Proceedings of IOE Graduate Conference, 2019-Summer
Peer Reviewed
Year: 2019 Month: May Volume: 6
ISSN: 2350-8914 (Online), 2350-8906 (Print)

An Approach for the Employee Face Recognition by RPN and Faster R-CNN Techniques

Rajan Gyawali [a], Dibakar Raj Pant [b]

[a, b] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] rajangyawali.np@gmail.com, [b] pdibakar@gmail.com

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 (CNN) and deep learning algorithms have shown greater performance in face recognition. In this article, region proposal network (RPN) is used to localize region of interests (faces) from the image and Faster R-CNN to output the region proposals' labels along with their associated bounding box. 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 accuracy of the model is 96.0% in recognition of Chokepoint employee dataset. The model is further tested with Nepal Telecom employee dataset 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
Face Recognition, Region Proposal Networks, Convolutional Neural Network, Faster R-CNN

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.

For increasing security concerns in big companies, the usage of face recognition system is increasing. Traditional recognition systems include RFID cards and GPS devices [2]. 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.

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.

2. Related Work

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.

CNN was proposed firstly by LeCun [3] and applied it on handwriting recognition. From his contributions, many scientists got true inspiration to work in this field. Krizhevsky et.al. [4] 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. [5] which showed a top-5 error of 7%.

Musab Coskun et. al. [6] 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. [7] 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%. The system didn't address the problems of pose variation and intensity variation.

Uijlings J.R.R et.al. [8] 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. [9] 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 [8], 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 [10] .

Shaoqing Ren et. al. [11] has further improved the results of selective search for object detection. The number of region proposals has been reduced to 300 from 2000. The model was tested on PASCAL dataset and the results are obtained faster than in R-CNN.

3. Methodology

In this article, faster R-CNN is used for face recognition. Faster R-CNN makes region proposals by neural networks [12].

3.1 Methodology

The general structure of employee face recognition process using faster R-CNN is as follows:

1. Image Datasets
2. Image Size Reduction by Lanczos Filter
3. CNN Feature Map by Visual Geometry Group
4. Regions Extraction by Region Proposal Network
5. Region of Interest Pooling
6. Classification by SoftMax

3.1.1 Image Datasets

WIDER Face dataset [13] and Chokepoint dataset [14] have been used. WIDER Face dataset is used to train the VGG16 network for face features extraction. Chokepoint dataset is used as employees faces for face recognition. Further Nepal Telecom employee dataset is recorded and used to check the consistency of the model.

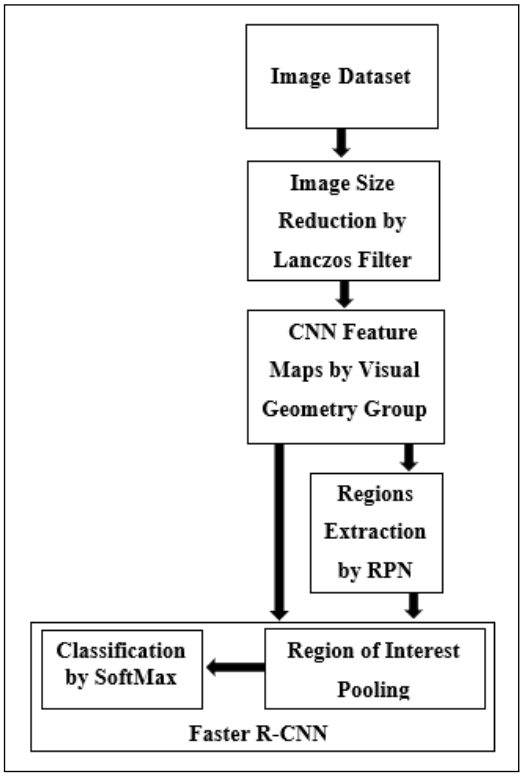

Figure 1: System Block Diagram

3.1.2 Image Size Reduction

CNN requires all of the input images to be of same size, thus image resizing is done. The image size of employees' dataset is 800x600 and that of recognized face is 128x128. For image resizing the Lanczos filter is used as it provides detail preservation and minimal generation of aliasing artifacts. The Lanczos filter is defined as:

$$L(x;n>0)=\begin{cases} \operatorname{sinc}(x)\cdot\operatorname{sinc}(\frac{x}{n}) & for \quad |x|\leq n \\ 0 & otherwise \end{cases} \tag{1}$$

3.1.3 Features Extraction by CNN

CNNs consist of filters or kernels or neurons that have learnable weights or parameters and biases. Each filter takes some inputs and does convolution. VGG16 architecture has been used for features extraction from images. A typical CNN consists of few layers of:

1. Convolutional Units
2. Rectified Linear Units (ReLU)
3. Pooling Units
4. Fully Connected Units

The primary purpose of Convolutional layer 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. A convolution operation is defined as:

$$y[m,n]=x[m,n]*h[m,n]=\sum_j\sum_i x[i,\ j]\,h[m-i,\ n-j] \tag{2}$$

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

ReLU 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 as:

$$f(x)=max(0,x) \tag{3}$$

where, x is the value in feature map.

Pooling layer reduces the dimensionality of each activation map and continues to have the most important information. This layer gains better generalization, faster convergence, robust to translation and distortion and usually placed between convolutional layers.

Fully Connected Layer 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 architecture is used. The convolutional feature maps produced by this network are passed as input to Region Proposal Network (RPN).

3.1.4 Region Proposals 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 [12] 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. 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.

The region proposal network consists of Anchor Generation Layer and Region Proposal Layer.

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 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. A total of W x H x 12 anchors are generated where W = w/16, H = h/16 and 16 is the sub sampling ratio. The anchors that lie outside of the image boundary have been excluded.

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 Proposal Layer, Anchor Target Layer and Proposal Target 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 and outputs the transformed bounding boxes by applying the regression coefficients.

Anchor target layer selects 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.

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 = Classification Loss + Bounding Box Regression Loss

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.

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.

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 deep convolutional network with several convolutions and max pooling layers.
2. An N x 5 matrix of representing a list of regions, 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 the fixed size 7x7. 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

This layer is the final layer of the network which classifies the employee face. 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 Softmax function squashes 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.

$$\sigma(z)_j = \frac{e^{z_j}}{\int_{k=1}^{N} e^{z_k}} \tag{4}$$

3.2 Evaluation Metrics

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

Table 1: Confusion Matrix for Model Evaluation

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision(Exactness) = \frac{TP}{TP + FP} \tag{6}$$

$$Recall(Completeness) = \frac{TP}{TP + FN} \tag{7}$$

4. Experiments and Discussion

The results of employee face recognition are shown. WIDER Face dataset is used for the purpose of face features extraction. Chokepoint employee dataset and Nepal Telecom employee dataset are used for employee face recognition.

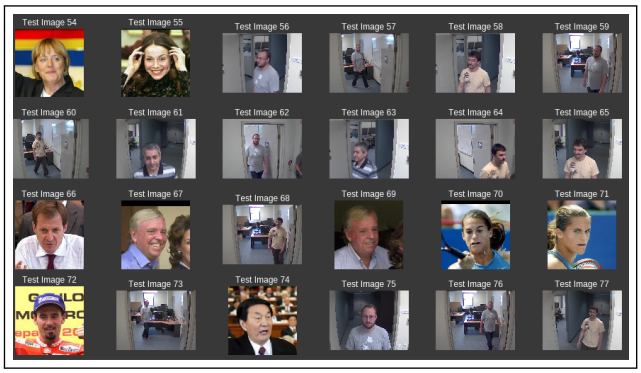

Figure 2: Test Samples of Chokepoint employee dataset

Figure 2 shows the test samples of employees from Chokepoint dataset and non-employees samples collected from web. Figure 3 shows the recognition results for these test samples. If the employee doesnot belong to Chokepoint employee dataset s/he is recognized as "Non-Employee".

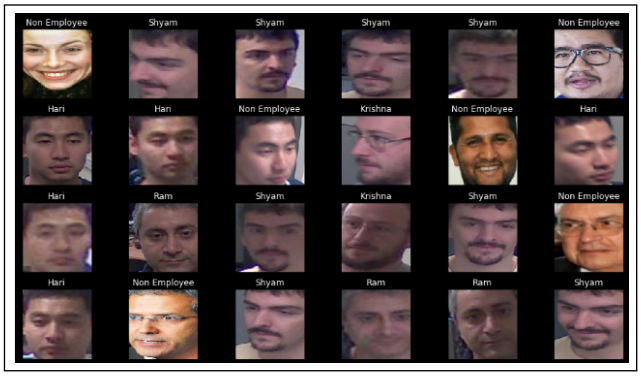

Figure 3: Recognition results of Chokepoint employee dataset. The unknown employees classified as "Non-Employee"

Figure 4: Test Samples of Nepal Telecom employee dataset

Figure 4 shows the test samples of employees from Nepal Telecom dataset and non-employees samples collected from web. Figure 5 shows the recognition results for these test samples. If the employee doesnot belong to Nepal Telecom employee dataset s/he is recognized as "Non-Employee".

Figure 5: Recognition results of Nepal Telecom employee dataset. The unknown employees classified as "Non-Employee"

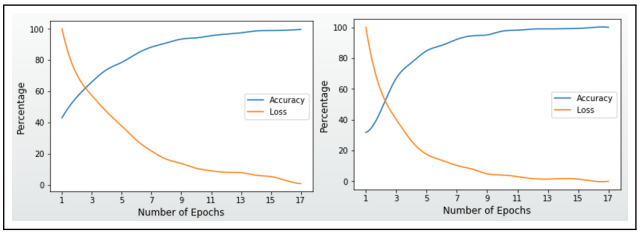

Figure 6: Loss/Accuracy percentage versus Number of Epochs in recognition of Chokepoint employee dataset (left) and Nepal Telecom employee dataset (right)

Figure 6 shows the accuracy/loss versus number of epochs plot in the recognition of employees from Chokepoint employee dataset and Nepal Telecom employee dataset.

5. Evaluation

The evaluation of the model is done by using Confusion matrix. The confusion metrics parameters are shown:

Table 2: Confusion Matrix for employees (E) and non-employees (NE) of Chokepoint dataset

Total Number	Predicted NE	Predicted E
Actual NE	23	3
Actual E	2	97

Table 3: Confusion Matrix for employees (E) and non-employees (NE) of Nepal Telecom dataset

Total Number	Predicted NE	Predicted E
Actual NE	24	1
Actual E	5	95

Table 2 and Table 3 shows the confusion matrix for employees and non-employees of Chokepoint dataset and Nepal Telecom dataset respectively. From Table 2, the accuracy, precision and recall values are 0.06, 0.920 and 0.885 respectively. From Table 3, the accuracy, precision and recall values are 0.952, 0.830 and 0.96 respectively for Nepal Telecom dataset.

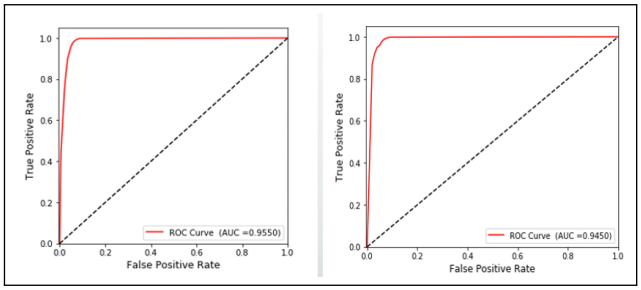

Figure 7: ROC plot between non-employees and employees of Chokepoint employee dataset (left) and Nepal Telecom employee dataset (right)

Figure 7 shows the ROC curve for two datasets and the curve gives a clear differentiation between non-employees and employees in both the dataset.

Table 4: Performance Comparison of Datasets

Dataset	Training Images	Accuracy (%)
Chokepoint	2131	96.0
Nepal Telecom	1682	95.2

Table 4 shows the performance comparison of two datasets in terms of accuracy.

6. Conclusion

In this article, an employee face recognition system using RPN and Faster R-CNN is applied on two different employee datasets. The accuracy of the model in classifying the employees from chokepoint dataset is 96.0%. The Nepal Telecom employees' dataset has been classified with 95.2% accuracy. The accuracy has decreased for this dataset because of the lesser number of training samples compared to that of Chokepoint dataset and the lower resolution of imaging device. However, the system shares the convolutional layers between the RPN and Faster R-CNN and is faster compared to other approaches as the number of test time for an image is 0.13 second approximately. Further, the comprehensive evaluation using ROC plot shows the classifier is able to differentiate between employees and non-employees.

References

[1] Li Meng Daniel Saez Trigueros. Face recognition: From traditional to deep learning methods. Oct. 2018.

[2] A. Anderla M. Arsenovic, S. Sladojevic and D. Stefanovic. Facetime - deep learning based face recognition attendance system. Sept. 2017.

[3] B. Boser Y. LeCun. Backpropagation applied to handwritten zip code recognition. Dec. 1989.

[4] Geoffrey E. Hinton Alex Krizhevsky and Ilya Sutskever. Imagenet classification with deep convolutional networks. 2012.

[5] Andrew Zisserman Karen Simonyan. Very deep convolutional networks for large scale image recognition. 2015.

[6] Ozal Yildirim Musab Coskun, Aysegul Ucar and Yakup Demir. Face recognition based on convolutional neural network. 2017.

[7] Karthikeyan Shanmugasundaram S Sharma, Santhees Kumar Ramasamy. FAREC - CNN based efficient face recognition technique using dlib. 2016.

[8] T. Gevers J. R. Uijlings, K. E. van de Sande and A. W. Smeulders. Selective search for object recognition. 2014.

[9] T. Darrell R. Girshick, J. Donahue and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014.

[10] C. K. I. Williams M. Everingham, L. Van Gool and A. Zisserman. The PASCAL visual object classes VOC challenge. 2010.

[11] R. Girshick Ren, K. He and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. 2015.

[12] E. Shelhamer J. Long and T. Darrell. Fully convolutional networks for semantic segmentation. 2015.

[13] Department of Information Engineering Multimedia Laboratory, The Chinese University of Hong Kong. WIDER FACE: A Face Detection Benchmark.

[14] Australia NICTA, Department of Broadband Communications. ChokePoint Dataset.