Identifying Identities using the CelebA Dataset and Machine Learning Models

Dataset: CelebA

The CelebA dataset is a commonly used library of various celebrity faces, in a variety of poses, expressions and does include images with accessories. There are 10,177 individual identities, with an average of 20 images per identity. With the dataset’s size, we first needed to cut down the data to a useable amount for this assignment.

Initially we chose to limit the total image numbers to 4000. However, this allowed for some identities to have a minimal number of images. To resolve we swapped to retrieving every image for an identity and getting X number of identities to reach our 4000 number. Once again, this still provided us of the issue that some of these identities could have to little images. Finally, we returned all identities with 30 or more images (max in data is 35 images per identity), and then limited the number of identities to allow our 4000 images.

Our data in the end compromised of two arrays: image data and identities data.

The images data are jpg images, which when imported are imported as a data frames. We specify a resolution of 48 x 48, and each pixel has a depth of 3 for each RGB colour. These properties (i.e. each pixel and the colour depth) are our features available to us for our data set.   
  
These features will allow our model to meet our target of enabling the prediction the identity of a person from a still image.

Several descriptive statistics and visualisations have been used in our code base to facilitate in us understanding the data.

Data Shapes: We first wanted to understand the shape of the image data (X) and the identities data (y). We also needed to understand the total number of samples, features and classes.

A black text on a white background

Description automatically generated

Figure

Colour Distributions: We also wanted to understand the HSV colour distributions for our images. We did this by charting out the Hue, Saturation and Value values into histograms. This allows us to understand the makeup of the dataset to avoid bias.

1. Hue (H) Histogram:
   * Hue represents the dominant colour in an image, often expressed as an angle in degrees on a colour wheel.
   * In our dataset can be seen to contain two strong peaks towards hue value 0 and 15.
2. Saturation (S) Histogram:
   * Saturation represents the intensity or vividness of colours in an image.
   * The Saturation histogram shows the distribution of colour vividness.
   * Due to the shape of this data with the high peaks being located at either end, saturation appears to be quite low or high, however high is less common.
3. Value (V) Histogram:
   * Value represents the brightness or lightness of colours in an image.
   * The Value histogram shows the distribution of brightness levels.
   * Peaks and valleys in this histogram can indicate variations in image brightness. For example, high peaks might correspond to well-lit images, while valleys could indicate darker or shadowy areas.
   * AS seen in the graph below, the peak is towards the end of the chart (near value 250) therefore the images tend to have high brightness and clarity.

A green line graph with numbers

Description automatically generated

Figure

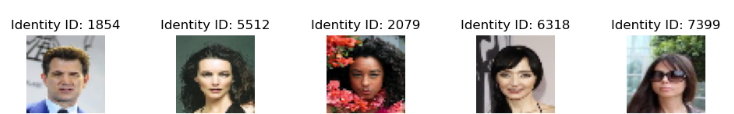
Sample Information: To verify our filtered data set contained data which would allow quality results the sample distributions were shown. This confirmed our data only included identities which had 30 or greater images and were randomly selected.

A bar code with text

Description automatically generated

Figure

Along side this information, we also displayed the first 20 images in our data set, to visualize what the images for understandability to the reader.



Figure

To ensure the data was viable to produce consistent results, a number of steps have been taken before we can train or evaluate any models.

Images were resized to a 48 x 48 dimension. This ensures all images have same dimensions which would be required for Neural Network models.

Colour Normalisation: The image values were divided to normalise pixel values in the range of [0,1]. This was performed for stabilising the training and performance.

Data Flattening – Some of the spot check models required the use of one-dimensional data. Therefore, data flattening was performed. Our data structure for these models transformed from 48x48x3 into 6912 attribute vector. This flattened data was not used in all models, only those which required it such as the LR, LDA and KNN).

Our image data can therefore be summarized as a collection of 48 x 48 images with 3 colour channels covering 3907 samples across 130 categories. The identities in our dataset are as highly populated (in terms of images per identity) as possible from the CelebA dataset. The images contain mainly colours at the white end of the colour spectrum but are not overly saturated and appear bright.

A model spot check which was acquired from the DataVzn\_walkthrough was used in our report to identify the accuracy of several possible models. This spot check has been run through multiple times during our testing. A variety of results have arisen when we differ the inputs. Two of the results of these can be seen below. The first image is that of the image structure described throughout this report. The second image shows a smaller data set and image size for comparison.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Figure - 48x48x3 - 3901 Samples (30 images +) Figure - 64x64x3 - 548 Samples (31+ Images)

From these results, we have chosen to focus attention on Logistic Regression, Linear Discriminant Analysis and Convolutional Neural Network model types. These three models have achieved the best results in multiple run-throughs, with differing inputs. The inputs we have changed include:

* Image Size
* Sample Size (Identities and therefore images)
* Sample Values (identities are chosen at random, each run-through)

The second run-through was run against a much smaller total data set with only 548 samples across 17 identities, however it was limited to the identities with more than 30 samples. This has seen a significant increase in accuracy levels for all models (other than CNN). A recommendation for future work will be to generate additional images for the training data for all identities with small sample sizes.