COMP90051 - Project group - Group 2

Thi Minh Hoai Bui – 1561309
Wei-Chiao Yang – 1213956
Ho Ting Poon – 1618009

How does the choice of feature extraction techniques affect the robustness of SVM, SVM-CNN, and Hybrid ViT on face verification?

Introduction

Face verification requires precise feature extraction to determine whether the faces in two images belongs to the same person. The effectiveness of various models in face verification largely depends on the feature extraction methods. Factors such as noise and occlusion lead to challenges to the robustness of methods (Singh et al., 2020). To investigate this, this research explores how different feature extraction techniques impact the performance and robustness of models including traditional Support Vector Machine (SVM), deep learning model of Support Vector Machine combines with Convolution Neural Network (SVM-CNN), and Hybrid Vision Transformer (Hybrid ViT) as modern transformers proposed in the research conducted by Phan et al. (2024). The models will be tested on both original data with aligned face preprocessing and distorted data with occlusion and noise, which provides valuable insights into the robustness and the ability to handle variations in face verification tasks.

Literature Review

Data set: We used the datasets called Labelled Faces in the Wild Home (LFW) provided by Huang, Ramesh, Berg, and Learned-Miller (2007) to conduct the face verification task as it captures people in real-life condition with various lighting, poses, colour, and background. This dataset contains more than 13,000 images with 5749 distinct people, and 1608 of them have more than one image.

Relevant Research: Face verification task has been evolved through different stages from traditional machine learning to more advanced techniques. In this project, SVM is used as the traditional model due to its strong performance in nonlinear classification tasks. Based on the research conducted by Guo et al. (2016), the combination of SVM and CNN has demonstrated the effectiveness of the advantages of CNN for feature extraction and SVM for classification for face recognition tasks and this finding inspired the adoption of the model of SVM-CNN. More recently, the Vision Transformer (ViT) has been found to be able to capture features more effectively by operating at the patch level instead of image level (Phan et al., 2024). The Hybrid ViT combining ViT with a pretrained CNN for patch embedding demonstrated enhanced robustness against variations in data, outperforming traditional CNN (Phan et al., 2024). Therefore, it serves as the advanced model in this project.
The two images undergo face detection and alignment preprocessing. After that, the performance of three models will be evaluated on the alignment faces dataset. To assess robustness of model, random noise or occlusion will be added to the dataset.

Methods

1. Preprocessing

The LFW dataset is pre-processed to create two distinct datasets:

- **Cropped and Aligned Dataset**: All images are processed using MTCNN to detect, crop, and align faces, ensuring consistent orientation and size for each face.
- **Distorted Dataset**: The aligned images are further augmented by randomly adding Gaussian noise or applying random-shaped occlusions to simulate challenging conditions such as noise and partial occlusions.

These two datasets allow us to systematically analyse the **robustness** of different models under varying conditions.

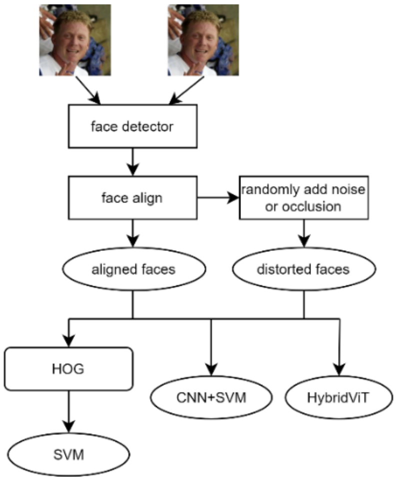


Figure 1. Method overview

2. Nested cross validation:

We use **20-fold nested cross validation** based on the predefined 10-fold splits in pairs.txt file (recommended by the LFW dataset authors). Each of the 10 folds is divided into 2 equal parts, ensuring that both subsets maintain a similar number of positive (matched) and negative (mismatched) pairs. This results in 20 balanced folds. This approach ensures:

- **Consistency**: Adheres to the dataset's standard evaluation protocol

- **Parallel execution**: By using fixed 20 folds, we can distribute the execution across multiple machines, running some folds on one machine and others on separate machines simultaneously.
- **Fair comparison**: Using the same splits for all three algorithms ensures a direct and unbiased comparison for their performance.

In each iteration, 1-fold is used for testing, and 19 folds are used for training. During training, hyperparameters are tuned using 0.632 sampling with replacement, repeated 5 times. The prediction results of all 20 folds are saved for further analysis.

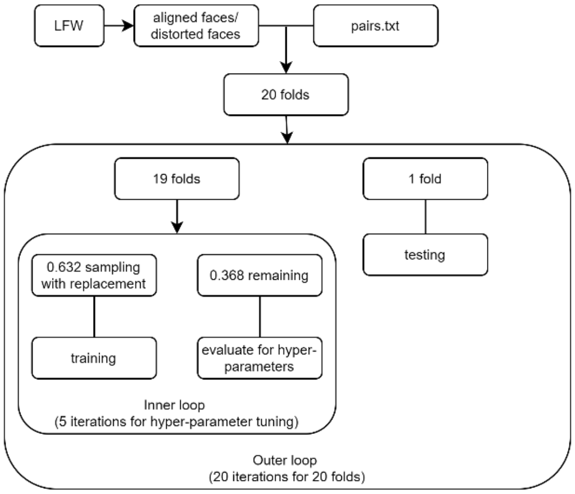

LFW
aligned faces/ distorted faces
pairs.txt
20 folds
19 folds
1 fold
testing
0.632 sampling with replacement
0.368 remaining
training
evaluate for hyper-parameters
Inner loop (5 iterations for hyper-parameter tuning)
Outer loop (20 iterations for 20 folds)

Figure 2. Nested cross validation

3. Algorithms

In our project, we chose **SVM**, **SVM-CNN** (Guo et al., 2016), and **Hybrid ViT** (Phan et al., 2024) to investigate the impact of different feature extraction techniques on face verification.

- **SVM** is a traditional and effective classifier. We use it as a baseline to compare against more complex methods. For the baseline SVM model, we use Histogram of Oriented Gradients (HOG) as the feature extraction technique to capture facial structure details. The SVM's regularization parameter **C** is tuned using three values: **0.001, 0.01, and 0.1** to identify the best setting. This ensures a balance between margin maximization and error minimization, allowing the model to generalize effectively.
- **SVM-CNN** is a deep learning model combining the strengths of SVM and CNN by using CNN for feature extraction and SVM for classification. In this project, a custom implementation of ResNet-50 architecture is used for feature extraction as it can balance computation cost and performance by capturing complex features and pattern compared to other ResNet. The output layer of CNN is removed after training and SVM will use features extracted from CNN to classify the pairs. The tuned hyperparameter is **kernel size** using three values: **1, 3, 5** in Bottleneck block to determine if a smaller kernel size that is able to capture finer details of image performs better than a larger kernel size capturing global features in face verification task.
- **Hybrid ViT**

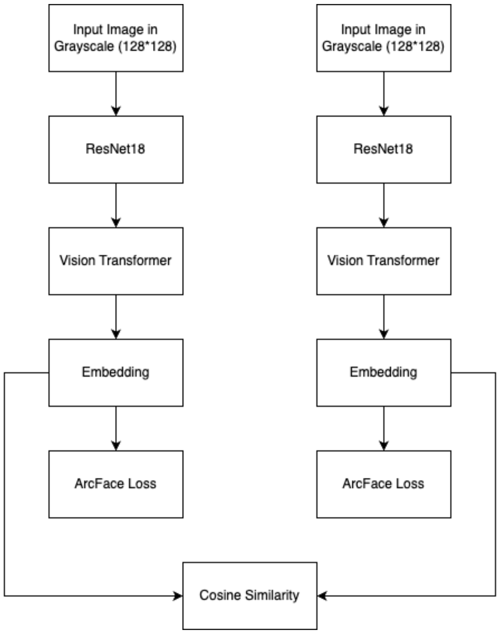


Figure 3. Hybrid ViT architecture

Hybrid ViT combines CNN and ViT in predicting similarity between two images. Pretrained ResNet18 is used for CNN, which converts the input image into a feature map that captures important facial features. Since the ViT requires tokens as input, CNN is needed to handle low-level feature extraction. The feature map generated by ResNet18 is reshaped into a sequence of tokens, which is used as the input for the transformer. A classification token is appended to the sequence of tokens, and positional embeddings are added to each token to provide information about their spatial relationships. The tokens are fed into the transformer, and the self-attention mechanism helps the transformer learn about the relationships between tokens. After learning, the transformer generates an informative representation (embedding) of the image. ArcFace Loss is used to optimize the created embeddings by enforcing an angular margin, which can help the models better distinguish between different classes (faces). For inference, cosine similarity is used to determine the similarity between two faces. A high similarity represents the two faces are likely to belong to the same person. To set a cosine similarity threshold for determining whether two faces belong to the same person, we observed the similarities of matched and mismatched image pairs. If the model is properly trained to capture the features of the input images, there should be a noticeable gap between the average similarity of matched images and that of mismatched images. Additionally, we analysed the minimums, maximums, and median of the similarities of matched and mismatched pairs to understand the distributions. The gap is used as a cosine similarity threshold to predict whether two faces belong to same person. The tuned hyper-parameter is the depth of transformer, which refers to the number of layers. Each layer consists of Multi-Head Self-Attention and Feed-Forward Neural Network. The values **2, 4 and 6** are used to evaluate the performance of Hybrid ViT under different depths.

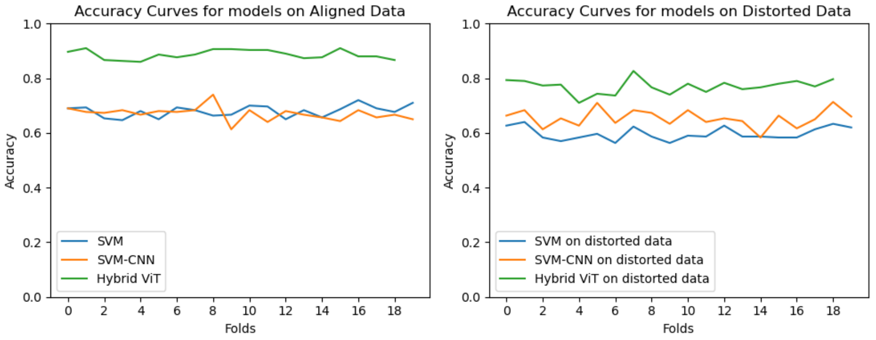
4. Robustness testing

To test the robustness of the three algorithms, we evaluate their performance on **distorted face images** generated by randomly adding one of:

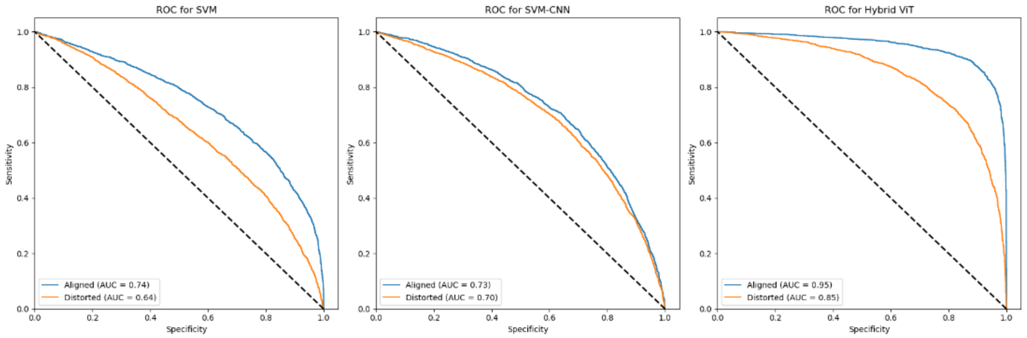
- **Gaussian Noise:** Simulates low-quality or degraded images by adding random pixel-level variations.
- **Random Occlusions:** Covers parts of the face with random shapes and sizes

By comparing the models' accuracy on these distorted images against the original aligned faces, we assess how well each feature extraction technique handles challenging conditions and maintains performance under different types of distortions.

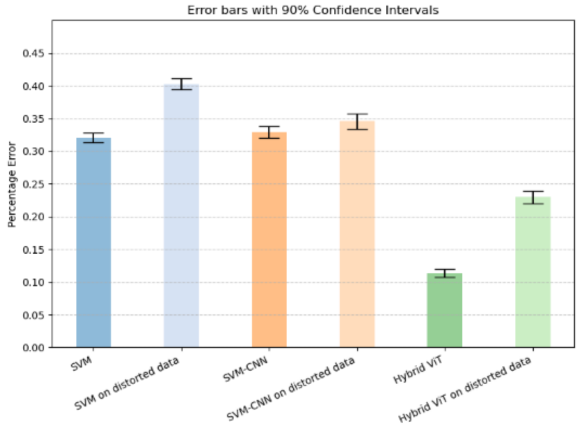
Result and Discussion



The accuracy curves show that Hybrid ViT consistently outperforms the other models across both aligned and distorted datasets. In terms of accuracy, Hybrid ViT demonstrates superior performance by maintaining over 80% accuracy on both datasets. This strong performance can be attributed to the complex architecture of Hybrid ViT, which combines both CNN and ViT for feature extraction (Phan et al., 2024). The CNN component excels at extracting local features and spatial hierarchies from images, capturing essential patterns such as edges and textures. Meanwhile, the ViT component leverages its self-attention mechanism to capture global relationships and long-range dependencies between image patches. This combination enables Hybrid ViT to capture a diverse set of features, making it highly effective even when dealing with challenging scenarios such as noisy or distorted data.



The Receiver Operating Characteristic (ROC) curves illustrate each model's ability to distinguish between classes by plotting sensitivity against specificity across varying thresholds. The Area Under the Curve (AUC) highlights overall model performance. Although Hybrid ViT achieves the best performance among the three models, it is still affected by distorted data, with AUC values dropping from 0.95 on aligned data to 0.85 on distorted data. SVM, as the baseline model, shows the poorest performance and high sensitivity to data changes, with an AUC of 0.64 under distortion. SVM-CNN, on the other hand, demonstrates better robustness, as its performance remains relatively stable with AUC values of 0.73 on aligned data and 0.70 on distorted data, indicating less sensitivity to noise compared to the baseline model.



This is further confirmed by examining **the error bars**, where the overlap between SVM-CNN on the two datasets is the largest. The error bar graph also reveals that SVM and SVM-CNN show comparable performance on the aligned faces dataset. However, SVM-CNN significantly outperforms SVM when making predictions on the distorted data. Hybrid ViT, on the other hand, achieves the best predictive performance overall, demonstrating its superior ability to capture complex features from images. However, the large gap in the error bars of Hybrid ViT between the two datasets indicates higher uncertainty when making predictions on the distorted data.

Discussion on feature extraction and robustness

All results from the experiments indicate that the HOG feature extraction used in SVM is highly sensitive to noise, whereas CNN-based feature extraction is more robust. This can be explained by the characteristics of HOG. HOG primarily captures edge and texture information; therefore, when noise or occlusion is added to the image, the local texture and edge patterns become disrupted, leading to poor performance. In contrast, CNNs have the ability to learn complex features from data through multiple layers of convolutional operations. The robustness of CNNs against noise can be attributed to their hierarchical structure, where initial layers capture simple features like edges and corners, and deeper layers progressively learn more abstract patterns like eyes, noses, etc. Additionally, CNNs leverage pooling layers, which down sample features, allowing them to focus on the most salient information and reduce sensitivity to small local changes, making them effective at handling noisy or partially occluded data.

Hybrid ViT, on the other hand, despite using a combination of CNN and ViT, still shows increased uncertainty observed in the distorted dataset. The superior performance on both datasets indicates that the combination of CNN and ViT is highly effective for feature extraction and capturing contextual information. However, the increased uncertainty may come from the final prediction process, where we use the embeddinges output output by the model from two images, and compute cosine similarity between them for classification. Cosine similarity measures the angle between feature vectors rather than their magnitude, making it highly sensitive to slight variations in the feature representation caused by noise or occlusion. Additionally, since the threshold for classification was manually set based on aligned data, it may not adapt well to the altered distribution of feature vectors in the distorted dataset, leading to inconsistent decision boundaries and higher uncertainty.

Conclusion and Future Work

In this study, we examined the impact of different feature extraction techniques on face verification models under varying conditions. Hybrid ViT, which combines CNN and Vision Transformer, consistently outperformed SVM and SVM-CNN on both aligned and distorted datasets, maintaining over 80% accuracy even under challenging conditions. Its superior performance is due to the complementary strengths of CNN's local feature extraction and the Vision Transformer's global contextual modelling. In contrast, SVM, using HOG features, showed the weakest performance and high sensitivity to noise, while SVM-CNN demonstrated better robustness. These results highlight that modern hybrid architectures like Hybrid ViT offer enhanced feature extraction in face verification tasks.

While Hybrid ViT shows the highest accuracy on both aligned and distorted datasets, it is still affected by noise, indicating that it is not the most robust model. One possible reason for this gap in performance on distorted data is the use of **fixed manual thresholds for cosine similarity** comparison, which may not be optimal for noisy datasets. In the future, we can focus on improving the comparison between embeddings by replacing cosine similarity with SVM in the final layer or experimenting with alternative classification methods to enhance robustness.

Bibliography

Guo, S., Chen, S., & Li, Y. (2016, 1-3 Aug. 2016). Face recognition based on convolutional neural network and support vector machine. 2016 IEEE International Conference on Information and Automation (ICIA),

Huang, G., Mattar, M., Berg, T., & Learned-Miller, E. (2007). Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. *Tech. rep.*

Phan, H., Le, C., Le, V., He, Y., & Nguyen, A. (2024). *Fast and Interpretable Face Identification for Out-Of-Distribution Data Using Vision Transformers*. https://doi.org/10.1109/WACV57701.2024.00618

Singh, R., Agarwal, A., Singh, M., Nagpal, S., & Vatsa, M. (2020). On the Robustness of Face Recognition Algorithms Against Attacks and Bias. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(09), 13583–13589. https://doi.org/10.1609/aaai.v34i09.7085