# Ensemble approach for Speech Analysis of AGE AND GENDER DETECTION

Name:p ajay kumar, k vinay reddy

Registration Number:11911903, 11901380

Section:KM044

Lovely Professional University, Punjab, India.

ajaychowdarypopuri@gmail.com
vinayreddykovvuri982@gmail.com

Abstract

Age and gender,two of the key facial attributes, play a very foundational role in social interactions, making age and gender estimation from a single face image an important task in intelligent applications, such as access control, human-computer interaction, lawenforcement,marketing intelligence and visual surveillance, etc.

Automatic age and gender classification has become relevant to anincre asing amount of applications, particularly since the rise of social platforms and social media. Nevertheless, performance of existing methods on real-world images is still significantly lacking, especially when compared to the tremendous leaps in performance recently reported for the related task of face recognition.Age and Gender Classification using Convolutional Neural Networks gives more accuracy compared to the previous one.

**Keywords:** Computer Vision, Opencv, Convolutional Neural Network , Gender Classification, Age classification

## INTRODUCTION

The task of detecting age and gender, however, is an inherently difficult problem, more so than many other computer vision tasks. The main reason for this difficulty gap lies in the data required to train these types of systems.

While general object detection tasks can often have access to hundreds of thousands or even millions of images for training, datasets with age and/or gender labels are considerably smaller, usually in the thousands or, at best, tens of thousands.

The reason is that to have tags for such images, we need to access the personal information of the subjects in the images. Namely, we would need their date of birth and gender, and in particular date of birth is infrequently published information.

Namely, we would need their date of birth and gender, and in particular date of birth is infrequently published information. Therefore, we have to settle for the nature of this problem that we are addressing and adapt network architectures and algorithmic approaches to deal with these limitations.

**Keywords**

**Computer Vision:**

**Computer Vision** is the field of study that enables computers to see and identify digital images and videos as a human would. The challenges it faces largely follow from the limited understanding of biological vision. Computer Vision involves acquiring, processing, analysing , and understanding digital images to extract high-dimensional data from the real world in order to generate symbolic or numerical information which can then be used to make decisions. The process often includes practices

like object recognition, video tracking, motion estimation, and image restoration.

**Opencv:**

**OpenCV** is short for Open Source Computer Vision. Intuitively by the name, it is an open-source Computer Vision and Machine Learning library. This library is capable of processing real-time image and video while also boasting analytical capabilities. It supports the Deep Learning frameworks TensorFlow, Caffe , and PyTorch.

**Convolutional neural network (CNN):**

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of Artificial Neural Network (ANN), most commonly applied to analyze visual imagery. They are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN), based on the shared-weight architecture of the convolution kernels or filters that slide along input features and provide translation equivariant responses known as feature maps. Counter-intuitively, most convolutional neural networks are only equivariant, as opposed to invariant, to translation. They have applications in image and video recognition, recommender systems,

image classification, image segmentation, medical image analysis, natural language processing, brain-computer interfaces, and financial time series.

**The CNN Architecture**

The convolutional neural network for this python project has 3 convolutional layers:

- Convolutional layer; 96 nodes, kernel size 7
- Convolutional layer; 256 nodes, kernel size 5
- Convolutional layer; 384 nodes, kernel size 3

It has 2 fully connected layers, each with 512 nodes, and a final output layer of softmax type.

To go about the python project, we'll:

- Detect faces
- Classify into Male/Female
- Classify into one of the 8 age ranges
- Put the results on the image and display it

**Related Work**

Before describing the proposed method we briefly review related methods for age and gender classification and provide a cursory overview of deep convolutional networks.

**2.1. Age and Gender Classification**

**Age classification**.

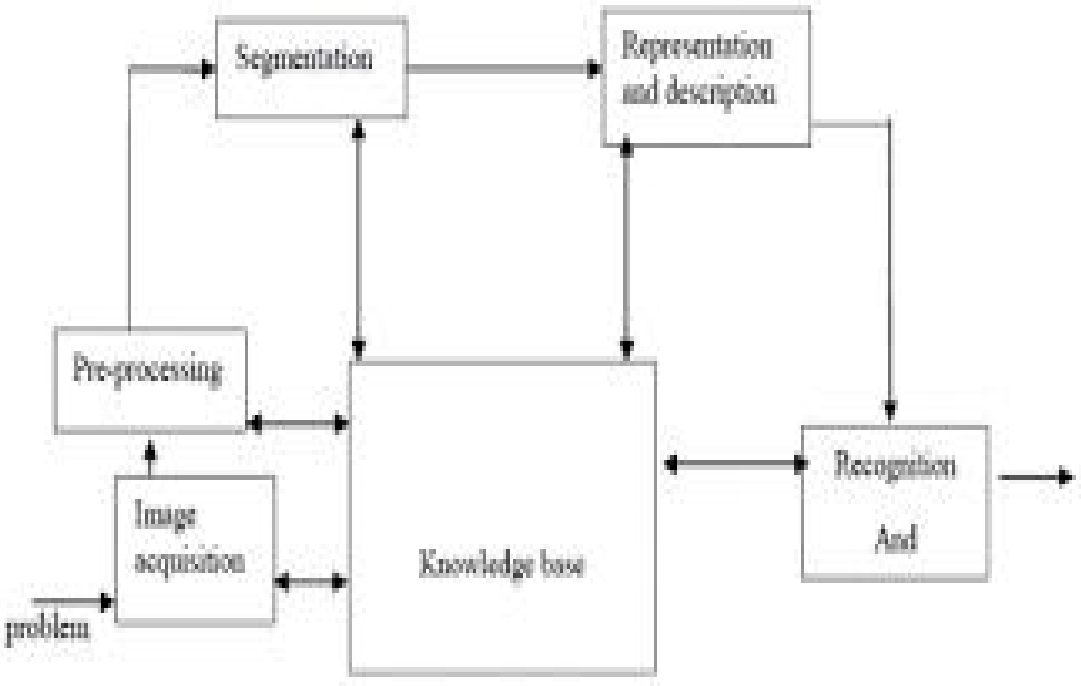
The problem of automatically extracting age related attributes from facial images has received increasing attention in recent years and many methods have been put fourth. A detailed survey of such methods can be found and, more recently. We note that despite our focus here on age group classification rather than precise age estimation (i.e., age regression), the survey below includes methods designed for either task.

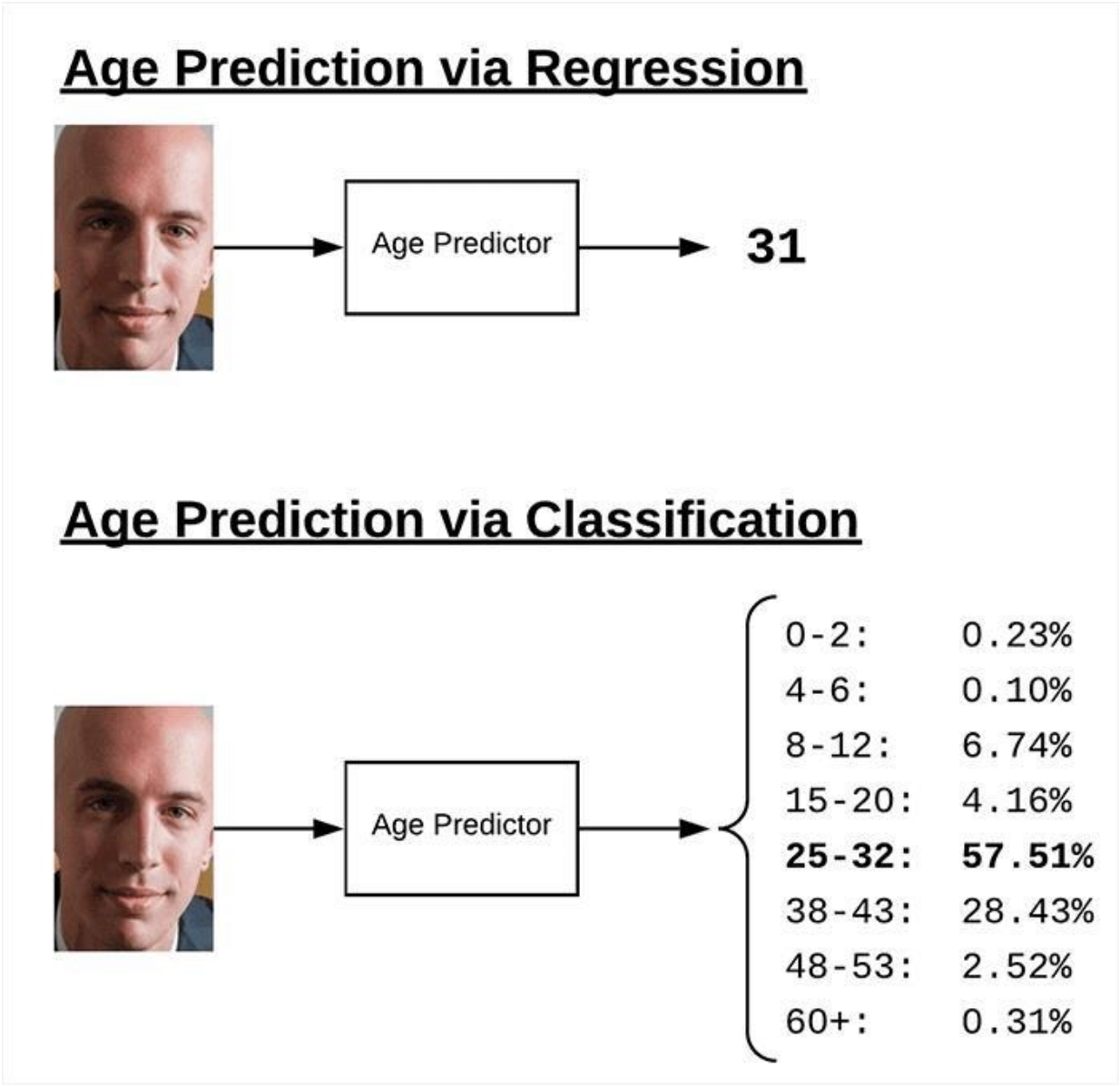
Early methods for age estimation are based on calculating ratios between different measurements of facial features . Once facial features (e.g. eyes, nose, mouth, chin, etc.) are localized and their sizes and distances measured, ratios between them are calculated and used for classifying the face into different age categories according to hand-crafted rules. More recently, uses a similar approach to model age progression in subjects under 18 years old. As those methods require accurate localization of facial features, a challenging problem by itself, they are unsuitable for in-the-wild images which one may expect to find on social platforms. On a different line of work are methods that represent the aging process as a subspace or a manifold. A drawback of those methods is that they require input images to be near-frontal and well-aligned. These methods therefore present experimental results only on constrained data-sets of near-frontal images. Again, as a consequence, such methods are ill-suited for unconstrained images. Different from those described above are methods that use local features for representing face images. In Gaussian Mixture Models were used to represent the distribution of facial patches. In GMM were used again for representing the distribution of local facial measurements, but robust descriptors were used instead of pixel patches. Finally,

instead of GMM, Hidden-Markov Model, super-vectors were used in  for representing face patch distributions.

An alternative to the local image intensity patches are robust image descriptors: Gabor image descriptors were used along with a Fuzzy-LDA classifier which considers a face image as belonging to more than one age class. In  a combination of Biologically-Inspired Features (BIF)  and various manifold-learning methods were used for age estimation. Gabor  and local binary patterns (LBP)  features were used in along with a hierarchical age classifier composed of Support Vector Machines (SVM) to classify the input image to an age-class followed by a support vector regression to estimate a precise age.

Finally, proposed improved versions of relevant component analysis and locally preserving projections. Those methods are used for distance learning and dimensionality reduction, respectively, with Active Appearance Models as an image feature.All of these methods have proven effective on small and/or constrained benchmarks for age estimation. To our knowledge, the best performing methods were demonstrated on the Group Photos benchmark. In state-of-the-art performance on this benchmark was presented by employing LBP descriptor variations  and a dropout-SVM classifier. We show our proposed method to outperform the results they report on the more challenging Adience benchmark, designed for the same task.






**Gender classification.**

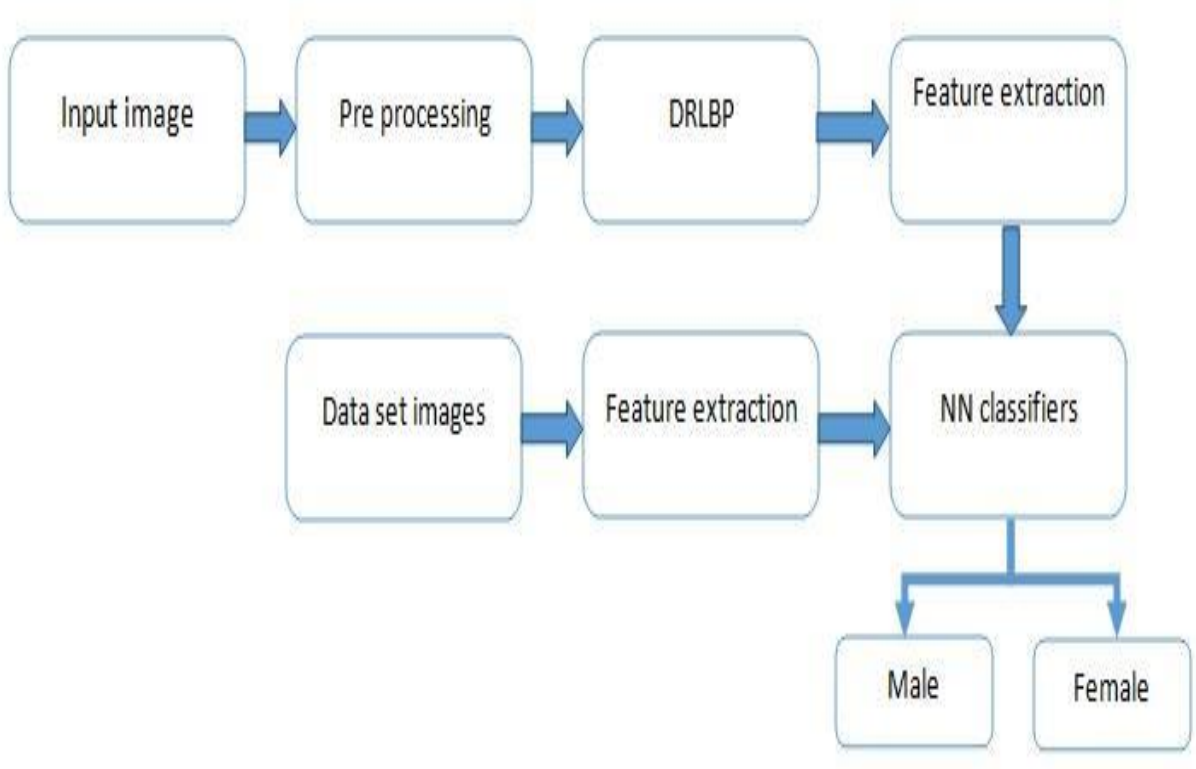
A detailed survey of gender classification methods can be found and more recently. Here we quickly survey relevant methods. One of the early methods for gender classification used a neural network trained on a small set of near-frontal face images. In the combined 3D structure of the head (obtained using a laser scanner) and image intensities were used for classifying gender. SVM classifiers were used by, applied directly to image intensities. Rather than using SVM, used AdaBoost for the same purpose, here again, applied to image intensities. Finally, viewpoint-invariant age and gender classification was presented. More

recently,used the Webers Local texture Descriptor for gender recognition, demonstrating near perfect performance on the FERET benchmark. In, intensity, shape and texture features were used with mutual information, again obtaining near-perfect results on the FERET benchmark.

Most of the methods discussed above used the FERET benchmark both to develop the proposed systems and to evaluate performances. FERET images were taken under highly controlled condition and are therefore much less challenging than in-the-wild face images. Moreover, the

results obtained on this benchmark suggest that it is saturated and not challenging for modern methods. It is therefore difficult to estimate the actual relative benefit of these techniques. As a consequence experimented on the popular Labeled Faces in the Wild (LFW) benchmark, primarily used for face recognition. Their method is a combination of LBP features with an AdaBoost classifier.

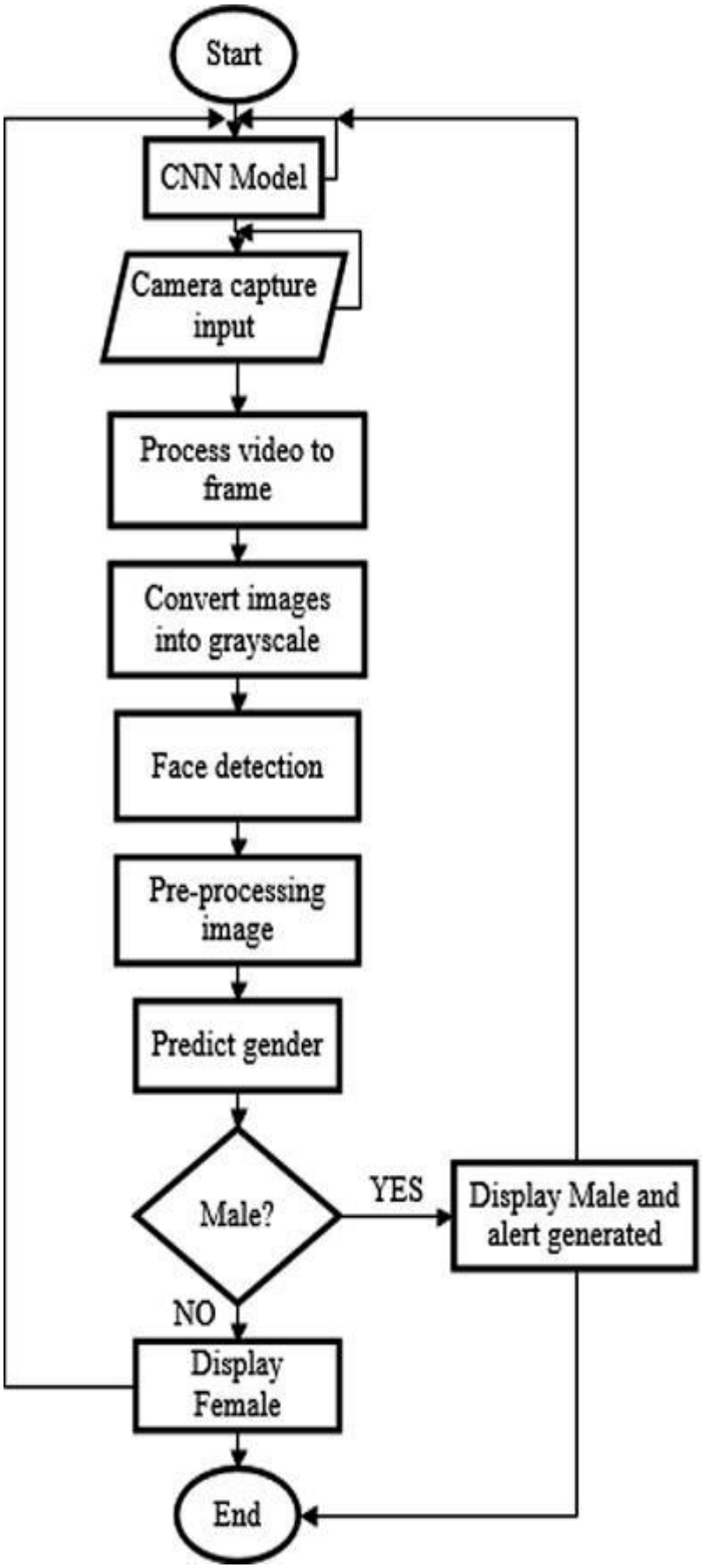
As with age estimation, here too, we focus on the Adience set which contains images more challenging than those provided by LFW, reporting performance using a more robust system, designed to better exploit information from massive example training sets.




**A CNN for age and gender estimation**

Gathering a large, labeled image training set for age and gender estimation from social image repositories requires either access to personal information on the subjects appearing in the images (their birth date and gender), which is often private, or is tedious and time-consuming

to manually label. Data-sets for age and gender estimation from real-world social images are therefore relatively limited in size and presently no match in size with the much larger image classification data-sets (e.g. the Imagenet dataset. Overfitting is common problem when machine learning based methods are used on such small image collections. This problem is exacerbated when considering deep convolutional neural networks due to their huge numbers of model parameters. Care must therefore be taken in order to avoid overfitting under such circumstances.


Start
CNN Model
Camera capture input
Process video to frame
Convert images into grayscale
Face detection
Pre-processing image
Predict gender
Male?
YES
Display Male and alert generated
NO
Display Female
End

### 3.1. Network architecture

Our proposed network architecture is used throughout our experiments for both age and gender classification. It is illustrated in Figure 2. A more detailed, schematic diagram of the entire network design is additionally provided in. The network comprises of only three convolutional layers and two fully-connected layers with a small number of This, by comparison to the much larger architectures applied. Our choice of a smaller network design is motivated both from our desire to reduce the risk of  as well as the nature of the problems we are attempting to solve: age classification on the Adience set requires distinguishing

Between eight classes; gender only two. This, compared to, e.g., the ten thousand identity classes used to train the network used for face recognition.

All three color channels are processed directly by the network. Images are first rescaled to 256 × 256 and a crop of 227 × 227 is fed to the network. The three subsequent convolutional layers are then defined as follows.  1. 96 filters of size 3×7×7 pixels are applied to the input in the first convolutional layer, followed by a rectified linear operator (ReLU), a max pooling layer taking the maximal value of 3 × 3 regions with two-pixel strides and a local response normalization layer [28].

2. The 96 × 28 × 28 output of the previous layer is then processed by the second convolutional layer, containing 256 filters of size 96 × 5 × 5 pixels. Again, this is followed by ReLU, a max pooling layer and a local response normalization layer with the same hyper parameters as before.

3. Finally, the third and last convolutional layer operates on the 256 × 14 × 14 blob by applying a set of 384 filters of size 256 × 3 × 3 pixels, followed by ReLU and a max pooling layer.

The following fully connected layers are then defined by:

4. A first fully connected layer that receives the output of the third convolutional layer and contains 512 neurons, followed by a ReLU and a dropout layer.

5. A second fully connected layer that receives the 512-dimensional output of the first fully connected layer and again contains 512 neurons, followed by a ReLU and a dropout layer.

6. A third, fully connected layer which maps to the final classes for age or gender. Finally, the output of the last fully connected layer is fed to a soft-max layer that assigns a probability for each class. The prediction itself is made by taking the class with the maximal probability for the given test image.

### 3.2. Testing and training

**Initialization**.

The weights in all layers are initialized with random values from a zero mean Gaussian with standard deviation of 0.01. To stress this, we do not use pre-trained models for initializing the network; the network is trained, from scratch, without using any data outside of the imagesand the labels available by the benchmark. This, again, should be compared with CNN

implementations used for face recognition, where hundreds of thousands of images are used for training. Target values for training are represented as sparse, binary vectors corresponding to the ground truth classes. For each training image, the target, label vector is in the length of the number of classes (two for gender, eight for the eight age classes of the age classification task), containing 1 in the index of the ground truth and 0 elsewhere.

**Network training**.

Aside from our use of a lean network architecture, we apply two additional methods to further limit the risk of overfitting. First we apply dropout learning (i.e. randomly setting the output value of network neurons to zero). The network includes two dropout layers with a dropout ratio of 0.5 (50% chance of setting a neuron's output value to zero). Second, we use data augmentation by taking a random crop of 227 × 227 pixels from the 256 × 256 input image and randomly mirror it in each forward-backward training pass. This, similarly to the multiple crop and mirror variations used.

Training itself is performed using stochastic gradient decent with image batch size of fifty images. The initial learning rate is e−3, reduced to e−4 after 10K iterations.

**Prediction**.

We experimented with two methods of using the network in order to produce age and gender predictions for novel faces:

• Center Crop: Feeding the network with the face image, cropped to 227 × 227 around the face center.

• Over-sampling: We extract five 227 × 227 pixel crop regions, four from the corners of the 256 × 256 face image, and an additional crop region from the center of the face. The network is presented with all five images, along with their horizontal reflections. Its final prediction is taken to be the average prediction value across all these variations.

We have found that small misalignments in the Adience images, caused by the many challenges of these images (occlusions, motion blur, etc.) can have a noticeable impact on the quality of our results. This second, over-sampling method, is designed to compensate for these small misalignments, bypassing the need for improving alignment quality, but rather directly feeding the network with multiple translated versions of the same face.

**Technical Details:**

**Local Response Normalization (LRN).**:

After the primary 2 pooling layers, there are local response normalization (LRN) layers. LRN could be a technique that was first introduced in as the way to assist the generalization of deep CNNs. The idea behind it's to introduce lateral inhibition between the various filters in a very given convolution by making them "compete" for big activations over a given segment of their input. This effectively prevents repeated recording of the identical information in slightly different forms between various kernels watching the identical input area and instead encourages fewer, more prominent, activations in some for a given area. If a(x,y) is that the

activation of a neuron by applying kernel i at position (x, y), then it's local response normalized activation b(x,y) is given by

$$b^i_{x,y} = a^i_{x,y} / \left( k + \alpha \sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} (a^j_{x,y})^2 \right)^{\beta}$$

where k,n,α, and β are all hyper-parameters. The parameter n is that the number of "adjacent" kernel maps (filters) over which the LRN is run, and N is that the total number of kernels therein given layer.

**Softmax**:

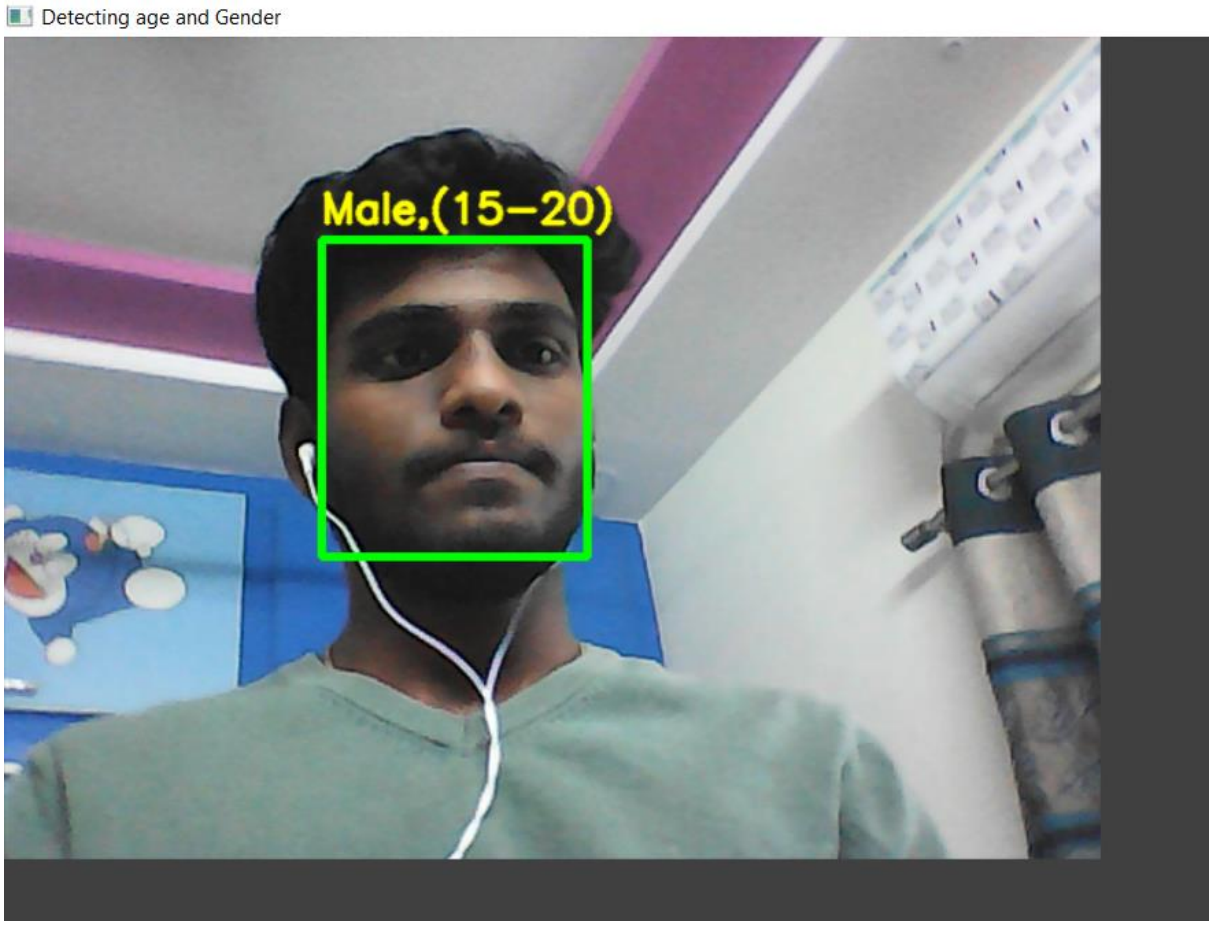
At the highest of the proposed architecture lies a softmax layer, which computes the loss term that's optimized during training and also the category probabilities during a classification. While some loss layers like multiclass SVM loss treat the output of the ultimate fully connected layer because the class scores, softmax (also called multinomial logistic regression) treats these scores because the unnormalized log probabilities of the classes. That is, if we've got zi is the score assigned to class i after the ultimate fully connected layer, then the softmax function is.

$$f_j(z) = \frac{e^{z_j}}{\sum_j e^{z_j}}$$

$$L_i = -log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

Because we would like to maximise the log likelihood of the proper class, the term we would like to reduce is that the negative log likelihood.

**Result:**


Detecting age and Gender
Male,(15–20)


**Conclusion:**

It has been observed and realized that the nature, behaviour and social interaction of people is greatly dominated by his/ her gender. Therefore an efficient gender recognition and classification system would play a pivotal role in enhancing the interaction between human and the machine .Moreover, there are several other applications where gender recognition plays a crucial role which includes biometric authentication, high technology surveillance and security systems, image retrieval, and passive demographical data collections. Identification of the gender and its classification based on the distinguishable characteristics between male and female facial image can be achieved easily by the human eye. However, machines cannot visualize this difference, hence the same task becomes difficult for the computer to

accomplish. Machines need meaningful data to perform gender classification. These data are usually the facial features based on which a computer classifies a facial image into either male or female .Gender Classification is a binary Classification problem. There exist several algorithms which have been already implemented to generate a solution to the stated problem. This study addresses the issue of gender classification and age detection of the identified gender using Support Vector Machine Classifier. Although the stated methodologies have been implemented on facial image data set and results are obtained with a level of accuracy, yet there are areas which are yet to be cultivated and where further enhancement can be achieved.

**References:**

1. T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. Trans. Pattern Anal. Mach. Intell., 28(12):2037–2041, 2006.

2. S. Baluja and H. A. Rowley. Boosting sex identification performance. Int. J. Comput. Vision, 71(1):111–119, 2007.

3. A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In Int.Conf. Mach. Learning, volume 3, pages 11–18, 2003.

4. W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recognition, 46(3):628–641, 2013.

5. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531, 2014

6 J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X.Chen, and W. Gao. Wld: A robust local image descriptor. Trans. Pattern Anal. Mach. Intell., 32(9):1705–1720, 2010

7 S. E. Choi, Y. J. Lee, S. J. Lee, K. R. Park, and J. Kim. Age estimation using a hierarchical classifier based on global and local facial features. Pattern Recognition, 44(6):1262–1281, 2011.

8 T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In European Conf. Comput. Vision, pages 484–498. Springer, 1998.

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

10 E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. Trans. on Inform. Forensics and Security, 9(12), 2014. 1, 2, 5,

11 Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. Trans. Pattern Anal. Mach. Intell., 32(11):1955–1976, 2010.

12 Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. Int. Conf. Multimedia, 10(4):578–584, 2008.

13 K. Fukunaga. Introduction to statistical pattern recognition. Academic press, 1991.

14 A. C. Gallagher and T. Chen. Understanding images of groups of people. In Proc. Conf. Comput. Vision Pattern Recognition, pages 256–263. IEEE, 2009.

15 F. Gao and H. Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. In Advances in biometrics, pages 132–141. Springer, 2009.

**GITHUB LINK:** https://github.com/ajaypopuri/Age-And-Gender-Detection