TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

THESIS NO: 073/MSI/611

Employee Face Recognition by Region Proposal Networks and Faster R-CNN

by

Rajan Gyawali

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
LALITPUR, NEPAL

MAY, 2019

Employee Face Recognition by Region Proposal Networks and Faster R-CNN

by

Rajan Gyawali

073/MSI/611

Thesis Supervisor

Dr. Dibakar Raj Pant

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Information and Communication Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Tribhuvan University

Lalitpur, Nepal

May, 2019

COPYRIGHT©

The author has agreed that the library, Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, may make this thesis freely available for inspection. Moreover, the author has agreed that the permission for extensive copying of this thesis work for scholarly purpose may be granted by the professor(s), who supervised the thesis work recorded herein or, in their absence, by the Head of the Department, wherein this thesis was done. It is understood that the recognition will be given to the author of this thesis and to the Department of Electronics and Computer Engineering, Pulchowk Campus in any use of the material of this thesis. Copying of publication or other use of this thesis for financial gain without approval of the Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus and author’s written permission is prohibited.

Request for permission to copy or to make any use of the material in this thesis in whole or part should be addressed to:

Head

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Pulchowk, Lalitpur, Nepal

DECLARATION

I declare that the work hereby submitted for Master of Science in Information and Communication Engineering (MSICE) at IOE, Pulchowk Campus entitled “Employee Face Recognition by Region Proposal Networks and Faster R-CNN” is my own work and has not been previously submitted by me at any university for any academic award.

I authorize IOE, Pulchowk Campus to lend this thesis to other institution or individuals for the purpose of scholarly research.

Rajan Gyawali
073/MSI/611
Date: May 2019

RECOMMENDATION

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled **“Employee Face Recognition by Region Proposal Networks and Faster R-CNN”**, submitted by **Rajan Gyawali** in partial fulfillment of the requirement for the award of the degree of **“Master of Science in Information and Communication Engineering”**.

....

Supervisor: Dr. Dibakar Raj Pant,

Associate Professor,

Department of Electronics and Computer Engineering,

Institute of Engineering, Tribhuvan University

....

External Examiner: Mr. Subash Dhakal,

IT Director,

Department of National ID and Civil Registration,

Ministry of Home Affairs

....

Committee Chairperson: Dr. Basanta Joshi,

Program Coordinator,

Information and Communication Engineering

Institute of Engineering, Tribhuvan University

Date: May, 2019

DEPARTMENTAL ACCEPTANCE

The thesis entitled “Employee Face Recognition by Region Proposal Networks and Faster R-CNN”, submitted by Rajan Gyawali in partial fulfillment of the requirement for the award of the degree of “Master of Science in Information and Communication Engineering” has been accepted as a bonafide record of work independently carried out by him in the department.

....

Dr. Surendra Shrestha

Head of the Department
Department of Electronics and Computer Engineering,
Pulchowk Campus,
Institute of Engineering,
Tribhuvan University,
Nepal.

ACKNOWLEDGEMENT

I owe a great deal to express thanks to **Dr. Dibakar Raj Pant** for his encouragement, suggestions and continuous guidance throughout the course of my thesis work.

I am very grateful to **Dr. Surendra Shrestha,** Head of Department of Electronics and Computer Engineering, Pulchowk Campus for his support.

I pay my sincere gratitude to **Dr. Basanta Joshi,** coordinator of MSc Information and Communication Engineering for his guidance and help during this research work.

I am highly indebted to **Prof. Dr. Shashidhar Ram Joshi, Prof. Dr. Subarna Shakya, Dr. Sanjeeb Prasad Panday** and **Dr. Aman Shakya** for their insights and opinions regarding the thesis work. I would like to express my heartily gratitude towards the Institute of Engineering, Pulchowk Campus along with all my respected teachers, my friends, my family for giving me continuous support for their invaluable help without whom I wouldn’t have come so far in my thesis.

Rajan Gyawali
073/MSI/611
Pulchowk Campus

ABSTRACT

Face recognition is becoming popular in companies, supermarkets, hospitals etc. for security systems, human machine interaction and video surveillances. Employee face recognition is required to differentiate between employees and non-employees. Face recognition is a challenging task. The traditional machine learning algorithms like Principal Component Analysis, Support Vector Machines, etc. rely on image-based features such as edges and texture descriptors. In the recent trends, the Convolutional Neural Networks and deep learning algorithms have shown greater performance in face recognition. This thesis work uses region proposal network (RPN) to localize region of interests (faces) from the image and uses Faster R-CNN to output the region proposals' labels and bounding box associated with them. The proposed system consists of three sections. The first section uses CNN for features extraction. From these features, the second section generates region proposals using RPN. The third section classifies these region proposals using faster R-CNN and the employee face is recognized. The recognized face has a size of 128x128. The accuracy of the model is 96.0% in recognition of employees from Chokepoint dataset. The model is further tested with recorded employees' dataset of Nepal Telecom and shows an accuracy of 95.2%. The performance of the proposed method is evaluated on these datasets using confusion matrix. Further, visual and comprehensive evaluation using receiver operating characteristics curve for these datasets shows a clear distinction between employees and non-employees.

Keywords: Employee Face Recognition, Region Proposal Networks, Convolutional Neural Network, Faster R-CNN

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
FAR	False Acceptance Rate
FN	False Negative
FP	False Positive
FRGC	Face Recognition Grand Challenge
GPU	Graphical Processing Unit
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
R-CNN	Region Convolutional Neural Network
ROC	Receiver Operating Characteristics
ROI	Region of Interest
RPN	Region Proposal Network
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VGG	Visual Geometry Group

CHAPTER 1

INTRODUCTION

1.1 Introduction

Face recognition is becoming popular in companies, supermarkets, hospitals etc. for security systems and video surveillances. The conventional technique for face recognition is by using biometrics but has some challenging issues when used in unconstrained environments due to high variability in head poses, aging, occlusions, illumination conditions and facial expressions.

Several traditional machine learning algorithms such as PCA, SVM [1] have been used to detect and recognize faces, however these methods relied on hand-crafted features such as edges and texture descriptors. In the recent trends, the convolutional neural networks and deep learning methods have showed improved performance for handwriting recognition, object recognition and face recognition.

In this thesis work, employee face recognition system is proposed to differentiate employees and non-employees in offices, companies and the areas requiring access control. The proposed system uses region proposal network and faster R-CNN. The RPN is used to localize the employees face from background and faster R-CNN for recognition.

Figure 1.1: A woman holds a tablet featuring biometric 3D facial recognition software [2]

1.2 Problem Statement

For increasing security concerns in big companies, the usage of face recognition system is increasing. Traditional recognition systems include RFID cards and GPS devices [3]. These systems have weaknesses. Employees could forget the RFID card or the location device and anyone else can use these devices creating a potential security issue. Face recognition system eliminates the weaknesses of such devices and provides flexible solutions.

Traditional face recognition system requires the input to be frontal face region and can't be used in surveillance environment. With variations in head pose, lighting conditions, facial expressions these algorithms can't provide greater accuracy. Convolutional neural networks outperformed the traditional machine learning algorithms in recognizing faces with greater accuracy. However, face recognition in surveillance environment requires the localization of face from the background and recognition in short time. This has initiated the need of localization and recognition of employees face in real time scenario. The proposed system localizes the employees face from the surveillance environment and gives identification to differentiate between employees and non-employees.

1.3 Objectives

The objectives of this thesis work are:

i. To localize employee face from background using Region Proposal Network.
ii. To recognize the localized faces using Faster R-CNN.
iii. To validate the recognized faces.

CHAPTER 2

LITERATURE REVIEW

A lot of research and work has been done in the field of object detection and recognition. Object recognition and face recognition seems like the same concept; however, face recognition task seems to be always challenging because most of the features in faces are same. With recent trends in deep learning, object recognition and face recognition tasks have been simplified.

2.1 Related Works

CNN was proposed firstly by LeCun [4] and applied it on handwriting recognition. From his contributions, many scientists got true inspiration to work in this field. Krizhevsky et.al. [5] achieved best results when they published their work in ImageNet Competition. In 2012, AlexNet significantly outperformed all the prior competitors and won the challenge by reducing the top-5 error from 26% to 15.3%. The second-place top-5 error rate, which was not a CNN variation, was around 26.2%.

The runner-up at the ILSVRC 2014 competition is a variant of CNN and is developed by Simonyan et.al. [6]. The model was trained on 4 GPUs for 2–3 weeks. It showed a top-5 error of 7%.

Musab Coskun et. al. [7] proposed a convolutional neural network for face recognition with number of convolutional layers. They have used Georgia Tech Database and showed that the approach has improved the face recognition performance with better recognition results.

Sharma S et.al. [8] published the CNN based efficient face recognition technique using Dlib. They have emphasized the importance of the face alignment, thus the accuracy and False Acceptance Rate (FAR) is observed. Their computational analysis has showed the better performance than other state-of-art approaches. The work has been done on Face Recognition Grand challenge (FRGC) dataset and giving accuracy of 96% with FAR of 0.1%. Their system didn’t address the problems of pose variation and intensity variation.

Uijlings J.R.R et.al. [9] used selective search for generating possible object locations for use in object recognition. They combined the features of exhaustive search and segmentation to generate possible object locations. It initializes small regions in an image and merges them with a hierarchical grouping. The detected regions are merged according to a variety of color spaces and similarity metrics. This algorithm generates high quality locations yielding 99% recall. However, the time cost of generating region proposals is higher in selective search.

R. Girshick et. al. [10] has introduced the new way for accurate object detection and semantic segmentation using the Region Proposals combined with CNN, called as R-CNN. The process has been divided into components, the region proposal step and the classification step. Using selective search [9], an altogether of 2000 different region proposals that have the highest probability of containing an object are extracted and fed into a trained CNN to extract a feature vector for each region. A set of linear support vector machines (SVM) has been used for the classification. The vector was also fed into a bounding box regressor to obtain the most accurate coordinates. This object detection algorithm had given a 30% relative improvement over the best previous results on PASCAL dataset [11].

Shaoqing Ren et. al. [12] has further improved the results of selective search for object detection. The number of region proposals has been reduced from 2000 in RCNN [10] to 300. The model was tested on PASCAL dataset and the results are obtained faster than R-CNN.

Employee face recognition also requires near real time recognition. The proposed model extracts the region proposals by neural network and uses faster R-CNN for recognition.

CHAPTER 3

METHODOLOGY

3.1 Methodology

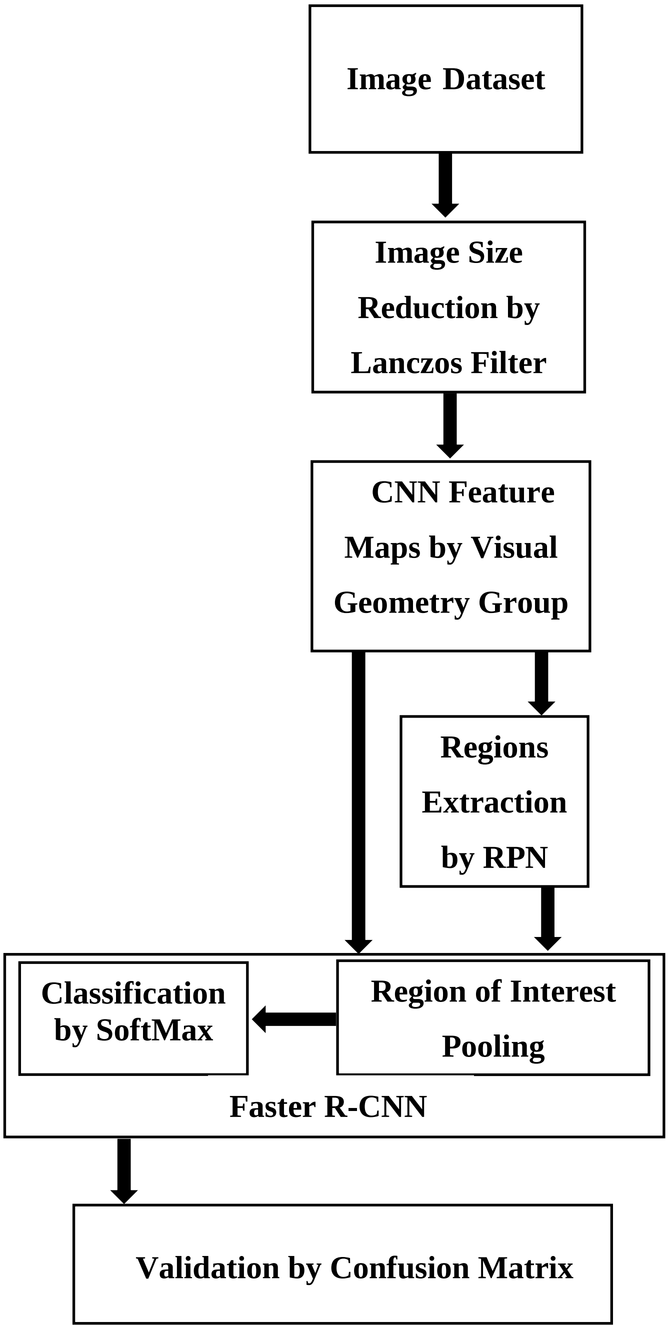


Image Dataset
Image Size Reduction by Lanczos Filter
CNN Feature Maps by Visual Geometry Group
Regions Extraction by RPN
Classification by SoftMax
Region of Interest Pooling
Faster R-CNN
Validation by Confusion Matrix

Figure 3.1: System Block Diagram

The system uses faster R-CNN for face recognition. In R-CNN [10], the region proposals are made by traditional methods, which costs a long time for preprocessing. Faster R-CNN makes region proposals by neural networks instead.

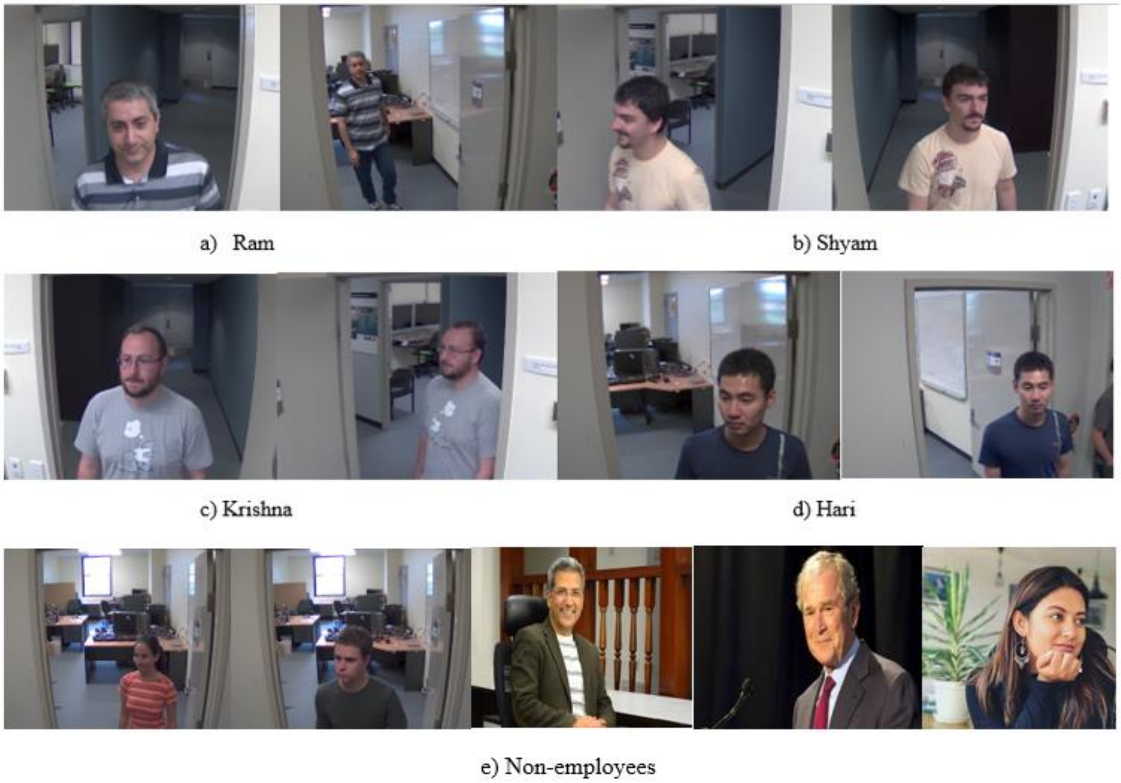
3.1.1 Datasets

WIDER Face dataset and Chokepoint dataset have been used. WIDER Face dataset has been used to train the VGG16 network for the purpose features extraction. Chokepoint dataset has been used as employees faces for the purpose of face recognition.

WIDER Face [14] dataset is a face detection benchmark dataset containing 32,203 images and labelled 393,703 faces. This dataset is open sourced by *Multimedia Laboratory, Department of Information Engineering, The Chinese University of Hong Kong*. The dataset consists of different images with variation in head pose, illuminations and facial expressions. This dataset is used for training purpose only to extract features of face and make network learn for face detection.

Figure 3.2: WIDER Face Dataset [11] to train RPN and faster R-CNN to learn features of face

ChokePoint [15] dataset has been used for employee's identification. This dataset has been open sourced for research purpose by *National ICT Australia Limited (NICTA)*. It consists images of different employees with variations in illumination conditions, pose, sharpness and alignment. The dataset has been used for the purpose of face recognition. Ram, Shyam, Krishna and Hari are the employees to be correctly recognized by the model and rest of the people have to be identified as "non-employee".



a) Ram

b) Shyam

c) Krishna

d) Hari

e) Non-employees

Figure 3.3: ChokePoint [15] dataset of employees and some non-employee dataset to recognize non-employee

For testing the consistency of the model, a separate dataset of employees recorded at Nepal Telecom is used. The images are captured by Pocophone smartphone with camera of 12Mp 1/2.55-inch sensor with 1.4µm pixels, f/1.9-aperture lens and frame rate of 30fps. The employee videos are taken in different angles and lighting conditions for a more diverse set of image representative of the employee's face.

Figure 3.4: Nepal Telecom employee dataset for the validation of model.

3.1.2 Image Size Reduction

Building an effective neural network model requires careful consideration of the network architecture as well as the input data format. The most common parameters are the number of images, image height, image width and number of channels. As CNN requires all of the input images to be of same size, image resizing is to be done. The image size of employees' dataset is 800x600 and to that of recognition network is 128x128. For image resizing the Lanczos filter [16] has been used as it provides detail preservation and minimal generation of aliasing artifacts.

The Lanczos filter is defined as below:

$$L(x; n > 0) = \begin{cases} sinc(x).\,sinc(\frac{x}{n}) & for \quad |x| \leq \, n \\ 0 & otherwise \end{cases} \tag{3.1}$$

Interpolation of a two-dimensional image *f* with a Lanczos filter of order *n* is performed with the following algorithm:

$$f(x,y) = \frac{1}{w} \sum_{i=-n+1}^{n} \sum_{j=-n+1}^{n} f(\lfloor x \rfloor + i, \lfloor y \rfloor + j). L(i - x + \lfloor x \rfloor; n). L(j - y + \lfloor y \rfloor; n) \quad (3.2)$$

where *(x, y)* are the coordinates of the interpolation point and $\lfloor \cdot \rfloor$ is the floor operator. The filter weight w is applied by division to preserve flux.

$$w = L(i - x + \lfloor x \rfloor; n). L(j - y + \lfloor y \rfloor; n) \quad (3.3)$$

3.1.3 Features Extraction by CNN

CNNs are the feed forward neural networks made up of many hidden layers. CNNs consist of filters or kernels or neurons that have learnable weights or parameters and biases. Each filter takes some inputs and does convolution. The components of CNN consist of following layers:

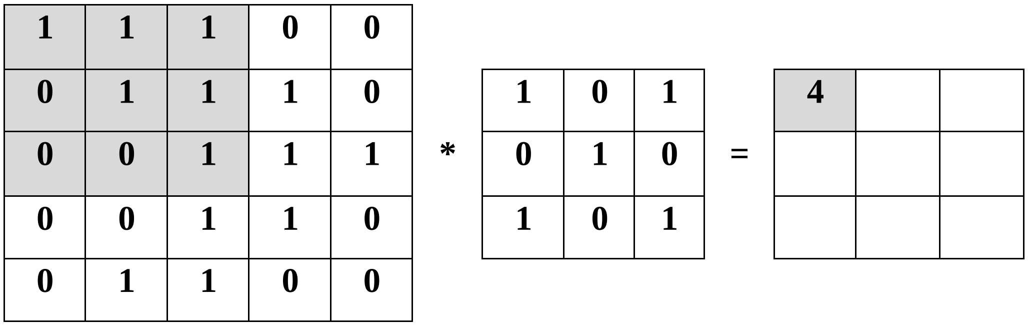
i. Convolutional Layer
ii. Rectified Linear Unit (ReLU) Layer
iii. Pooling Layer
iv. Fully Connected Layer

Convolutional Layer

Convolutional layer is the core building block of a convolutional network that performs most of the computational heavy lifting. Its primary purpose is to extract features from the input data which is an image. Convolution preserves the spatial relationship between pixels by learning features using small squares of input image. This produces a feature map or activation map in the output image and after then feature maps are fed as input data to the next convolutional layer. A convolution is done by multiplying a pixel's and its neighboring pixels color value by a matrix (kernel). The convolution formula is defined as below:

$$y[m,n] = x[m,n] * h[m,n] = \sum_{j} \sum_{i} x[i,j] h[m-i, n-j] \quad (3.4)$$

where, *y* is the convolved feature map, *x* is the input image and *h* is a kernel.



$$\begin{bmatrix}1&1&1&0&0\\0&1&1&1&0\\0&0&1&1&1\\0&0&1&1&0\\0&1&1&0&0\end{bmatrix} * \begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix} = \begin{bmatrix}4& & \\ & & \\ & & \end{bmatrix}$$

Figure 3.5: Illustration of Convolution operation in image processing

ReLU Layer

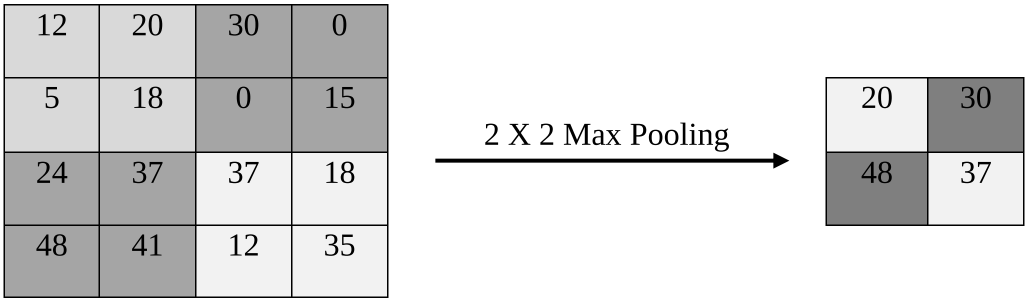
It is a non-linear operation similar to the rectification. It is an element wise operation that reconstitutes all negative values in the feature map by zero. The equation of ReLU operation is defined below:

$$f(x) = max\ (0, x) \tag{3.5}$$

where, *x* is the value in feature map.

Pooling Layer

Pooling layer reduces the dimensionality of each activation map and continues to have the most important information. The input images are divided into a set of non-overlapping rectangles. Each region is down-sampled by a non-linear operation like average or maxima. This layer gains better generalization, faster convergence, robust to translation and distortion and usually placed between convolutional layers.



$$\begin{bmatrix}12&20&30&0\\5&18&0&15\\24&37&37&18\\48&41&12&35\end{bmatrix} \xrightarrow{\text{2 X 2 Max Pooling}} \begin{bmatrix}20&30\\48&37\end{bmatrix}$$

Figure 3.6: Illustration of Max Pooling in image processing

Fully Connected Layer

This indicates that every filter in the previous layer is connected to every filter in the next layer. The output from the convolutional, pooling and ReLU layers are embodiments of high-level features of the input image. Using fully connected layer employs these features for classifying the input image into various classes based on training set.

Fully connected layer is the final pooling layer feeding the features to a classifier that uses Softmax activation function.

For the purpose of features extraction from the input images, a variant of Convolutional Neural Network called VGG16 [17] architecture has been used. VGG16 usually refers to a convolutional network for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG), which achieved very good performance on the ImageNet dataset.

For extracting features, the last fully connected layers have been excluded. In VGG16 architecture [17], there are 13 convolutional layers with ReLU activation function with stride of 1 and 5 MaxPooling layers with stride of 2. All the Conv layers have one padding and MaxPooling layers have zero padding. The feature map of last Convolutional Layer with feature map 512 is taken and fed to region proposal network. As the SoftMax output of head network is not taken it is not necessary to convert the images to size of 224x224.

The first few layers of the network learn general features like edges and color blobs. The later layers learn more problem specific features. The convolutional feature maps produced by this head network are passed as input to Region Proposal Network (RPN).

3.1.4 Regions Extraction by RPN

Face recognition from faces with background needs face segmentation. To localize the face, selection of sub-regions (patches) of the image is required before applying the recognition algorithm. Generation of these smaller sub-regions is done by use of Region Proposal Network.

The region proposal network [13] takes the feature maps provided by head network through a convolutional layer followed by ReLU activation. This convolutional layer has 512 channels as input and 512 channels as output as shown in Appendix D.

This output is run through two (1,1) kernel convolutional layers to produce background/foreground class scores and probabilities and their corresponding bounding box regression coefficients.

The main task of RPN network is to produce promising RoIs and that of classification network is to assign object class scores to each RoI. Therefore, training this network requires corresponding ground truth annotations i.e. the coordinates of the bounding boxes around the faces present in an image. The ground truth comes from the image dataset. The annotation file in the dataset contains the coordinates of the bounding box and the respective class label for each object present in an image.

For the working of Region Proposal Network, following layers are required:

a. Anchor Generation Layer

b. Region Proposal Layer

3.1.4.1 Anchor Generation Layer

This layer produces a set of bounding boxes (anchors) of varying sizes and aspect ratios. These anchors must be spread through the image and agnostic to the content of image. These anchors must enclose the foreground objects (faces) but most of the anchors won't. The goal of the RPN network is learning to identify the anchors enclosing the faces and calculate target regression coefficients. The identified anchor is transformed to a better bounding box fitting the face more closely.

Anchors with scales of {4, 8, 16, 32} and aspect ratios of {0.5, 1, 2} are used. This gives a total of 12 anchors for each grid in the image. Thus, there will be a total of W x H x 12 anchors where $W = w/16$ and $H = h/16$ where 16 is the sub sampling ratio and is equal to stride length. The anchors that lie outside of the image boundary have been excluded.

3.1.4.2 Region Proposal Layer

The inputs to proposed system are the "region proposals" that produce a sparse or a dense set of features. In this approach a sliding window technique is used to generate a set of dense candidate regions and the Region Proposal Network is used to rank these region proposals according to the probability of a region containing faces. The region proposal layer has to identify the background and foreground anchors and transform the foreground anchors by applying a set of regression coefficients to make them fit the face boundary.

The region proposal layer consists of "Region Proposal Network", Proposal Layer, Anchor Target Layer and Proposal Target Layer.

a) Proposal Layer

The proposal layer takes the anchor boxes produced by the anchor generation layer and reduces the number of anchors by applying non-maximum suppression based on the foreground scores. It also generates transformed bounding boxes by applying the regression coefficients generated by the RPN [13] to the corresponding anchor boxes. The detail methodology of region proposal layer is shown in Appendix C.

b) Anchor Target Layer

The goal of this layer is to select promising anchors that can be used to train the RPN network to distinguish between foreground and background regions and generate good bounding box regression coefficients for the foreground boxes.

Bounding Box Regression Coefficients

R-CNN [10] produces good bounding boxes that closely fit face boundaries. Bounding boxes are defined by the coordinates of top left corner, width and height. R-CNN tweaks these coordinates, width and height by applying a set of regression coefficients. Let the coordinates of the top left corner of the target and original bounding box be denoted by T_x, T_y, O_x, O_y respectively and the width/height of the target and original bounding box be denoted by T_w, T_h, O_w, O_h respectively. Then the regression coefficients are calculated as

$$t_x = \frac{T_x - O_x}{O_w} \tag{3.6}$$

$$t_y = \frac{T_y - O_y}{O_h} \tag{3.7}$$

$$t_w = log\frac{T_w}{O_w} \tag{3.8}$$

$$t_h = log\frac{T_h}{O_h} \tag{3.9}$$

RPN Loss

RPN loss is formulated to encourage the network to classify anchors as background or foreground and transform the foreground anchor to fit the face region more closely. RPN loss is the summation of classification loss and bounding box regression loss.

$$RPN\ Loss = Classification\ Loss + Bounding\ Box\ Regression\ Loss \qquad (3.10)$$

The classification loss uses cross entropy loss to penalize the incorrectly classified boxes and regression loss uses a function of the distance between the true regression coefficients and the regression coefficients predicted by the RPN [13].

Classification Loss:

The classification loss is the cross-entropy loss and is calculated as:

$$-\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right) \qquad (3.11)$$

where, *M* is the number of classes, y is a binary indicator (0 or 1) if class label *c* is the correct classification for observation *o* and *p* is the predicted probability observation *o* is of class *c*.

Bounding Box Regression Loss:

$$L_{loc} = \sum\nolimits_{u \in all\ foreground\ anchors} l_u \qquad (3.12)$$

Sum over the regression losses for all foreground anchors is calculated. Doing this for background anchors doesn’t make sense as there is no associated ground truth box for a background anchor.

$$l_u = \sum\nolimits_{i \in x,y,w,h} smooth_{L1}\left(u_i(predicted) - u_i(target)\right) \qquad (3.13)$$

The difference between the predicted (by the RPN) and target (calculated using the closest ground truth box to the anchor box) regression coefficients is taken to calculate regression loss. The smooth L1 function is defined as follows:

$$smooth_{L1}(x) = \begin{cases} \frac{\sigma^2 x^2}{2}, & \|x\| < \frac{1}{\sigma^2} \\ \|x\| - \frac{0.5}{\sigma^2}, & otherwise \end{cases} \qquad (3.14)$$

To calculate the RPN loss the classification score for the anchor boxes and target regression coefficients for the foreground anchor boxes are required.

A single ground-truth box may assign positive labels to multiple anchors. A negative label to a non-positive anchor will be assigned if its IoU ratio is lower than 0.3 for all ground-truth boxes. Anchors that are neither positive nor negative do not contribute to the training objective.

In anchor target layer all these parameters are calculated. The anchor boxes that lie within the image span are selected. The good foreground boxes are selected by first computing the intersection over union (IoU) of all anchor boxes with all ground truth boxes in the image.

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{3.15}$$

From IoU information two types of boxes are marked as foreground boxes:

a) For each ground truth box, all foreground boxes that have the maximum IoU overlap with the ground truth box.
b) Anchor boxes whose maximum overlap with sum ground truth box exceeds a threshold.

Boxes whose overlap are less than a negative threshold, are labelled as background boxes. Boxes that are neither foreground nor background are labelled as “don’t care”. These boxes are not used in RPN loss calculation. The thresholds to select anchor as a foreground box and background box are taken 0.7 and 0.3 respectively. Between these values anchors are marked as “don’t care”.

The inputs to anchor target layer are outputs from RPN (predicted foreground/background class labels, regression coefficients), anchor boxes (generated by anchor generation layer) and ground truth boxes. Similarly, the outputs are good foreground/background boxes with labels and target regression coefficients.

c) Proposal Target Layer

The proposal target layer selects promising ROIs from the list of ROIs output by the proposal layer. These promising ROIs are used to perform RoI pooling from the feature maps produced by the head layer and passed to the rest of the network that calculates predicted class scores and box regression coefficients.

The proposal target layer starts with the ROIs computed by the proposal layer. Using the max overlap of each ROI with all ground truth boxes, it categorizes the ROIs into background and foreground ROIs. Foreground ROIs are those for which max overlap exceeds a threshold (foreground threshold of 0.5). Background ROIs are those whose max overlap falls between the lower threshold and upper threshold of 0.1 and 0.5 respectively.

The bounding box target regression targets are computed between each ROI and the closest matching ground truth box (this includes the background ROIs also, as an overlapping ground truth box exists for these ROIs also).

The inputs to this layer are RoIs produced by the proposal layer and ground truth information. Similarly, the outputs produced by this layer are foreground/background RoIs that meet the overlap criteria and class specific regression coefficients for the RoIs.

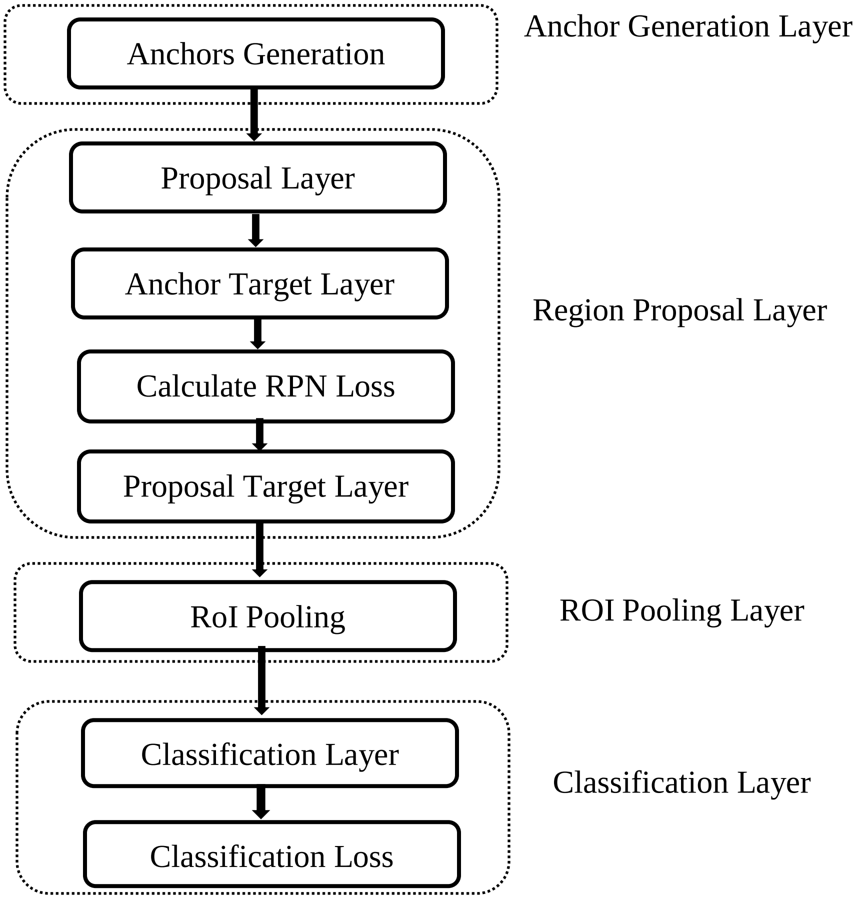


Figure 3.7: Different Layers in Region Proposal Network and Faster R-CNN

3.1.5 Region of Interest Pooling

The main purpose of RoI pooling is to speed up the training/testing time and to train the whole system from end-to-end.

The regions corresponding to the promising ROIs produced by proposal target layer are extracted from the convolutional feature map produced by the head network. The extracted feature maps are then run through the rest of the network to produce object class probability distribution and regression coefficients for each ROI.

RoI pooling layer takes two inputs:

1. A fixed-size feature map obtained from a last convolutional layer of VGG16 network i.e. head network.
2. An *N* x 5 matrix of representing a list of Regions of Interest, where N is a number of RoIs. The first column represents the image index and the remaining four are the coordinates of the top left and bottom right corners of the region.

For every Region of Interest from the input list, it takes a section of the input feature map that corresponds to it and scales it to some pre-defined size (i.e., 7×7). The scaling is done by:

1. Dividing the region proposal into equal-sized sections (the number of which is the same as the dimension of the output).
2. Finding the largest value in each section.
3. Copying these max values to the output buffer.

The result is that from a list of rectangles with different sizes, a list of corresponding feature maps with a fixed size are calculated. The dimension of the RoI pooling output doesn't actually depend on the size of the input feature map nor on the size of the region proposals.

3.1.6 Softmax Classification

The Softmax function gives the outputs of each unit to be between 0 and 1. It also divides each output such that the total sum of the outputs is equal to 1.

Mathematically, the Softmax function is shown below, where z is a vector of the inputs to the output layer (having 10 output units, then there are 10 elements in z). And again, j indexes the output units, so j = 1, 2, ..., K.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}} \tag{3.16}$$

The RoI pooling layer takes the ROI boxes output by the proposal target layer and the convolutional feature maps output by the "head" network and outputs square feature maps.

These features are now used for classification. The proposed method takes the feature map for each proposal, flattens it and uses two fully-connected layers of size 4096 with ReLU activation.

- A fully-connected layer with N+1 unit where N is the total number of classes and that extra one is for the background class.
- A fully-connected layer with 4N units to have a regression prediction; Δ*centerx*, Δ *centery*, Δ*width*, Δ*height* for each of the N possible classes are required.

The employee face is recognized with a corresponding bounding box.

3.1.7 Validation by Confusion Matrix

For the validation and performance evaluation of the model, Confusion Matrix is used. From the confusion matrix accuracy, precision and recall are calculated.

Confusion Matrix gives a matrix as output and describes the complete performance of the model.

Table 3.1: Confusion Matrix for binary classification

	Predicted True	Predicted False
Actual True	TP	FN
Actual False	FP	TN

- True Positives (TP): The cases in which predicted value is True and the actual output is also True.

- True Negatives (TN): The cases in which predicted value is True and the actual output is False.
- False Positives (FP): The cases in which predicted value is True and the actual output is False.
- False Negatives (FN): The cases in which predicted value is False and the actual output is True.

The Confusion Matrix Parameters for model evaluation are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.17}$$

$$Precision\ (Exactness) = \frac{TP}{TP+FP} \tag{3.18}$$

$$Recall\ (Completeness)\ = \frac{TP}{TP+FN)} \tag{3.19}$$

$$True\ Positive\ Rate\ (TPR) = Sensitivity = \frac{TP}{TP+FN} \tag{3.20}$$

$$False\ Positive\ Rate\ (FPR) = 1 - Specificity = \frac{FN}{TN+FN} \tag{3.21}$$

3.2 Tools Used

Python has been used for programming. The open source libraries like pytorch, matplotlib, numpy, opencv and keras are used.

For training in GPU environment, the open source Google Colab platform has been used with GPU support of Tesla K80 GPU having 13GB RAM and 360GB storage.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results

Employee face recognition system has been developed using the Region Proposal Network and Faster R-CNN. The system recognizes the employees face and outputs his/her name. If the test image doesn't belong to the employees' dataset, she/he is labelled as "Non-Employee". From the Chokepoint dataset [14], employees namely, "Ram", "Shyam", "Krishna" and "Hari" have been taken and the model is trained. To differentiate between employees and non-employees, a separate class called "Non-Employee" has been created to train the model.

Figure 4.1 shows the face detected output of employees. Using WIDER Face dataset [14] the system has been trained to detect the faces of employees along with their corresponding bounding box. The main purpose of using WIDER face dataset is to extract features for the face as employee dataset size is low and doesn't have ground truth annotations. The employee dataset is passed through the faster R-CNN to localize the faces from background and corresponding identification.

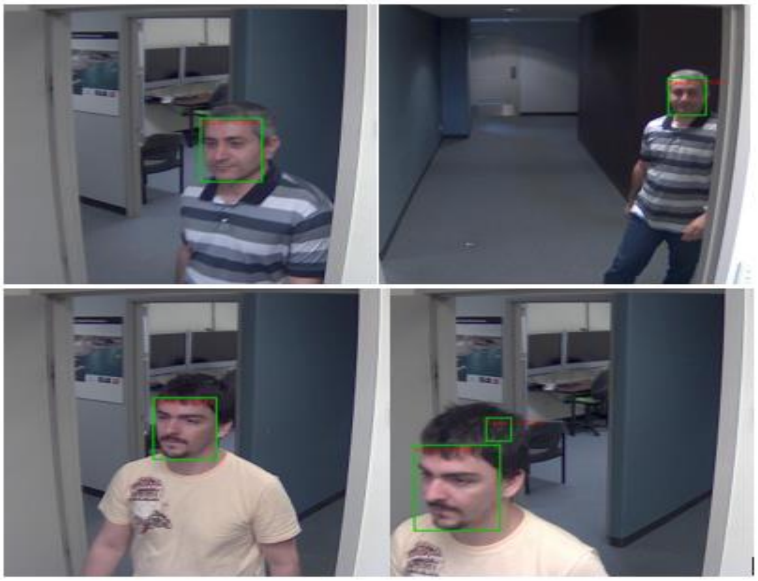


Figure 4.1:Employees' face detection with corresponding bounding box

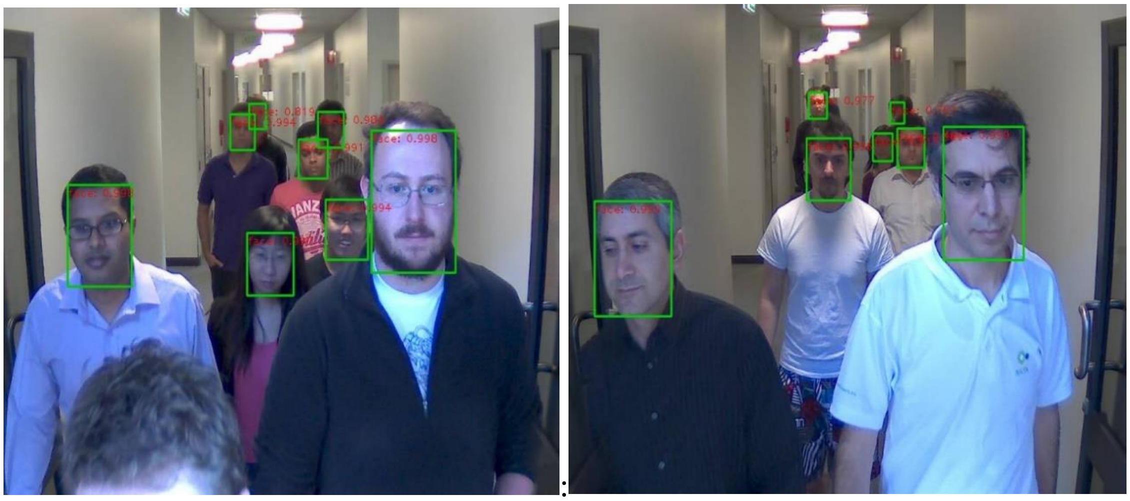
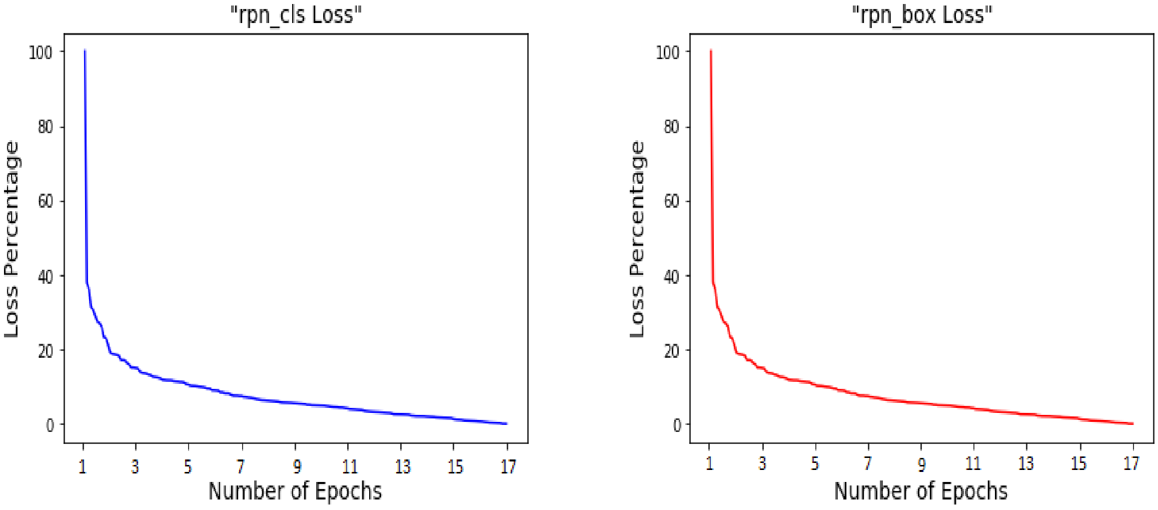


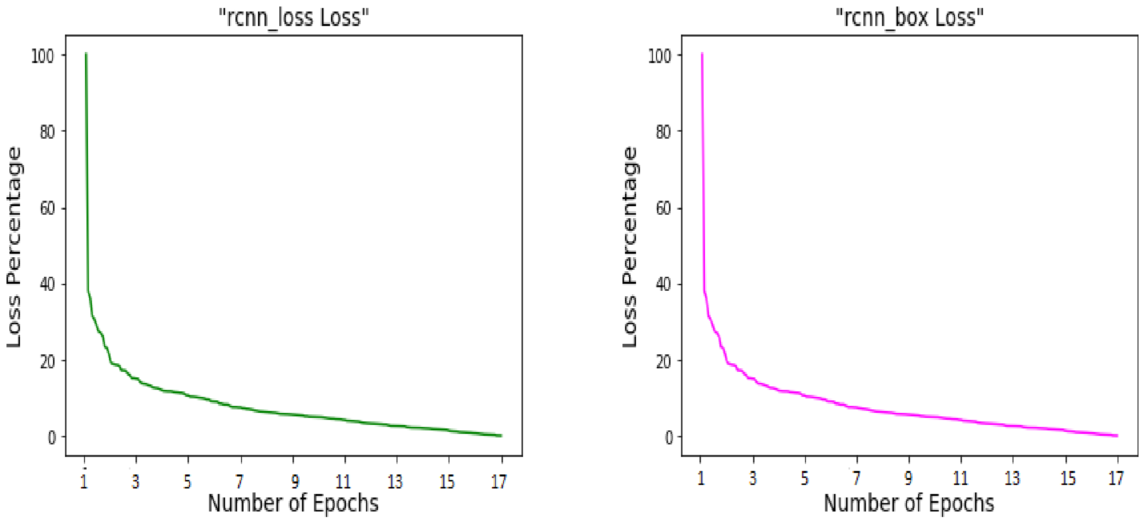
Figure 4.2: Face Detection results for multiple employees from Chokepoint dataset. The results are the output of the model trained on WIDER Face dataset.

The WIDER face dataset contains rich annotations, including occlusions, poses, event categories, and face bounding boxes. The proposed system uses this dataset just to train the RPN and faster R-CNN for feature learning. Once the face features are learned the model works well for employee dataset.



a)

b)



c) d)

Figure 4.3: a) Classification loss in training of WIDER Face dataset in RPN. b) Bounding box regression loss in RPN. c) Classification loss in faster R-CNN. d) Bounding box regression loss in faster R-CNN.

Figure 4.3 shows the loss values at different stages of the network. The *rpn_cls* loss and *rcnn_cls* loss is the cross-entropy loss. The *rpn_box* loss and *rcnn_box* loss is the regression loss calculated by smooth L1 function. These loss values are during the training of the network by WIDER Face dataset. This dataset is used for training purpose only to extract face features.

Figure 4.4 shows the test dataset for recognition of employees and non-employees. The employee images are from Chokepoint dataset and non-employee images are randomly collected from internet. For recognition of non-employee random images of male, female, children and adult from web are taken to train the model. Figure 4.5 shows the recognition results of test dataset. When the test image is from Chokepoint dataset the model outputs his/her identification with name. If the test image doesn't belong to employee dataset, the model outputs him/her as "non-employee".

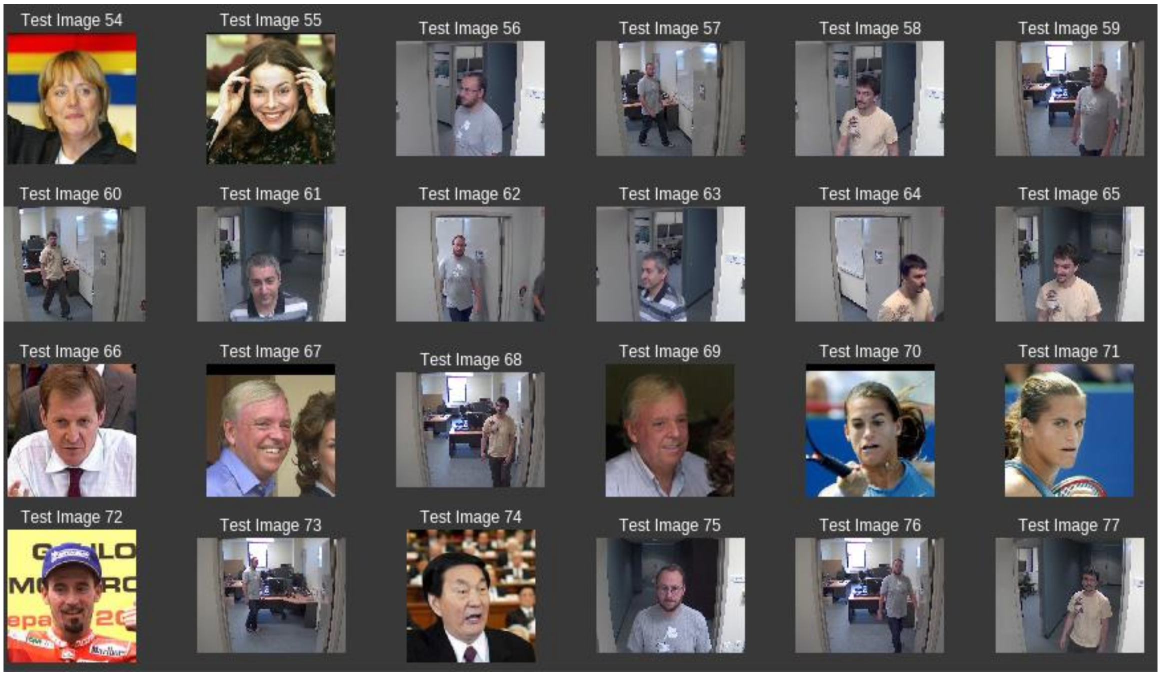


Figure 4.4: Test Dataset containing employees from Chokepoint dataset and non-employee images collected from internet

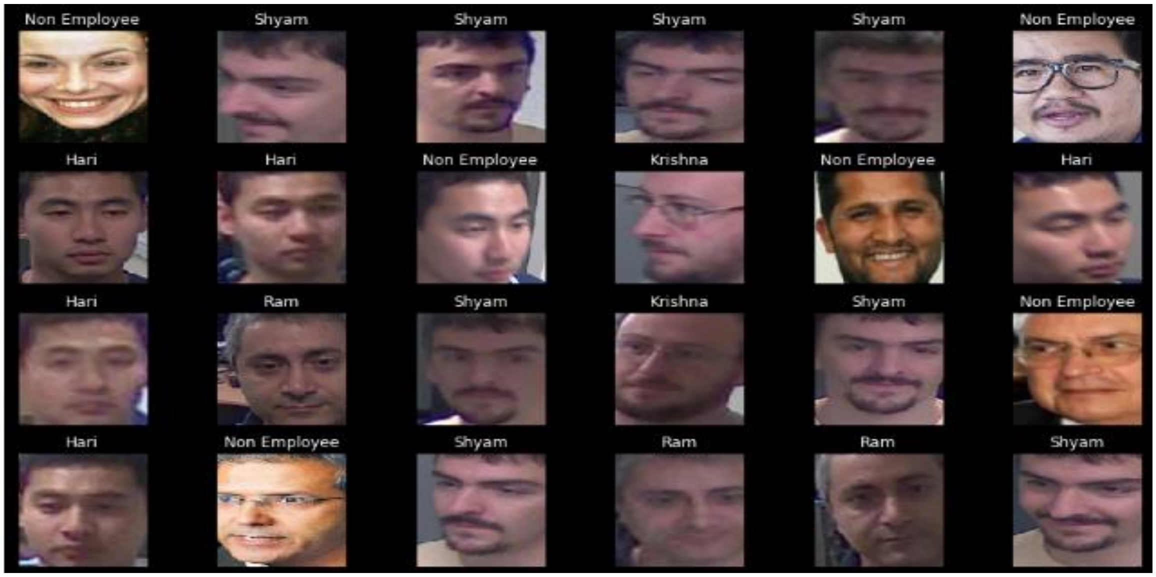


Figure 4.5: Recognition results with their identification. The unknown employees are classified as “Non-Employee”.

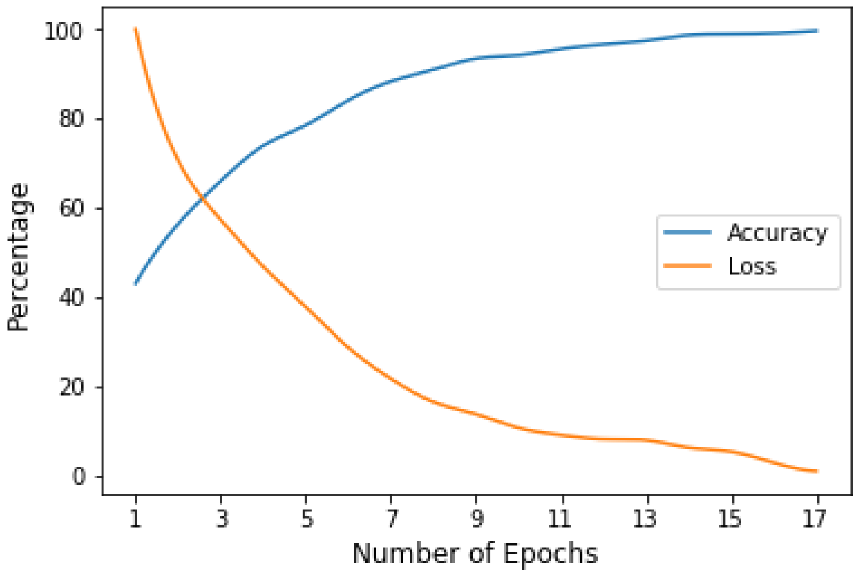


Figure 4.6: Loss/Accuracy vs Epochs in Recognition of Chokepoint Employee Dataset recognizing employees and non-employees

Figure 4.6 shows the loss percentage and accuracy percentage with respect to number of epochs. The loss function in this classification is the cross-entropy loss. The accuracy of the model in recognizing employees versus non-employees is 96.0% for the test dataset. The model recognizes people outside from Chokepoint employees' dataset as "Non-Employee".

To check the consistency of the model, a separate dataset is created at Nepal Telecom's office. Considering the employee dataset, the same model is trained. To recognize non-employees, a non-employee training dataset is used. An accuracy of 95.2% is achieved in differentiating employees and non-employees.

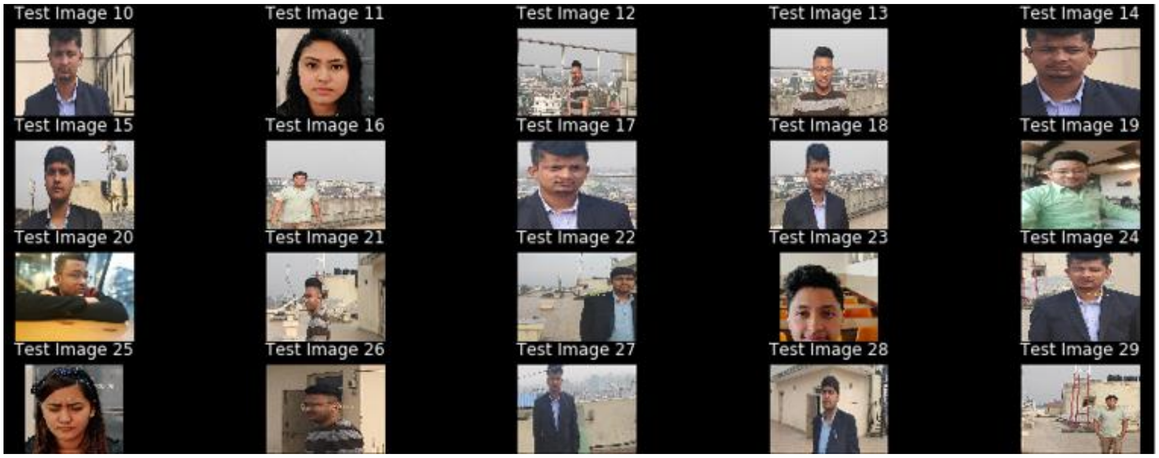


Figure 4.7: Nepal Telecom Employees’ test dataset samples

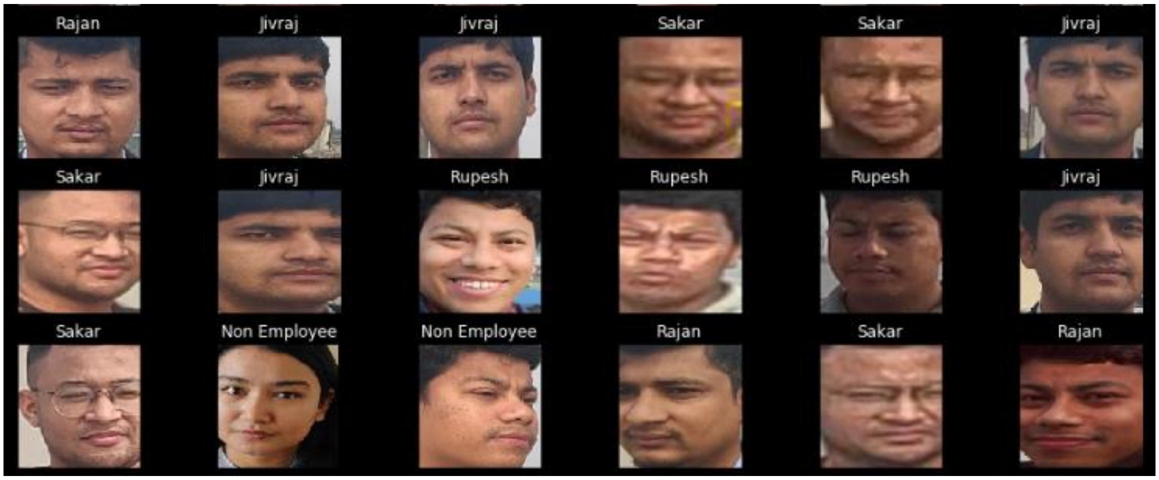


Figure 4.8: Nepal Telecom Employees' dataset recognition results

Figure 4.7 shows the Nepal telecom Employees’ test dataset samples. Employees namely “Jivraj”, “Rajan”, “Rupesh” and “Sakar” are used to test the model. Figure 4.8 shows the recognition results of these employees. The plot of loss and accuracy with respect to number of epochs is shown in Figure 4.9 The test accuracy has been slightly decreased due to the lesser number of training samples, low resolution imaging device.

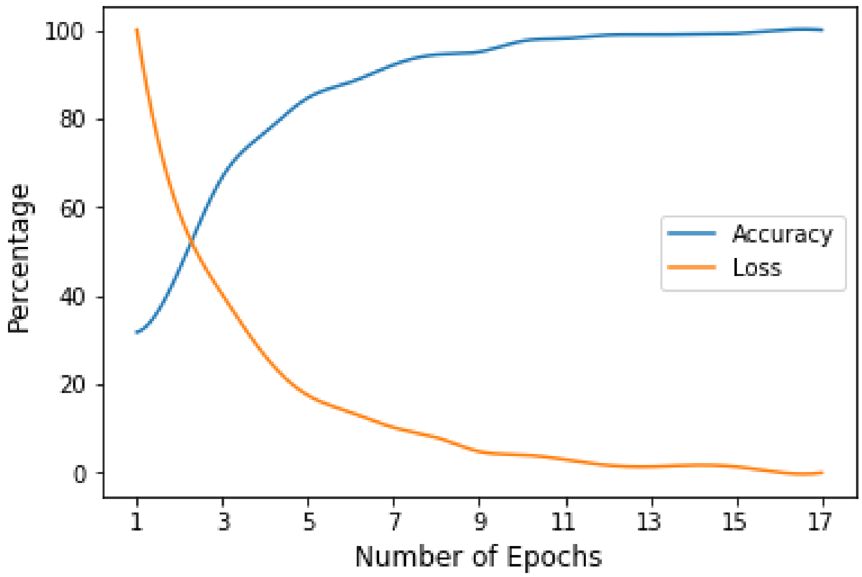


Figure 4.9: Loss/Accuracy vs Epochs in Recognition of office Employee Dataset recognizing employees and non-employees

Table 4.1 provides the test accuracy of Chokepoint dataset and Nepal Telecom employees' dataset. The number of test images used are 125 in each dataset.

Table 4.1: Comparison of employee dataset

Dataset	Number of Training Images	Accuracy (%)
Chokepoint Employees dataset with employees and non-employees.	2131	96.0
Nepal Telecom Employees dataset with employees and non-employees.	1682	95.2

Table 4.2 shows the comparison of different CNN based face recognition models in terms of accuracy. The accuracy of proposed method for face recognition has increased using region proposal network and Faster R-CNN.

Table 4.2: Comparison of different CNN based Face Recognition Models

CNN based Face Recognition Model	Accuracy
Face Recognition Based on Convolutional Neural Network on Georgia Tech Database [7]	94.8%
CNN Based Efficient Face Recognition Technique using Dlib on FRGC Dataset [8]	96.0%
Convolutional Neural Network for Face Recognition with Pose and Illumination Variation on FERET Dataset [18]	85.13%
Proposed Method on Chokepoint Dataset	96.0%

4.2 Evaluation Metrics

The evaluation of the proposed model is done by using confusion matrix. Table 4.3 shows the confusion matrix for employees from Chokepoint dataset and non-employees while Table 4.5 shows the precision and recall values for each class.

Table 4.3: Confusion Matrix for employees of Chokepoint dataset

		Predicted Class				
		Ram	Shyam	Krishna	Hari	Non-Employee
Actual Class	Ram	23	0	1	1	0
	Shyam	0	24	0	0	1
	Krishna	0	1	23	0	1
	Hari	0	3	0	21	1
	Non-Employee	1	1	0	0	23

Table 4.4: Confusion Matrix as a binary classification for employees from Chokepoint dataset and non-employees

		Predicted Class	
		Non-Employee	Employee
Actual Class	Non-Employee	23	3
	Employee	2	97

Table 4.5 shows employees “Ram” and “Krishna” has been classified with greater accuracy. The precision value of employee “Shyam” is less compared to others while recall value of employee “Hari” is less compared to others. Precision value shows the exactness of classifier and recall value shows the completeness of classifier.

Table 4.5: Precision and Recall values for employees from Chokepoint dataset and non-employees

Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score
Ram	0.976	0.920	0.0198	0.958	0.920	0.939
Shyam	0.952	0.960	0.0104	0.828	0.960	0.889
Krishna	0.976	0.920	0.0198	0.960	0.920	0.939
Hari	0.960	0.840	0.0390	0.955	0.840	0.894
Non-Employee	0.960	0.885	0.0300	0.920	0.885	0.902

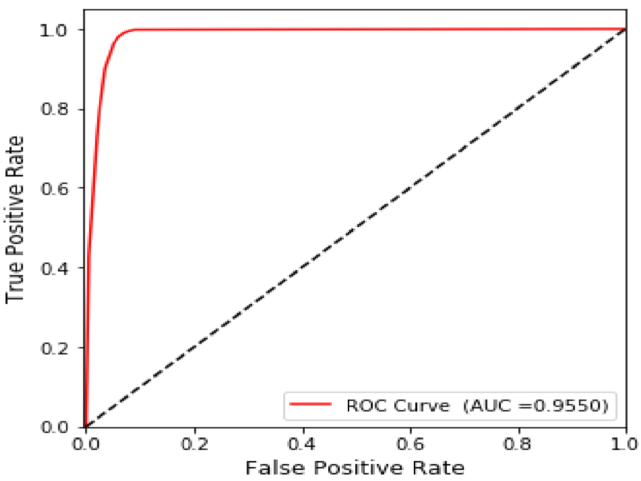


Figure 4.10: ROC curve differentiating employees/non-employees of Chokepoint dataset

Figure 4.10 shows the receiver operating characteristics of the classifier differentiating employees and non-employees. ROC curve the measure of separability between the classes. The higher the value of AUC, higher is the separability between classes. The ROC curve has guaranteed the validity of the classifier as the area under the curve for all classes is almost near to unity.

Table 4.6 shows the confusion matrix for employees of Nepal Telecom. Table 4.8 shows the precision and recall values for each class.

Table 4.6: Confusion Matrix for Nepal Telecom Employees' Dataset

		Predicted Class				
		Jivraj	Rajan	Rupesh	Sakar	Nonemployee
Actual Class	Jivraj	22	0	3	0	0
	Rajan	0	23	0	2	0
	Rupesh	1	1	20	0	3
	Sakar	0	0	0	23	2
	Nonemployee	1	0	0	0	24

Table 4.7: Confusion Matrix for employees and non-employees of Nepal Telecom Employees' dataset

		Predicted Class	
		Non-Employee	Employee
Actual Class	Non-Employee	24	1
	Employee	5	95

Table 4.8 shows the high precision values for employees "Jivraj", "Rajan" and "Sakar" compared to employee "Rupesh". This indicates these employees have been classified with greater accuracy.

Table 4.8: Precision and Recall for Nepal Telecom Employees' Dataset

Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score
Jivraj	0.960	0.88	0.0297	0.916	0.88	0.898
Rajan	0.976	0.92	0.0198	0.958	0.92	0.939
Rupesh	0.936	0.80	0.0290	0.869	0.80	0.833
Sakar	0.968	0.92	0.0200	0.920	0.92	0.920
Non-Employee	0.952	0.96	0.0104	0.830	0.96	0.890

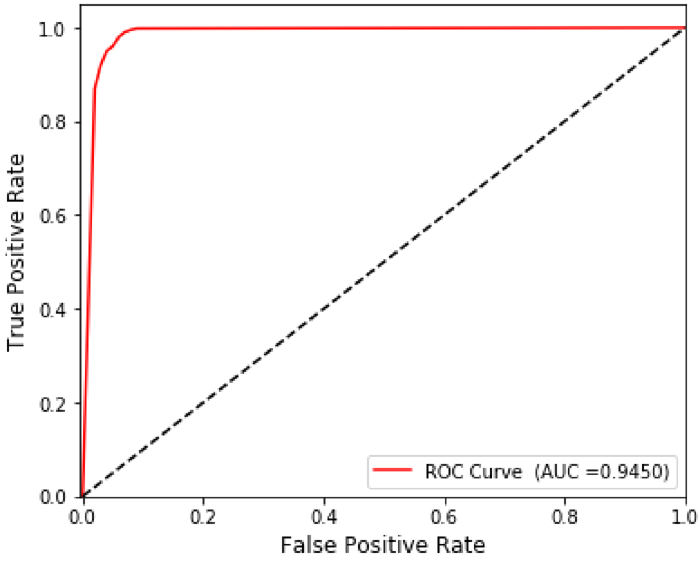


Figure 4.11: ROC curve differentiating Nepal Telecom Employees and Non-employees

Figure 4.11 shows the ROC plot between employees and non-employees of Nepal Telecom employees' dataset. The AUC value of 0.9450 indicates the classifier has a greater degree of separability between employees and non-employees of Nepal Telecom.

4.3 Discussion

The usage of employee images from standard datasets gives promising results in employee face recognition. This thesis work shows the improved performance in face recognition compared to other state-of-art technologies. The method uses region proposal network for regions extraction and faster R-CNN for classification of these region proposals. The region proposal network localizes the faces from the background with less proposals compared to traditional methods like selective search. The desired region proposals are pooled and fed to faster R-CNN for classification.

The performance of the proposed method is demonstrated by use of two different datasets. The first dataset is the Chokepoint dataset which is a standard dataset. Another employee dataset is recorded at Nepal Telecom to check the consistency of the model. The Nepal Telecom employee dataset is recorded with variations in pose and lighting conditions for a more diverse set of images representative of the employee's face. An employee face recognition model has been developed and the dataset is tested. The obtained results are validated from the evaluation metrices of confusion matrix. The accuracy of the Nepal Telecom employee dataset has slightly decreased because of lesser number of training images and low resolution of the imaging device. The comprehensive evaluation using receiver operating characteristics clearly shows the classifier is able to differentiate between employees and non-employees.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Employees' faces are recognized by region proposal network and faster R-CNN. The model is tested by the usage of two datasets. Chokepoint dataset is the standard employee dataset used for training and testing of the model. To check the consistency of the model, separate employee dataset is recorded at Nepal Telecom. The obtained results are validated from the evaluation metrices of confusion matrix. From the confusion matrix the accuracy of the classifier in differentiating employees and non-employees is 96.0% and 95.2% for Chokepoint dataset and Nepal Telecom employee dataset respectively. The precision and recall values for these datasets are 0.920, 0.885 and 0.830, 0.96 respectively. For the comprehensive analysis of the proposed method, the receiver operating characteristics (ROC) curve has been plotted. The area under the curve (AUC) value has been calculated. The AUC values are 0.9550 and 0.9450 respectively. The AUC values are closer to unity which shows the classifier is distinctly able to differentiate between employees and non-employees.

The proposed method shares the convolutional layers between the RPN and Faster R-CNN detector and is faster compared to other approaches as the number of test time for an image is 0.13 second approximately only.

5.2 Limitations

While the training of CNN models shows exceptional performance with large dataset but they are not suitable for learning from datasets with few samples. Similarly, the change in physical appearances like wearing of spectacle, long beard keeping, the model may not recognize the employee correctly.

5.3 Recommendation

To counter the problem of learning face representation from a smaller dataset, a new approach can be defined where the training dataset is augmented with synthetically generated samples by adding Gaussian or Poisson noise. This actually improves the generalization power of CNNs and adds robustness against overfitting to the model. To improve face recognition performance and keep pace with the natural changes of face and look, the system can store mathematical representation over time. Upon successful recognition, the system may use the newly calculated mathematical representation and add to the training dataset. If no new representation is found the data may be discarded. As CNN considers only the spatial features, an LSTM can be combined with CNN for consideration of temporal features.

REFERENCES

[1] Daniel Saez Trigueros, Li Meng, " Face Recognition: From Traditional to Deep Learning Methods", *School of Engineering and Technology, University of Hertfordshire Hatfield,* Oct. 2018

[2] [Online]. Available: https://www.vox.com/future-perfect/2019/4/27/18518598/ai-facial-recognition-ban-apple-amazon-microsoft

[3] M. Arsenovic, S. Sladojevic, A. Anderla, D. Stefanovic, "FaceTime - Deep Learning based Face Recognition Attendance System",*International Symposium on Intelligent Systems and Informatics*, Sep. 2017

[4] Y. LeCun, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Comput,* vol. 1, pp. 541-551, Dec. 1989.

[5] I. S. a. H. G. A. Krizhevsky, "ImageNet Classification with Deep Convolutional Networks," *Adv. Neural Inf. Process. Syst.,* pp. 1-9, 2012.

[6] A. Z. Karen Simonyan, "Very Deep Convolutional Networks for Large Scale Image Recognition," in *ICLR,* 2015.

[7] A. U. O. Y. Y. D. Musab Coskun, "Face Recognition Based on Convolutional Neural Network," *IEEE,* 2017.

[8] K. S. S. K. R. Sharma S, "FAREC - CNN Based Efficient Face Recognition Technique using Dlib," in *Advanced Communication Control and Computing Technologies (ICACCCT)* , 2016.

[9] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” *International Journal of Computer Vision (IJCV),* 2013.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in IEEE Conference on *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes (VOC) Challenge.”, *IJCV,* 2010.

[12] Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015.

[13] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] [Online]. Available: http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/

[15] [Online]. Available: http://arma.sourceforge.net/chokepoint/

[16] [Online]. Available: https://pillow.readthedocs.io/en/3.0.x/releasenotes/2.7.0.html

[17] [Online]. Available: http://www.robots.ox.ac.uk/~vgg/research/very_deep/

[18] AR Syafeeza, M Khalil-Hani, SS Liew, and R Bakhteri, “Convolutional neural network for face recognition with pose and illumination variation.,” *International Journal of Engineering & Technology (0975-4024)*, vol. 6, no. 1, 2014.

APPENDIX A

More Results on recognition of Chokepoint Employees’ Dataset

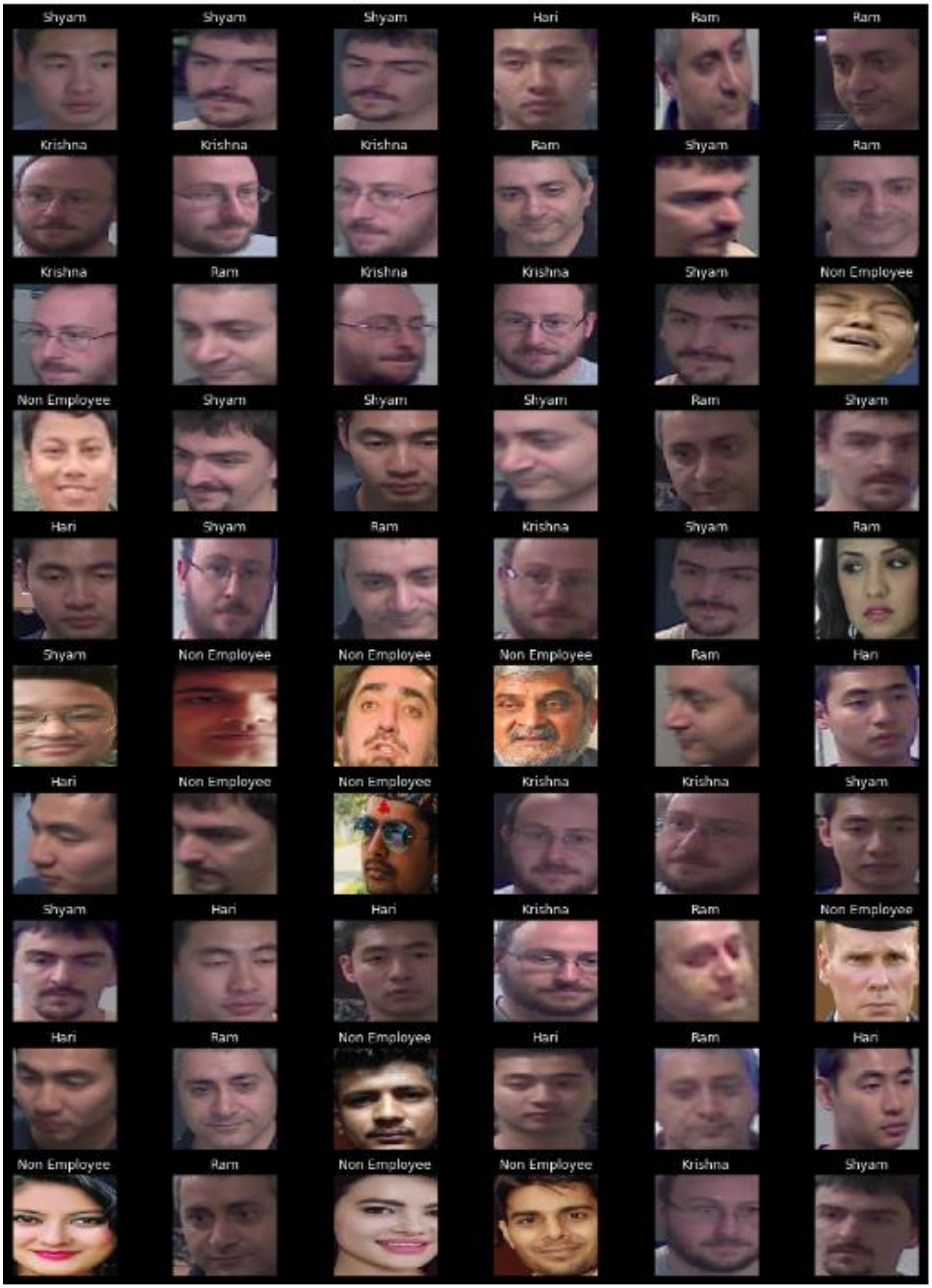


Figure A.1: More recognition results on Chokepoint Employees’ dataset

APPENDIX B

More Results on recognition of Nepal Telecom Employees’ Dataset

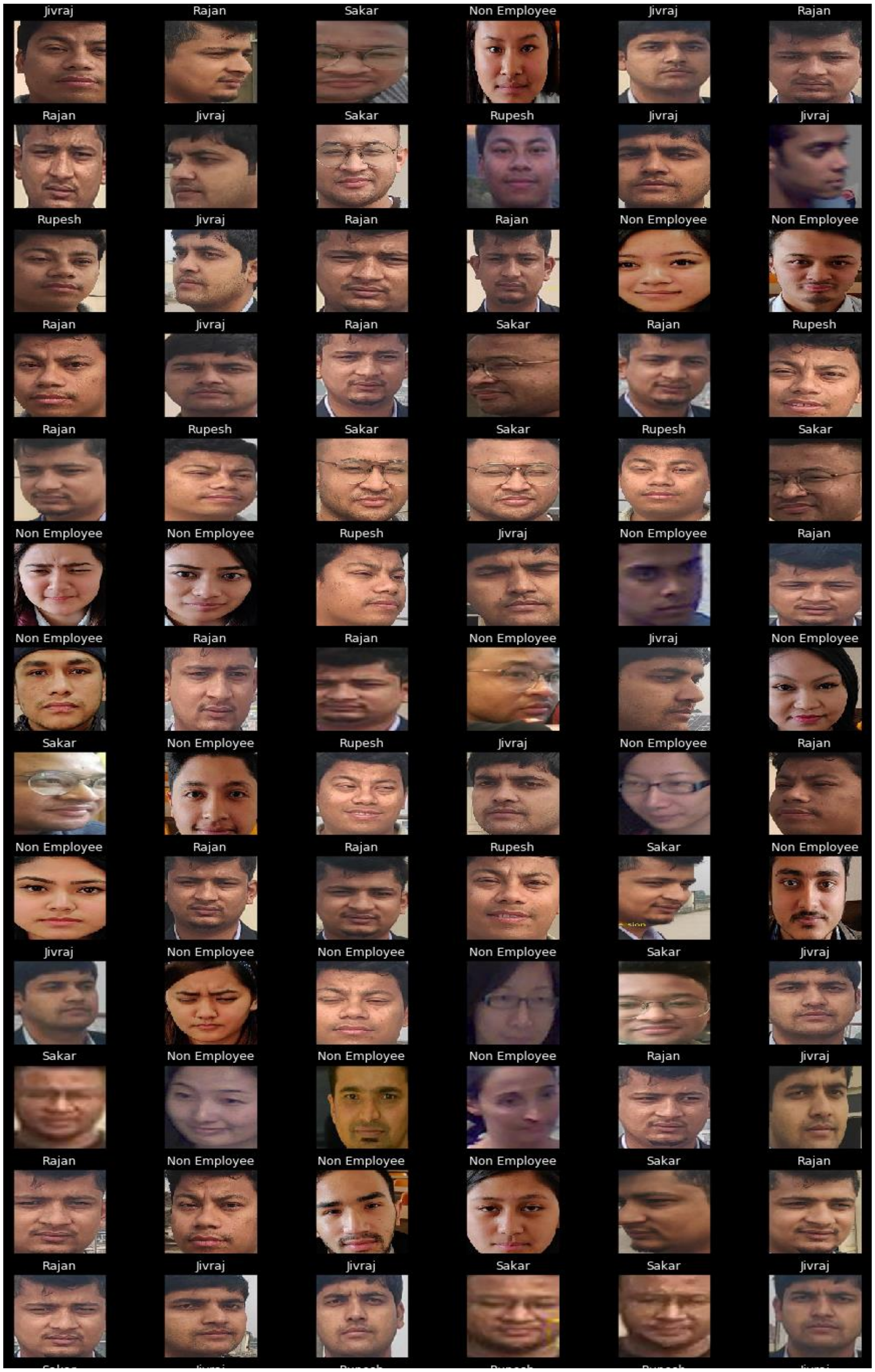


Figure B.1: More recognition results on Nepal Telecom Employees’ dataset

APPENDIX C

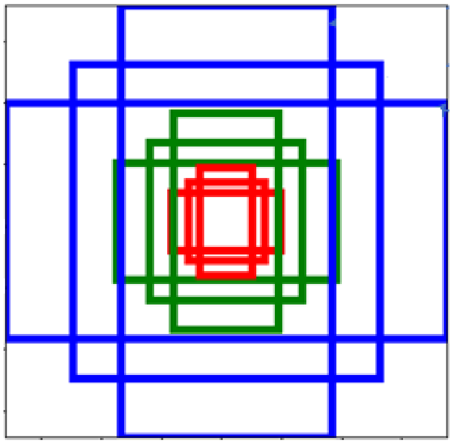
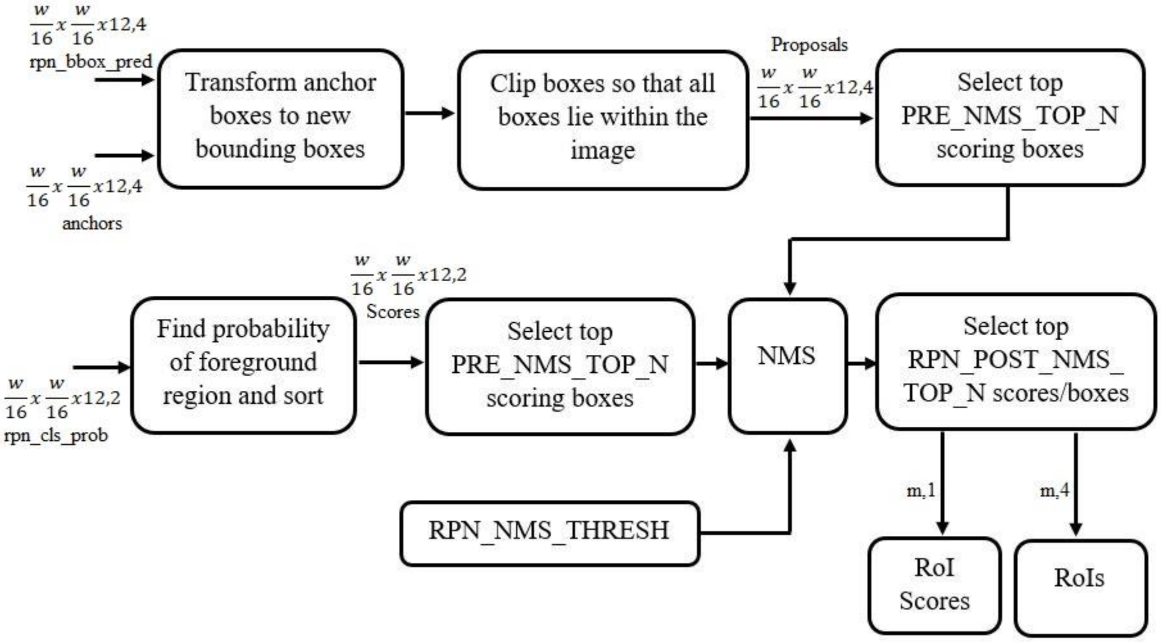


Figure C.1: Anchors of different sizes and scales



$\frac{w}{16}x\frac{w}{16}x12,4$
rpn_bbox_pred
$\frac{w}{16}x\frac{w}{16}x12,4$
anchors
Transform anchor boxes to new bounding boxes
Clip boxes so that all boxes lie within the image
Proposals
$\frac{w}{16}x\frac{w}{16}x12,4$
Select top PRE_NMS_TOP_N scoring boxes
$\frac{w}{16}x\frac{w}{16}x12,2$
Scores
Find probability of foreground region and sort
$\frac{w}{16}x\frac{w}{16}x12,2$
rpn_cls_prob
Select top PRE_NMS_TOP_N scoring boxes
NMS
Select top RPN_POST_NMS_ TOP_N scores/boxes
RPN_NMS_THRESH
m,1
m,4
RoI Scores
RoIs

Figure C.2: Working of Region Proposal Layer

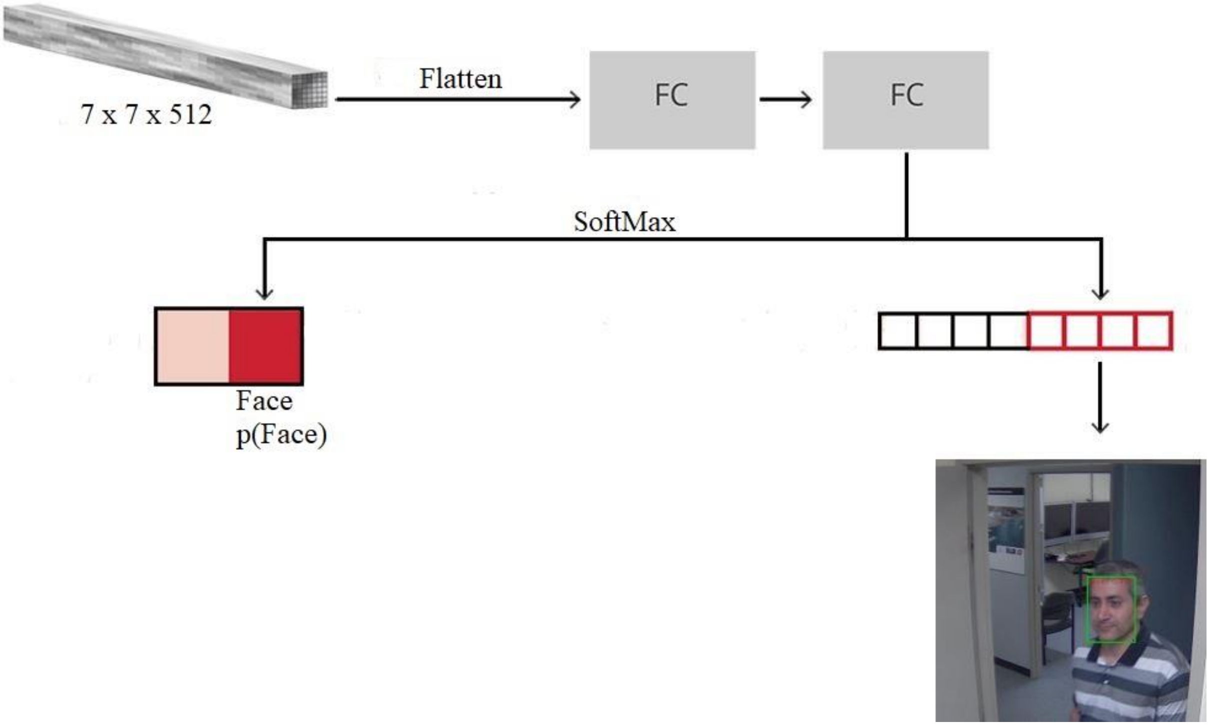


Figure C.3: Classification using Softmax Activation

APPENDIX D

Convolutional Layers in Region Proposal Network

RPN_Conv *= Conv2D (in_channels = 512, out_channels = 512, kernel_size = 3,*

stride = 1, padding = 1, activation = relu, bias = True) (D.1)

nc_score_out *= len(anchor_scales) * len(anchor_ratios) * 2* (D.2)

RPN_cls_score *= Conv2D (512, nc_score_out, kernel_size = 1, stride= 1, padding = 0)*

(D.3)

nc_bbox_out *= len(anchor_scales) * len(anchor_ratios) * 4* (D.4)

RPN_bbox_pred *= Conv2D (512, nc_bbox_out, kernel_size= 1, stride= 1, padding= 0)*

(D.5)

The proposal layer structure is as below:

RPN_proposal *= _ProposalLayer (stride = 16, anchor_scales, anchor_ratios)* (D.6)