time in milliseconds, respectively. The test is run in a Python Anaconda setting with 32GB RAM and a 2.9 GHz Intel Core i9 processor. We've come to pose all of our results here after conducting many studies with the Neural networks. Both experiments were run on an AMD Magny Cours with dual 12 cores as well as an NVIDIA Tesla C2050 GPU with 14 Stream CPUs (32 cores each). Both 115 characteristics were used to build and prepare the Convolutional algorithm. Fallback settings Neurotic for Keras modeling were used. Using Keras technology, learning was restricted to 100 iterations with the external variable of cross validation. This means that the template is really only equipped until the rating on the test dataset does not worsen, without curse of dimensionality the training examples.

1. The generalization ability of the models on video data with strong motion blur.

As human detection is included in the challenge of object detection, it is possible that the models can also to some extent handle data from Tobii Pro Glasses 2 videos, despite the differences between the two datasets. Therefore, we first evaluate the performance of pre-trained models on both datasets. The evaluated models include YOLO, SSD300 (SSD with input size 300 × 300) and SSD512 (SSD with input size 512 × 512).

2. Fine-tuning models with Pascal VOC 2012 dataset.

Pascal VOC has larger set of images with more complete annotation. In contrast, although Shopping dataset is extracted from 12 videos with 2,914 frames, the amount can still be insufficient for a deep-learning network, not to mention the annotations are only limited to human objects.

Taking the advantage of fine-tuning technology introduced in 3.4, we first of all fine-tune our models on Pascal VOC with “person” related data.

We also use data augmentation method during the training phase. Data augmentation is largely used to increase the size of dataset. Typical methods include image cropping, flipping, rotating, etc. It is also said that data augmentation increases the performance of models, as well as helps generalization .In this thesis, data augmentation not only enlarges the size of the dataset, but also helps narrow down the differences between Pascal VOC and video data. Aside from image cropping and flipping, we also simulate motion blur in images. As the videos are recorded by a head device which enables movement in all directions, we adopt 4 motion blur directions–up-down, left-right, and two diagonals. The ratio of original: flipped: blurred: cropped images is approximately 2: 2: 2: 1. Furthermore, experiments are designed to test several detailed factors of the networks, according to the architectures and observed results.

4.2 FINE-TUNING ON SHOPPING DATASET

In the previous sections, our experiments mainly focus on training with Pascal VOC dataset, only with technical tricks to modify the data or the model to have a better generalization on video data with motion blur. However, if the dataset is large enough, training directly on Shopping dataset should be the most ideal way. In this section, we fine-tuned SSD300 (opponent-class) and YOLO 9 9×4, to study how fine-tuning the models with video data can improve the performance. Shopping dataset contains 2,914 frames from 12 videos, in which 500 are left out for validation and evaluation. As we choose to fine-tune on SSD300 with opponent classes, negative samples are required. As a result, we also take false detection results from a poorly trained YOLO network as negative samples to enlarge the annotation. It is obvious that fine-tuning on Shopping dataset improves the performance on video data, as expected. The mAP (mean average Precision) increased by 11% and 14% for YOLO and SSD respectively. However, we also notice that after fine-tuning with Shopping dataset, the generalization ability on Pascal VOC data drops quickly, by 11% and 15% respectively.

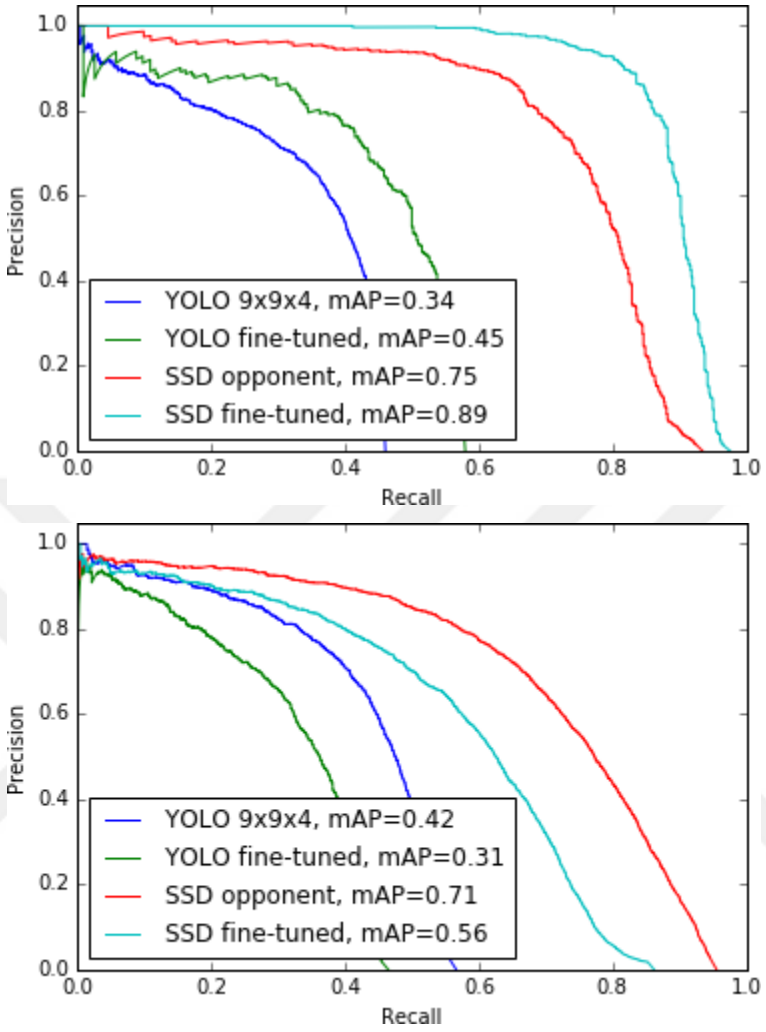

Figure 4.1: mAP of models before and after fine-tuning on Shopping dataset.

4.3 SSD PERFORMANCE ON DIFFERENT SCALES AND LOCATIONS OF OBJECTS

SSD has clear definitions on scales and aspect ratios, which give a better performance on different scales of objects than YOLO. However, still it shows different ability at different scales and locations. In this experiment, we plot the centers of ground truth bounding boxes as well as their scales, normalized to [0, 1]. Then we evaluate the SSD models and record the positive detection results, i.e. the predictions matching ground truth with IoU (Intersection over union) > 0.5, along with their confidence level. Those ground truths not able to be detected are set with confidence level 0.0. Specially, according to our project target–human detection, we here place more importance on precision. As a result, further tests will be based on SSD300 model trained with opponent classes. With this model, the results are shown in Figure 4.2 and 4.3.

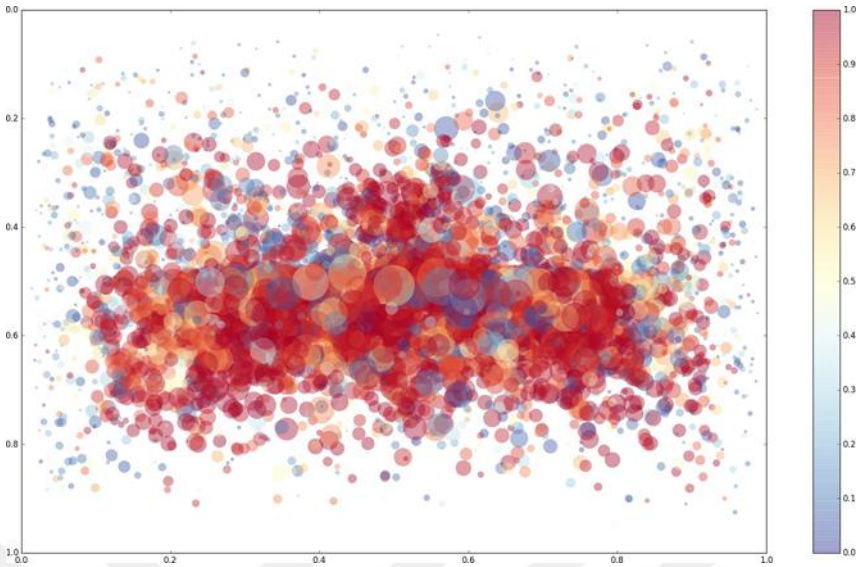

Figure 4.2: on Pascal VOC 2007 test

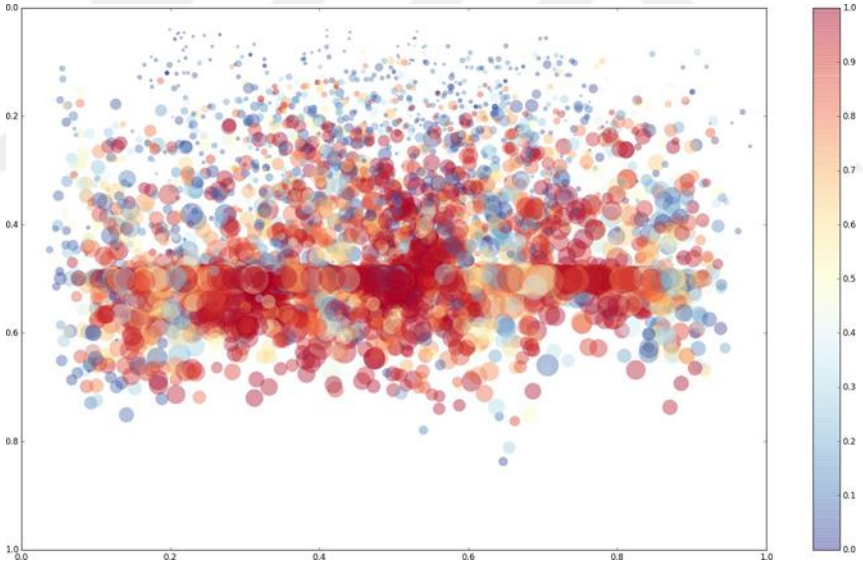

Figure 4.3: on Shopping dataset

Figure 4.2 and Figure 4.3 shows the SSD prediction confidences on different location of images. Each point indicates the position of the center of a ground truth bounding box. The diameters of the points show the scale of the object, i.e. a larger point indicates a larger scale. The color bar is corresponding to the confidence level (threshold) of detection from the SSD300 model trained with opponent classes, from scale 0.0 to 1.0. When setting a threshold, if the confidence level of a bounding box is lower than the threshold, this detection is a false negative. With this figure, it can be clearly shown with a threshold, which of these objects can be detected.

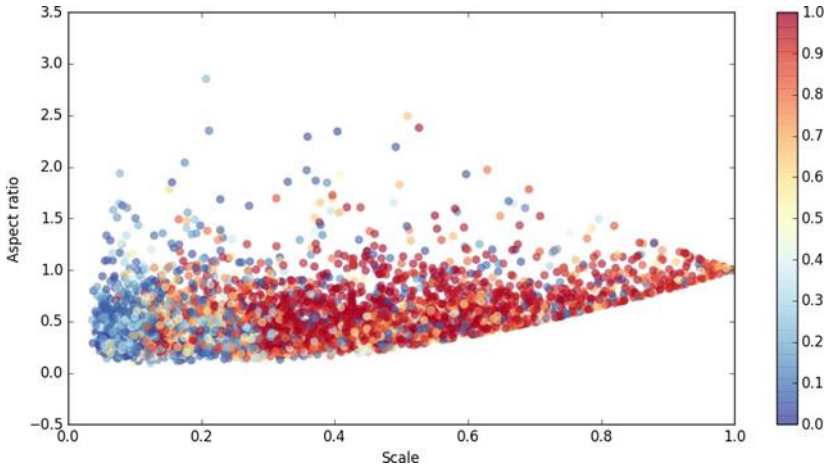

Figure 4.4: Pascal VOC 2007 test

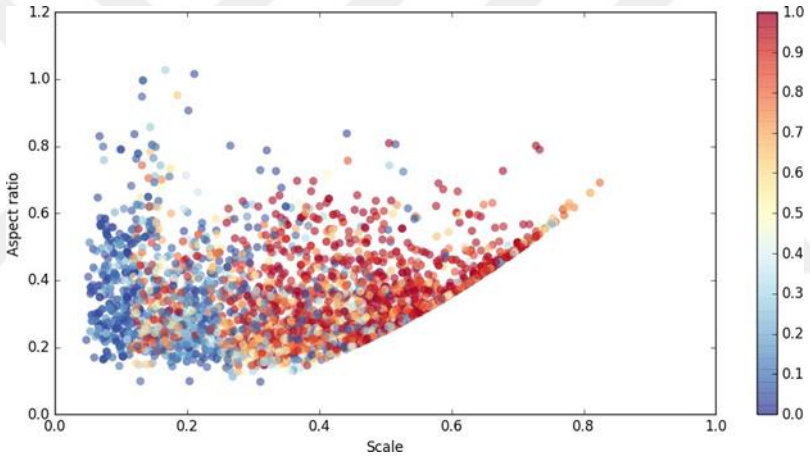

Figure 4.5: Shopping dataset

Figure 4.5 and 4.6 SSD prediction confidences on different scales of objects. Each point Indicates a ground truth bounding box. The definition of color bar is the same in Figure 4.2 and figure 4.3. From both of the figures we can have a rough estimate about how recall rate should be at different threshold levels (shown by color bar). From 4.3, it can be seen that usually the objects at the edge of the images also have small scales, and the bounding boxes with larger scales are mostly located at the middle of the images. It is more clear in Figure 4.4 and Figure 4.5 that SSD has a better performance when the scale of an object is larger. Especially if the threshold is chosen as 0.4, all the cold-colored bounding boxes will be missed in detection, in which the objects with small scales counts for the most part. We further tested whether the outputs of the network show a similar pattern to ground truth. We collected all the bounding boxes predicted on Shopping dataset.

Without filtering predictions by threshold, we get the result shown in Figure 4.4 and figure 4.5. It is obvious that the clusters are quite similar to pre-defined scales and aspect ratios. It is notable that there are also some clusters between the scales. They correspond to the extra scale for each layer at aspect ratio 1, calculated by $\sqrt{(s_k s_{(k+1)})}$.

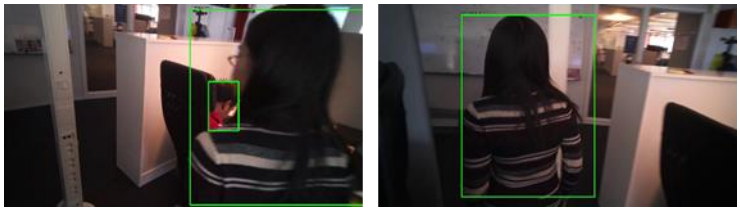

Figure 4.6: Samples of detection result from dataset. Video 1.

The results are shown in Figure 4.6 and 4.7. As can be seen in theses samples, the model is quite tolerant to motion blur, and has the ability to detect part of human. It has a good performance especially when the human is at the foreground of the frame and has clear differences with background. The Figure 4.7 shows the human detection on video stream.

Figure 4.7: Samples of detection result from Office dataset. Video 2.

4.3.1 Negative samples

Meanwhile, negative detection results still occur. The negative results mainly include 3 types: wrong detection (false positive), missed detection (false negative), and large variance on location and size of bounding boxes. The results are shown as below:

1. Wrong detection (false positive)

Some samples of wrong detection results are shown in Figure 4.8.

Figure 4.8: Samples of wrong detection results.

It is shown that wrong detection is more likely to happen when other objects have similar characters to human. For example: hanging clothes. As clothes always appear to be part of human in human detection, hanging clothes tends to have similar texture and shape with human. It is not surprising the model can detect clothes as human when the model is also able to handle the case when there is only part of a human appears in the image. Other frequently occurred wrong detection results include cash machines, pot plants and blurring objects. In the last case, there shows a pattern that when the background consists of objects with noisy colors and gradients, it may detect the background as human. It is also a common case that the model may miss the objects in the image. objects with a small scale are hard to detect for SSD models. Especially the video data from the head-mounted camera have large resolution, so that when they are resized to 300×300, humans in background may only count for a few pixels. Similarly, a human at the corner or the edge of an image is less likely to be detected since the humans are less likely to be entire, which adds up the difficulty to detection. On the other hand, the performance is also not satisfactory when the human has a similar pattern as the background or when motion blur is severe. It is also notable that SSD doesn't ensure to detect the same object during a continuous series of frames. This result requires further investigation into the network, which is not included in the thesis. Figure 4.9 shows the detection time of SSD and YOLO.

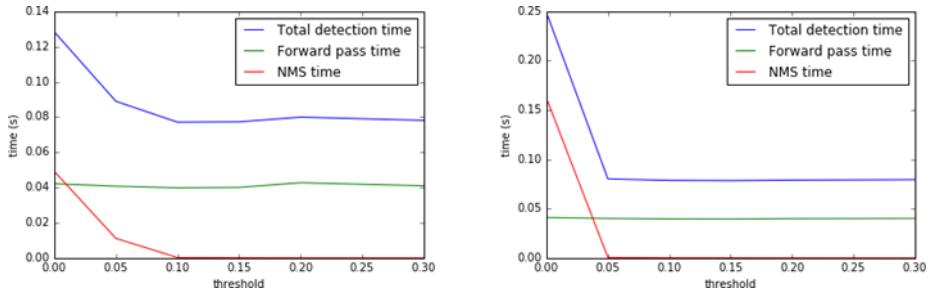

Figure 4.9: The graph shows the time duration for the detection of humans using SSD and YOLO

The speed of SSD and YOLO are almost at a same level, where SSD outperforms YOLO a little with 0.077s against 0.079s at threshold 0.2. The speed of forward pass for both SSD and YOLO are stable at about 0.04s. The time for non-maximum suppression plays very little part when the threshold is large enough to filter out most of the unreliable results. Other time-consuming functions including reading and resizing image, etc. As a result, both of the models have ability to carry out real-time detection on GPU.

5. DISCUSSION

5.1 DISCUSSION

Human detection Tracking in neural network applications are simply limitless. What does this mean for the future of artificial intelligence? It means that more efficient algorithms for training neural networks need to be created [42]. One such algorithm, fastprop, is seeing usage. Fastprop is based around the instantaneous relationship between weights and outputs as the network is being propagated. Although still in experimental stages, tests have been promising [43]. Of course, it is possible to use other algorithms for training neural networks; backpropagation is generally agreed, however, to be the best class of algorithms to train neural networks. However, backpropagation is not without its faults. One glaringly obvious problem is that backpropagation requires a large amount of training data. This training data also must be labeled: this combination presents a tedious problem for interns and grad students alike. annotating these images is incredibly time consuming. This problem can be solved by using what's known as unsupervised learning. Other criticisms of the backpropagation algorithm include sheer amount of computational power needed to perform these some of the problems. It may not be the most efficient use of processing power depending on the situation. Interesting problems in the future of neural networks are limitless. Google tore out the guts of its most important features: image search to voice recognition, recreating them from the ground up with machine learning. Computer vision is an incredibly important field that will have implications from self-flying planes to Snapchat filters to allowing the blind to see. The possibilities are endless but remember that neural networks are just a tool, not the end-all be-all of artificial intelligence. Similarly, backpropagation is not the only way to train neural networks. One should not plan on throwing algorithms at a problem until it works but consider each problem carefully. Neural networks are just a tool in the toolbox of programmers.

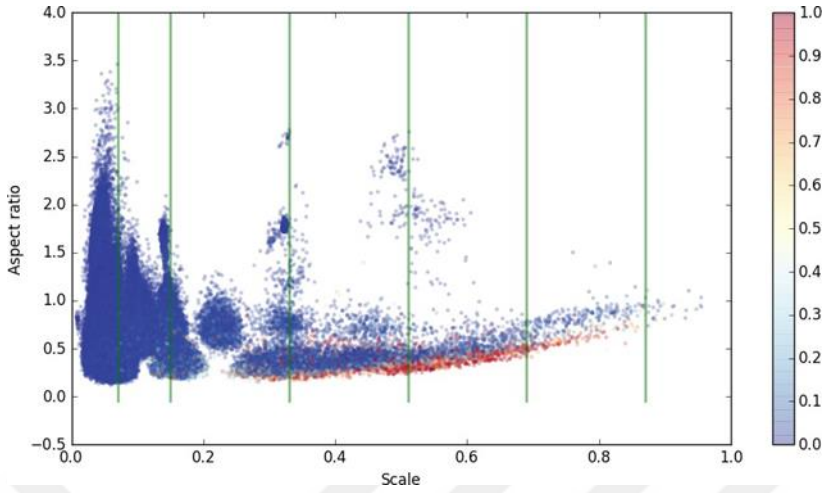

Figure 5.1: All predictions without selection by threshold. Each of the points is a prediction of bounding boxes. The green lines show the pre-defined scales in SSD model. The red points indicate true-positive predictions while the blue points indicate false-positive predictions.

As the color shows, most of the predictions are at low confidence level. Many of them actually are not aligned with common aspect ratios in Shopping dataset. However, after filtering by threshold 0.4, we get Figure 5.1, which shows a high accuracy of the model. Meanwhile, almost all of the predictions with redundant aspect ratios are filtered out. Therefore, it further proves that the aspect ratios larger than 1 defined in original SSD models are redundant,

In this thesis Firstly, this Algorithm will fill the gap of research where a separate model for each Human detection device in the network has been created. The motivation is that training, deploying and maintaining separate models for each device could quickly become burdensome in big networks especially for tracking systems. Some details, such as activation function choice and optimization algorithm choice are this work's original contributions, as they were not described in detail in the referred work [39]. Additionally, this work will employ a feature selection mechanism and will predict the accuracy of feature selection parameters efficiently. This will allow for assessment of artificial neural network method accuracy on smaller feature subsets and for making a fairer comparison on human detection and classification of different Machine Learning algorithms which are used in it.

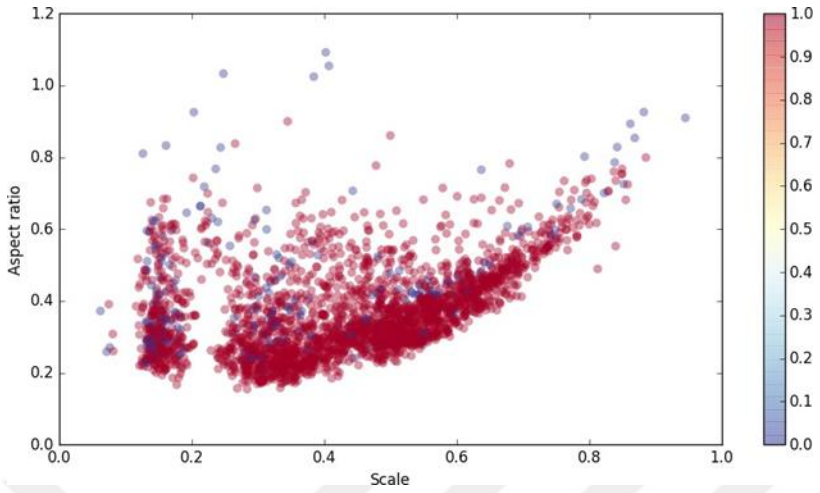

Figure 5.2: Precision of predictions with threshold 0.4. Each of the points is a prediction of bounding boxes. The red points indicate true-positive predictions while the blue points indicate false-positive predictions.

5.2 ROLE OF CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network algorithm is a classification algorithm that classifies data into groups that augments the utility for researchers [40]. Using CNN we can split data according to its classification. CNN work well on data that is continuous too, because if the regularization parameters are set high, it shall still form a very thin hyper plane to set the data even when it is continuous. Since we have lots of data and it is arranged in a continuous manner after implementing CNN on the data set, it shows a run-time accuracy of 85 percent in 4.55s. Similarly, CNN is tested on the network layer data where it classifies by creating a hyper plane to separate attacking and non-attacking classes. Since the computational power required on the actual data set was high, we ran it on a pre-processed data set and where it concluded with an accuracy of 88.8 percent but took 5.5s to finish. A convolutional neural network was used, and it determines a universal indicator from latent regression attributes using inter testing. The neuronal method is used to build the machine, which sets a likelihood threshold level that ultimately defines the two groups. In 2.17s, this algorithm achieves a 91.6 percent accuracy. The aim of this research was to demonstrate that machine learning algorithms are used to recognize humans, and these processes are based to datasets from various sources, with a neural net trained for every dataset collected. Another interesting aspect of the studied approach is that the accuracy of the neural network seems to stay relatively constant with varying number of features. If such a system were to be

implemented in real life, with just one neural network for a whole (sub) network, and not separate models for each video dataset, then the size of the neural network model, which increased quadratically with the number of features used. Deformable Part-based Model (DPM) [43] was the original condition approach for identification. The approach can be thought of as an expansion of directed gradients linear regressions (HOG). Both a rough global representation of the input frame and six sections of the object in a better resolution are used to rate the expected items. HOG is used to define growing data. In this way, the HOG multi-model will address the issue of shifting points of view. Through preparation, a transient SVM (latent SVM) is used to reduce the detection method to a class label. Part places are viewed as a latent constructs. Because of its sturdiness, the approach had a huge impact. This thesis adopts single-frame detection for humans in video streams from head-mounted cameras, i.e. the detection is only based on the content of video at each frame and no temporal information is included. In this way, it can provide a prior study for the human tracking problem. On the other hand, owing to the promising development in CNNs, we strive to research on how robust different structures of CNNs (You look only once (YOLO) and Single Shot Multibox Detector (SSD) perform on challenging videos with fast motion and strong blur, and how the structures may influence the results.

6. CONCLUSION

6.1 CONCLUSION

In this novel research work, we have developed a technique for the Detection of Humans in Video Streams Using Convolutional Neural Networks. The techniques we used in this thesis are YOLO and SSD, and along with these two techniques CNNs would perform on human recognition on recordings catching characteristic vision. With a superior planned structure, SSD beats YOLO on each angle, including accuracy and review, assembly speed and speculation capacity. In the interim, SSD has a decent exhibition when preparing different classes, and most of the times outflanks the first model prepared with 20 classes, the technique we used is YOLO and YOLO's preparation results are unsuitable when just single-class object is prepared. With further trials on SSD models, it shows that preparation with rival classes makes the system more strong at accuracy rate. Interestingly, preparing with just one class will in general have a superior review, yet way bound to deliver bogus identifications than the model prepared with rival classes. It is likewise evident that SSD is fit to learn obscured designs well overall, coming about to a superior mAP on Shopping dataset, which additionally demonstrates that information expansion is significant for this situation. This research intent to train the proposed system with Convolutional neural network on any open-source dataset of video objects. And for the methodology of proposed system, CNN would be utilized. CNN method is one of the most advance and efficient method of getting better results by combination of different agents to perform any task on set of rules. The Convolutional neural network Libraries will be used in python. Dataset for video-based human detection will be procured from any open-source repository like Kaggle like UCL or Open Sets and the name of the datasets are Pascal VOC 2007 and shopping dataset. For the implementation of this thesis, Python would be forwarded in use as it is the most efficient tool for processing large scale data with more accuracy. Many existing deep learning methods like convolutional neural network (CNN) or deep CNN (CNN) are capable to cover only one part that is appearance based feature extractions. We have achieved the accuracy of 91.6%.

6.2 FUTURE RECOMMANDATION

There are as yet numerous perspectives that can be improved later on. In this proposition we just contemplated two ordinary instances of CNNs for the detection of humans in video streams.

Be that as it may, numerous papers are seeking to improve the models. It is additionally valuable if the video datasets can be amplified and have explanations of higher caliber. Until further notice, the outcomes indicated that tweaking on Shopping dataset to a great extent hurts the speculation capacity of human. The explanation isn't sure, which can either came about because of the contrasts between the Pascal VOC and Shopping dataset, or basically from the nature of Shopping dataset as it is excessively little. At long last, the examination is primarily centered around the exhibition of the models. It ought to be helpful if some pre-or post-handling of pictures can be added to the model. Constrained by computational assets, models are not prepared with huge scope. In spite of the structures, a few contrasts between our preparation and the ones in unique paper incorporates: a littler bunch size, shorter preparing time. The cluster size we utilized for YOLO and SSD are 25 and 15 separately, against 128 and 64 in unique paper. The overfitting perception doesn't show the model is likewise overfitting to preparing information. Subsequently, it is conceivable to have better outcomes on the off chance that we train the models with bigger emphases.

REFERENCES

[1] Jeon, E.S.; Kim, J.H.; Hong, H.G.; Batchuluun, G.; Park, K.R. Human detection based on the generation of abackground image and fuzzy system by using a thermal camera. Sensors 2016, 16, 453.

[2] Li, J.; Gong, W. Real time pedestrian tracking using thermal infrared imagery. J. Comput. 2010, 5, 1606–1613.

[3] Chen, Y.; Han, C. Night-time pedestrian detection by visual-infrared video fusion. In Proceedings of the 7th World Congress on Intelligent Control and Automation, Chongqing, China, 25–27 June 2008; pp. 5079–5084.

[4] Huang, K.; Wang, L.; Tan, T.; Maybank, S. A real-time object detecting and tracking system for outdoor night surveillance. Pattern Recognit. 2008, 41, 432–444. [CrossRef]

[5] Nazib, A.; Oh, C.-M.; Lee, C.W. Object detection and tracking in night time video surveillance. In Proceedings of the 10th International Conference on Ubiquitous Robots and Ambient Intelligence, Jeju island, Korea, 31 October–2 November 2013; pp. 629–632.

[6] Tau 2. Available online: http://www.flir.com/cores/display/?id=54717

[7] Ge, J.; Luo, Y.; Tei, G. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. IEEE Trans. Intell. Transp. Syst. 2019, 10, 283–298.

[8] Gonzalez, R.C.; Woods, R.E. Digital Image Processing, 3rd ed.; Prentice Hall: New Jersey, NJ, USA, 2010.

[9] Coltuc, D.; Bolon, P.; Chassery, J.-M. Exact histogram specification. IEEE Trans. Image Process. 2006, 15, 1143–1152. [CrossRef] [PubMed]

[10] Coltuc, D.; Bolon, P. Strict ordering on discrete images and applications. In Proceedings of the IEEE International Conference on Image Processing, Kobe, Japan, 24–28 October 1999; pp. 150–153.

[11] Jeon, E.S.; Choi, J.-S.; Lee, J.H.; Shin, K.Y.; Kim, Y.G.; Le, T.T.; Park, K.R. Human detection based on the generation of a background image by using a far-infrared light camera. Sensors 2015, 15, 6763–6788. [CrossRef] [PubMed]

[12] Jen, T.-C.; Hsieh, B.; Wang, S.-J. Image contrast enhancement based on intensity-pair distribution.cIn Proceedings of the IEEE International Conference on Image Processing, Genova, Italy, 11–14 September 2005; pp. 1–4.

[13] Ke, S.R.; Thuc, H.; Lee, Y.J.; Hwang, J.N.; Yoo, J.H.; Choi, K.H. A review on video-based human activity recognition. Computers 2013, 2, 88–131.

[14] Dawn, D.D.; Shaikh, S.H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis. Comput. 2016, 32, 289–306. [CrossRef]

[15] Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. Image Vis. Comput. 2017, 60, 4–21. [CrossRef]

[16] Kumar, S.S.; John, M. Human activity recognition using optical flow based feature set. In Proceedings of the 2016 IEEE International Carnahan Conference on Security Technology (ICCST), Orlando, FL, USA, 24–27 October 2016; pp. 1–5.

[17] Guo, K.; Ishwar, P.; Konrad, J. Action recognition using sparse representation on covariance manifolds of optical flow. In Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, USA, 29 August–1 September 2010; pp. 188–195.

[18] Niu, F.; Abdel-Mottaleb, M. HMM-based segmentation and recognition of human activities from video sequences. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; pp. 804–807.

[19] Raman, N.; Maybank, S.J. Activity recognition using a supervised non-parametric hierarchical HMM. Neurocomputing 2016, 199, 163–177. [CrossRef]

[20] Liciotti, D.; Duckett, T.; Bellotto, N.; Frontoni, E.; Zingaretti, P. HMM-based activity recognition with a ceiling RGB-D camera. In Proceedings of the ICPRAM—6th International

Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017.

[21] Ma, M.; Fan, H.; Kitani, K.M. Going deeper into first-person activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1894–1903.

[22] Lipton, A.J.; Fujiyoshi, H.; Patil, R.S. Moving target classification and tracking from real-time video. In Proceedings of the IEEE Workshop on Applications of Computer Vision, Princeton, NJ, USA, 19–21 October 1998; pp. 8–14.

[23] Oren, M.; Papageorgiou, C.; Sinha, P.; Osuna, E.; Poggio, T. Pedestrian detection using wavelet templates.In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17–19 June 1997; pp. 193–199.

[24] Viola, P.; Jones, M.J.; Snow, D. Detecting pedestrians using patterns of motion and appearance. In Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 734–741.

[25] Mikolajczyk, K.; Schmd, C.; Zisserman, A. Human detection based on a probabilistic assembly of robust part detectors. Lect. Notes Comput. Sci. 2004, 3021, 69–82.

[26] Arandjelovi´c, O. Contextually learnt detection of unusual motion-based behaviour in crowded public spaces. In Proceedings of the 26th Annual International Symposium on Computer and Information Science, London, UK, 26–28 September 2011; pp. 403–410.

[27] Martin, R.; Arandjelovi´c, O. Multiple-object tracking in cluttered and crowded public spaces. Lect. Notes Comput. Sci. 2010, 6455, 89–98.

[28] Khatoon, R.; Saqlain, S.M.; Bibi, S. A robust and enhanced approach for human detection in crowd. In Proceedings of the International Multitopic Conference, Islamabad, Pakistan, 13–15 December 2012; pp. 215–221.

[29] Rajaei, A.; Shayegh, H.; Charkari, N.M. Human detection in semi-dense scenes using HOG descriptor and mixture of SVMs. In Proceedings of the International Conference on Computer and Knowledge Engineering, Mashhad, Iran, 31 October–1 November 2013; pp. 229–234.

[30] Lee, J.H.; Choi, J.-S.; Jeon, E.S.; Kim, Y.G.; Le, T.T.; Shin, K.Y.; Lee, H.C.; Park, K.R. Robust pedestrian detection by combining visible and thermal infrared cameras. Sensors 2015, 15, 10580–10615. [CrossRef]

[31] Batchuluun, G.; Kim, Y.G.; Kim, J.H.; Hong, H.G.; Park, K.R. Robust behavior recognition in intelligent surveillance environments. Sensors 2016, 16, 1010. [CrossRef] [PubMed]

[32] Serrano-Cuerda, J.; Fernández-Caballero, A.; López, M.T. Selection of a visible-light vs. thermal infrared sensor in dynamic environments based on confidence measures. Sensors 2014, 4, 331–350. [CrossRef]

[33] Fukui, H.; Yamashita, T.; Yamauchi, Y.; Fujiyoshi, H.; Murase, H. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In Proceedings of the IEEE Intelligent Vehicles Symposium, Seoul, Korea, 28 June–1 July 2015; pp. 223–228.

[34] Angelova, A.; Krizhevsky, A.; Vanhoucke, V.; Ogale, A.; Ferguson, D. Real-time pedestrian detection with deep network cascades. In Proceedings of the 26th British Machine Vision Conference, Swansea, UK, 7–10 September 2015; pp. 1–12.

[35] Komagal, E.; Seenivasan, V.; Anand, K.; Anand raj, C.P. Human detection in hours of darkness using Gaussian mixture model algorithm. Int. J. Inform. Sci. Tech. 2014, 4, 83–89. [CrossRef]

[36] Pawłowski, P.; Piniarski, K.; D ˛abrowski, A. Pedestrian detection in low resolution night vision images. In Proceedings of the IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications, Pozna ´n, Poland, 23–25 September 2015; pp. 185–190.

[37] Wang, W.; Zhang, J.; Shen, C. Improved human detection and classification in thermal images. In Proceedings of the IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 2313–2316.

[38] Wang, W.; Wang, Y.; Chen, F.; Sowmya, A. A weakly supervised approach for object detection based on soft-label boosting. In Proceedings of the IEEE Workshop on Applications of Computer Vision, Tampa, FL, USA, 15–17 January 2013; pp. 331–338.

[39] Li, W.; Zheng, D.; Zhao, T.; Yang, M. An effective approach to pedestrian detection in thermal imagery. In Proceedings of the International Conference on Natural Computation, Chongqing, China, 29–31 May 2012; pp. 325–329.

[40] Neagoe, V.-E.; Ciotec, A.-D.; Barar, A.-P. A concurrent neural network approach to pedestrian detection in thermal imagery. In Proceedings of the International Conference on Communications, Bucharest, Romania, 21–23 June 2012; pp. 133–136.

[41] Olmeda, D.; Armingol, J.M.; Escalera, A.D.L. Discrete features for rapid pedestrian detection in infrared images. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; pp. 3067–3072.

[42] Lin, C.-F.; Lin, S.-F.; Hwang, C.-H.; Chen, Y.-C. Real-time pedestrian detection system with novel thermal features at night. In Proceedings of the IEEE International Instrumentation and Measurement Technology Conference, Montevideo, Uruguay, 12–15 May 2014; pp. 1329–1333.

[43] Olmeda, D.; Premebida, C.; Nunes, U.; Armingol, J.M.; Escalera, A.D.L. Pedestrian detection in far infrared images. Integr. Comput. Aided Eng. 2013, 20, 347–360.