Detecting Demographics of People in Images Using Computer Vision

Authors Aarushi Dua, Ganesh Arkanath, Sai Teja Burla, Rasika Muralidharan.
https://github.iu.edu/napoom/demographics-detection

A. Abstract

We propose a single-shot custom-made model that analyzes the demographic features of a human face in an image. Specifically, we classify an image based on gender, age, and ethnicity. We used two types of facial features - geodesic distances and Local Binary Pattern (LBP) which are fed into a deep neural network in which we include two skip block connections to increase the computational speed and for a better spatial understanding of the model. In our model with LBP and geodesic distances, we received an accuracy for gender as 83.03%, a mean absolute error (MAE) of 0.7735 for age, and an accuracy of 46.12% for race. We also tested our model without LBP as a feature and got 86.02%, 0.7111, and 55.8% as our final values for gender accuracy, age MAE and race accuracy respectively.

B. Introduction

In recent years, the analysis of human faces has emerged as a prominent and captivating field of research within the realm of computer vision. Understanding and extracting valuable insights from facial attributes, such as age, gender, race, and facial expressions, hold immense potential for numerous applications across various domains, including security, marketing, healthcare, and more. The ability to automatically analyze these demographics from facial images using computer vision techniques presents a compelling avenue for advancing the state-of-the-art in this field.

Motivated by the potential impact of facial analysis and its diverse range of applications, this research project aims to develop a single-shot algorithm that analyzes human faces in images to classify it according to gender, age, and ethnicity. Our objective is to leverage the power of image processing and pattern recognition to accurately and efficiently determine the demographic attributes of individuals depicted in facial images. The proposed algorithm will utilize advanced feature extraction techniques, such as Local Binary Patterns (LBP), to capture intricate facial textures and patterns indicative of age, gender, race, and geodesic distances that capture facial landmarks. These features will be subsequently fed into robust deep neural networks (DNN) classifiers to accurately predict the age, gender, and race of individuals in images. Through our project, we want to answer the following questions:

a) Does the facial textures extracted using LBP impact the classification of age and ethnicity?

b) How do the geometric features computed using the dlib library with deep learning, help in gender classification?

c) How effective is deep learning-based simultaneous classification of age, gender, and race using facial images?

C. Background and Related Work

C.1. Background

Demographic analysis for facial images is the process of using facial recognition technology to identify the demographic characteristics of an individual based on their facial features. This can include identifying age, gender, ethnicity, and other physical attributes. The process of demographic analysis typically involves using machine learning algorithms to analyze large datasets of facial images, and training the algorithms to recognize patterns and characteristics associated with different demographic groups. The algorithms can then be used to classify new images based on these learned patterns, and to provide estimates of the demographic characteristics of individuals in the images.

C.2. Related Work

Some of the earliest works in this area start from the 1970s with the Facial Action Coding System and Eigenface. While these are not works that use computer vision to analyze whether a face in an image is male or female or a specific ethnicity, these are early pieces of work that analyzed features of a face.

A well-known approach to analyzing features of a face for classification of age and gender is geodesic distance. Geodesic distance in faces refers to the distance between two points on a face along the surface of the face. This distance is measured along the shortest path on the face, rather than in a straight line through space. The geodesic distance is useful in facial recognition and analysis because it takes into account the curvature and topology of the face, which can affect how different features are perceived. The paper, “Gender and race Classification using geodesic distance measurement” [7] uses this method along with PCA and classifier like SVM and KNN to find age, gender, and ethnicity. Another known method of this kind of feature extraction is through Local Binary Patterns (LBP) as done in paper [5]. Other methods like Binarized Statistical Image Features (BSIF) is used in paper [3]. LBPs are widely used because of it’s simplicity, efficiency and robustness. Paper like [8] use variations of LBP to improve the model perfor-

mance.

Deep learning methods are common and popular for a problem statement like this. One such paper is [2] that uses a Deeply-Supervised Attention Network for the problem. Further, paper [6] implements a single-shot methodology. Paper [1] explores methods of combining multiple features in deep neural network architecture.

D. Dataset

A large-scale, varied, and balanced facial recognition dataset called FairFace was released in 2019 by University of Washington academics. The dataset was created to address some of the biases and restrictions of existing facial recognition datasets, which have come under fire for being unrepresentative of underrepresented groups and lacking in diversity.

Over 108,000 facial images of people from various ages, genders, and ethnic backgrounds can be found in the FairFace dataset. The photographs were gathered from publicly accessible sites like Flickr and Google photographs, and demographic characteristics like ethnicity, gender, and age were carefully annotated on each one.

The FairFace dataset's emphasis on capturing the variation in facial appearances among each demographic group is one of its distinctive characteristics. With 47% male photos and 53% female images, the FairFace dataset offers a good gender balance.

Numerous studies have examined the fairness and accuracy of facial recognition algorithms across various demographic groups using the FairFace dataset. One of the main conclusions from these studies is that current facial recognition algorithms frequently exhibit prejudice and underperform on underrepresented groups, such as women and people with darker skin tones.

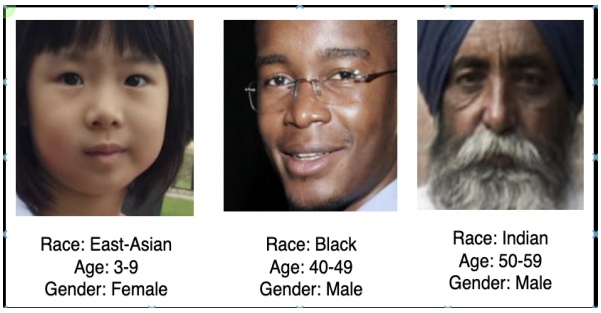

Figure 1. Sample Images from FairFace

E. Methodologies

E.1. Feature Extraction

To evaluate, categorize, and recognize the content of an image, it is necessary to find and pick out relevant visual patterns, structures, and qualities from the image. This process is called Feature Extraction.

In this project we are using feature extraction to find and mark important points on a human face to make use of them in our final model to predict the humans age, gender and race. For performing this, a variety of techniques and methodologies were employed and experimented with. They are as follows:

- SIFT (Scale Invariant Feature Transform)
- ORB (Orient Fast or Rotated Brief)
- FAST (Feature from Accelerated Segment Test)

Through our experimentation with the above 3 techniques, we noticed that even with hyperparameter tuning these methods was not giving us clear and useful points on a human face to use in our model. In searching for an alternative, we came across DLIB's inbuilt function called 'shape_predictor' which employs the use of a pre-trained model file 'shape_predictor_68_face_landmarks.dat' that contains pre-trained weights that helped us find landmark points on a human face.

The output we have at this point consists of points that are marked which recognizes the shape of the The output we have at this point consists of points that are marked which recognizes the shape of Eyebrows, Eyes, Nose, Mouth, Jaw in the face.

We use these points to find seven geodesic distances on the face.

Geodesic Distances

1. Eye Distance – The distance between the outer corners of both the eyes
2. Eyebrow Gap – The gap between both the eyebrows
3. Nose Height – The distance between the top point and the bottom point recognized as a part of the trail of the nose
4. Face Width – The distance between the extreme points on either side of the face
5. Eye Height – The distance between the top and bottom lid of the eye
6. Lip Width – The distance between both the corners of the lips
7. Jaw Angle – The angle that is made by the points recognized as a part of the Jaw

LBP

To capture the discriminative information related to age, gender, and race, we applied Local Binary Patterns (LBP)

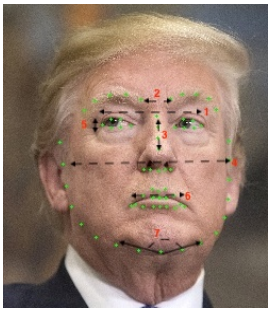

Figure 2. Geodesic distances between facial landmarks

as these attributes often manifest as variations in facial textures and patterns. LBP is a feature extraction method used in computer vision to locate textures in images. Texture features are necessary to detect changes in color and brightness across the face. Local Binary Patterns (LBP), which compares the intensity of surrounding pixels in an image, can be used to extract these properties. [5] We used skimage library to apply LBP and below is the parameter which we experimented with like number of points and radius.

Parameters: skimage.feature.local_binary_pattern(image, n_points, radius, METHOD)

The parameters n_points and radius determine the behavior of the Local Binary Patterns (LBP) algorithm.

n_points: This parameter specifies the number of sampling points to be used around each pixel.

radius: The radius parameter determines the radius of the circular neighborhood around each pixel.

By setting n_points = 8 and radius = 1 in our case, we instructed the LBP algorithm to compute the LBP code for each pixel by comparing the intensity value of the central pixel with its 8 neighboring pixels located at 1 pixel.

Figure 3. Facial texture extracted using LBP

F. Model Building

F.1. Experimentation Process

Through literature study, we found that ResNet50V2 and EfficientNetV2M architectures are useful for feature extraction in images. We used the pre-trained weights from the ImageNet dataset in addition to dense layers at the end of the network to fine-tune it to detect demographic data from faces. This did not yield satisfactory results, and hence we decided to create our own custom model from scratch. This yielded much better results and had lower execution time.

The goal with the custom model was to be able to pass image/LBP data as one of the inputs and run it through a convolutional block; whose output would be concatenated with the extracted geometric facial features (taken as the second input) after flattening. The network would then contain a block of dense layers before being branched to three simultaneous output layers. We started out by building a convolutional block containing 3 convolutional layers, where each convolutional layer was run with L2 regularization and followed by a 2x2 maxpooling layer. The output was then flattened and concatenated with the geometric facial features. This was then passed through a 5 dense layers before obtaining the output. Using this network, two primary observations were made - race classification was quite weak compared to age and gender; and the model was overfitting for gender.

By increasing the number of convolutional layers to 4 and increasing the number of dense layers to 8, we noticed better performance in the model. However, the model was learning extremely slowly. This was attributed to the vanishing gradient problem encountered amidst the array of dense layers that were present. To solve this, we made use of residual connections from [cite resnet paper here] and implemented residual connections between the output of the convolutional block and the output of the first set of 5 dense layers, combining the two before being passed to the second set of dense layers. Noticing an improvement in the learning capacity of the model, we implemented additional dense layers and another skip connection.

While implementing the skip connection, we were faced with two choices - implement a skip connection by taking the output of the global average pooling layer and reshaping it using a dense layer or, use a 1x1 convolution to obtain the required projection from the output of the last convolutional layer and pass it to a global average pooling layer before concatenating the geometric facial features. We ran experiments with both implementations and found out that the second approach produced better results.

The final model now contains 4 convolutional layers, followed by global average pooling in two dimensions to get a linear output, followed by concatenation of the geometric facial features. This is further connected to a dense block, and a skip connection passes the input data and adds it to the output of the first dense block. This is fed to a second dense block, whose output is again added with the first dense block's input using another skip connection. This output is passed through another 3 dense layers before obtain-

ing the output.

After some amount of hyperparameter tuning, we noticed that learnability of the model was still lacking. This is where we decided to remove L2 regularization from convolution and replaced it with a dropout layer having a rate of 0.15.

F.2. Custom Model

The model takes in two inputs, an image/LBP input in 2 dimensions, and another linear input having 7 geometric facial features of the image/LBP being used. The image/LBP is passed through 4 convolutional blocks for feature extraction. Each block performs a 3x3 Convolution, followed by a Dropout with rate 0.15 and 2x2 MaxPooling. Each convolutional block has a different number of filters, and the number of filters increases in sequential order to achieve various levels of abstraction.

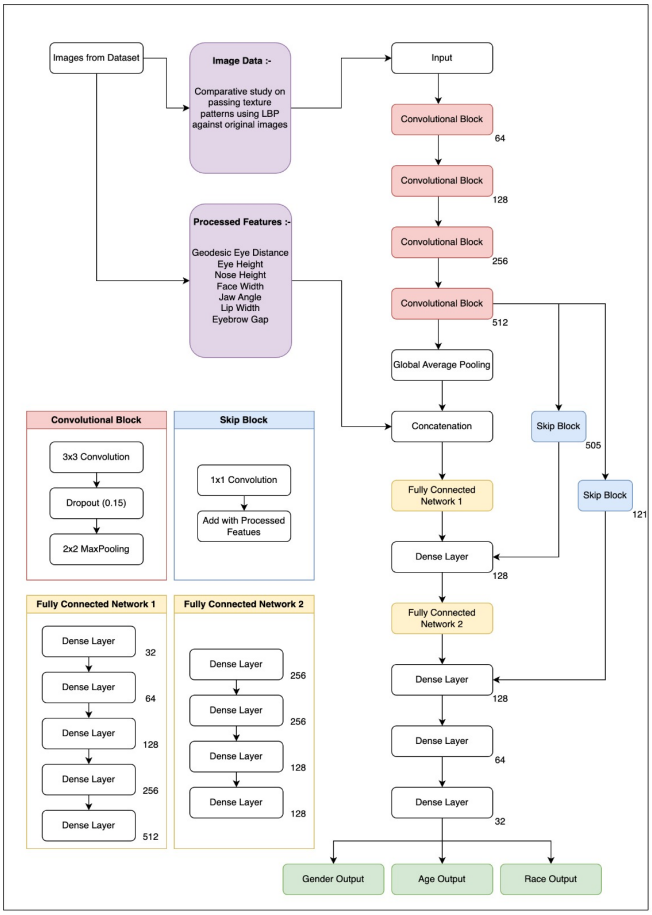

Figure 4. Model Architecture

The output of this is passed to a global average pooling layer. After global average pooling, the linear input of the geometric facial features are appended. This is passed as the input to the first dense block, which consists of 5 dense layers of varying width. The output of this dense layer is added with the input features using a skip connection.

The skip connection is implemented as skip block, where the first layer is a 1x1 convolution to project the output of the final convolutional block to the required number of filters, which is then passed to a global average pooling layer before concatenating with the geometric facial features.

This is passed as input to the second dense block, containing 4 dense layers of varying filter sizes. Its output is also added with input features using a second skip connection. The output of the above is finally passed through 3 dense layers before obtaining the output using three separate output layers, one each for gender, age, and race.

G. Results

Using our custom model, we ran an experiment comparing the usage of the original image as the input to the convolutional block and usage of texture patterns obtained from the image using LBP as the input to the convolutional block. Higher accuracy for gender and race, and lower MAE for age, was observed when passing the original image as an input to the convolutional block.

The Figure 5 shows the run where the convolutional block received the original image as the input. It can be clearly observed that the training and validation plots for each of the three measures, age, gender and race, are consistent. The difference between the training accuracy/MAE and the validation accuracy/MAE is not large for any of the measures, indicating that the model is generalizing well over the data. The validation curves for each of the measures have plateaued after a certain number of epochs but haven't become worse (as per their measurement criteria), suggesting that the training data is not overfit. A peak accuracy of 86.02% is achieved for gender, 56.16% for race, and best MAE for age at 6.8 years.

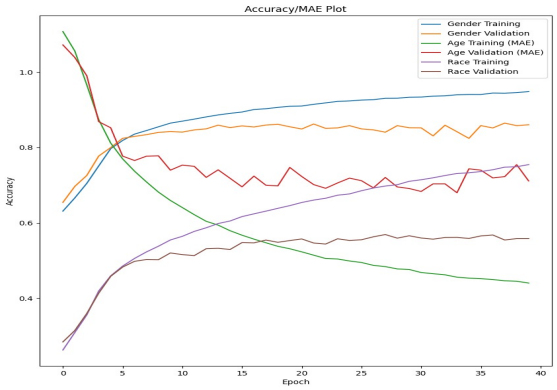

Figure 5. Model without LBP

The Figure 6 shows the run where the convolutional block received texture patterns of the image extracted using LBP as the input. It can be observed that there is a massive gap between the age and validation curves, suggesting that the model is being overfit and does not generalize well at all. We can also notice some major dips in the validation curves, suggesting overfitting on the training data. Peak performance of this model is capped at 82.87% for gender, 48.9% for race, and an MAE of 7.1 years on age. These results may not seem very bad at first, but noticing the overfitting nature of the graph, it can be deduced that this model is not practical for generalizing from the training data.

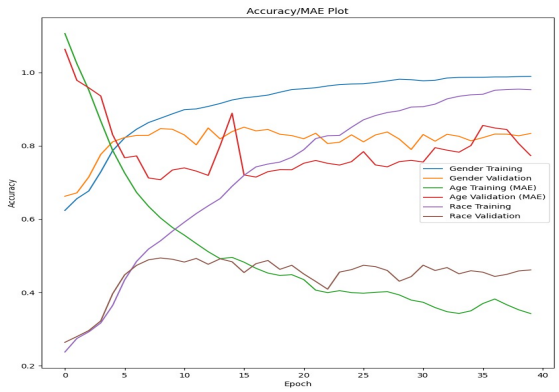

Figure 6. Model with LBP

A graphical comparison of these numbers indicating the performance of each model can be seen in Figure 7. The model taking the original image as input clearly outperforms the model taking LBP as input in all of the measures, i.e., higher race accuracy, higher gender accuracy, and lower age MAE.

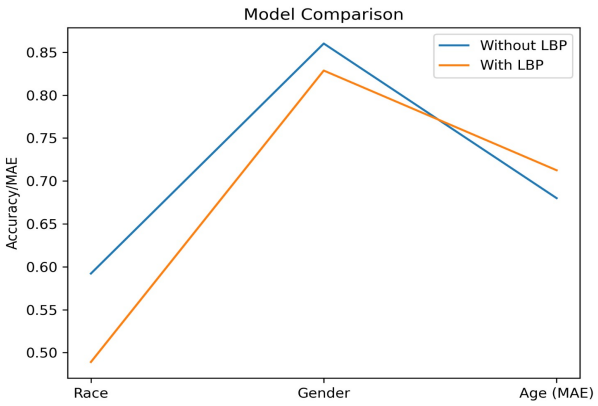

Figure 7. Model Comparison

H. Discussions

We found that our custom model gave us better than a pre-trained model such as ResNext50V2 or EfficientNetV2M. Through several experiments, we created a single-shot model with four convolutional layers with a dropout rate of 0.15, eight dense layers, and two skip connections. We tested this model with input feature combinations as follows:

- LBP and geodesic features
- Only Geodesic feature

The results of the experiments with the above mentioned models are shown in Table 1.

Table 1. Model Result Comparision

Model	Gender Acc	Age MSE	Race Acc
LBP + Geodesic	83.03	0.7735	46.12
Only Geodesic	86.08	0.711	55.8

As observed, across all evaluation metrics, we get better results with just geodesic features. We believe that LBP and geodesic features don't give us good enough reasons because the LBP features are not augmented to the facial image and hence the information is too sparse for the model to get any useful information.

The geodesic features with our deep neural network give us a strong accuracy for gender classification. However, race classification accuracy isn't good enough with just geodesic features either. We suspect this is because geodesic features do not have enough discriminatory power to allow to classify race accurately enough. We also suspect that our model needs more complexity to be able to classify race more accurately. Creating a model that has greater power to analyze local and spatial information might be useful.

The single-shot method seems to produce decent results but requires more work to get worthwhile results for race classification. This could be due to the fact that the model complexity isn't strong enough and that we do not have enough features for our model to classify all variables effectively enough.

H.1. Limitations

- We used the FairFace dataset for our project which included images of human faces from a frontal angle and also non-frontal angles. At this time, we were unable to use non-frontal angles as our feature extractors, we were unable to pick up features from it. This reduced how much data we could work with. From a real-time image processing perspective, it is essential to be able to detect images from a non-frontal angles.

- The FairFace dataset came with 7 categories of ethnicity. However, each image was classified into only one ethnicity. In a 21st century world, we must be mindful of nuances these and find datasets or create one that allows for multiple ethnicities for one face where true.

- Within the current scope of the project, we have implemented feature extraction with LBP and geodesic distances. We found that LBP wasn't as useful as we hoped and hence as part of future work we will find more features.

- Our images from the dataset consisted of one human per image. For real-world or real-time detection, it would be ideal to have multiple humans in an image to detect and classify.

H.2. Future Works

1. **Additional facial features**: Through the course of this project, we found a few facial feature extraction methods but have found only geodesic distances to be useful so far. We hope to find some other facial features to add to our model in order to get better spatial information and thereby increase the model's accuracy.

2. **Addressing Bias**: As we are working with human images and particularly gender classification and race classification, it is essential that we monitor bias and how we can mitigate as much it as possible. Through the timeline of this project, we haven't addressed this, but it is an important aspect to consider and hence will be a focal point for future work

3. **Explainability**: While deep learning gives good-generalized results, it happens to be a blackbox. As we are working if sensitive information, it is crucial for us to have a better understanding of why our model works in a certain way.

4. **Data Augmentation**: As we only considered frontal images, we want our model to be able to capture information of any angle.

5. **Real-Time application**: A model is only as good as its application. We will aim to improve the model to work on low-resolution images for frontal and non-frontal images and for multiple humans within an image. Furthermore, we'd like to make the model deployment onto lightweight devices as well for ease of usage.

I. Conclusion

Our model is able to classify an image based on three demographic attributes: age, gender, and race effectively. Our computational time with our final model was 45 mins (for 40 epochs). We saw that LBP did in fact have an effective on our model output though it was not positive. Dlib's geodesic features with deep learning proved to be useful and effective. Our single-shot attempt was excellent for gender classification, above average for MAE of age and poor for race detection.

References

[1] Zaheer Abbas. Joint demographic features extraction for gender, age and race classification based on cnn. *International Journal of Advanced Computer Science and Applications*, 10(12):460–467, 2019.

[2] Yingruo Fan. Facial expression recognition with deeply-supervised attention network. *IEEE Transactions on Affective Computing*, 13(2):1057–1071, 2022.

[3] Rishi Gupta. Identification of age, gender, race smt (scare, marks, tattoos) from unconstrained facial images using statistical techniques. *International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 2018.

[4] Kaiming He. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[5] Di Huang. Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews*, 41(6):765–781, 2011.

[6] Anh-Thu Mai. Real-time age-group and accurate age prediction with bagging and transfer learning. *International Conference on Decision Aid Sciences and Application (DASA)*, 13(2):27–32, 2021.

[7] Zahraa Shahad Marzoog. Gender and race classification using geodesic distance measurement. *Indonesian Journal of Electrical Engineering and Computer Science*, 27(2):820–831, 2022.

[8] Ahmed Abdulateef Mohammed. Robust single-label classification of facial attributes. *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 651–656, 2017.