IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Ling-Yun, Huang (230048952) - PG
- **Google Drive folder:** [redacted]

Data

The provided dataset contains two subsets, one is for training, another one is for testing, this will remain unseen until the best is selected in each implemented method and use to test each implemented method's performance. The train dataset has 376, 1940, 48 samples in "no mask", "mask proper wear", and "mask without proper wear" respectively. For the unbalanced classes, might cause some problem in the later model training.

The video is from the Taiwan Centres for Disease Control and Prevention, and it mainly promotes precautions for taking taxis during the epidemic[1]. The images in the video cover scenarios in which individuals are not wearing masks, are wearing masks correctly, and are wearing masks incorrectly, and are suitable for testing the predictive ability of the model.

Implemented methods

Before the training process, I preprocessed the provided dataset by resizing each image to 32x32 pixels and normalizing each pixel value to fall within the range of [0,1]. Additionally, I divided the training dataset into 80% for training and 20% for validation purposes. The three distinct methods implemented for this task are "HOG + SVM", "SIFT + MLP", and "CNN". For each implementation, I experimented with various hyperparameters using the validation subset to select the best performance based on the highest weighted F1-score.

In "HOG + SVM", HOG (Histogram of Oriented Gradients) is a global image descriptor which capture the global structures across entire image, and SVM (support vector machine) is a supervised machine learning algorithm. After extracting information from the image using HOG, these features are fed into an SVM classifier, which learns to distinguish between different classes based on these features. In this coursework, I experimented with different kernel ("rbf", and "poly") and different values of C (0.01, 0.1, 1, and 10) settings in SVM and selected the model with the highest weighted F1-score.

In "SIFT + MLP", SIFT (Scale-Invariant Feature Transform) is a local image descriptor used to detect keypoints and extract features from images. After extracting features with SIFT, k-means is applied to cluster similar descriptors together, generating a set of representative visual words or codewords. These representations are used as input to a MLP (multilayer perceptron) to train to recognize patterns and make predictions base on them. In this coursework, I experimented with different hidden layers (one or two), and different numbers of neurons (15, 30, 60, 120) to select the model with the highest weighted F1-score.

CNN (convolutional neural network) is a deep neural network specifically designed to learn the spatial features from image data. It consists of convolutional layers combined with max pooling, a fully connection layer and an output layer. Before the training process, the images are normalized using means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. Additionally, 10% of the data is set aside for monitoring the model during training to prevent overfitting. If the model fails to show improvement in 5 epochs, the training is stopped. In this coursework, I experimented with different convolutional layers (one or two), different kernel size (3, 5) with different numbers of filters (6, 12) to select the model with the highest weighted F1-score.

For the video data, I employed a face detector based on a pre-trained MTCNN model[2]. This detector identifies faces in each frame. Subsequently, I used previously trained models to classify individuals' mask-wearing status: whether they are not wearing masks, wearing masks correctly, or wearing masks incorrectly.

Results

The model selection in each method:

In each implementation, I experimented with different hyperparameters, which are displayed in Table 1. The table highlights the highest weighted F1-score, indicating the chosen model.

- In HOG + SVM, the 'poly' kernel generally outperforms the 'rbf' kernel. However, with the 'rbf' kernel, increasing the value of C above 1 significantly improves results compared to smaller values.
- In SIFT + MLP, the results are similar across different configurations. Any observed differences are more likely due to variations in the initial weights and biases chosen for the network.
- In CNN, with two convolution layers has slightly better performance compared to a single layer, although the difference is not particularly significant.

Table 1: The weighted F1-score results in each model.

HOG + SVM			SIFT + MLP	
kernel	C value	Weighted F1	hidden layers	Weighted F1
'rbf'	0.01	0.7250	[15]	0.7808
'rbf'	0.1	0.7250	[30]	0.7845
'rbf'	1	0.8391	[60]	0.7751
'rbf'	10	0.8522	[120]	0.7832
'poly'	0.01	0.8116	[15,15]	0.7935
'poly'	0.1	0.8403	[30,30]	0.7832
'poly'	1	0.8448	[60,60]	0.7763
'poly'	10	0.8448	[120,120]	0.7750

CNN			
Convolutional layers	filters	kernel size	Weighted F1
1	6	3	0.9192
1	6	5	0.9186
1	12	3	0.9277
1	12	5	0.9096
2	6	3	0.9118
2	6	5	0.9377
2	12	3	0.9213
2	12	5	0.9355

The results on provided test dataset:

Figure 1 illustrates the results of three methods on the provided test dataset. This includes metrics such as accuracy, precision, recall, and F1-score, as well as confusion matrices depicting the classification performance of each method.

- The performance of the models on the test dataset aligns with the validation results obtained during training, indicating consistency in their predictive capabilities.
- The impact of the unbalanced dataset is evident in the results, with models tend to predict the largest class (wearing mask correctly) more accurately than the smaller classes. This imbalance causes challenges, particularly for the smallest class (wearing mask incorrectly), which struggles to be adequately learned by the models. Even the best-performing model, CNN, demonstrates significant misclassification rates for smallest class, highlighting the difficulty of learning from imbalanced data.
- Both HOG + SVM and SIFT + MLP models perform similar misclassification patterns, with half of the samples from class 0 (no mask) being misclassified as class 1 (wearing mask correctly), and most samples from class 2 (wearing mask incorrectly) being misclassified as class 1 (wearing mask correctly). This trend underscores the models' tendency to favour the majority class.

HOG + SVM			
Class	Precision	Recall	F1-score
0	0.53	0.47	0.50
1	0.90	0.95	0.92
2	0.67	0.21	0.32
Accuracy: 0.86; Weighted F1-score: 0.85			

SIFT + MLP			
Class	Precision	Recall	F1-score
0	0.38	0.41	0.39
1	0.88	0.91	0.90
2	0.00	0.00	0.00
Accuracy: 0.82; Weighted F1-score: 0.80			

CNN			
Class	Precision	Recall	F1-score
0	0.84	0.94	0.89
1	0.98	0.97	0.97
2	0.56	0.53	0.54
Accuracy: 0.95; Weighted F1-score: 0.95			

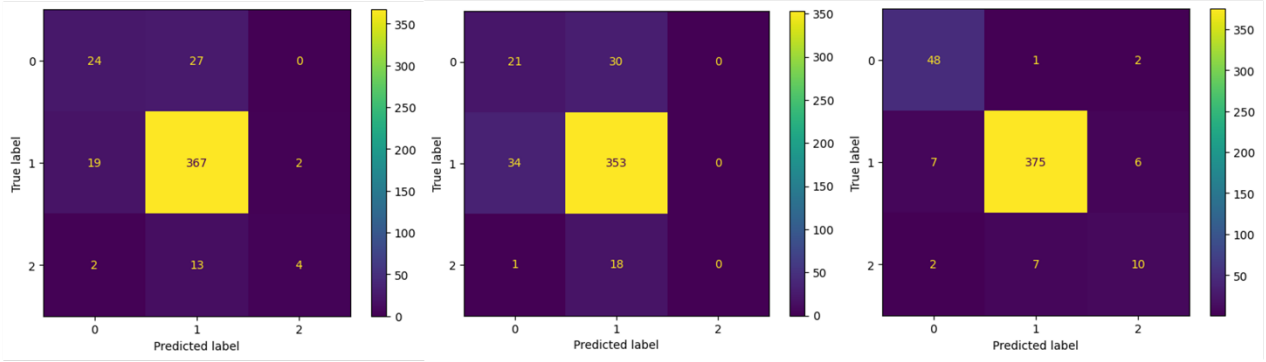

Figure 1. Performance Comparison of Three Methods on the Provided Test Dataset

Figure 2. displays two images from each class alongside their true labels and predictions made by three different methods.

- The outcomes correspond with the quantitative observations, as most predictions in the “Mask” group are accurate.
- Within the “No mask” group, only half of the predictions are correct in the first two methods, whereas CNN achieves a higher accuracy.
- In “Mask incorrect” group, neither of the first two methods has correct predictions, while CNN achieves a 50% accuracy rate.

Figure 2. Comparison of True Labels and Predictions from Three Methods

The trained models predict on the video:

The trained models were used to predict on the video dataset, but the results didn't match expectations.

- When comparing the three models, CNN showed more consistent predictions with faces that had minimal movement. However, the other two models gave different predictions even with small shifts in facial positions.
- Surprisingly, although CNN performed well in predicting “No mask” group with high accuracy in provided test dataset, it struggled with correctly identifying clear faces without masks in the video. This suggests that the model might have learned other features rather than just focusing on the presence of nose and mouth.

Figure 3. Examples of Predictions on Video Frames

References

[1] “Epidemic control and taxis,” Taiwan CDC. Accessed: Apr. 18, 2024. [Online]. Available: https://www.cdc.gov.tw/Advocacy/SubIndex/2xHloQ6fXNagOKPnayrjgQ?diseaseId=N6XvFa1YP9CXYdB0kNSA9A

[2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” *IEEE Signal Process Lett*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.