Detection of Gender, Age and Emotion of a Human Image using Facial Features

Sidharth Nair[1], Dipesh Nair[2], Gautam Nair[3], Anoop Pillai[4] and Prof Sujith Tilak[5]

[1-5]*Department of Computer Engineering, PCE, Navi Mumbai, India - 410206*

Abstract— *The main motive is to develop an automatic age and gender estimation method towards human faces which will continue to possess an important role in computer vision and pattern recognition. Apart from age estimation, facial emotion recognition also plays an important role in computer vision. Non-verbal communication methods such as facial expressions, eye movement and gestures are used in many applications of human computer interaction. In order to create computer modeling of humans age, gender and emotions a plenty of research has been accomplished. But it is still far behind the human vision system. In this project, we propose a Convolutional Neural Network (CNN) based architecture for age & gender classification. The architecture is trained to label the input images into 8 labels of age and 2 labels of gender. Our approach shows better accuracy in both age and gender classification compared to classifier-based methods. In order for computer modeling of human's emotions we are planning to predict human emotions using deep CNN and observe how emotional intensity changes on a face from low level to high level of emotion. By using the preprocessing algorithm Viola-Jones we extracted features of the image which are fed as an input to CNN. With a proper user interface, the result of the prediction is revealed.*

Keywords— Face Detection, Viola Jones, Face Recognition, Deep CNN

1. Introduction

Age estimation from face images plays an important role in human and computer vision which has many applications in for example, forensics or social media. It can determine the prediction of other biometrics of human and facial attributes tasks such as gender, ethnicity, hair color and expressions.

Large amount of research has been conducted to determine age estimation using facial features. Different public standard datasets which can be used for real age estimation which permit public performance comparison of the proposed methods.

As a result, a lot of active research has been, with several recent works utilizing the concept of Convolutional Neural Networks (CNNs) for extraction of features and inference. The facial expressions can be recognized using non-verbal communication between humans, along with the interpretation of facial expressions have been widely studied [1]. Facial expression plays an important role in human interaction, Facial Expression Recognition (FER) algorithm with the help of computer vision which helps in applications such as human-computer interaction and data analytics [2].

2. Literature Survey

We began our background search with research papers and blog posts online, related to our topic. The research paper details:

A new framework for facial expression recognition using an attentional convolutional network has been developed. Attention plays an important role in detecting facial expressions, which can then enable neural networks having less than 10 layers to compete with much deeper networks for emotion recognition presented in [1].

Face recognition was achieved successfully but they are affected by illumination, pose, facial expression, face containing eyebrows, nose, mouth length, face local points using Dlib in opencv [2].

In their model, during preprocessing Adaboost method is used to remove irrelevant features and Viola Jones algorithm is used to extract Haar like features which are given as input to CNN model for processing.[5].

Their deep model is trained on a large dataset of four million images for the task of face recognition. Above model serves as the backbone to our facial attribute recognizers and is used to fine-tune networks for four tasks: apparent age estimation, gender recognition and emotion recognition. From different sources images have been collected which are used for different tasks. Over 4 million images of more than 40,000 people are collected for facial recognition. Every image is labelled according to gender and the data part is annotated with emotion. These images are later trimmed using a semi-automated

process with a team of human annotators in the loop.

The images are then pre-processed next to extract the faces and align them. The aligned images are then fed to our proprietary deep network for training. [6].

Summary of Related Work

The summary of methods used in literature is given in Table 1.

Literature	CNN Model	SVM Classifier	Hybrid
Vladimir khryashchev , Alexander Ganin ,Olga Stepanova ,Anton Lebedev et al. 2016 [1]	No	Yes	No
Xiaofeng Wang , Azliza Mohd Ali, Plamen Angelov et al. 2017 [2]	Yes	Yes	Yes
Gil Levi , Tal Hasneer et al. 2018 [3]	Yes	No	No
Shivam Gupta et al. 2018 [4]	Yes	Yes	No
D DPribavkin, P Y Yakimov et al. 2019 [5]	Yes	No	Nos

Table 1. Summary of literature survey

The overview of comparison of different parameters are given in Table 2.

Literature	Age	Gender	Emotion
Vladimir Khryashchev, Alexander Ganin ,Olga Stepanova	Yes	No	No
Xiaofeng Wang, Azliza Mohd Ali, Plamen Angelov et al. 2017 [2]	Yes	Yes	Yes
Gil Levi , Tal Hasneer et al. 2018 [3]	Yes	Yes	No
Shivam Gupta et al. 2018 [4]	No	No	No
D D Pribavkin, P Y Yakimov et al. 2019 [5]	Yes	Yes	Yes

Table 2. Summary of literature survey

3. Proposed Work

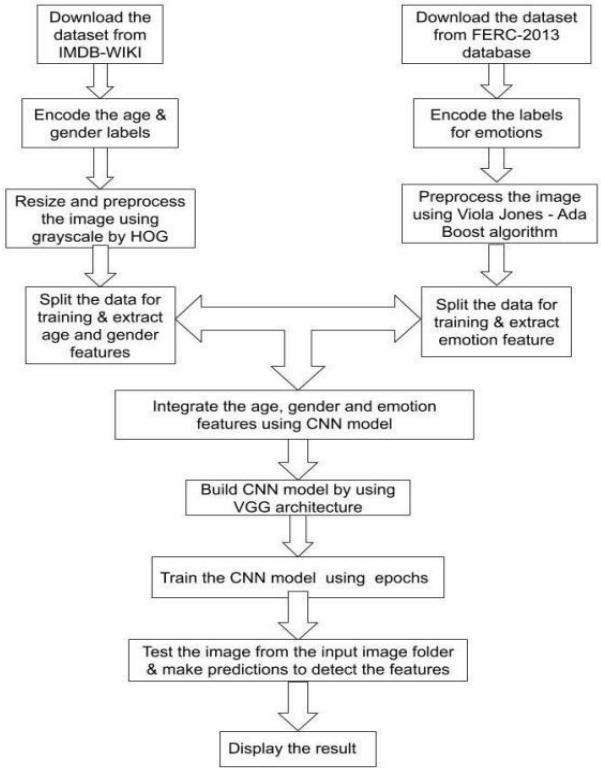

Fig. 1 Proposed System Architecture

For *Age and Gender Detection*, Deep EXpectation (DEX) – is used for age estimation which can be seen in image classification [5, 32, 47] and object detection [19] fuelled by deep learning. From the deep learning concept we learn four key ideas that we apply to our solution: (i) the deeper

the neural networks (by sheer increase of parameters / model complexity) the better is the capacity to model highly non-linear transformations - with some optimal depth on current architectures; (ii) the larger and more diverse the datasets used for training, the better the network learns to generalize and the more effective it becomes to over-fitting; (iii) the alignment of the object in the input image impacts the overall performance; (iv) when the training data is small that is when we must fine-tune a network pre-trained for comparable inputs and goals which would benefit us from the transferred knowledge.

We always start by rotating the input image at different angles to detect the face with the highest score. We then align the face using the angle and crop it for the further steps. This is a simple and effective procedure which does not involve facial landmark detection. We use deep VGG-16 architecture for our Convolutional Neural Network (CNN). We start from pre-trained CNNs on the large ImageNet dataset to classify images such that (i) it helps us by discriminating 1000 object categories in images by the representations learned, and (ii) to obtain a meaningful representation and a smooth and warm start for further fine-tuning on relatively smaller face datasets. Adjusting the CNNs on facial images with age annotations is an important step for superior performance, because the CNN adapts to best fit the particular data distribution and perform effective age estimation. Due to the shortage of facial images with apparent age annotation, we explore the benefit of adjusting over crawled Internet face images with available age. We compute 523,051 face images from the IMDb and Wikipedia websites to form IMDB-WIKI - our new dataset. It is the largest publicly available dataset with gender and age annotations.

While age estimation is expounded to regression problem, we go further and cast the age estimation as a multi-class classification of age bins followed by a softmax expected value refinement.

Our main contributions are as follows:

1. The IMDB-WIKI dataset which is the largest dataset with real age and gender annotations
2. A novel regression formulation is used with deep classification followed by expected value refinement
3. The DEX system, which is the winner of the LAP 2015 challenge on apparent age estimation

We have the tendency to then introduce our IMDB-WIKI dataset for age estimation that provides a more elaborated analysis of the projected DEX system, then apply the method and reports the result of standard age estimation datasets.

On the opposite hand, for *Emotion Detection & Classification*, we've evaluated and tested completely different preprocessing techniques and several other model architectures, ultimately developing a custom CNN model capable of achieving near-state-of-the-art accuracy of 70.47% on the FER-2013 test set. For preprocessing, we experimented with centering and scaling data. We later on found, that subtracting the mean is generally more helpful for the train distribution from all sets before training/evaluating. For implementing data augmentation: we randomly rotate, shift, flip, crop, and sheer our training images. This yielded about a 10p.p. increase in accuracies. We have implemented several CNN architectures from different papers applying emotion recognition to these and other datasets. Finally, what yielded the best performance was our custom developed CNN architecture. It is very difficult to analyze errors in neural networks

We analyzed our errors across different classes, as well as by visual inspection of images we classified correctly and incorrectly. One early observation was that we fail much more at certain emotions, and that we were failing to classify images which were necessary to rely on fine details in the images (e.g., small facial features or curves). Due to this, we have increased the number of layers and decreased filter sizes to increase the number of parameters in our network, so that it had a clear effect in allowing us to fit the dataset better. This led to the problem of overfitting, which we later addressed by using dropout, early stopping at around 100 epochs, and augmenting our training set. Given this, we could only start learning training set noise after achieving approx. 70% dev set accuracy; this is clear from plotting accuracy during training. Finally, this leaves us with some suggestions for future work, which focuses largely on enabling increased parameterization of the network.

We used OpenCV's Haar cascades to detect and extract a face region from a webcam video feed, then classified it using our CNN model. We discovered that it's best to neither subtract the training mean nor normalize the pixels within the detected face region before classifying it. During Real-time classification it exposed our model's strengths: neutral, happy, surprised, and angry were generally well-detected. Illumination was a very important factor in the model's performance. This suggests that our training set may not truthfully represent the distribution of emotion during less brightness on screen.

4. Implementation Details

The implementation detail is given during this section.

In the planned design, shown in Figure 1, we have the tendency to begin with downloading the image-set from a dataset referred to as IMDB WIKI as a result of it being the largest publicly available dataset with gender and age labels

for training. Simultaneously, image-set is downloaded from a dataset referred to as FERC-2013.

A. *Download the Dataset:*

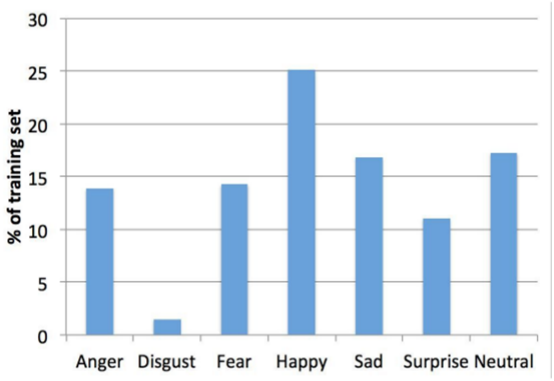

Fig. 2 Distribution of different emotions across the dataset

B. ***Function to encode the labels:*** After downloading the dataset, we assign labels to this image-set using one-hot encode because CNN does not work with categorical data- variables that contain label values rather than numeric values. To overcome this problem of CNN, we use a one-hot encoding algorithm to convert these label values into numeric values which will be easily processed by CNN.

C. ***Resizing and preprocessing the data:*** Resizing of all the images to fixed pixel 256x256 values are done and preprocessing is done by converting them into grayscale using HOG (Histogram of Gradient) algorithm. For emotions, image preprocessing is done by using Viola Jones - AdaBoost algorithm to extract haar like features specially used for detecting emotions. Viola-Jones takes an ensemble approach. What which means is that Viola-Jones uses many alternative classifiers, each staring at a distinct portion of the image. Every individual classifier is weaker (less correct, produces more false positives, etc) than the ultimate classifier as a result of it's taking in less information. The image is reshaped in such a way that it only considers the facial features of the image.

D. ***Extracting and integrating the features:*** After preprocessing, the facial features such as eyebrows and distance between them, nose, mouth length, and face landmark points are extracted using the DLIB library which is present in OpenCV. Then age, gender and emotion features are integrated as one for training.

E. ***Training and testing:*** The CNN model is built by using VGG-16 architecture. The CNN model is then trained using epochs, where each epoch contains a certain number of training images. To remove distorted and unwanted images, the loss Gauss function is used. For testing, the input image is given by the user. The model makes the predictions to estimate age, gender and emotion of that input image by comparing with the trained images.

F. ***Output:*** In the Output phase, we apply the same feature extraction process to the new images and we pass the features to the trained machine learning algorithm to predict the label.

5. Requirement Analysis

The implementation detail is given in this section.

5.1 Software

Prerequisites are:

1. Keras2 (with TensorFlow backend)
2. OpenCV
3. Python 3.5 (TensorFlow not supported in higher versions)
4. NumPy
5. TensorFlow
6. h5py (for Keras model serialization)

5.2 Hardware

Intel core processor with high GPU power & frequency

5.3 Dataset

1. IMDB-WIKI – for age and gender detection.
2. FER 2013 – for emotion detection

ACKNOWLEDGMENT

We would prefer to specify our deepest appreciation to all or any people who provided us the chance to complete this report. A special feeling, we tend to offer to our project manager, Prof. Rupali Nikhare, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project particularly in putting this on ink report. Many thanks go to our guide, Prof. Sujit Tilak who has invested his full effort in guiding the team in achieving the goal. Last but not least, many thanks to our Head of Department of Computer Engineering, Prof. Sharvari Govilkar for this opportunity. We have to appreciate the guidance given by other supervisors as well as the panels particularly in our project presentation that has improved our presentation skills thanks to their comments and advice.

REFERENCES

1. Shervin Minaee, Amirali Abdolrashidi, Deep - Emotion: Facial Expression recognition using attentional convolutional network, 2019 3rd IEEE International Conference on Cybernetics (CYBCONF)doi:10.1109/cybconf.2017.7985780
2. Alen Salihbasi and Tihomir Orehovacki, Development of android application for gender, age and face recognition using opencv, MIPRO 2019, ISSN: 1057-7149, 20 May 2019
3. Jiu-Cheng Xie, Chi-Man Pun, 'Age and Gender Classification using Convolutional Neural Networks, IEEE Workshop on Analysis and Modelling of Faces & Gestures', IEEE Transactions on Information Forensics and Security, ISSN: 1556-6013,07 March 2019
4. Gil Levi and Tal Hassner,' Facial Emotion Analysis using Deep Convolutional Neural Network', 2017. International Conference on Signal Processing and Communication (ICSPC), Coimbatore, 2017.
5. Mostafa Mohammadpour, Seyyed Mohammed. R Hashemi, Hossein Khaliliardali, Mohammad. M Alyan Nezhadi,'Facial Emotion Recognition using Deep Convolutional Networks', 2017 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 22 December, 2017.
6. Dehghan, Afshin & G. Ortiz, Enrique & Shu, Guang & Zain Masood, Syed. (2017).DAGER: Deep Age, Gender and Emotion Recognition Using Convolutional Neural Network.