Module: Computer Vision
Candidate Number: 283127
Name: Afnan Ahmed

Locating Facial Landmark through CNN model and HOG Feature Extractor

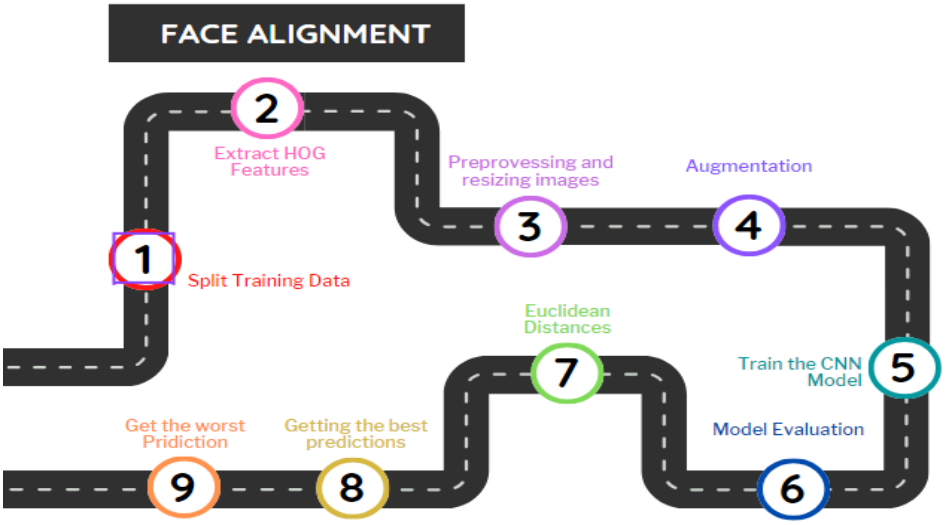
INTRODUCTION

The primary objective of this assignment is to design, build, test and critique a system capable of performing a face alignment task. Which involves locating the facial landmark around the faces, eyes, nose and lips and chin. To implement an accurate and robust facial recognition system I came up with the idea of using a feature descriptor called 'Histogram of oriented gradients' (HOG) to get the image features and use those features to train an image through convolutional neural network (CNN). CNN is a very powerful tool used in image recognition and processing due to its ability to recognize pattern in images. The secondary task of this assignment is to do some modification around the lips and the eyes. We use a pre-trained model and various technique like HOG and CNN to get the best landmark predication accuracy.

We will first create our face alignment model and go through each and every step for face alignment like Preprocessing, Feature Extraction and then Prediction Model. We will be doing Quantitively and Qualitative analysis and put a discuss the failure cases and limitation. Then on second part we will me implementing lip/Eye modification system.

METHODOLOGY

Overview of Face Alignment Model:



A. Preprocessing

At first stage data the training data (images and points) are slip into training and validation sets using a **train_test_split** from scikit-learn. This is common practice to evaluate model on unseen data during training. As we know that images are too much and it will take lot of time to process so we resize the images ensuring uniform input dimension while normalization help with model training efficiency. The images are resized to 128x128 pixels.

B. **HOG Feature Extraction**

HOG features are a best way to describe structure of an image based on distribution of edge directions. The images are divided into small cells and compiling a histogram of gradient direction. In the context of face detection, HOG features provide a robust way to detect facial features based on the structural edges and the efficiency and effectiveness of HOG make it suitable for real-time face detection applications. [1].

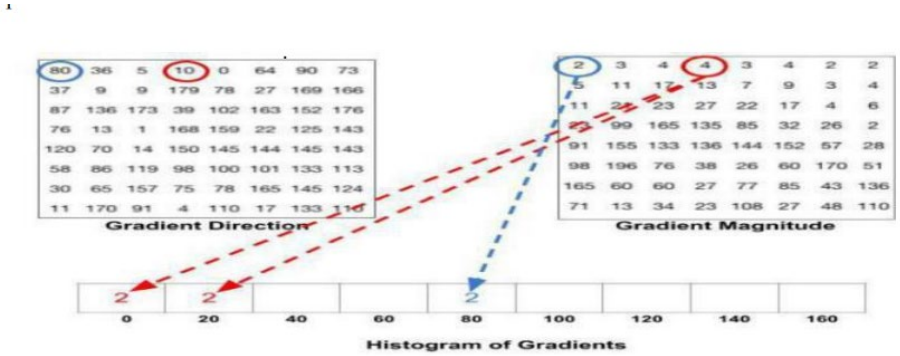


Figure 1. Bin Selection

I will be showing the HOG processed images to help in understanding how the feature features represent the underlying the content of images in Figure 2.

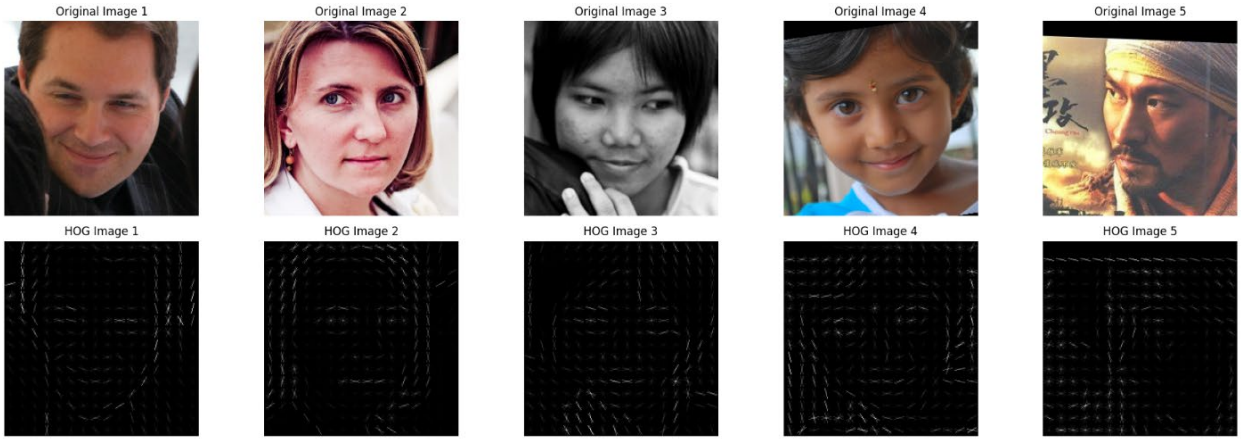


Figure 2. Original image and their HOG version

C. **Data Augmentation**

This technique is widely used in image processing to prepare a robust dataset for training. The earliest demonstrations showing the effectiveness of Data Augmentations come from simple transformations such as horizontal flipping, color space augmentations, and random cropping [2]. We have use one of the geometric transformations: Horizontal Flipping. The keypoints are also transform according to image transformation and finally they are normalized which is very important step to maintain consistency and stability in datasets. Figure 3. Below shows the augmented images.

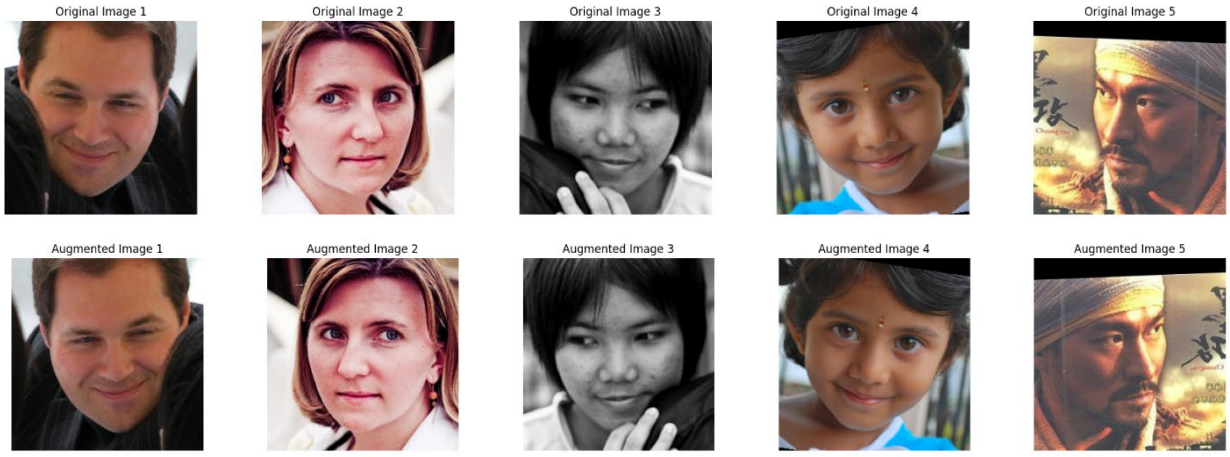


Figure 3. Original image and its corresponding augmented images

D. **Creating a CNN Model**

We have built a CNN model with layers specifically suited for image data. This includes convolutional layers ('Conv2D') that extract image feature through filters, and pooling layers ('MaxPooling2D') that reduces the dimensionality when retaining features from the image. We have created two model architecture one is **Image-Based CNN** and another is **HOG-Based CNN**. The CNN model is we design is look much similar to image in Figure 4.

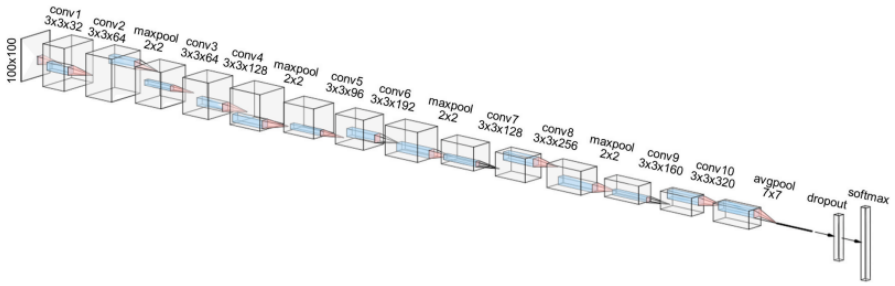


Figure 4. CNN architecture

The loss function use here is Mean Squared Error (MSE) which measure the average of squares of differences between predicted and ground values. We also use ImageDataGenerator to augment image data by applying various transformation. This aim to prevent overfitting and allow model to generalize better on unseen image data. Due to heavy usage of RAM and limitations in Colab we have set the number of epochs to 20 and uses a smaller batch to better generalize. During training we find the validation loss and validation mean absolute error for both simple and hog-based CNN.

```
Validation Loss: 0.0018657727632671595
Validation MAE: 0.031033378094434738

Validation Loss with HOG: 0.0019003474153578281
Validation MAE with HOG: 0.03155200555920601
```

E. **Calculating Euclidean Distances**

We have computed the Euclidean distances between the pairs of predicted and ground truth points which we will be using in quantifying the accuracy of predictions. We have plotted an image with landmarks and I will be sharing the best and the worst prediction of CNN model with their corresponding Euclidean distances.

```
Simple CNN: [35.81390172 34.20947651 34.14782853 ... 11.91326568 11.20401849
  9.16396764]
18/18 [==============================] - 0s 2ms/step
Euclidean Distances for CNN with HOG: [33.96545538 33.00030982 33.2844236  24.25973238
13.23150709]
```

Qualitative Analysis

Figure 5. Show the best and the work prediction on validation sets

We also have prediction on test images as well shown below Figure 6.

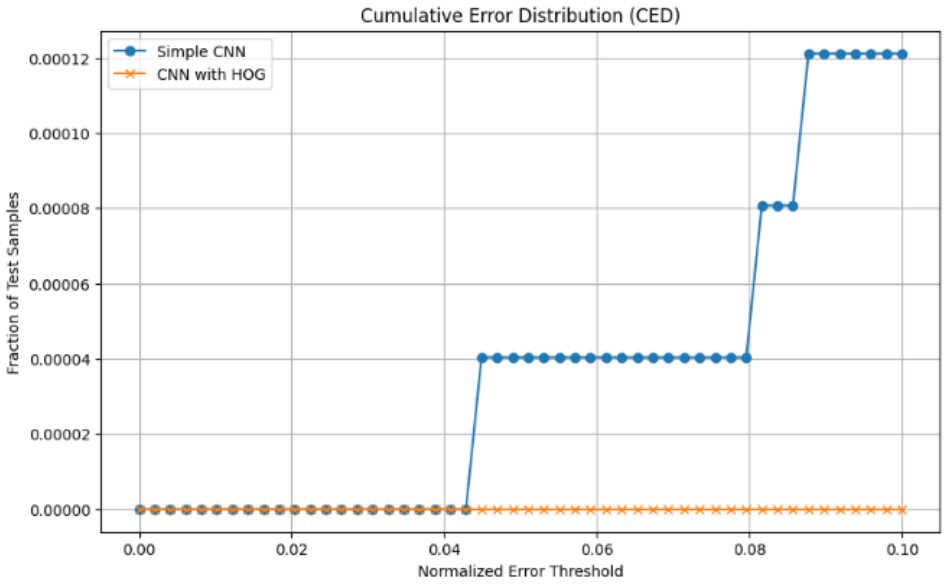
Figure 6. Test_Images Outcome and it's predictions

EXPERIMENTS AND RESULTS/FINDINGS

A. **Quantitative Results**

The first result we find is Cumulative Error Distribution which help in visually comparing the accuracy and reliability of different model. It also provides insight on images which meets the degree of precision requirement.

Figure 7. CED for Simple CNN and CNN_HOG



The graph shows simple CNN perform better than CNN with HOG. The simple CNN achieves lower normalized errors for a higher fraction of test samples, indicating better accuracy and robustness in

predicting facial landmarks. The flat line for the CNN with HOG suggests that this model struggles to make accurate predictions in this context

Another analysis is the performance of simple CNN and CNN with HOG over multiple training epochs, using the average Euclidean distances as the performance metrics. We can see in graph CNN with HOG have slower improvement rate whereas simple CNN improvement suggesting more effective at learning task at hand. But overall distance in CNN_HOG is lesser. See figure 8.

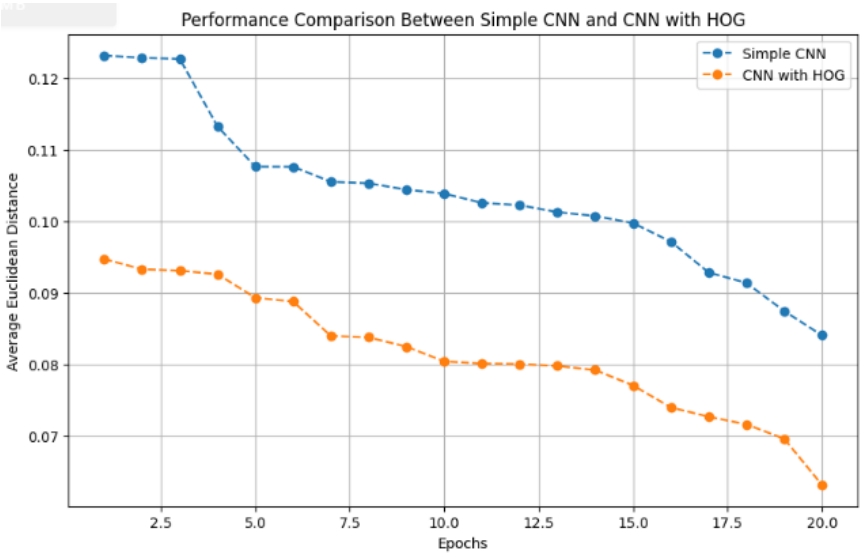
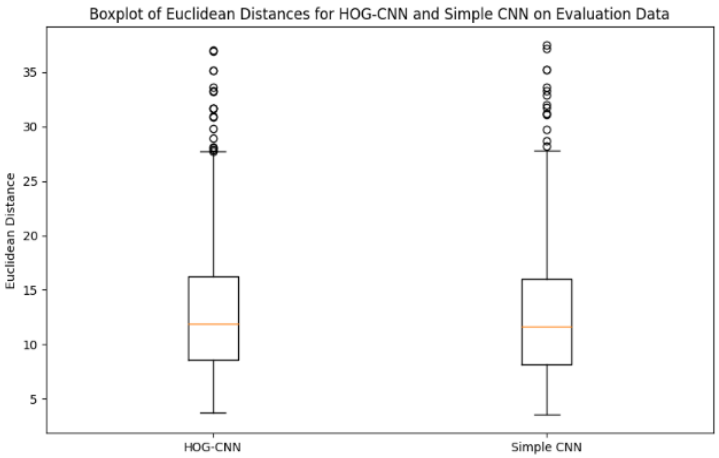


Figure 8. Performance comparison between Simple CNN and CNN with HOG

The below plots below show the comparing two model shows both are consistent in performance



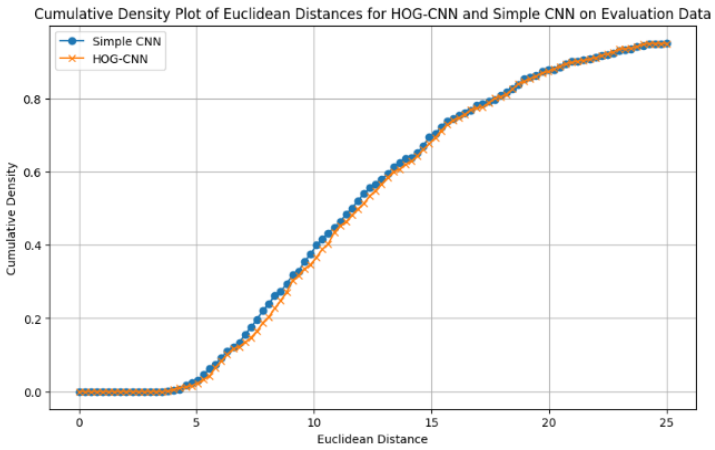


Figure 9. Comparing distances between predicted and truth point and CDF of Euclidean distances

B. <u>Detailed Analysis of Failure Cases</u>

Success: In majority of cases both simple CNN and HOG-CNN model perform reasonably well due the image being clear and unobstructed view of face. Best case is where distance is low as shown in Figure 5.

Failure: High Euclidean Distances, Obstruction, pose variations, poor lighting and unusual facial expressions.

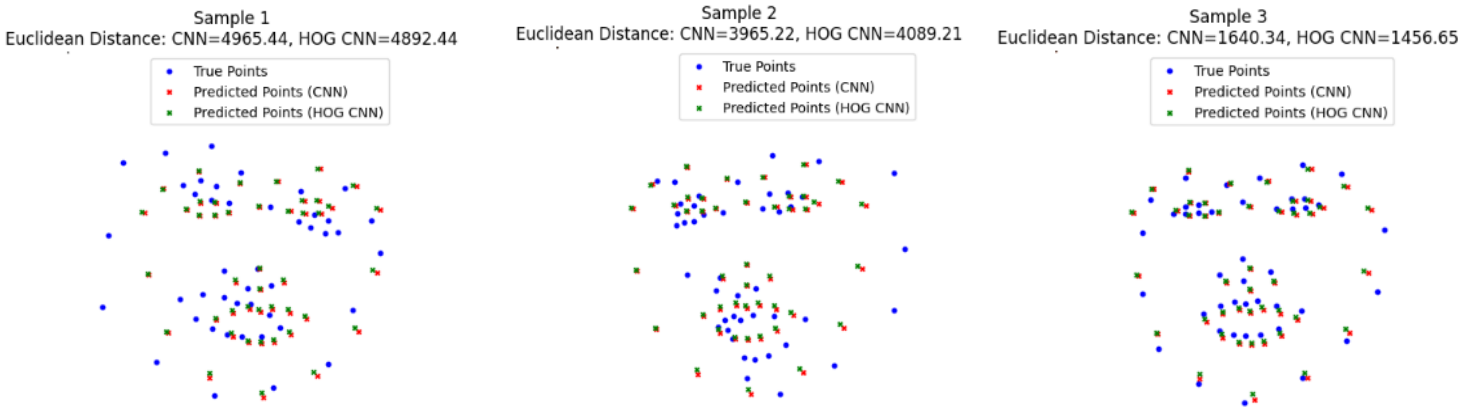


Figure 10. High Euclidean Distances in some bad images

Solutions: To deal with those type of images we have to implement advance preprocessing technique that normalize the lighting and reduces shadows in image. Increase number of epochs in training model and get more complex CNN model. Apply post-processing like spatial constraints to refine predictions.

Lips / eyes Color Modification

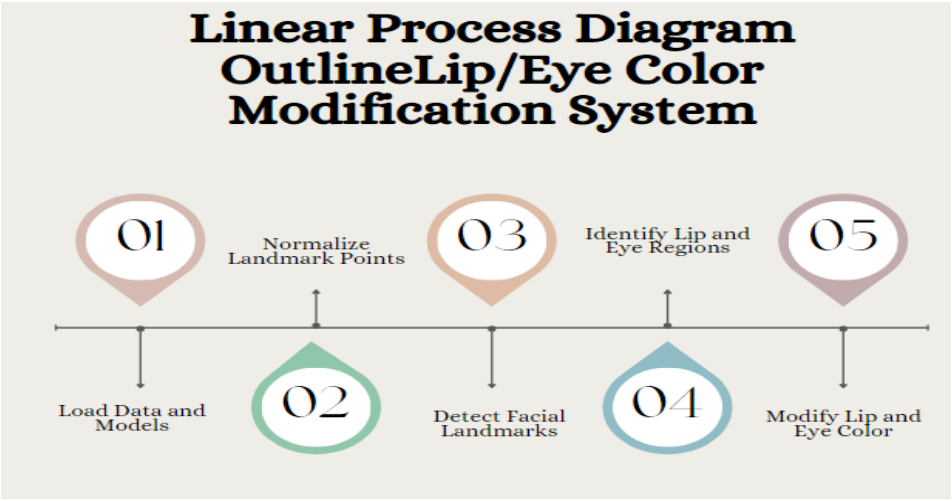


Fig 11. Flowchart for facial landmarks coloring

The first step is loading our pre-trained CNN model. This will also load its images and landmarks on that image. We normalize the points using **'MinMaxScaler'** to resized image scale. Extracted the HOG feature

form image and then predict the points and rescale image back to original. Based on detected landmarks the right eye, left eye, outer lips and inner lips points are found and the mask is created using identified regions. And then we use blending technique to combine original image with colored region back. Euclidean distances are calculated and image are displayed as shown below in figure 11.

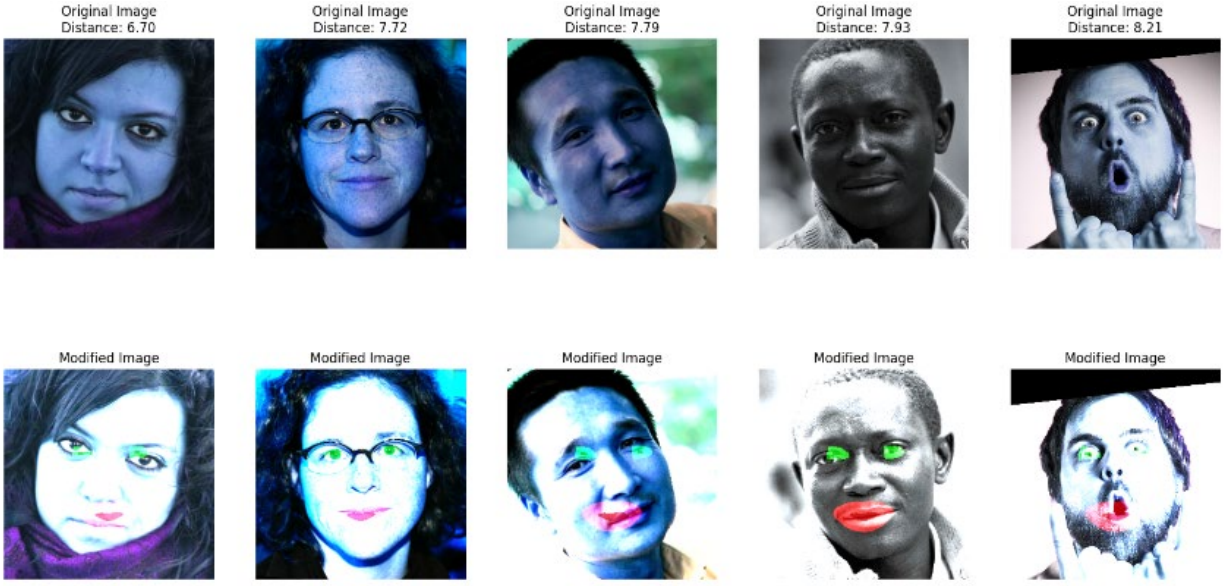


Figure 12. Best modifications with their corresponding Euclidean distance

I found that there are some inaccuracies in landmark detection that's why modification on RHS is misplaced. Color blending was difficult with different colors of skin tones. We need to more robust method to detect facial landmarks so it performs well under any kind of pose or lighting. Adaptive color selection technique so modification appears natural. Analyzing video frames could help in selecting most suitable frames for modification where facial features are visible clearly.

REFERENCES:

[1] 'Face recognition using hog feature extraction and SVM classifier' (2020) *International Journal of Emerging Trends in Engineering Research*, 8(9), pp. 6437–6440. doi:10.30534/ijeter/2020/244892020.

[2] Shorten, C. and Khoshgoftaar, T.M. (2019) 'A survey on Image Data Augmentation for Deep Learning', *Journal of big data*, 6(1), pp. 1–48. doi:10.1186/s40537-019-0197-0.

[3] Tavolara, T.E. *et al.* (2021) 'Identification of difficult to intubate patients from frontal face images using an ensemble of deep learning models', *Computers in biology and medicine*, 136, pp. 104737–104737. doi:10.1016/j.compbiomed.2021.104737.

[4] Loos, A. and Ernst, A. (2013) 'An automated chimpanzee identification system using face detection and recognition: Doc 95', *EURASIP journal on image and video processing*, 2013, pp. 1-. doi:10.1186/1687-5281-2013-49.