IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Baibhav Datta (230059246) - PG
- **Google Drive folder:** https://drive.google.com/drive/folders/1Tjw7s7wkT19Nq6cYsNxlRokIApLzR_5F?usp=sharing

Data

The dataset employed for the project encompasses images portraying human faces.Each image is accompanied by a label denoted as an integer from 0 to 2. Specifically, 0 signifies the absence of a mask, 1 denotes the correct wearing of a mask, and 2 indicates the improper wearing of a mask. These labels are stored in accompanying txt files. The dataset originally has the following statistics-
Train: 1915 Training Images (1549 with mask, 310 without mask, 56 improper mask)
Test:458 Test Images (388 with mask, 51 without mask, 19 improper mask)

Additionally, the dataset for Face Covering Detection in a video was recorded by me, ensuring instances for each case: no mask, mask worn correctly, and mask worn improperly.

Implemented methods

For the classification of human face images into three distinct categories—wearing a mask, not wearing a mask, and improperly wearing a mask—I employed three distinct models: Resnet-50 CNN, HOG + Support Vector Machine (SVM), and HOG + Multi-Layer Perceptron (MLP).

In selecting these models for this use case, several factors were considered. The Resnet-34 CNN is renowned for its effectiveness in image classification tasks, leveraging its deep architecture to extract intricate features from images. This model was chosen due to its ability to capture complex patterns in facial images, making it suitable for discerning between the different mask-wearing states.

On the other hand, HOG (Histogram of Oriented Gradients) combined with SVM and MLP offers a more traditional approach to image classification. HOG is adept at capturing shape and gradient information from images, which can be particularly useful in scenarios where deep learning models might not be feasible due to limited computational resources or dataset size. By combining HOG with SVM and MLP, we leverage the strengths of both feature extraction and classification algorithms, thereby enhancing the model's overall performance.

HOG+SVM & HOG+MLP

For the SVM and MLP models, data preparation involved partitioning the dataset into training and validation sets, extracting HOG features from the images. During the feature extraction phase, essential characteristics of the face images necessary for categorising them into the three classes are obtained. This stage holds significance as its efficiency enhances the recognition accuracy and minimizes misclassification. This study utilises a HOG-based feature extraction method. Each facial image undergoes conversion to grayscale and resizing to dimensions of 128 x 64 pixels before HOG feature extraction. The proposed HOG feature is derived by employing a cell size of 16 x 16 dimensions and a cell_per_block setting of 2 × 2 for the input facial image.

Before training the final SVM and MLP models, an iterative process was undertaken to identify the optimal hyper-parameters which involved testing different hyper-parameter values on the validation set in order to maximise model accuracy. The chosen hyper-parameters were then used to train the models, followed by evaluation using metrics such as accuracy, F1 score, recall, precision, and confusion matrix.

Resnet-50 CNN

For the Resnet-50 CNN model, the data preprocessing involved splitting the training set into training and validation subsets, followed by directory restructuring to facilitate data loading using the torch dataloader function. Data augmentation and normalization were performed to enhance model robustness and generalization.

I employed the pre-trained ResNet-50 model, which was trained on the ImageNet dataset with over 1 million images across 1000 classes. Utilising these pre-trained weights accelerates learning for our task. Upon loading the ResNet-50 model, I replaced its final layer with a new fully-connected Linear layer. Finally, training commenced on our training dataset.

A dedicated function was defined to train the pre-trained Resnet-34 model. The iterative training method spans over 15 epochs. The training dataset is utilized to train the model, while the validation dataset is employed for evaluation. Throughout each epoch, we compute the cross-entropy loss on both the training and validation sets. For optimization, we employ the Stochastic Gradient Descent (SGD) optimizer with a learning rate set at 0.001. I have also used StepLR as the learning rate scheduler, which reduces the learning rate by a factor of gamma every N epochs.

Functions to visualise predictions

Specific functions were created to assess the performance of various models on the test dataset. One such function, named MaskDetection, was developed to visualise the predictions generated by different models on the test data. This function facilitates the selection of four random images from the test set, displaying them alongside the corresponding predictions made by the models as well as the actual labels for comparison.

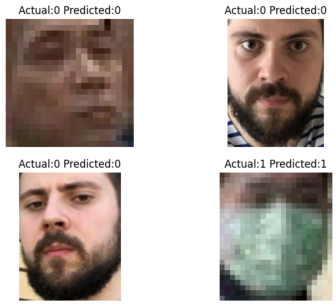
The MaskDetectionVideo function is designed to accept the filename of a video as input. It then loads the video, predicts the classes for each frame using our top-performing model, and saves 400 frames of the video annotated with the predictions.

Results

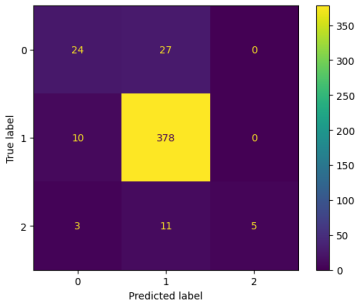
The Results section of the project report presents the outcomes of the implemented classification models. It includes examples of test images alongside their ground-truth labels and the corresponding predictions made by the models. Additionally, qualitative assessments such as accuracy metrics and confusion matrices are provided to evaluate the performance of the models. This section aims to offer a comprehensive overview of how well each model performs in classifying human face images into the specified categories of wearing masks, not wearing masks, and improperly wearing masks.

HOG+SVM

Accuracy on test data-89%

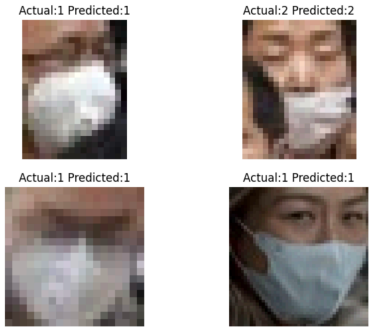


	precision	recall	f1-score	support
0	0.65	0.47	0.55	51
1	0.91	0.97	0.94	388
2	1.00	0.26	0.42	19
accuracy			0.89	458
macro avg	0.85	0.57	0.63	458
weighted avg	0.88	0.89	0.87	458

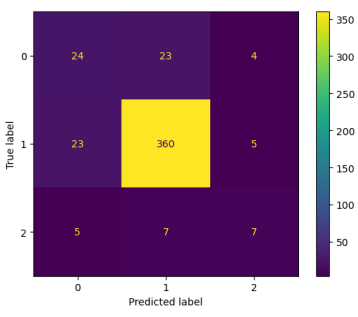


HOG+MLP

Accuracy on test data-85%



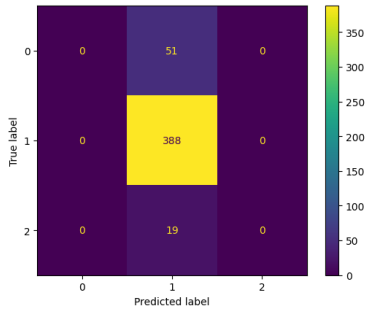
	precision	recall	f1-score	support
0	0.46	0.47	0.47	51
1	0.92	0.93	0.93	388
2	0.44	0.37	0.40	19
accuracy			0.85	458
macro avg	0.61	0.59	0.60	458
weighted avg	0.85	0.85	0.85	458



ResNet-50 CNN

Accuracy on test data-85%

	precision	recall	f1-score	support
0	0.00	0.00	0.00	51
1	0.85	1.00	0.92	388
2	0.00	0.00	0.00	19
accuracy			0.85	458
macro avg	0.28	0.33	0.31	458
weighted avg	0.72	0.85	0.78	458



Results Discussion

After analysing the evaluation metrics and confusion matrices of the three models, I have determined that the HOG+MLP model outperforms the others. Although the accuracy of predictions on the test set for the HOG+SVM model is the highest at 89%, the HOG+MLP and ResNet-50 models achieve an accuracy of 85%. Upon examining the confusion matrices, it became evident that all models struggle with predicting class-2 samples. However, the ResNet-50 model fails to predict class-2 correctly entirely, while the HOG+MLP model performs better in this aspect compared to the HOG+SVM model. Therefore, based on these observations, I conclude that the HOG+MLP model is the best performing among the three.

The factors that may have contributed to SVM and MLP models outperforming ResNet-50 in face mask detection are:

- *Model complexity: ResNet-50's complexity may lead to overfitting with a small dataset, while simpler models like SVM and MLP generalise better.*
- *Overfitting: ResNet-50's large parameter count can cause overfitting with small datasets, unlike SVM and MLP.*
- *Feature extraction: SVM and MLP effectively utilise features like HOG, more suitable for face mask detection than ResNet-50's learned features.*

To showcase the outcomes generated by employing the best-performing model, i.e. HOG+MLP on the real-world video footage, below are some screenshots from the video annotated with the predicted class labels.