A Project Report

On

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

Submitted in partial fulfillment of the requirement of
University of Mumbai for the Degree of

Bachelor of Engineering
In
Computer Engineering

Submitted By

Dipesh Nair

Sidharth Nair

Anoop Pillai

Gautam Nair

Supervisor
Prof. Sujit Tilak

Department of Computer Engineering

PILLAI COLLEGE OF ENGINEERING

New Panvel – 410 206

UNIVERSITY OF MUMBAI

Academic Year 2019 – 20

DEPARTMENT OF COMPUTER ENGINEERING
Pillai College of Engineering
New Panvel – 410 206

CERTIFICATE

This is to certify that the requirements for the BE Synopsis entitled '**Detection of Gender, Age and Emotion of a Human Image Using Facial Features**' have been successfully completed by the following students:

Name	**Roll No.**
Dipesh Nair	B822
Gautam Nair	B823
Sidharth Nair	B830
Anoop Pillai	B847

in partial fulfillment of Bachelor of Engineering of Mumbai University in the Department of Computer Engineering, Pillai College of Engineering, New Panvel – 410 206 during the Academic Year 2018 – 2019.

Supervisor
(Prof. Sujit Tilak)

Head of Department
Dr. Sharvari Govilkar

Principal
Dr. Sandeep M. Joshi

DEPARTMENT OF COMPUTER ENGINEERING

Pillai College of Engineering

New Panvel – 410 206

SYNOPSIS APPROVAL FOR B.E

This project synopsis entitled "**Detection of Gender, Age and Emotion of a Human Image Using Facial Features**" by **Dipesh Nair, Sidharth Nair, Anoop Pillai and Gautam Nair** are approved for the degree of B.E. in **Computer Engineering**.

Examiners:

1. _________________

2. _________________

Supervisors:

1. _________________

2. _________________

Chairman:

1. _________________

Date:

Place:

Declaration

We declare that this written submission for B.E. Declaration entitled "**Detection of Gender, Age and Emotion of a Human Image Using Facial Features**" represent our ideas in our own words and where others' ideas or words have been included. We have adequately cited and referenced the original sources. We also declared that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any ideas / data / fact / source in our submission. We understand that any violation of the above will cause for disciplinary action by the institute and also evoke penal action from the sources which have thus not been properly cited or from whom paper permission have not been taken when needed.

Project Group Members:

Dipesh: ____________________________

Sidharth: ____________________________

Anoop: ____________________________

Gautam: ____________________________

Date:

Place:

Table of Contents

Abstract

Developing an automatic age and gender estimation method towards human faces continues 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 human vision system. In this project we propose a convolutional neural network (CNN) based architecture for joint age-gender classification. The architecture is trained to label the input images into age and 2 types of gender. Our approach shows improved accuracy in both age and gender classification compared to classifier-based methods. In order to computer modeling of human's emotion we are predicting human emotions using deep CNN and how emotion intensity changes on a face from low level to high level of emotion. By, using preprocessing algorithm Viola-Jones we extracted features of the image which are fed as an input to CNN. With a proper user interface result of the prediction is revealed.

List of Figures

Chapter 1

Introduction

1.1 Fundamental

Age estimation from a single face image is an important task in human and computer vision which has many applications such as in forensics or social media. It is closely related to the prediction of other biometrics and facial attributes tasks such as gender, ethnicity, hair color and expressions. A large amount of research has been devoted to age estimation from a face image under its most known form - the real, biological, age estimation. This research spans decades as summarized in large studies [42, 2, 9, 25, 20]. Several public standard datasets [2, 42, 44] for real age estimation permit public performance comparison of the proposed methods. In contrast, the study of apparent age, that is the age as perceived by other humans, is at the beginning.

Consequently, there has been active research in this field, with several recent works utilizing Convolutional Neural Networks (CNNs) for feature extraction and inference. Being able to recognize facial expressions is key to non-verbal communication between humans, and the production, perception, and interpretation of facial expressions have been widely studied [1]. Due to the important role of facial expressions in human interaction, the ability to perform Facial Expression Recognition (FER) automatically via computer vision enables a range of novel applications in fields such as human-computer interaction and data analytics [2].

1.2 Objective

To study the recommendation techniques and identify their limitations that may help to suggest a hybrid approach which may overcome the drawbacks existing methods.

a) To understand the method of feature extraction for content based recommendation system and collaborative filtering that may help user for decision making.

b) To identify evaluation metrics used for performance analysis of different recommendation systems.

1.3 Scope

With the recent rapid emergence of the intelligent applications there is a growing demand for automatic extraction of biometric information from face images or videos. Applications where growing demand for automatic extraction of biometric information from face images or videos. Applications include: (i) access control, e.g., restricting the access of minors to sensible products like alcohol from vending machines or to events with adult content; (ii) human-computer interaction (HCI), e.g., by a smart agent estimating the age of a nearby person or an advertisement board adapting its offer for young, adult, or elderly people, accordingly; (iii) law enforcement, e.g., automatic scanning of video records for suspects with an age estimation can help during investigations; (iv) surveillance, e.g., automatic detection of unattended children at unusual hours and places; (v) perceived age, e.g., there is a large interest of the general public in the perceived age (c.f. howhot.io), also relevant when assessing plastic surgery, facial beauty product development, theatre and movie role casting, or human resources help for public age specific role employment.

Chapter 2

Literature Survey

In this chapter the relevant techniques in literature is reviewed. It describes various techniques used in the work, and to identify the current literature on related domain problem, identify the techniques that have been developed and present the various advantages and limitations of these methods used extensively in literature.

2.1 Related Works

a) For Age and Gender:

While almost all literature prior the LAP 2015 challenge focuses on real (biological) age estimation from a face image, Han et al. [26] provide a study on demographic estimation in relation to human perception and machine performance. In the next, we briefly review the age estimation literature and describe a couple of methods that most relate with our proposed method. We refer to [42, 20, 14, 26, 2, 9] for broader literature reviews. Most of the prior literature assumes a normalized (frontal) view of the face in the input image or employ a face pre-processing step such that the face is localized and an alignment of the face is determined for the subsequent processing steps. Generally, the age estimators work on a number of extracted features, feature representations and learn models from training data such that to minimize the age estimation error on a validation data. The whole process assumes that the train, validation, and test data have the same distribution and are captured under the same conditions.

FG-NET [42] and MORPH [44] datasets with face images and (real) age labels are the most used datasets allowing for comparison of methods and performance reporting under the same benchmarking conditions. We refer to [42] for an overview of research (365+ indexed papers) on facial aging with results reported on FGNET dataset.

A large number of face models has been proposed. We follow the taxonomy from [20] and mention: wrinkle models [33], anthropometric models [11, 33, 43], active appearance models (AAM) [6], aging pattern subspace [18], age manifold [13, 23, 21], biologically-inspired models (including biologically-inspired features (BIF) [24]), compositional and dynamic models [54, 49], local spatially fexible patches [56], and methods using fast Fourier transform (FFT) and genetic algorithm (GA) for feature extraction and selection [15], local binary patterns (LBP) [58], Gabor filters [16]. Recently, the convolutional neural networks (CNN) [35], biologically inspired, were successfully deployed for face modeling and age estimation [53, 36, 52].

The age estimation problem can be seen as a regression [13] or as a classification problem up to a quantization error [34, 18]. Among the most popular regression techniques we mention Support Vector Regression (SVR) [8], Partial Least Squares (PLS) [17], Canonical Correlation Analysis (CCA) [27], while for classification the traditional nearest neighbor (NN) and Support Vector Machines (SVMs) [7].

In the next we select a couple of the representative (real) age estimation methods. Yan et al. [55] employ a regressor learning from uncertain labels, Guo et al. [23] learn a manifold and local SVRs, Han et al. [26] apply age group classification and within group regression (DIF), Geng et al. [18] introduce AGES (AGingpattErn Subspace), Zhang et al. [61] propose a multi-task warped Gaussian process (MTWGP), Chen et al. [4] derive CA-SVR with a cumulative attribute space and SVR, Chang et al. [1] rank hyperplanes for age estimation (OHRank), Huerta et al. [29] fuse texture and local appearance descriptors, Luu et al. [38] use AAM and SVR, while Guo and Mu [22] use CCA and PLS. Recently, Yi et al. [59] deployed a multi-scale CNN, Wang et al. [53] used deep learned features (DLA) in a CNN way, while Rothe et al. [46] went deeper with CNNs and SVR for accurate real age estimation on top of the CNN learned features.

b) For Emotion Detection:

In FER2013 challenge of the ICML 2013 Representation Learning [4], Tang introduced a CNN jointly learned with linear support vector machine (SVM) for facial expression recognition [11]. With a simple CNN and a SVM instead of softmax classifier, the model outperformed the others and won the first place in the challenge. Inspiring by the success of GoogLeNet [8], Mollahosseini et al. proposed an architecture containing four Inception modules [12]. However, their research cannot lead to a better performance on the FER2013dataset.

In 2016, Zhou et al. proposed the multi-scale CNNs[13]. This model consists of three other networks with different input sizes. In addition, they used late fusion technique to get the final classification results. By combining multiple CNNs and modifying the loss function, Yu et al. obtained a higher accuracy compared to the previous approaches [14]. Similarly, Kim et al. introduced a multiple CNNs for facial expression recognition in the wild [15]. Another multi-scale CNN was proposed by Wang et al. [16]. In this work, the authors use the entire feature maps in the network for classification. However, using all generated features without selection may reduce the overall performance due to trivial information in shallow layers of the network. To the best of our knowledge, this work is the current state-of-the-art method on the FER2013 dataset.

Chapter 3

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

3.1 Overview

3.1.1 Proposed System Architecture

Gender and age recognition are using Levi and Hassner's Caffe model. The model is trained on the Adience collection of unfiltered faces for gender and age classification and contains 26.580 images of a total of 2.284 subjects. The source for the photos in Adience collection are the Flickr.com albums, produced by automatic upload from smartphones that are later publicly available through the Creative Commons (CC) license. All images were manually labeled for age and gender using both the images themselves and using any available contextual information.

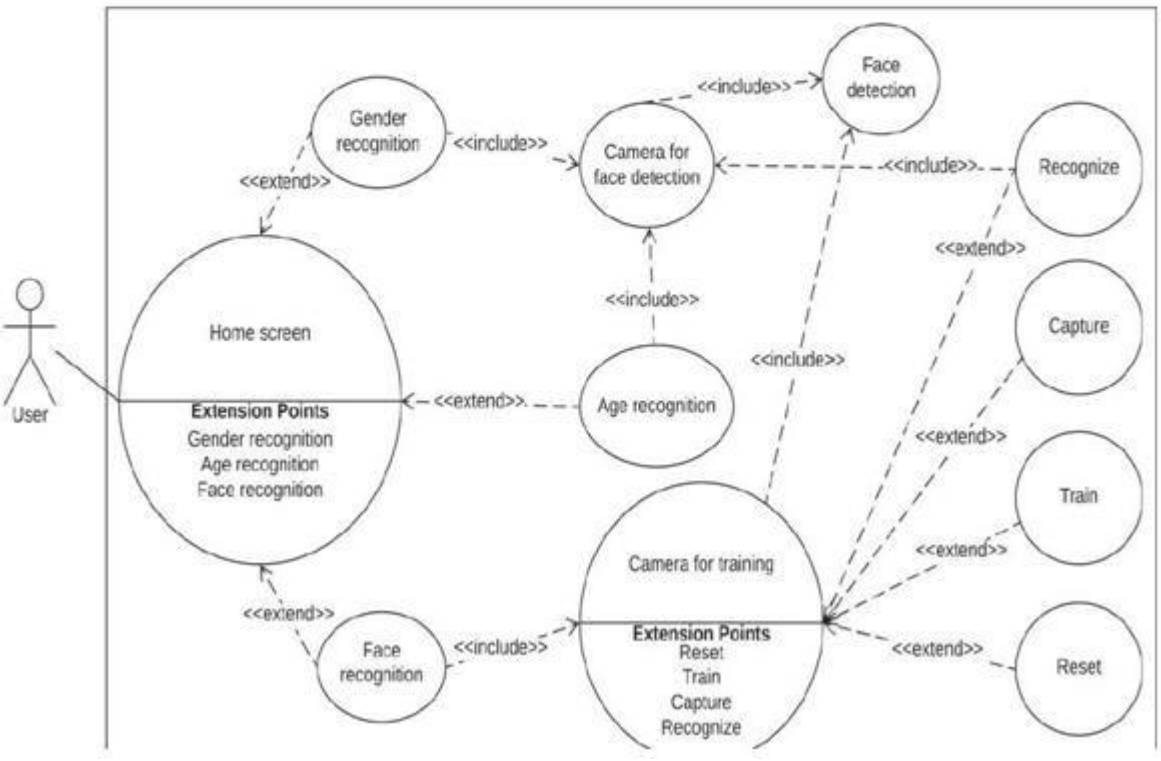

Fig. 3.1 Existing System Architecture used for Gender, Age and Face Recognition

The convolutional neural network contains three convolutional layers, each followed by a rectified linear operation and a pooling layer. The first convolutional layer contains 96 filters of 7x7 pixels, the second convolutional layer contains 256 filters of 5x5 pixels, the third and final convolutional layer contains 384 filters of 3x3 pixels. Finally, two fully-connected layers are added, each containing 512 neurons. At the end, the result is obtained from fully-connected layers, which is the class attribute, in this case gender or age, to which the input image belongs. The model was implemented with OpenCV module for Deep Neural Networks, DNN. The gender recognition result is described as the output value of 0 or 1, where 0 indicates a male person, and the value 1 indicates a female person. In age recognition, the result of the age estimation is described in the form of output values from 0 to 7, where each value represents a particular age group; 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+.

The mobile application uses the LBPH face recognition model that is created using a static FaceRecognizer class located in the OpenCV face analysis module, Face. After creating a face recognition model, it is possible to use its methods. One of the methods is a train() method that as a parameter requires a set of images to be used to train the model and the corresponding labels of these images. After training a face recognition model, the model is saved in the device memory of the mobile device by the write() method. The result is a file in XML format that contains all of the extracted features of the face images over which the training was performed. The file is saved in the directory FacePics. After face recognition model training has been completed, by pressing the Recognize button on the mobile device display, the face recognition activity will launch. When the face recognition activity is started, in the background is loaded trained face recognition model in XML format. Trained face recognition model is used to recognize detected faces on a mobile device display.

3.1.2 Existing System Architecture

In the proposed architecture, shown in Figure 1, we start with downloading the image-set from a dataset called IMDB WIKI because it is the largest publicly available dataset with gender and age labels for training. Simultaneously, image-set in downloaded from a dataset called FERC-2013. 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 one-hot encoding algorithm to convert these label values into numeric values which will be easily processed by CNN.

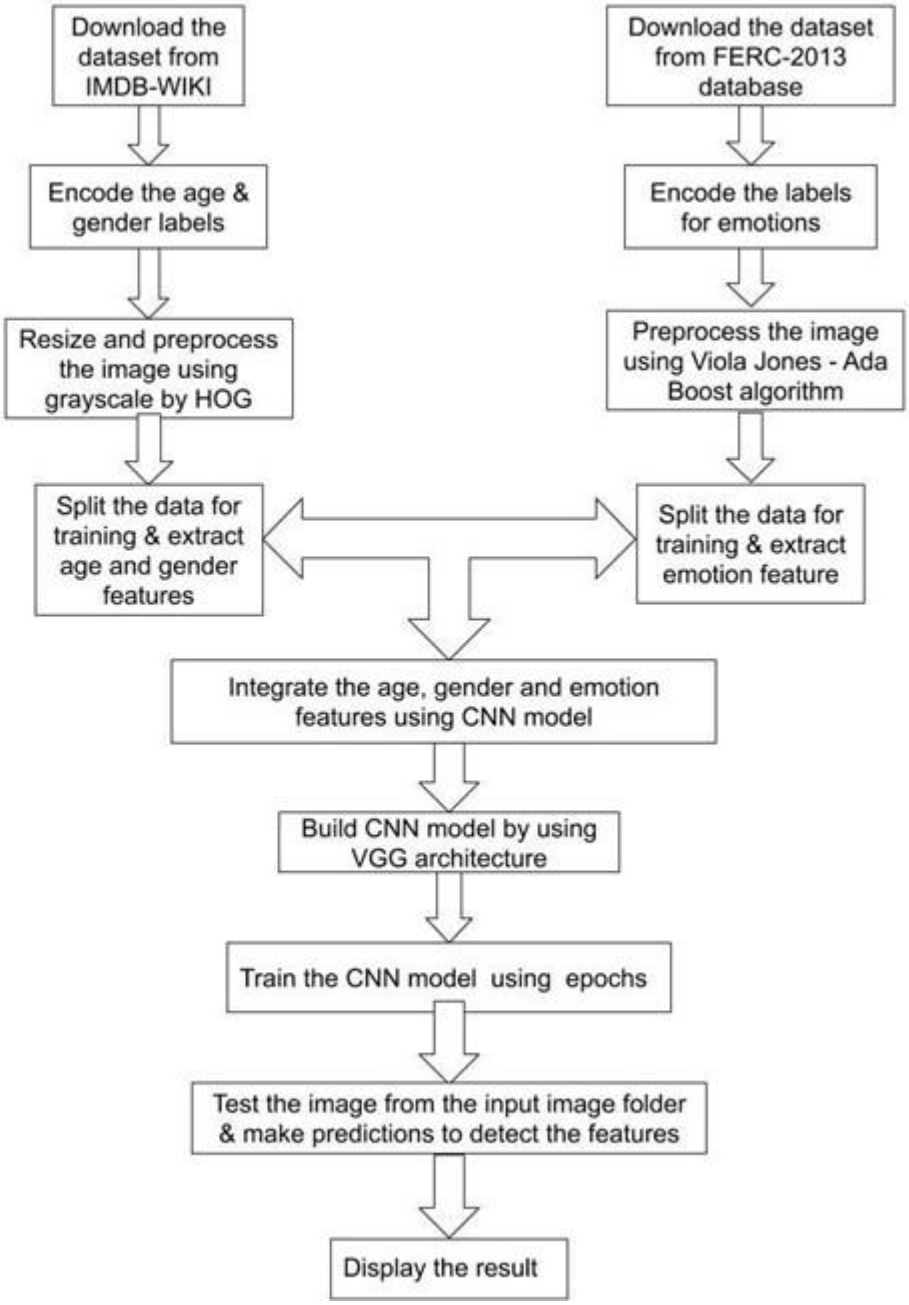
Download the dataset from IMDB-WIKI
Encode the age & gender labels
Resize and preprocess the image using grayscale by HOG
Split the data for training & extract age and gender features
Download the dataset from FERC-2013 database
Encode the labels for emotions
Preprocess the image using Viola Jones - Ada Boost algorithm
Split the data for training & extract emotion feature
Integrate the age, gender and emotion features using CNN model
Build CNN model by using VGG architecture
Train the CNN model using epochs
Test the image from the input image folder & make predictions to detect the features
Display the result

Fig. 3.2 Proposed system architecture for age, gender & emotion identification

After assigning labels to the image dataset, we proceed to the next step i.e., resizing the image to 256x256 pixels and preprocessing by converting them into grayscale using HOG (Histogram of Gradient) algorithm which counts occurrences of gradient orientation in localized portions of an image. This method is similar to edge orientation histograms, scale-invariant feature transform descriptors and shape contexts, but in this, it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy.

The resizing is done using lossy compression technique because this would extract all the necessary features from the images such as, face structure, beard hair, shape of eyebrows along with its width, etc. After resizing the image to the specified size we preprocess the resized image into grayscale using the luminance algorithm. In this algorithm, is based on the fact that cone density in the human eye is not uniform across colors. Humans perceive green more strongly than red, and red more strongly than blue. This makes sense from an evolutionary biology standpoint-much of the natural world appears in shades of green, so humans have evolved greater sensitivity to green light. Because humans do not perceive all colors equally, ‘the average method’ of grayscale conversion (the easiest technique) is inaccurate. Instead of treating red, green and blue qually, a good grayscale conversion algorithm will weight each color based on how the human perceives it. The formula for the ‘Luminance’ algorithm is given as:

$$\text{Gray} = (\text{Red} * 0.2126 + \text{Green} * 0.7152 + \text{Blue} * 0.0722)$$

For emotions, image preprocessing is done by using Viola Jones - Ada Boost algorithm to extract haar like features specially used for detecting emotions. Viola-Jones takes an ensemble approach.What that means is that Viola-Jones uses many different classifiers, each looking at a different portion of the image. Each individual classifier is weaker (less accurate, produces more false positives, etc) than the final classifier because it is taking in less information. The image is reshaped in such a way that it only considers the facial features of the image.

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

After resizing and converting the image into a grayscale image, 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 make the predictions to estimate age, gender and emotion of that input image by comparing test image with trained images. 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. In the prediction 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.

3.2 Implementation

3.2.1 Methodology

(a) For Age and Gender Detection

Datasets

IMDB-WIKI. We introduce a new dataset for age estimation which we name IMDB-WIKI. To the best of our knowledge this is the largest publicly available dataset for age estimation of people in the wild containing more than half a million labeled images. Most face datasets which are currently in use (1) are either small (i.e. tens of thousands of images) (2) contain only frontal aligned faces or (3) miss age labels. As the amount of training data strongly affects the accuracy of the trained models, especially those employing deep learning, there is a clear need for large datasets. For our IMDB-WIKI dataset we crawl images of celebrities from IMDb 1 and Wikipedia 2. For this, we use the list of the 100,000 most popular actors as listed on the IMDb website and automatically crawl from their profiles date of birth, name, gender and all the images related to that person. Additionally, we crawl all profile images from pages of people from Wikipedia with the same meta information. For both data sources we remove the images that do not list the year when it was taken in the caption. Assuming that the images with single faces are likely to show the celebrity and that the year when it was taken and date of birth are correct, we are able to assign to each such image the biological (real) age.

Especially the images from IMDb often contain several people. To ensure that we always use the face of the correct celebrity, we only use the photos where the second strongest face detection is below a threshold. Note that we cannot vouch for the accuracy of the assigned age information. Besides incorrect captions, many images are stills from movies - movies that can have extended production times. Nonetheless for the majority of the images the age labels are correct. In total IMDB-WIKI dataset contains 523,051 face images: 460,723 face images from 20,284 celebrities from IMDb and 62,328 from Wikipedia. Only 5% of the celebrities have more than 100 photos, and on average each celebrity has around 23 images. We make the dataset publicly available at http://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/. We also release pre-trained models. Note that this dataset can also be used for gender classification. We provide the entire image, the location of the face, its score and the score of the second most confident face detection.

Our approach

Deep EXpectation (DEX) – to age estimation is motivated by the recent advances in fields such as image classification [5, 32, 47] or object detection [19] fuelled by deep learning. From the deep learning literature we learn four key ideas that we apply to our solution: (i) the deeper the neural networks (by sheer increase of parameters / model complexity) are the better the capacity to model highly non-linear transformations - with some optimal depth on current architectures as [28] suggests; (ii) the larger and more diverse the datasets used for training are the better the network learns to generalize and the more robust 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 the best is to fine-tune a network pre-trained for comparable inputs and goals and thus to benefit from the transferred knowledge.

We always start by first rotating the input image at different angles to then pick the face detection [41] with the highest score. We align the face using the angle and crop it for the subsequent steps. This is a simple butrobust procedure which does not involve facial landmark detection. For our convolutional neural networks (CNNs) we use the deep VGG-16 architecture [48]. We always start from pre-trained CNNs on the large ImageNet [47] dataset for image classification such that (i) to benefit from the representation learned to discriminate 1000 object categories in images, and (ii) to have a meaningful representation and a warm start for further re-training or fine-tuning on relatively small(er) face datasets.

Fine-tuning the CNNs on face images with age annotations is a necessary step for superior performance, as the CNN adapts to best fit to the particular data distribution and target of age estimation. Due to the scarcity of face images with (apparent) age annotation, we explore the benefit of fine-tuning over crawled Internet face images with available (biological, real) age. We crawl 523,051 face images from the IMDb and Wikipedia websites to form IMDB-WIKI - our new dataset which we make publicly available. Fig. 4 shows some images. It is the largest public dataset with gender and real age annotations. While age estimation is a 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, the largest dataset with real age and gender annotations
2. A novel regression formulation through a deep classification followed by expected value refinement
3. The DEX system, winner of the LAP 2015 challenge [10] on apparent age estimation

This work is an extended and detailed version of our previous LAP challenge report paper [45]. We now officially introduce our IMDB-WIKI dataset for apparent age estimation, provide a more in depth analysis of the proposed DEX system, and apply the method and report results also on standard real age estimation datasets.

Age estimation

We employ a convolutional neural network (CNN) to predict the age of a person starting from a single input face image. This takes an aligned face with context as input and returns a prediction for the age. The CNN is trained on face images with known age.

CNN architecture

Our method uses a CNN with the VGG-16 [48] architecture (cf. Fig. 2 (4)). Our choice is motivated (i) by the deep but manageable architecture, (ii) by the impressive results achieved using VGG-16 on the ImageNet challenge [47], (iii) by the fact that as in our case the VGG-16 architecture starts from an input image of medium resolution (256 × 256), (iv) and that pretrained models for classification are publicly available allowing warm starts for training.

The VGG-16 architecture is much deeper than previous architectures such as the AlexNet [32] with 16 layers in total, 13 convolutional and 3 fully connected layers. It can be characterized by its small convolutional filters of 3x3 pixels with a stride of 1. AlexNet in comparison employs much larger filters with a size of up to 11 × 11 at a stride of 4. Thereby each filter in VGG-16 captures simpler geometrical structures but in comparison allows more complex reasoning through its increased depth. For all our experiments we start with the convolutional neural network pre-trained on the ImageNet images, the same models used in [48]. Unless otherwise noted, we fine-tune the CNN on the images from the newly introduced IMDB-WIKI dataset to adapt to face image contents and age estimation. Finally, we tune the network on the training part of each actual dataset on which we evaluate. The fine-tuning allows the CNN to pick up the particularities, the distribution, and thebias of each dataset and thus to maximize the performance.

The pre-trained CNN (with VGG-16 architecture) for the ImageNet classification task has an output layer of 1000 softmax-normalized neurons, one for each of the object classes. In contrast, age estimation is a regression and not a classification problem, as age is continuous rather than a set of discrete classes. For regression we replace the last layer with only 1 output neuron and employ an Euclidean loss function. Unfortunately training a CNN directly for regression is relatively unstable as outliers cause a large error term. This results in very large gradients which makes it difficult for the network to converge and leads to unstable predictions.

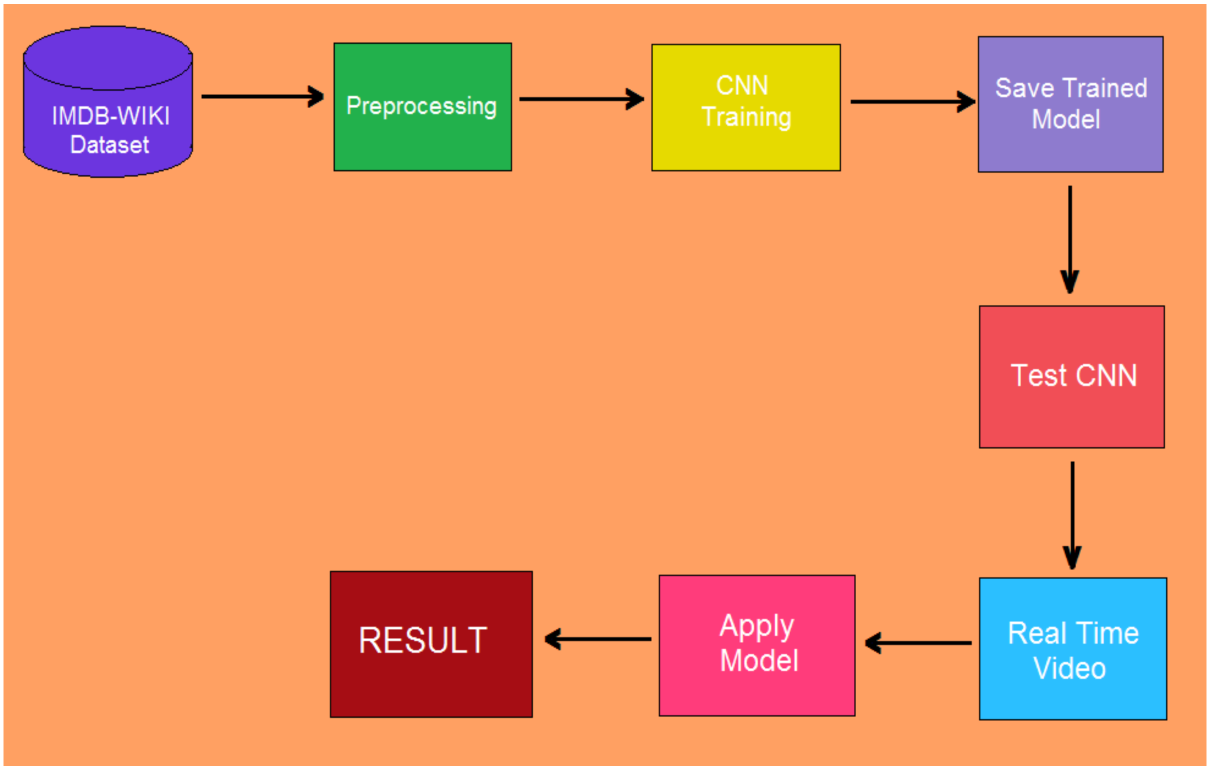

Fig. 3.3 Proposed Architecture for Age and Gender Detection

(b) For Emotion Detection

Dataset

For facial expression or emotion, we used FER-2013 dataset, which consists of about 37,000 well structured 48 × 48 pixel gray-scale images of faces. The images are processed in such a way that the faces are almost centered and each face occupies about the same amount of space in each image. Each image has to be categorized into one of the seven classes that express different facial emotions. These facial emotions have been categorized as: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise and 6=Neutral. Figure 1 depicts one example for each facial expression category. In addition to the image class number (a number between 0 and 6), the given images are divided into three different sets which are training, validation, and test sets. There are about 29,000 training images, 4,000 validation images, and 4,000 images for testing.

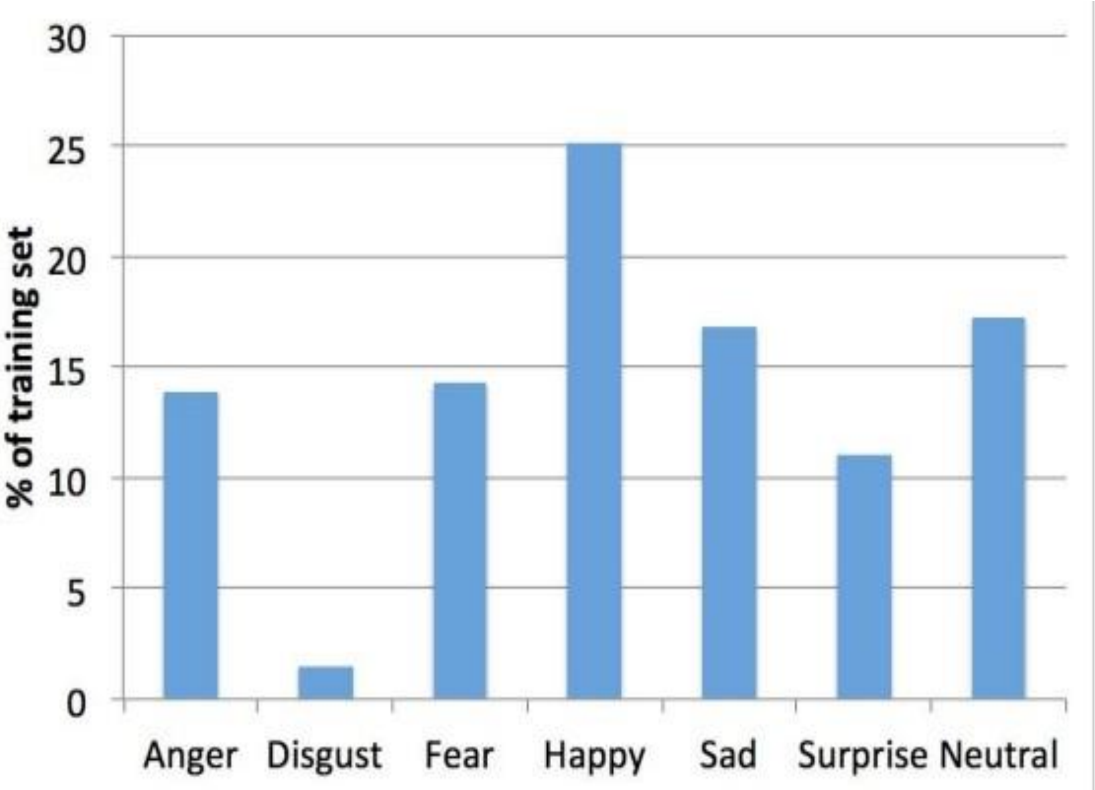

Fig. 3.4 Distribution of different emotions across the FER-2013 dataset

CNN

We evaluated both a variety of preprocessing techniques as well as several model architectures, ultimately developing a custom CNN model capable of attaining near-state-of-the-art accuracy of 70.47% on the FER-2013 test set. For preprocessing, we experimented with centering (i.e., subtracting mean) and scaling data. We found it generally helpful to subtract the mean found in the train distribution from all sets before training/evaluating. We also implemented data augmentation: we randomly rotate, shift, flip, crop, and sheer our training images. This yielded about a 10 p.p. increase in accuracies. We implemented several CNN architectures from papers applying emotion recognition to these and other datasets. Ultimately, what yielded the best performance was our custom developed CNN architecture (left). Analyzing error in neural networks is infamously difficult. We analyzed our error 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 where it was necessary to rely on fine details in the images (e.g., small facial features or curves).

Due to this, we increased the number of layers and decreased filter sizes to increase the number of parameters in our network, which had a clear effect in allowing us to fit the dataset better. This led to some overfitting, which we addressed by using dropout, early stopping around 100 epochs, and augmenting our training set. Given this, we only start learning training set noise after achieving approx. 70% dev set accuracy; this is clear from plotting accuracy during training. This leaves us with some suggestions for future work, which largely focus on enabling increased parameterization of the network.

Real-Time Classification

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 found it best to neither subtract the training mean nor normalize the pixels in the detected face region before classifying it. Real-time classification better 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 out training set may not truthfully represent the distribution of lighting conditions.

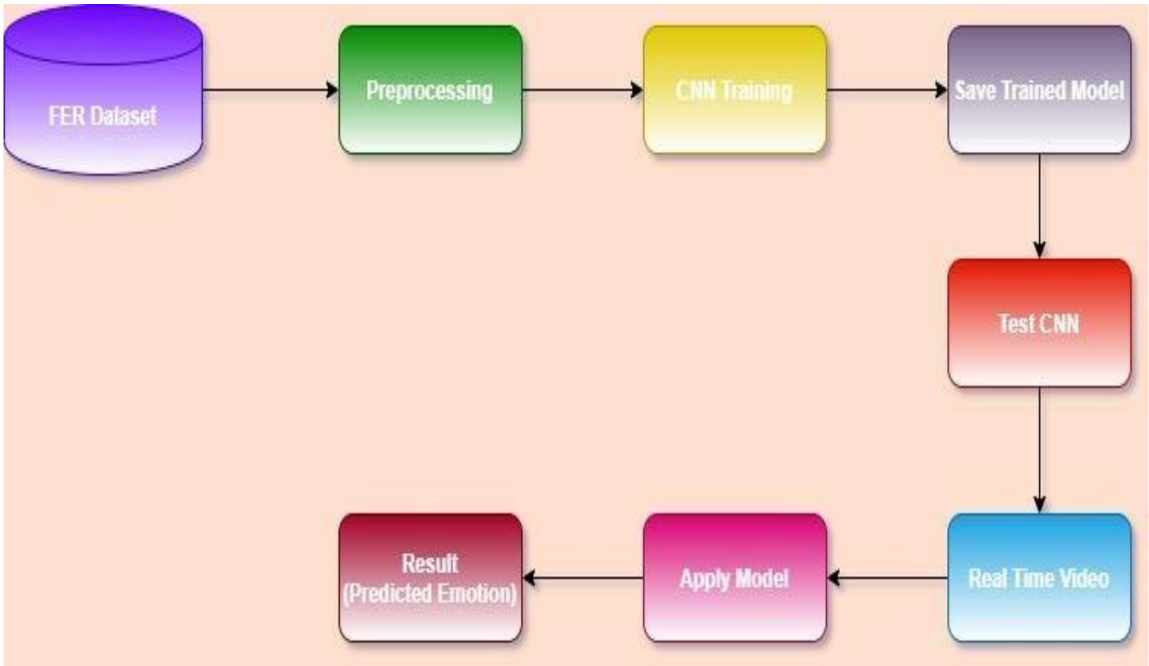

Fig. 3.5 Proposed Architecture for Emotion Detection

Chapter 4
Result and Discussion

When the program is executed, the GUI as shown in Fig.4.1 appears, which consist of two options –Age /Gender and Emotion.

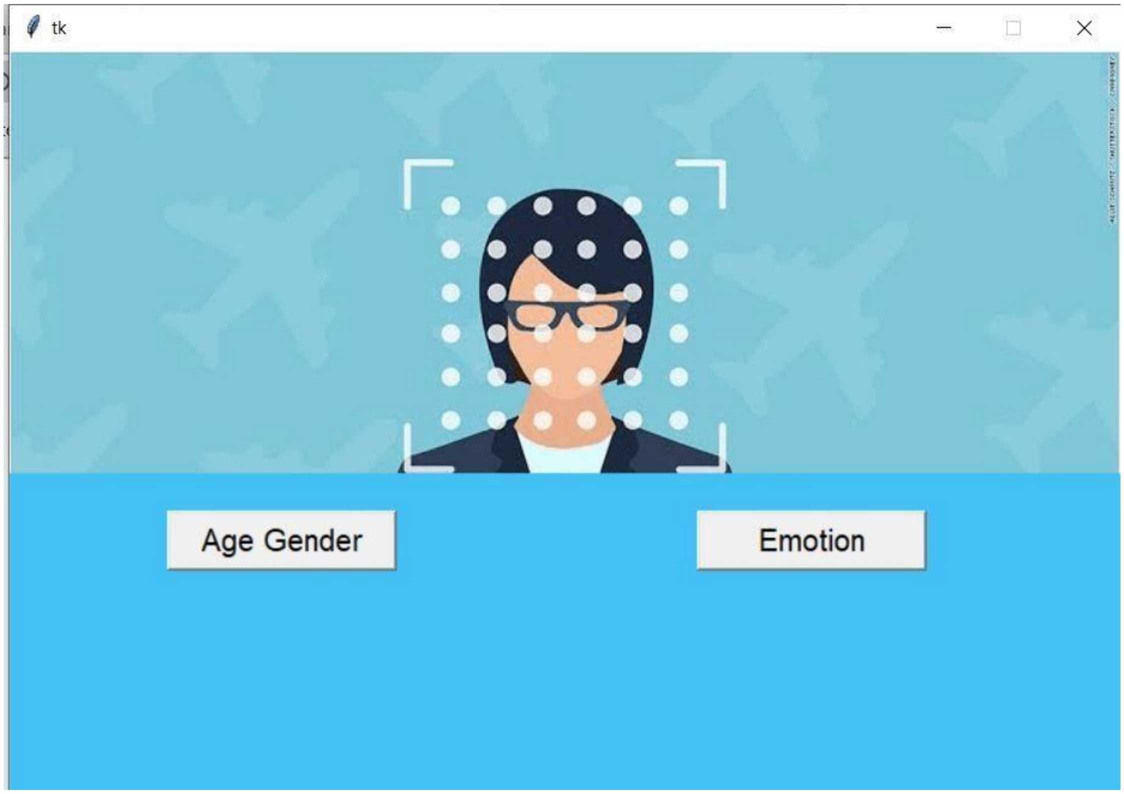

Fig. 4.1 Basic GUI model of the Age, Gender and Emotion Detection

After clicking one of them will take to the real time video (camera/webcam) where the detection will be shown.

Fig 4.2 represents the implementation of Age/Gender detection. It shows how a dimension is created over the face of the person in real time. Detection of emotion can also be known by the user just by clicking the 'Emotion' button, which is represented in Fig. 4.3

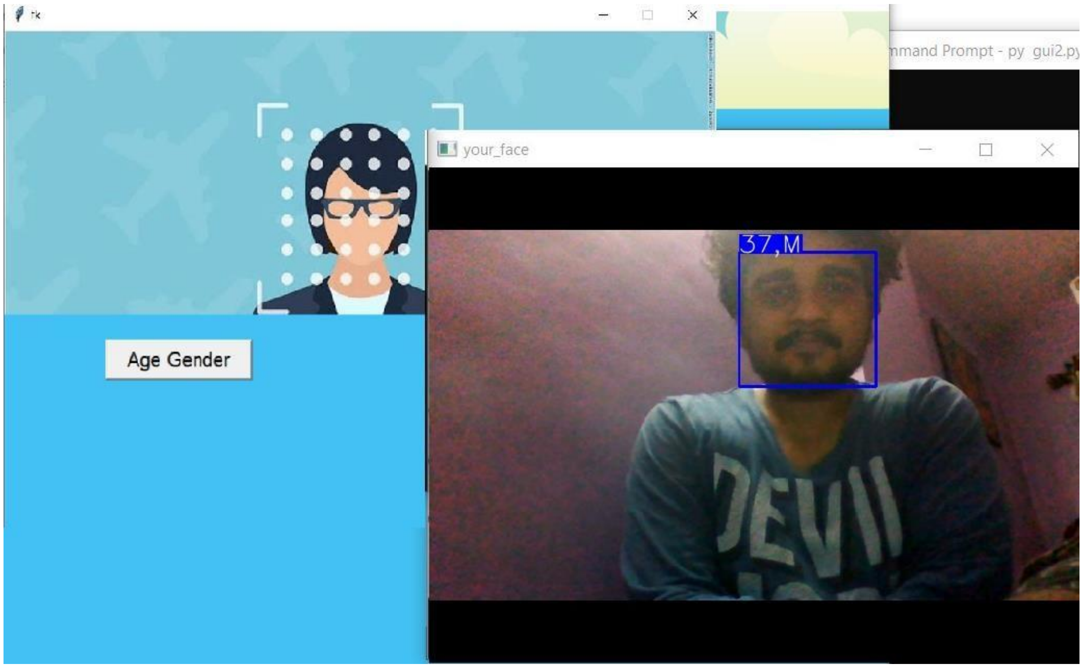

Fig. 4.2 Representation of Age and Gender Detection in real time

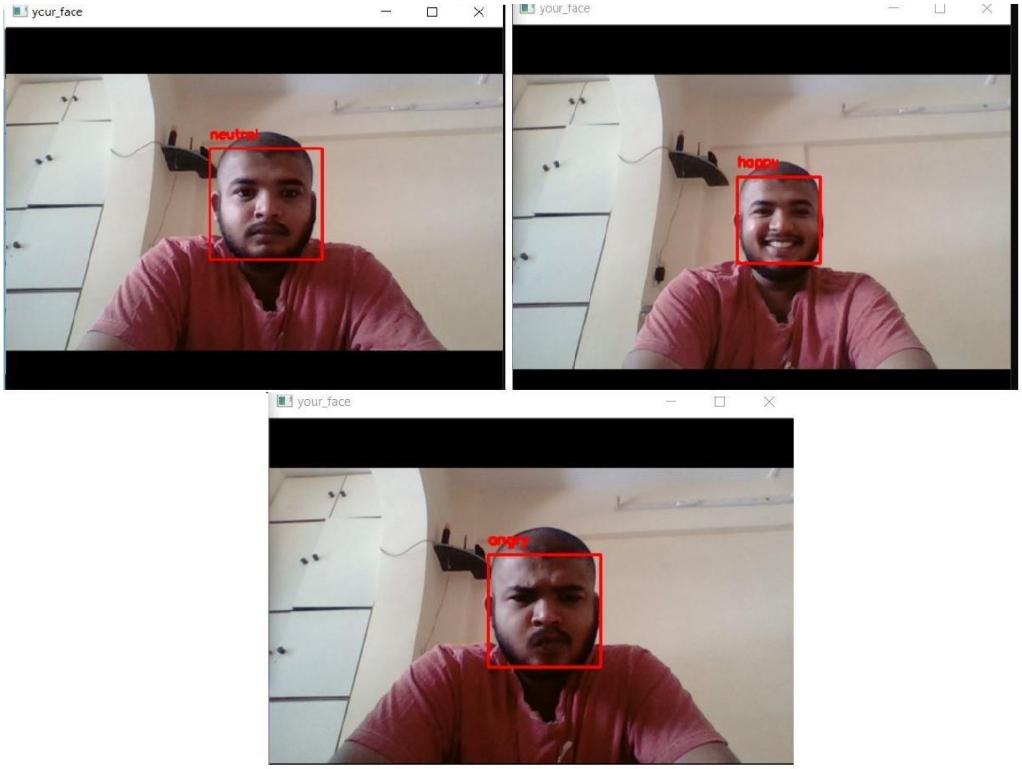

Fig. 4.3 Representation of Emotion Detection in real time

Chapter 5

Conclusion and Future Scope

Our Deep EXpectation (DEX) formulation builds upon a robust face alignment, the VGG-16 deep architecture and a classification followed by a expected value formulation of the age estimation problem. Another contribution is IMDB-WIKI, the largest public face images dataset to date with age and gender annotations. We validate our solution on standard benchmarks and achieve state-of-the-art results. If the real age estimation research spans over decades, the study of apparent age estimation or the age as perceived by other humans from a face image is a recent endeavor. The key factors of our solution are: deep learned models from large data, robust face alignment, and expected value formulation for age regression.

Apart from the age and gender detection, we believe there are two major areas of focus that would improve our real-time emotion recognition system. First, we suggest fine tuning the architecture of the CNN used for the model to fit perfectly with the problem at hand. Some examples of this fine tuning include finding and removing redundant parameters, adding new parameters in more useful places in the CNN’s structure, adjusting the learning rate decay schedule, adapting the location and probability of dropout and experimenting to find ideal stride sizes.

A second area of focus lies in adapting the datasets to more closely reflect real-time recognition conditions. For example, simulating low light conditions and “noisy” image backgrounds, could help the model become more accurate in real-time recognition. Additionally making sure that the distribution of models in the training dataset accurately reflects the distribution of subjects that the system will see when running in real-time.

References

[1] Chang KY, Chen CS, Hung YP (2011) Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[2] Chen BC, Chen CS, Hsu WH (2015) Face recognition and retrieval using cross-age reference coding with crossage celebrity dataset. IEEE Transactions on Multimedia 17(6):804–815

[3] Chen JC, Patel VM, Chellappa R (2016) Unconstrained face verification using deep CNN features. In: IEEE Winter Conference on Applications of Computer Vision (WACV)

[4] Chen K, Gong S, Xiang T, Change Loy C (2013) Cumulative attribute space for age and crowd density estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[5] Ciregan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[6] Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 23(6):681–685

[7] Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20(3):273–297

[8] Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: Advances in Neural Information Processing Systems 9, pp 155–161

[9] Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security 9(12):2170–2179

[10] Escalera S, Fabian J, Pardo P, Baro X, Gonzalez J, Escalante HJ, Misevic D, Steiner U, Guyon I (2015) Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In: IEEE International Conference on Computer Vision (ICCV) Workshops

[11] Farkas LG, Schendel SA (1995) Anthropometry of the head and face. American Journal of Orthodontics and Dentofacial Orthopedics 107(1):112–112

[12] Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and

Machine Intelligence (TPAMI) 32(9):1627–1645

[13] Fu Y, Huang TS (2008) Human age estimation with regression on discriminative aging manifold. IEEE Transactions on Multimedia 10(4):578–584

[14] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),vol. 37, no. 6, pp. 1113–1133, 2015.

[15] M. V. B. Martinez, "Advances, Challenges, and Opportunities in Auto-matic Facial Expression Recognition," in Advances in Face Detection and Facial Image Analysis. Springer, 2016, pp. 63–100.

[16] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza,B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou,C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor,M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra,J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges inrepresentation learning: A report on three machine learning contests,"Neural Networks, vol. 64, pp. 59–63, 2015.

[17] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.

[18] K. He, X. Zhang, haoqing Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CoRR, vol. 1512, 2015.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[20] Y. Tang, "Deep Learning using Support Vector Machines," in International Conference on Machine Learning (ICML) Workshops, 2013.

[21] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in ACM International Conference on Multimodal Interaction (MMI),

2015, pp. 435–442.

[22] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, “Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach,” in IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2016, pp. 48–57.

[23] Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, “Video and Image Based Emotion Recognition Challenges in the Wild: EmotiW 2015,” in ACM International Conference on Multimodal Interaction (ICMI), 2015, pp. 423–426.

[24] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, “Hierarchical committee of deep convolutional neural networks for robust facial expression recognition,” Journal on Multimodal User Interfaces, pp. 1–17, 2016.

[25] Mollahosseini, D. Chan, and M. H. Mahoor, “Going Deeper in Facial Expression Recognition using Deep Neural Networks,” CoRR, vol. 1511, 2015.

[26] Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

Acknowledgement

I would like to express my deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude we give to our project manager, Prof. Rupali Nikhare, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project especially in writing this report.

Many thanks go to our guide, Prof. Sujit Tilak who has invested her 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 supervisor as well as the panels especially in our project presentation that has improved our presentation skills thanks to their comments and advice.

Dipesh Nair
Sidharth Nair
Anoop Pillai
Gautam Nair