CITY
UNIVERSITY OF LONDON
EST 1894

IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Fatema Fidvi (230024865) - PG
- **Google Drive folder:** https://drive.google.com/drive/folders/12IJhAG9m_yc868fenUiSwuO3AoxfYUyG?usp=sharing

Data

The dataset provided for this study includes 2,393 labeled images in the training set and 458 images in the test set, displaying significant variations in image size and marked class imbalance. In terms of class distribution, approximately 83% of the training images are labeled as mask worn (Label 1), 16% as mask not worn (Label 0), and merely 3% as mask worn improperly (Label 2). Image resolutions range broadly from 16x11 to 269x340 pixels, with the majority of the images falling within the 32x32 and 64x64 pixel range. This variability necessitates preprocessing to standardize input dimensions and ensure consistent data quality. To further optimize model training, the training dataset is split into training and validation sets using an 80:20 ratio.

To evaluate the real-world applicability of our top-performing mask detection model, a 2-minute and 30-second test video was selected from YouTube [1]. Chosen for its high resolution (1280x720 pixels) and clear visibility of individuals wearing masks in various states, the video serves as an excellent basis for assessing the model's effectiveness. With 3,628 frames and BGR color format, this video provides a thorough test of the model's ability to detect masks in real-time and adapt to new environments. This choice ensures we can assess the model's accuracy while maintaining computational efficiency, by selecting a video that is long enough to be informative yet concise enough for practical testing.

Implemented methods

For mask classification task, Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) with Bag of Visual Words (BOVW) were selected as feature descriptors, both using Support Vector Machines (SVM) as their classifiers. These methods were chosen to compare the performance of local versus global feature descriptors on our dataset. SVMs were preferred due to their efficiency and lower risk of overfitting compared to more complex models like MLPs, as noted in lecture slides 5 & 6.

Additionally, I explored transfer learning with pre-trained CNN models MobileNet and ResNet50. MobileNet was chosen for its computational efficiency and status as an earlier CNN architecture, providing a benchmark. In contrast, ResNet50 was selected for its robust performance and relevance in recent image classification tasks. This approach facilitated a detailed comparison of traditional methods and modern CNNs, highlighting their applicability to face mask detection.

- **Preprocessing Steps:**

To ensure compatibility across all models, images were resized to 128x128 pixels, despite most originals being smaller. This standardization allows both traditional and CNN models to process inputs uniformly, facilitating consistent comparisons.

- *For SVM-based models*: Images were converted to grayscale and normalized using the /255 approach to reduce computational demands, as color information is not crucial for mask detection and normalization helps in processing pixel values within a standard scale.

- *For CNN models*: Images were ensured to be in RGB format and were normalized using the `preprocess_input` function from ResNet50, applicable for both MobileNet and ResNet50 architectures, ensuring that the inputs are standardized for optimal performance. The labels were also one-hot encoded.

- **Data Augmentation for Class Imbalance:**

Significant class imbalance was addressed through targeted data augmentation applied only to the training set, leaving the validation and test sets unchanged to accurately evaluate model performance for all models. Techniques such as rotations, shifts, shears, zooms, and flips were applied randomly: class 0 augmented by 4 folds and class 2 by 20 folds. This approach did not equalize the classes but significantly reduced the disparities, bringing them into closer alignment and enhancing the generalizability of the models.

- **Feature Extraction:**

- *HOG:* Adjustments in pixels per cell were made to balance computational efficiency and performance, with decisions informed by empirical testing.

- *SIFT+BoVW:* The number of clusters (k) for BOVW was determined from an elbow curve, which guided the creation of effective histograms for classification.

- **SVM Models with HOG and SIFT+BOVW:**

- Baseline and Optimization: Initial models used a linear kernel to establish a baseline performance on the validation set. Subsequent tuning was performed using grid search to explore various kernels, and the parameters C and gamma, optimizing these based on validation results.

- **CNN Models – Transfer Learning:**

- *MobileNet:* The training involved iterative refinement, adjusting layers and experimenting with different combinations of neurons and layers in each iteration. The optimal setup was selected based on its performance in validation metrics, ensuring the model was finely tuned to our specific task.

- *ResNet50:* Employed a straightforward transfer learning strategy, fine-tuning a predetermined set of layers once to suit our specific classification task.

- **Training Process:**

All models underwent rigorous training on the augmented dataset, closely monitored by validation metrics to prevent overfitting. Hyperparameters were finely adjusted according to these metrics to ensure robust model performance across unseen data. This approach provided a detailed comparison of various classification strategies, highlighting the advantages and limitations of each within the context of face mask detection. It allowed us to assess the effectiveness of local versus global feature extraction and the advantages of CNN model architectures in real-world applications.

- **Video "In the wild":**

To extend our study from static images to dynamic real-world scenarios, face mask detection on a video "in the wild" was implemented. This process involved using the MTCNN model to detect faces in video frames and classify them based on the presence and correctness of mask wearing. The system utilizes the ResNet50 model, fine-tuned on our image dataset. Video frames undergo preprocessing similar to our CNN models (resizing, normalizing, and RGB conversion) to maintain consistency with our image data handling. To manage real-time processing demands, temporal smoothing is applied to stabilize predictions across frames, and a frame-skipping strategy is used to maintain efficiency. Annotations, based on the model's predictions, are displayed directly on the video frames. This approach demonstrates the practical applicability and robustness of our best trained model in everyday scenarios.

Results

Table 1: provides the quantitative details of each model with their results on test set and their speed and size

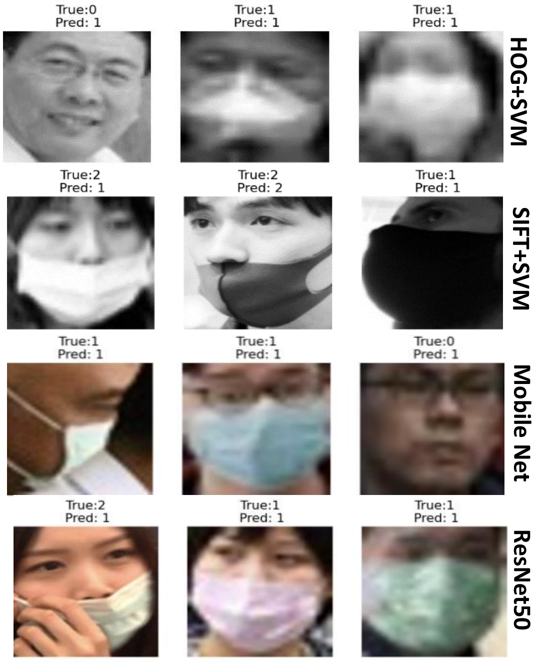
Model	Precision			Recall			F1-Score			Accuracy	Macro-avg F1 score	Model Size	Training Time
	0	1	2	0	1	2	0	1	2				
HOG+SVM	0.36	0.88	0.50	0.31	0.93	0.11	0.33	0.90	0.17	0.82	0.47	1.03GB	30 min
SIFT+BoVW+SVM	0.30	0.95	0.08	0.57	0.75	0.21	0.39	0.84	0.11	0.71	0.45	2MB	3 min
MobileNet	0.88	0.96	0.23	0.75	0.94	0.42	0.81	0.95	0.30	0.90	0.69	26.3MB	1.2 hrs
ResNet50	0.85	0.96	0.47	0.88	0.96	0.42	0.87	0.96	0.44	**0.93**	0.76	93.6MB	3.8 hrs

I. The **Quantitative Analysis and Discussion** of the test results for the four models shows varied performance across different metrics displayed in table 1.

- My **HOG+SVM** model achieved an overall accuracy of 82% on the test set, using a detailed 4x4 pixel per cell ratio in feature extraction. This detailed capture of gradients required the SVM to manage a larger array of support vectors, which not only lengthened training time but also significantly increased the model's size to 1.03GB (highlighted in table 1). This highlights the challenges of using this global feature descriptor efficiently. While my model maintained a decent overall accuracy, it struggled significantly with complex classifications, such as detecting masks worn improperly, achieving low recall (0.11) and F1-score (0.17). Further increasing the pixel density degraded my SVM performance, underscoring the need for carefully considered feature extraction settings to balance performance and efficiency effectively.
- In my **SIFT+BoVW+SVM** implementation, the model effectively recognized 'No Mask' cases with a recall of 0.57 and showed better detection of 'Mask Worn Improperly' with a recall of 0.21, outperforming the HOG+SVM model's recall of 0.11 for this class. Despite precision challenges, SIFT was more adept at identifying specific mask-wearing errors. Its compactness (2MB) and rapid training (3 minutes) stemmed from using MiniBatch K-Means clustering with 70 clusters, which streamlined features and reduced data complexity. However, the SIFT model struggled with complex features in resized and pixelated images, leading to a low macro-average F1 score of 0.45. This reflects the limitations of local feature descriptors in processing diverse image conditions efficiently.

- My **MobileNet** model showed a strong performance for the 'Mask' class with a precision of 0.96 and a recall of 0.94, outperforming traditional SVM-based methods. Initially, it struggled due to its less complex architecture, which made capturing detailed features challenging. Through iterative tuning and adjustments, I improved its accuracy, making it efficient for real-time applications. The model's relatively modest size (26.3MB) and moderate training time (1.2 hours) further establish it as a viable option that provides good results without extensive computational demands. This balance makes MobileNet particularly suitable for environments where computational resources are limited.
- My **ResNet50** stood out with the highest overall accuracy (0.93) and macro-avg F1 score (0.76), highlighted in table 1, indicating its robust performance across all classes. It excelled particularly in class 1 with precision and recall both at 0.96. However, its larger size (93.6MB) and longer training time (3.8 hours) reflected its computational demands, which are justified by its superior performance, especially in more challenging scenarios involving mask detection.

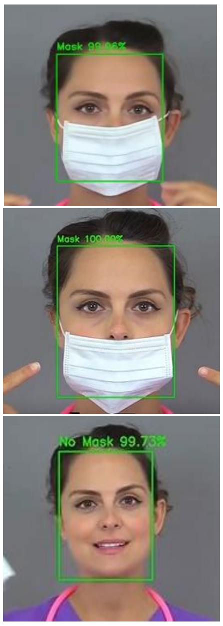
II. The **Qualitative Assessment and Discussion of** the 4 models on the test images with their true and predicted labels:



- In the images classified by **HOG+SVM** model, there's a noticeable difficulty in correctly predicting class 0, with instances where masks are absent but predicted as present. On the other hand, for pixelated images where masks are worn, the model accurately identifies class 1. This suggests that while the model tends to predict conservatively, it handles lower image resolutions effectively, which is crucial in real-life scenarios with variable image qualities.
- The **SIFT+BoVW+SVM** model presents a varied performance in the displayed images. It successfully identifies class 2 when the image is clear, but in cases of lower image quality, it mistakes class 2 for class 1. This inconsistency highlights the model's sensitivity to image clarity, implying its potential reliability in optimal conditions but also its vulnerability to variations in image quality.
- In the images, **MobileNet** accurately identifies class 1. However, it often misclassifies class 0 as class 1, showing a bias towards detecting masks even when they are not present. This suggests the model may be overfitting on mask detection and could benefit from calibration to reduce false positives.
- In the displayed images, **ResNet50** reliably identifies class 1 (Mask), but it occasionally confuses class 2 (Mask Worn Improperly) with class 1. Despite its difficulty with class 2, ResNet50 excels at correctly classifying class 0 and 1, outperforming other models with high accuracy rates of 88% and 96%, respectively.

In the **Qualitative Assessment of the test video**, ResNet50, while effective in static image testing, did not identify any instances of class 2 (Mask Worn Improperly), despite clear examples in the video. This indicates that the model's 42% accuracy for this class from image testing did not translate well to the more complex video environment. However, ResNet50 accurately differentiated between class 0 (No Mask) and class 1 (Mask) throughout the video, demonstrating its effectiveness in more straightforward scenarios. This consistent performance is highlighted in video snapshots where the model's predictions are clearly annotated, showcasing its reliability in recognizing correctly worn masks and no mask scenarios.

Comparison: ResNet50 outperformed the other models with the highest overall accuracy and macro-average F1 score, demonstrating robustness across classes, especially in recognizing correctly worn masks. While all models showed some ability to identify properly worn masks, traditional models like HOG+SVM and SIFT+BoVW+SVM struggled with accuracy and consistency in identifying improperly worn masks, a challenge less pronounced but still present in CNN models like MobileNet and ResNet50, which showed better discrimination but were not infallible, particularly in dynamic video scenarios.



Video Snapshots

References

[1] https://youtu.be/etZK-GrUYgM?si=hxSy6AFCoYH5tmSj