IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Saundarya Sutar (S230044065) - PG

Data

The dataset consisted of training and testing sets with image files and corresponding labels. The training set has 2,394 face images, and the testing set had 458. Labels categorized images into three classes: 0 for no mask, 1 for correct mask usage, and 2 for improper usage. This aimed to assess model performance across different mask-wearing scenarios. Before training and evaluation, images underwent preprocessing steps, including resizing to 128x128 pixels, grayscale conversion, pixel normalization to [0, 1] range, and adding a channel dimension for grayscale images. The pre-processed images and labels were then used for training and evaluating the face mask detection models.

Implemented methods

The SVM classifier was selected for its ability to effectively model complex decision boundaries, making it well-suited for the multi-class face problem. HOG features were chosen to capture edge and gradient information, beneficial for object detection tasks. SIFT descriptors, robust to scale and rotation variations, when dealing with face images captured from different angles and distances. And then, the lightweight and efficient MobileNetV2 CNN architecture was implemented, as its computational efficiency.

The first model was an SVM classifier trained on HOG features from images, with data augmentation for classes representing no mask (0) and improperly worn masks (2). HOG features were standardized, and a grid search tuned the SVM's hyperparameters. Class imbalances were handled via a weighted scheme. The optimized HOG+SVM model and scaler were saved.

The second model used a BoVW representation of SIFT descriptors as input for an SVM classifier. After augmenting classes for (0) no mask and (2) improperly worn masks, SIFT descriptors were extracted and clustered via k-means to create a visual vocabulary. BoVW histograms of visual word occurrences formed the feature vectors. A grid search tuned the SVM's hyperparameters, handling class imbalances. Evaluation metrics were computed, and the optimized SIFT+SVM model, k-means model, and scaler were saved.

The third model employed transfer learning, fine-tuning the pre-trained MobileNetV2 CNN. Data augmentation techniques were applied for class 0 and class 2. The pre-trained MobileNetV2, excluding top layers, was loaded with frozen weights. New top layers with a softmax output were added. The model was trained with a ModelCheckpoint callback, and evaluation metrics were computed on both training and validation sets. Visualizations included a confusion matrix heatmap and loss/accuracy plots and classification report metrics. Visualizations were generated, including a confusion matrix heatmap and training/validation loss and accuracy plots.

We applied our best model to a short video downloaded from the web using a function called MaskDetectionVideo. This function implements a facial mask detection system for video processing. It comprises multiple functions tailored for different tasks within the pipeline. The preprocess_face function resizes the input face region of interest (ROI) to the model's expected size, converts pixel values to float32, normalizes them to the range [-1, 1], and adds a batch dimension. The classify_mask function predicts the mask status of the preprocessed face ROI using a trained model and returns the corresponding mask label. The update function iterates through video frames, detects faces using a face detector, preprocesses face ROIs, classifies mask labels, and annotates frames with bounding boxes and mask labels. Finally, the MaskDetectionVideo function processes the entire video, randomly selecting frames for display, and applying the update function to each selected frame, resulting in a visual output showing annotated faces with their mask status.

Results

The HOG_SVM model, optimized with parameters C=10, gamma='scale', and an RBF kernel, achieved perfect training accuracy, signifying complete accuracy on the training dataset. It exhibited a robust validation accuracy of 87.57%, demonstrating its effectiveness on unseen data. However, while the model showed strong precision and recall for most classes, it struggled particularly with class 2, which had significantly lower recall. This was reflected in the validation performance, where the confusion matrix and classification reports highlighted the model's tendency to misclassify class 2. On the test dataset, the overall accuracy slightly decreased to 85.81%. Notably, the performance for class 2 dropped further, indicating challenges in consistently identifying this less represented class. The model maintained high accuracy for class 1 across both datasets, but the difficulties with class 2 underscored potential issues with class imbalance and generalization to less frequent categories.

The SIFT_SVM model showed a high training accuracy of 94.18%, but its performance dropped in validation and test phases with accuracies around 67%. It handled classes 0 and 1 reasonably well in validation, with both precision and recall above 70%, but struggled significantly with class 2, showing only 41% precision and 48% recall. Test results mirrored these trends; although class 1 maintained high precision at 91%, its recall dropped to 72%. Class 0 and class 2 exhibited poor precision at 24% and 9% respectively, indicating a high rate of false positives for class 0 and overall inefficacy in class 2 identification. These results highlight the model's challenges with generalization and class imbalance, particularly for the underrepresented class 2.

The CNN model exhibits strong performance in both training and validation phases, achieving an impressive training accuracy of 95.61% and a slightly lower but still commendable validation accuracy of 92.90%. The confusion matrix on the validation set illustrates the model's ability to correctly classify most instances across the three classes, with particularly high precision and recall for class 1. However, it struggles with class 2, as indicated by its lower precision, recall, and F1 score. Despite these challenges, the model maintains a solid overall weighted average F1-score of 92% on the validation set. On the test set, the model maintains a high accuracy of 93.89%, with the confusion matrix revealing similar trends as seen in the validation phase. Class 2 continues to be the most challenging to classify, with the lowest F1-score of 27%, indicating room for improvement in recognizing this class. Overall, the model demonstrates strong generalization capabilities, showcasing its effectiveness in accurately classifying unseen data.

Model	Test Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Precision (Class 2)	Recall (Class 2)	F1-Score (Class 2)
HOG_SVM	0.858	0.5	0.49	0.5	0.9	0.94	0.92	0.8	0.21	0.33
SIFT_SVM	0.677	0.24	0.57	0.34	0.91	0.72	0.8	0.09	0.16	0.12
CNN - Mobilenet	0.939	0.87	0.88	0.87	0.95	0.98	0.97	1.0	0.16	0.27

And after that, we applied the best model on the video and show cased the video frames, we used a facial mask identification system to provide a series of frames with visual annotations that show whether people are wearing masks correctly, not at all, or not at all. The system is applied to a video. You can see in (Fig 1,2) the outcomes are displayed as updated frames with a label indicating the mask state above each detected face contained in a green bounding box.

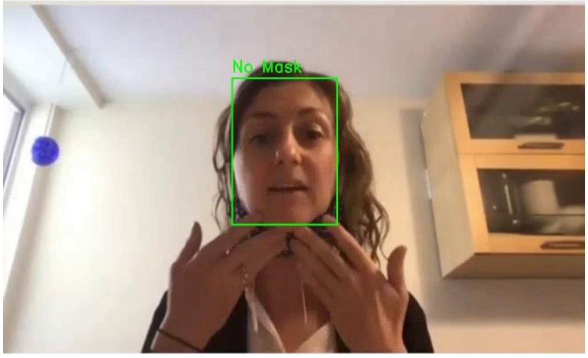

Fig: 1

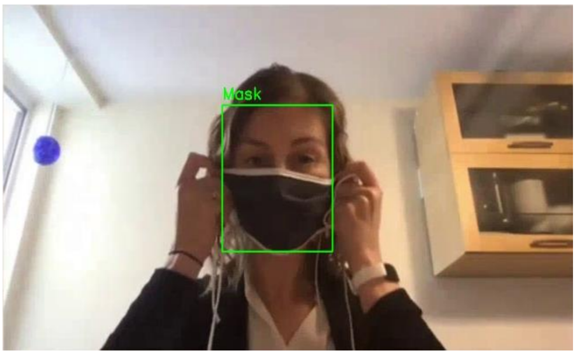

Fig: 2

The results offer a straightforward and easily understood assessment of mask-wearing behavior among individuals captured in the video frames. While the model was set to process the full video, we encountered an issue with video playback. Although the video processing proceeded without error, the video itself did not display. Various methods to present the video were attempted, yet none succeeded in rendering the visual output.

References

[1] Computer Vision Lab 5

[2] Computer Vision Lab 6

[3] Computer Vision Lab 9