Computer Vision for Estimation of Age and Prediction of Gender

Colin Manuel Fernandes
Project Supervisor : Prof Stefano Ferrari

Course Assesment Project for
Intelligent Systems - Prof. Vincenzo Piuri
Informatics for industrial applications and robotics - Prof. Fabio Scotti

914777
Universita Degli Studi Di Milano

May 2018

Abstract

With the introduction of Convolutional Neural Networks in computer vision, Image analysis has found deeper dimensions[1]. CNN combined with Image processing can achive prediction of Age and Gender from image of a subject [2]. In the following project I provide comparable results using such a process. Facial Features are extracted from a labeled image dataset using ImageNet networks like AlexNet, Vgg, and GoogLeNet, later these features called feature map are used as an input to a simple neural network to classify these images on a scale of 0 to 120 in case of age or binary in case of Gender. Finally the network object is saved and used in an application that predicts Age and Gender usings snapshots from the Webcam.

Contents

1 Introduction

Age and Gender are the primary parameters in identification of a user. Its uses are important in field of Security, Social Networking etc. In the recent times with increasing demand for managing security aspects, identifying of users on physical parameters like age, gender, emotion with minimal information is challenging. Social Networking and e-commerce companies invest huge amount of time and resources to estimate Age and Gender to provide appropriate suggestions of products and services to their users. The progress in the field of computer vision has been amplified with introduction of neural networks. Convolutional Networks have introduced a new paradigm in the field of image analysis. Profiling of users, using images or video footage's can be achieved with the help of image processing and neural networks.
In my current project classification of Age and Gender is achieved by using the pre-trained convolutional Neural Networks trained which are built for object detection for the ImageNet challenge. Facial features are found to be ideal inputs as facial skin is exposed in general cases which can produce a clear distinction to map a function for the requirement. This function is approximated using a Convolutional operation at various layers in a neural network. Each layer has specific weights trained to identify certain features in an image. The weights in ImageNet CNN's used in the following project are trained on data-sets having millions of images to identify 1000 different objects. The different layers are trained to identify simple to complex features in an image. The features range from simple like edges and corners in an image to complex objects in final layers like limbs of different dog species. Since the initial layers identify basic features we exploit this property to extract features of a face image from the pretrained weights of the network. Such collection of features is called the feature map. Post extraction of feature maps a Feed Forward Neural Network (FFNN) is trained to classify age and gender based on a labeled dataset. To have a sufficient training data for FFNN I use feature maps extracted from a publicly available face dataset. I have used WIKI dataset which has age and gender labels. These images are cropped to contain only faces in them. Images which have incorrect labels have been discarded to improve the accuracy of the network. Later these are fed to a CNN to generate features. These features are used to train and test the accuracy FFNN with a validation procedure.The Trained FFNN is saved and used to classify the images for age and gender in an application using the snapshot of webcam as input.

2 Experimentation

2.1 Programing Environment and Programing Language

2.1.1 MATLAB 2017b

Toolbox/Libraries

- Computer Vision System Toolbox
- Neural Network Toolbox
- MatConvNet Toolbox (External)

MatConvNet v25 The toolbox is provided by vlfeat.org[3]. Matconvnet uses C++ compiler so Visual studio or any other C++compiler needs to be installed with the toolbox. Please note depending on the versions of C++ and matlab changes need to be done on the vl_compilenn function of matconvnet toolbox. (Fun Fact: Setting up matconvnet functions takes longer time than training neural network). Matlab has inbuilt Neural Network Toolbox for deep learning models from which GoogLeNet and AlexNet have been used. Vgg-Face trained exclusively on face images is not available with the matlab inbuilt toolbox.

2.2 Image Dataset

Publicly available face image datasets have been selected. The labels for age and gender can be found using the details of the images.

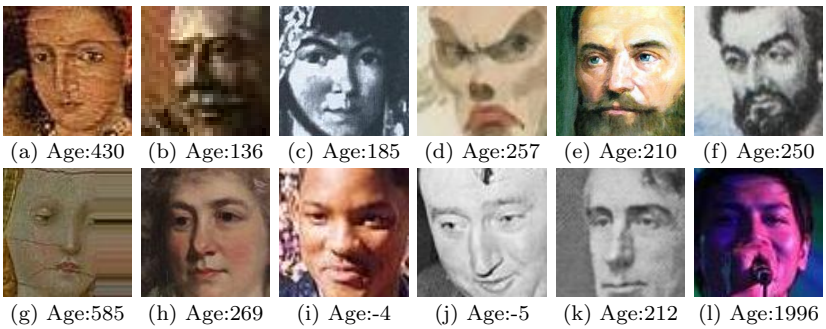
2.2.1 WIKI

WIKI dataset[4] 52K+ images in total. All images have been crawled from the profile images from the pages of people from Wikipedia with the attribute information given below.

- dob: date of birth (Matlab serial date number)
- photo_taken: year when the photo was taken
- full_path: path to file
- gender: 0 for female and 1 for male, NaN if unknown
- name: name of the celebrity
- face_location: location of the face. To crop the face in Matlab run
  $img(face\_location(2) : face\_location(4), face\_location(1) : face\_location(3), :))$
- face_score: detector score (the higher the better). Inf implies that no face was found in the image and the face_location then just returns the entire image
- second_face_score: detector score of the face with the second highest score. This is useful to ignore images with more than one face. second_face_score is NaN if no second face was detected
  The age of a person can be calculated based on the date of birth and the time when the photo was taken (note that we assume that the photo was taken in the middle of the year):
  $[age,\ ] = datevec(datenum(wiki.photo\_taken, 7, 1) - wiki.dob);$

2.3 Preprocessing and Data Cleansing

After the first complete execution for classification on the images for age and gender, it was discovered that there was a huge difference with prediction and actual. Post investigation I could locate the wrongly predicted images as below. Main issues incorrectly labeled data and incidentally these images were not found to be captured images of identities rather paintings or animations. These issues were handled by excluding the images using the below mentioned techniques.



(a) Age:430 (b) Age:136 (c) Age:185 (d) Age:257 (e) Age:210 (f) Age:250

(g) Age:585 (h) Age:269 (i) Age:-4 (j) Age:-5 (k) Age:212 (l) Age:1996

Figure 1: Images with incorrect labels

2.3.1 Exclusion of incorrectly labeled data

Age in WIKI dataset was calculated by subtracting date of picture taken with date of birth of the identities. Few calculated ages were in the negative range, and other identities were extremely high like over 120- 1900 (almost immortal). By performing a primary investigation online longest-lived person on was 122 years. Hence in order avoid misclassification. Data for Age above 117 have been

removed. Adding to that images labels which are not available have been removed. The accuracy of the network can improve by simply not training the weights with incorrect data.

2.3.2 Face Detection

The images contained regions other than the faces or even none in some cases. Face recognition technique has been employed to detect the faces in the pictures and crop the bounding boxed region as the processed image. This is necessary for the network to learn features only from the skin of the face and not bias over the rest of the region. This is accomplished using matlab's inbuilt computer vision system toolbox. The toolbox has function for Viola -Jones Algorithm[5] to segment out only the face from the picture.

(a) Lateral (b) Lateral (c) Multiple (d) cropped (e) None (f) Irrelevant

Figure 2: Images with Non frontal Faces

Viola -Jones Algorithm The algorithm detects the face using features extracted using Harr filters. To compute these features rapidly at many scales the integral representation of images is used. Integral representation is also a novel contribution from the developers of the algorithm[5]. Only important features are chosen using the AdaBoost algorithm. Further combining successively more complex classifiers in a cascade structure which dramatically increases the speed of the detector by focusing on the promising regions of the image.

2.3.3 Input Resizing

All images have been resized to match input format of the Neural Networks. This is done with each file before the features are extracted

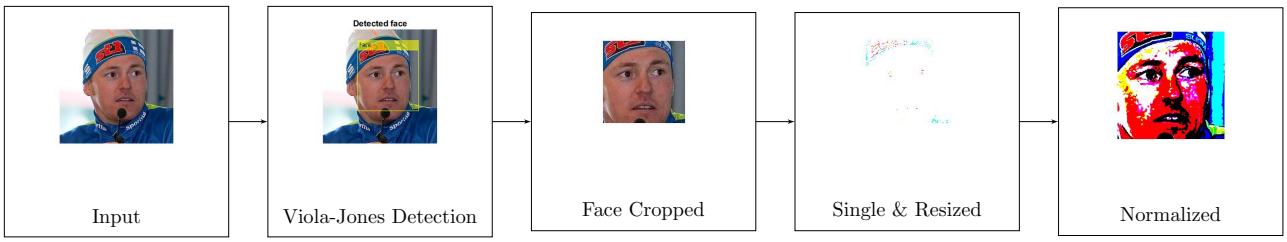


Figure 8: Data Processing

2.4 Convolutional Neural Networks and Feature map Extraction

CNN use convolution operations[1] using a filters of different sizes over the image in each of the layers of the network. This filter is iterated over the image to extract certain features spanning the filter area. The values in the filter are trained using back propagation algorithm and gradient decent for biases. Although the pre-trained networks have trained weights for specific classification, Age and gender classification can be achieved by using the training from the initial layers in the network. The weights in initial layers are trained to detect edges, shapes, color etc., Parts of objects are detected in the final layers of the network. This property is used to filter the image through existing weights and finally extract the map from the fully connected layers. The resulting vectors becomes the input attributes to classify using simple neural network.

2.4.1 Alexnet

Alexnet famously won the ImageNet challenge in 2012 with a huge margin. Alexnet was developed and trained by Alex Krizhevsky et al.[6], contains 8 trained layers out of which 5 are convolutional and three fully connected. The main highlights for the network is using Relu instead of Tanh to add non-linearity, use dropout instead of regularization to deal with overfitting and overlap pooling to reduce the size of the network. The network accepts an input of $224 \times 224 \times 3$ input image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolutional layer used the output of the first convolutional layer and filters it with 256 kernels of size $5 \times 5 \times 48$. The third, fourth and the fifth convolutional layers are connected without any intervening pooling or normalization layers. The fourth convolutional layer has 348 kernels of size $3 \times 3 \times 192$, and the fifth convolutional layer has 256 kernels of size $3 \times 3 \times 192$. The fully connected layers have 4096 neurons each. We use the first fully connected layer to extract features.

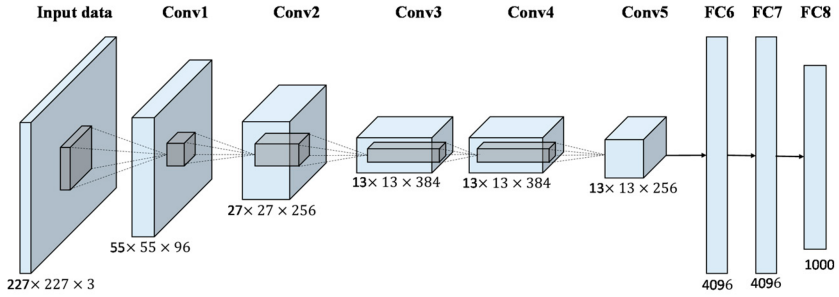


Figure 9: Alexnet Architecture

2.4.2 Vgg-Face

Vgg Face model was introduced by Omkar Parkhi et al.[7] which is a CNN model and has specifically trained weights for face recognition using 2.6 million face images and 2622 subjects. The developers extracted face images from the IMDB and Wikipedia websites. Hence a pretrained model can extracts features from the face invariably if provided with face images. The CNN accepts a face image of dimensions of $224 \times 224 pixels$ with the average face image subtracted. The architecture consists of 11 blocks, each block is consisting of convolutional layer (linear transformation), rectification layer (nonlinear operation) and pooling (max pooling). The last 3 layers are the fully connected layers which are same as the convolutional layers before it, but having the same size of output as input. All convolutional layers are followed by the ReLU Operation. The first 31 layers are freezing to obtain different levels of abstraction for an input image ranging from low level representation in the initial layers like corners and edges, to mid-level representations in intermediate layers and high-level representation in the last layers. The fully connected layers range from 32 to 36th layer. To use pretrained VGG-Face for feature extraction we extract 4K dimensional activation vector from one of these layers. In my approach the features extracted from the fully connected which is layer no 33- the 'ReLu' layer after the first fully connected layer.

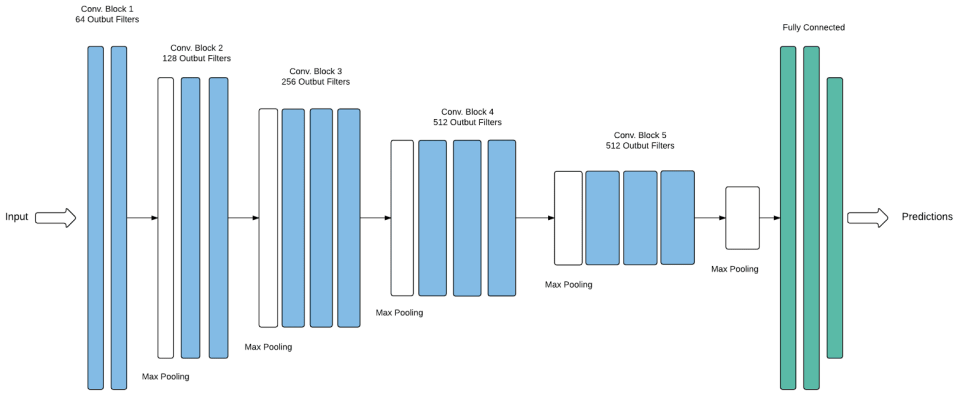


Figure 10: VGG Architecture

2.4.3 GoogLeNet

Unlike the previous networks with linear convolution layers, GoogLeNet[8] avoids choosing between the different size of convolutional layers, the network uses $3\times3, 5\times5, 1\times1$ convolution layers and pooling in a single layer. All these layers are concatenated and fed as input for the next layer. Such a layer called the Inception layer Inspired by the above meme form the movie Inception. the idea of movie is based on the dreaming within the dream, thus going into deeper levels of dreams. Similarly the network has many such Inception layers. The size of the receptive field of the network is $224 \times 224 \times 3$(RGB) with mean subtraction. The network is 22 level deep and has more than 150 layers. The ImageNet competition was won with an ensemble of 7 such networks. Unlike VGG and Alex there are 2 Fully connected layers can be found in the mid the network architecture. I have extracted the feature map from the layer before the SoftMax input.

(a) We Need to Go Deeper-Meme

Figure 11: Christopher Nolan's - Inception

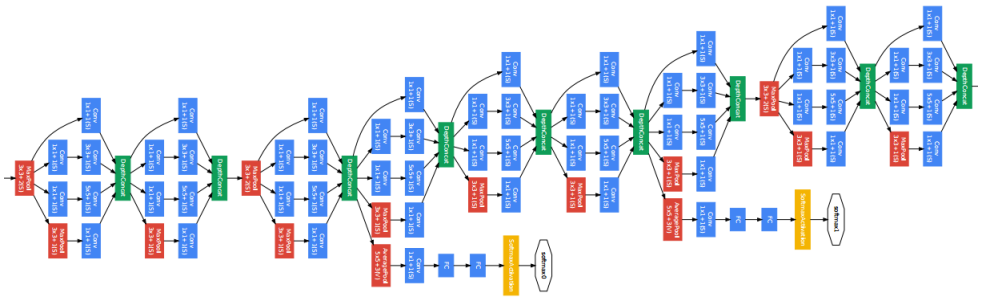


Figure 12: Google Architecture

2.5 Dimensionality Reduction

2.5.1 Principal Component Analysis

PCA is a mathematical way that transforms number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. Dimensionality reduction can be achieved with the reduction of variables. PCA is achieved by decomposing the eigen vectors and eigen values of the co-variance matrix of the dimensions. These dimensions are later arranged in the decreasing order of the eigen values which gives the importance of the dimensions. We build the training set file using these top 500 features out of 4096 .

2.6 Feed Forward Neural Network

The goal of the project is to Classify to age or gender on a feed forward network. The FFNN predicts the age or gender by classifying the values using the features extracted from above layers.

2.6.1 Validation

Cross validation was performed on the data-set by holding out 10% for testing. This configuration is repeated 10 iterations and results are averaged.

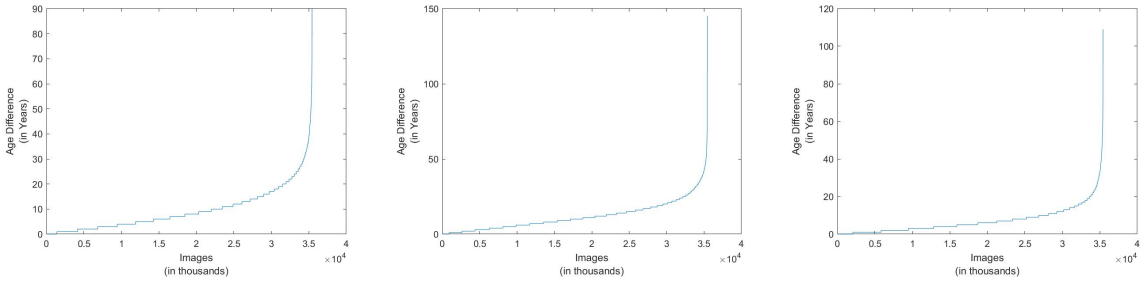
3 Observations

3.1 Age

3.1.1 Classification Results

FFNN Classification Result				
CNN Model	Mean Absolute Error $mean(Actual - Predictions)$	Classification Error Rate $\left(\frac{length(Actual!=Predictions)}{length(Actual)}\right)$ %	Sample Size #	Exact Predictions Average
AlexNet	9.4	96.0	3543	139.7
Vgg-Face	7.0	94.1	3543	200
GoogLeNet	11.81	97.1	3543	89.9

3.1.2 Plots



(a) AlexNet (b) GoogLeNet (c) Vgg-Face

Figure 13: Age Difference Classification

3.1.3 Error Distribution across Age For VGG-FACE Model

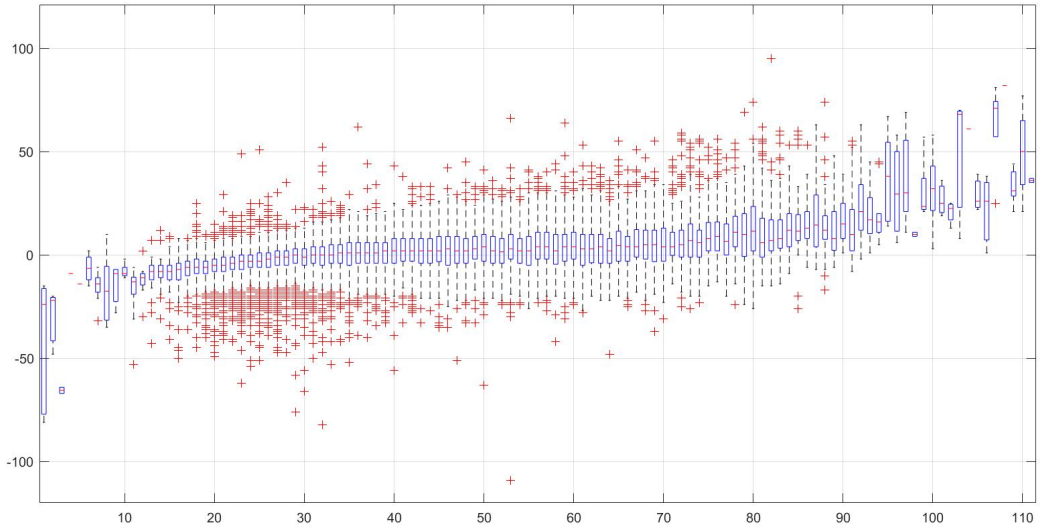


Figure 14: Distribution Error

Actual:82 Predicted:13 | Actual:1 Predicted:82 | Actual:112 Predicted:30 | Actual:32 Predicted:114 | Actual:61 Predicted:29 | Actual:77 Predicted:46 | Actual:86 Predicted:47

Figure 15: Incorrectly labeled Faces Images for Age difference greater than 80

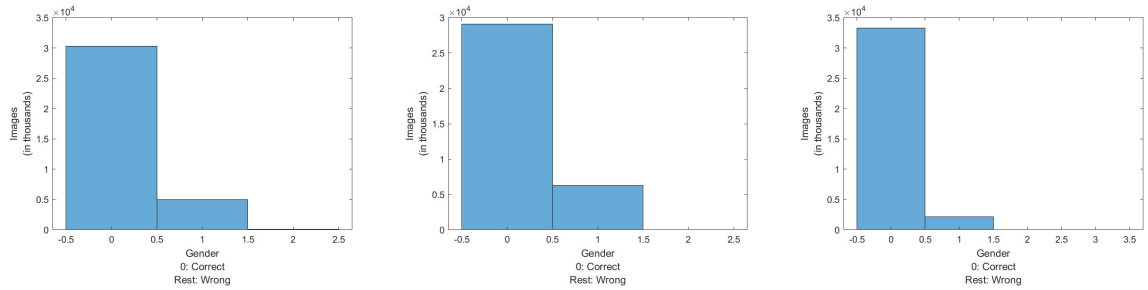
3.1.4 Misclassified Images

3.2 Gender

3.2.1 Classification Results

FFNN Classification Result				
CNN Model	Mean Absolute Error $mean(Actual - Predictions)$	Classification Error Rate $\left(\frac{length(Actual != Predictions)}{length(Actual)}\right)\%$	Sample Size #	Exact Predictions Average
AlexNet	0.14	14.4	3543	3032.2
Vgg-Face	0.06	6.11	3543	3326.4
GoogLeNet	0.17	17.8	3543	2912.1

3.2.2 Plots



(a) AlexNet (b) GoogLeNet (c) Vgg-Face

Figure 16: Gender Difference Classification

3.2.3 Confusion Matrix

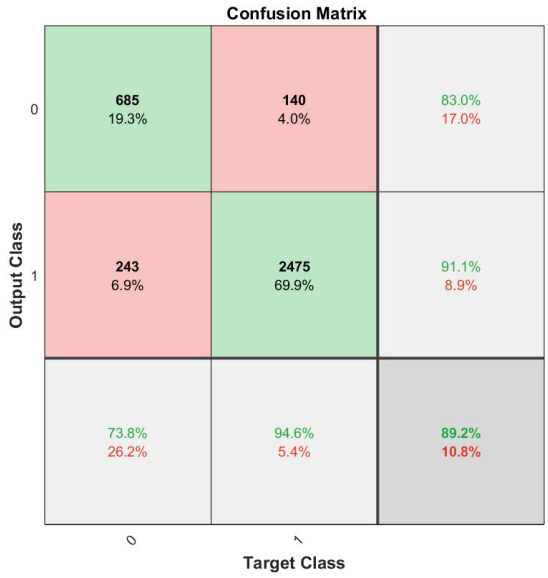


Figure 17: Confusion Plots

3.2.4 Misclassified Images

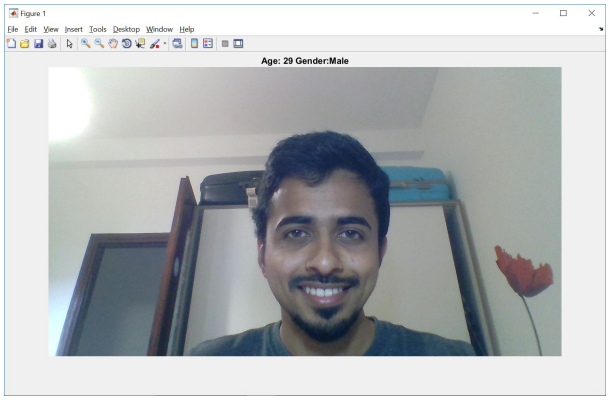
Most images are occluded, or the visible features are covered like the eyes in the 6th picture

Female Female Female Male Male Male Female

Figure 18: Incorrectly Classified Faces for Gender

4 Prototype of Working application on live Feed Captured Through Webcam

Using all the images from the data-set and all the 4K extracted features for each of those images, I have trained and saved two feed-forward network objects as 'AgeNet' and 'GenderNet'. Frames of live feed are captured using webcam. The Face region is detected on the frames using Viola-Jones detector and later cropped for the feature extraction. The features are extracted using the VGG-Face CNN. These features are used as input for AgeNet and GenderNet to classify for Age and Gender, which are displayed in the title of the image.



(a) Actual Age:30, Gender: Male

Figure 19: Application Snap Shot (Age and Gender in the Figure Title)

5 Conclusion

In the literature [2] the authors are able to achieve a best MAE of 3.30 for Age Estimation when tested on a different data-set. With a quick investigation we could see that most of the misclassified images were either paintings, incorrect labels or the person seemed younger than recorded. VGG-Face classifies the Age and Gender close to accuracy as the network is already trained on face images than the other networks. But the other networks are also comparable to the latter even though the networks are trained on object detection. AlexNet and GoogleNet is faster than the VGG in extracting features. GoogLeNet has a feature set of 1K dimensions, where as AlexNet and VGG has 4K dimensions

6 Future Work

6.1 Ideal Training Dataset With image processing and Semantic Segmentation

Even though we filtered all images using Viola-Jones algorithm, there were few images that even crossed though the filters like the paintings as seen in the misclassification sections above. These images affect the accuracy of the model. Such images need to be identified and excluded. This can be done by perfecting the face recognition algorithms only for images having faces excluding paintings later use semantic segmentation[9] to mask all regions other than face using image processing techniques in the image. The Segmented face area is provided as a training input to the neural networks. The semantic Segmentation can be achieved using trained neural networks for image segmentation like SegNet.

6.2 Training weights of deep learning network

Training a deep Neural Network weights using only face image dataset would result in model learning only the required features for face images.

6.3 Using Non Frontal images in the dataset

Not all environments are coorperative while collecting data. In such cases it would require the setup to estimate age and gender when the user is not completely facing the camera. Such a setup could be implemenented by using accurate datasets from non frontal faces.

6.4 Tracking the identified user

Futher use the optimised Neural Network to track user using Kanade-Lucas-Tomasi (KLT) algorithm or cam shift in a live video environment.

References

[1] A. NG, "Convolutional Neural Networks (Course 4 of the Deep Learning Specialization)," 2017. [YouTube Source].

[2] A. Anand, R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, and F. Scotti, "Age estimation based on face images and pre-trained convolutional neural networks," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7, Nov 2017.

[3] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," 2015.

[4] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision (IJCV)*, July 2016.

[5] Matlab, "vision.CascadeObjectDetector System object."

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] U. Güçlü, Y. Güçlütürk, M. Madadi, S. Escalera, X. Baró, J. Gonzàlez, R. van Lier, and M. A. J. van Gerven, "End-to-end semantic face segmentation with conditional random fields as convolutional, recurrent and adversarial networks," *CoRR*, vol. abs/1703.03305, 2017.