Final Report: EEE4610-01

Relative Embedding for Periocular and Face Recognition: Conditional Multimodal Biometrics

Report Advisor: Prof. Andrew Beng Jin Teoh,
Dr. Jae Woo Park

December 2022

Han Jun Ko

School of Electrical and Electronic Engineering
College of Engineering
Yonsei University

Contents

Abstract

The Covid-19 pandemic has compelled people to wear masks, which undermines the performance of conventional face recognition systems. In this context, recognizing the periocular region of the human face has become a critical issue in biometric identification. This paper briefly presents the methodologies of periocular recognition, the scope of periocular recognition, and the fusion of periocular with full-face recognition. In addition, the main technique employed in face recognition is end-to-end deep face recognition, which consists of three core components: face detection, face alignment, and face representation. Finally, this paper introduces a new type of multimodal biometrics, termed Conditional Multimodal Biometrics (CMB), which is based on end-to-end deep face recognition. The proposed method performs four modes of recognition: periocular-to-periocular, face-to-face, periocular-to-face, and vice versa. A novel loss function, called CMB loss, is defined to simultaneously pull intra-subject embeddings closer and push inter-subject embeddings apart in the periocular-to-face (and vice versa) setting. The software implementation section explains how similarity is calculated for identification and verification tasks, and introduces the basic concept of KL divergence, an alternative similarity metric to cosine similarity.

Keywords: Covid-19, periocular recognition, end-to-end deep face recognition, Conditional Multimodal Biometrics, CMB loss, cosine similarity, KL divergence

Chapter 1

Introduction

Biometrics is a methodology that differentiates individuals through their physical, behavioral, or other distinctive traits. As the Covid-19 pandemic has compelled people to wear masks, biometric systems must take additional factors into consideration. One key issue is that recognizing the area around the eyes, known as the periocular region, has become increasingly important in face recognition.

This paper presents the methodologies of periocular recognition, the scope of periocular recognition, and the integration of periocular with full-face recognition. In addition, the main technology for face recognition is end-to-end deep face recognition, which is composed of three essential components: face detection, face alignment, and face representation.

Finally, this study introduces a new type of multimodal biometrics, termed *Conditional Multimodal Biometrics* (CMB), which is based on end-to-end deep face recognition. The CMB framework performs four types of matching: periocular-to-periocular, face-to-face, periocular-to-face, and vice versa. Beyond conventional loss functions, a novel loss function called CMB loss is employed to simultaneously pull intra-subject embeddings closer and push inter-subject embeddings apart in the periocular-to-face (and vice versa) mode.

The software implementation section further explains how similarity is calculated for identification and verification tasks. It also covers the basic concept of KL divergence, an alternative similarity metric, as well as methods for improving computational efficiency when applying KL divergence.

Chapter 2

Computer vision for masked faces

2.1 Introduction

As Covid-19 spread rapidly across the world, wearing a mask became necessary for everyone. In this context, the performance of face recognition systems based on computer vision has been significantly degraded, since the facial parts covered by a mask are crucial for recognition. As a result, the periocular region, referring to the area around the eyes, has become critical for distinguishing individuals in biometric systems. During the pandemic, periocular biometrics has been widely adopted to complement conventional face recognition due to its advantages such as stability and robustness.

Figure 2.1: Difference between naked faces and masked faces

2.2 Different Method

Periocular biometrics is typically conducted using two main approaches. The first is to estimate and reconstruct the entire face from the periocular area. This method detects a person's eye region, estimates the complete face, and generates the reconstructed face over the occluded part. The second approach, which is the basis of this project, is to compare faces using only the periocular region. In this case, a person's eyes and the surrounding area are directly matched against those of others, treating the periocular features as a biometric identity.

The process of periocular recognition generally involves several steps:

1. **Acquisition:** Capture periocular images (or videos) using a camera.

2. **Pre-processing:** Improve image quality through normalization.

3. **Localization:** Detect and extract the periocular region, also known as the region of interest (ROI), often using deep learning techniques.

4. **Feature extraction:** Identify distinctive features within the localized ROI using global, local, or deep learning approaches.

5. **Post-processing:** Refine the feature vectors to improve accuracy and computational efficiency, for example through Principal Component Analysis (PCA).

6. **Matching:** Compare processed features and calculate similarity scores using various similarity metrics.

Periocular recognition can be used on its own, but combining periocular and full-face modalities, as emphasized in this project, provides a more reliable approach for human recognition. Since periocular features localize and extract information from a limited facial region, they can enhance the overall performance of face recognition. This multimodal strategy is particularly useful in challenging conditions, such as when a person changes facial expressions quickly, when images are captured at very close range, or when faces are covered by masks, scarves, or helmets.

Chapter 3

End-to-End Deep Face Recognition

3.1 Introduction

As previously mentioned, this project employs a deep learning approach, specifically end-to-end deep face recognition. The system is composed of three main elements: face detection, face alignment, and face representation. Since the model was trained on 2D images, the explanations in this chapter will be focused on the two-dimensional setting.

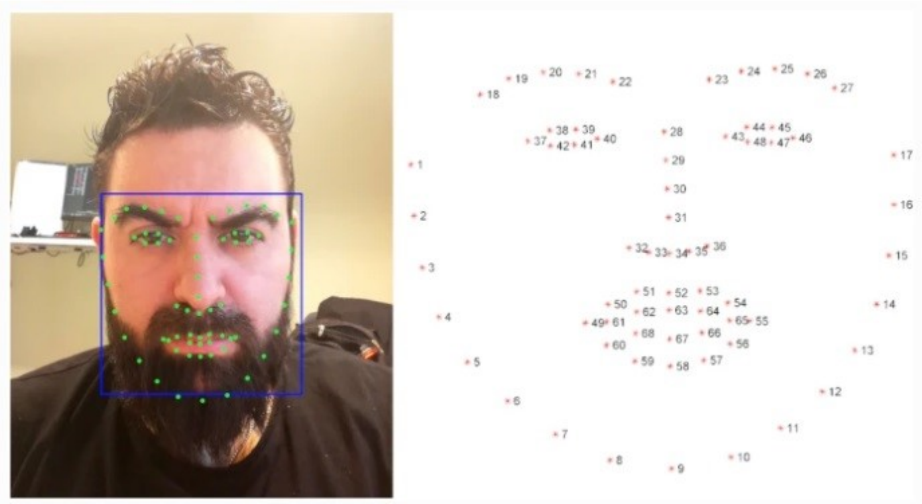

Figure 3.1: Example of landmark locations of a picture

3.2 Face Detection

The first stage of the system is face detection, which identifies human faces within a given image. Once an image is provided, the algorithm attempts to detect all human faces, generates bounding boxes around them, and outputs their coordinates along with a confidence score. Several approaches are commonly used in face detection:

- **Multi-stage methods:** These output multiple candidate bounding boxes and refine them in subsequent stages. The first step generates a sliding window with possible boxes, while the next step eliminates false positives and improves the remaining boxes.

- **Single-stage methods:** These directly classify and generate boxes from feature maps without proposals. While faster, they generally provide lower accuracy compared to multi-stage methods.

- **Anchor-based methods:** Widely used in practice, these place predetermined anchor boxes on the feature map and iteratively refine them.

- **Anchor-free methods:** These detect objects without anchors. While more general, they often require improvements in stability and false positive reduction.

- **Multi-task learning methods:** These perform face detection jointly with related tasks, such as gender classification or landmark localization.

The performance of face detection methods is usually evaluated using the Average Precision (AP) metric and the corresponding precision-recall curve. To compute these values, the *Intersection over Union* (IoU) is first calculated. IoU measures the overlap between a predicted bounding box (B_p) and the ground-truth bounding box (B_{gt}):

$$\mathrm{IoU} = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}. \tag{3.1}$$

A detection is considered a true positive (TP) if the IoU is greater than a given threshold, and a false positive (FP) otherwise. By calculating TP and FP over varying thresholds, the precision-recall curve is obtained. The AP is then computed as the mean of precision values across all recall levels.

In addition to AP, the Receiver Operating Characteristic (ROC) curve is also widely used for evaluation.

3.3 Face Alignment

The second component of end-to-end face recognition is face alignment. Face alignment adjusts detected face regions to a standardized frame in order to improve subsequent facial representation. The most widely used method is landmark-based alignment, which applies spatial transformation by referencing facial landmarks. Landmarks represent key points of facial features and serve as baselines for recognition.

Landmark-based alignment can be categorized into three approaches:

- **Coordinate regression:** Treats landmarks as numerical coordinates and estimates the nonlinear mapping from input images to landmark positions.

- **Heatmap regression:** Produces likelihood response maps for each key point instead of direct coordinates.

- **3D model fitting:** Reconstructs a 3D face model from a 2D image to determine landmark locations.

An alternative method, known as landmark-free alignment, uses a spatial transformer network without explicitly depending on facial landmarks.

The most common evaluation metric is the *Normalized Mean Error* (NME). Let M be the number of landmarks, p_k the predicted coordinate of landmark k, g_k the ground-truth coordinate of landmark k, and d a normalization distance (e.g., interocular distance). Then the NME is defined as:

$$\text{NME} = \frac{1}{M} \sum_{k=1}^{M} \frac{\|p_k - g_k\|_2}{d}. \tag{3.2}$$

Another evaluation metric is the *Cumulative Error Distribution* (CED), which represents the distribution function of the NME. The *Area Under Curve* (AUC) is also widely used and can be written as:

$$\text{AUC}_\alpha = \int_0^\alpha f(e)\, de, \tag{3.3}$$

where e denotes the normalized error, $f(e)$ the CED function, and α a designated error threshold.

3.4 Face Representation

The final component of end-to-end face recognition is face representation, which extracts discriminative features from aligned images. These features are embedded into a feature space where embeddings of the same identity are located close together, while embeddings of different identities are far apart.

Network Architectures

Popular backbone architectures include AlexNet, GoogleNet, FaceNet, and ResNet. Specialized models such as MobileFaceNet further optimize feature extraction. For instance, MobileFaceNet replaces the global average pooling layer of MobileNet with a global depth-wise convolution layer, improving the expressiveness of output features.

Training Supervision

Training supervision plays a crucial role in shaping feature embeddings. Three general strategies are used:

- **Classification scheme:** Trains the network as a classification task with identities as classes. The softmax loss is defined as:

  $$L_{\text{softmax}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{W_{y_i}^\top x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^\top x_i + b_j}}, \tag{3.4}$$

  where B is the batch size, C the number of classes (identities), x_i the feature of sample i, y_i the ground-truth label, and W_j, b_j are the weight and bias parameters.

- **Feature embedding scheme:** Focuses on relative distances between embeddings. The most common is the *triplet loss*:

  $$L_{\text{triplet}} = \sum_{i=1}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+, \tag{3.5}$$

  where $f(x)$ is the embedding function, x_i^a the anchor, x_i^p a positive sample (same identity), x_i^n a negative sample (different identity), and α a margin.

- **Hybrid methods:** Combine classification and embedding-based schemes to exploit the strengths of both.

3.5 Identification vs Verification

Before introducing the evaluation metrics, it is useful to distinguish the two main tasks in face recognition: identification and verification.

Face Identification

Face identification determines which identity in a large gallery set corresponds to a given query image (probe). This is a one-to-N task, where N is the number of gallery identities. The gallery consists of faces with known labels, while the probe set contains faces to be identified.

During computation, a similarity score is obtained between each probe and gallery sample. If the similarity exceeds a predefined threshold, the system accepts the match; otherwise, it rejects it. Identification can be evaluated in two settings:

- **Closed-set:** Every probe belongs to one of the gallery identities.
- **Open-set:** Some probes may not exist in the gallery.

For closed-set identification, the most common metric is the *Cumulative Match Characteristic* (CMC). The CMC curve shows the probability that the correct identity is found within the top-k ranked results. Formally, the rank-k identification rate is given by:

$$\mathrm{IR}(k) = \frac{\text{Number of correctly identified probes within rank } k}{\text{Total number of probes}}.$$

Face Verification

Face verification, on the other hand, determines whether two given faces belong to the same person. This is a one-to-one task. A pair of probe and gallery images is compared, and the system decides whether they match based on a threshold.

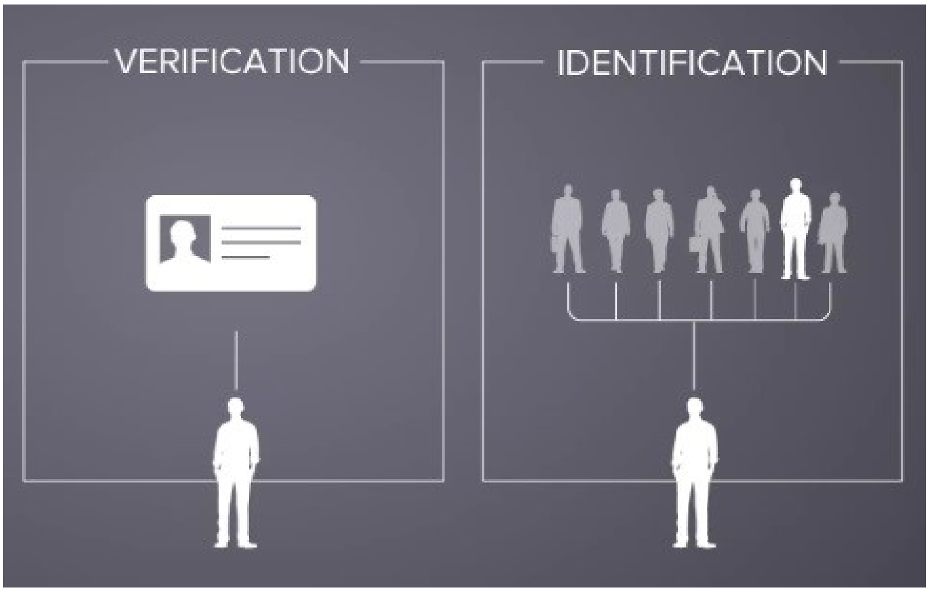

Figure 3.2: identification vs verification

In verification, the following outcomes are possible:

- **True Acceptance (TA):** Same identity and similarity $\geq$ threshold.
- **False Rejection (FR):** Same identity but similarity $<$ threshold.
- **True Rejection (TR):** Different identities and similarity $<$ threshold.
- **False Acceptance (FA):** Different identities but similarity $\geq$ threshold.

A widely used metric for verification is the *Equal Error Rate* (EER), which occurs when the *False Acceptance Rate* (FAR) equals the *False Rejection Rate* (FRR). That is,

$$\text{EER:} \quad \text{FAR} = \text{FRR}.$$

Other common metrics include the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC). In ROC analysis, the *True Acceptance Rate* (TAR) is defined as:

$$\text{TAR} = 1 - \text{FRR},$$

and plotted against the FAR to visualize verification performance.

Chapter 4

CMB for Periocular and Face Recognition

4.1 Introduction

Among multimodal biometric approaches, this project employs *Conditional Multimodal Biometrics* (CMB). The term "conditional" is used because one biometric modality influences the other to enhance feature discrimination. The proposed model is implemented using end-to-end deep learning with multiple biometric modalities and is referred to as *CMB-Net*.

The backbone network $f(\cdot)$ has shared parameters ϕ for both face and periocular inputs, which serve as predictor variables. The backbone is based on MobileFaceNet, with fully connected layers of equal size applied at the end of the last feature map, since the face and periocular inputs differ in image dimensions. In this project, the embedding dimension for both face and periocular features is set to 512.

4.2 CMB Loss Function

Beyond the traditional classification losses described earlier in end-to-end deep learning, a new loss function called *CMB loss* is introduced with regularization. The goal of CMB loss is to reduce the distance between embeddings of the same identity across different modalities (intra-subject, inter-modality) while increasing the distance between embeddings of different identities (inter-subject, inter-modality).

Formally, the CMB loss can be expressed as:

$$L_{\text{CMB}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big((z_i \cdot z_i^{+})/\tau\big)}{\exp\big((z_i \cdot z_i^{+})/\tau\big) + \sum_{j\neq i} \exp\big((z_i \cdot z_j^{-})/\tau\big)} + \lambda\mathcal{R}, \tag{4.1}$$

where:

- z_i denotes the embedding of a sample,
- z_i^{+} the intra-subject embedding (same identity in another modality),
- z_j^{-} inter-subject embeddings (different identities),
- τ is the temperature parameter,
- $\lambda\mathcal{R}$ represents a regularization term.

The extracted embeddings are then used for both identification and verification. For identification, embeddings from the probe set are compared to those of the gallery set, and similarity is measured using cosine similarity. A k-fold cross-validation is applied, and the average rank-1 identification rate (IR) is calculated.

For verification, four random gallery images are selected and compared to a probe image. Verification performance is evaluated using metrics such as the ROC curve, EER, and AUC.

Chapter 5

Software Implementation

5.1 Introduction

As previously mentioned, this project investigates four modalities: periocular-to-periocular (p2p), face-to-face (f2f), periocular-to-face (p2f), and face-to-periocular (f2p). The identification and verification tasks for these modalities are implemented using logit vectors derived from the network.

5.2 Identification

Each dataset is divided into gallery and probe sets. The subfolders named *gallery* are typically used for gallery samples, while subfolders such as *probe1*, *probe2*, or *occlude* are used for probes. However, in this project, the role of gallery and probe is interchangeable: a subfolder designated as “probe” may serve as the gallery, and vice versa. Thus, gallery and probe sets are randomly assigned from available images.

For evaluation, the Cumulative Match Characteristic (CMC) curve is used. After extracting logit vectors for gallery and probe sets, similarity scores are obtained by matrix multiplication. For each probe, the top-k most similar gallery samples are identified. A Boolean operation between predicted and true labels is then used to calculate the CMC at each rank k. Formally, the rank-k identification rate is defined as:

$$\mathrm{IR}(k) = \frac{1}{N}\sum_{i=1}^{N} \nVdash \big[y_i \in \text{Top-}k(\hat{y}_i)\big],$$

where N is the number of probes, y_i the true label, and $\hat{y}_i$ the predicted ranking list. The average of IR values across all folds is reported.

5.3 Verification

Verification also requires multiplication of logit vectors, resulting in a similarity score matrix. If the folder contains n identities, the resulting matrix is of size $(4n \times 4n)$, since four images per identity are considered. The diagonal block submatrices (size 4×4) correspond to same-identity comparisons, with similarity scores typically close to 1.

A label matrix is defined such that elements corresponding to same-identity pairs are 1 and all others are 0. Using this label matrix, the similarity values are separated into positive and negative sets. These values are then passed to the ROC function (e.g., in `sklearn`), producing the False Positive Rate (FPR) and True Positive Rate (TPR). The Equal Error Rate (EER) is computed at the operating point where FPR = 1 − TPR.

5.4 KL Divergence and Computation Improvement

While cosine similarity is the default similarity measure, p2f and f2p tasks were also evaluated using Kullback–Leibler (KL) divergence, a measure of the difference between two probability distributions. For two discrete distributions P and Q with probability mass functions p and q, the KL divergence is:

$$D_{\mathrm{KL}}(P \| Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{5.1}$$

A naive implementation computes KL divergence pairwise using for-loops, which is computationally expensive. For example, identification using cosine similarity takes about two hours, whereas the same task with KL divergence was estimated to take more than one day.

To reduce computation time, matrix broadcasting is used to compute KL divergence values simultaneously, rather than iterating through pairs. This optimization reduces the runtime by approximately an order of magnitude (around 10 times faster), making KL-based evaluation feasible for larger datasets.

Chapter 6

Datasets

This project employed six types of image datasets: *ethnic*, *pubfig*, *facescrub*, *imdb_wiki*, *ar*, and *ytf*. Each dataset contains face and periocular images from diverse individuals, including Caucasian, African, and Asian subjects.

The first four datasets (*ethnic*, *pubfig*, *facescrub*, and *imdb_wiki*) contain both face and periocular images. The *ar* dataset includes specialized subfolders with modified images:

- **gallery:** Original facial images of subjects.
- **blur:** Images degraded with blurring effects.
- **occlude:** Images partially covered with black squares.
- **scarf:** Images covered with a red scarf over the lower half of the face.

The final dataset, *ytf* (YouTube Faces), contains a large number of images per identity. However, its resolution is relatively low compared to the other datasets, which makes recognition more challenging.

In summary, these datasets were chosen to evaluate the robustness of the proposed method under diverse conditions, including ethnic variability, image degradation, occlusion, and low resolution.

Chapter 7

Results

7.1 Periocular-to-Periocular

Table 7.1 summarizes the performance of the periocular-to-periocular modality for each dataset. The results show that most datasets achieve high performance, except for *ytf*, where the low image resolution results in a substantial drop in accuracy.

Table 7.1: CMC and EER for periocular-to-periocular (p2p) modality

Dataset	CMC (%)	EER (%)
ethnic	95.594	6.119
pubfig	97.449	5.564
facescrub	97.401	2.835
imdb_wiki	83.904	6.588
ar	96.045	7.128
ytf	63.490	15.041

7.2 Face-to-Face

Table 7.2 presents the performance of the face-to-face modality. As with the periocular-only results, most datasets show strong performance, while *ytf* performs worse due to lower resolution.

Table 7.2: CMC and EER for face-to-face (f2f) modality

Dataset	CMC (%)	EER (%)
ethnic	96.525	5.321
pubfig	98.157	4.654
facescrub	97.917	2.428
imdb_wiki	86.461	5.748
ar	92.503	7.672
ytf	71.272	13.719

7.3 Periocular-to-Face

The periocular-to-face (p2f) modality results are shown in Table 7.3. As before, *ytf* shows a lower performance compared to other datasets.

Table 7.3: CMC and EER for periocular-to-face (p2f) modality

Dataset	CMC (%)	EER (%)
ethnic	95.309	5.718
pubfig	97.456	5.094
facescrub	97.392	2.714
imdb_wiki	83.454	6.062
ar	87.640	8.865
ytf	65.000	14.339

7.4 Face-to-Periocular

Table 7.4 shows the results for the face-to-periocular (f2p) modality. The performance trends are consistent with those observed in other modalities.

Table 7.4: CMC and EER for face-to-periocular (f2p) modality

Dataset	CMC (%)	EER (%)
ethnic	95.400	5.717
pubfig	97.050	5.094
facescrub	97.058	2.714
imdb_wiki	82.742	6.062
ar	90.989	8.865
ytf	63.816	14.339

7.5 KL Divergence

The results for KL divergence–based similarity are presented in Tables 7.5 and 7.6, corresponding to temperature parameters $T = 1$ and $T = 2.5$, respectively. While CMC values remain similar, the EER shows noticeable variation depending on both the similarity metric and the temperature parameter. Generally, cosine similarity outperforms KL divergence, and lower temperature values yield better EER.

Cosine similarity works better here mainly because the network is trained to separate identities using inner products between embeddings (see the CMB loss), so cosine similarity directly matches that geometry, is scale-invariant, and is robust to variations in feature norm. In contrast, KL divergence assumes the vectors are proper probability distributions, requires extra normalization, is asymmetric, and is very sensitive to small noisy components, which distorts the learned embedding space and leads to worse verification performance (higher EER).

Table 7.5: CMC and EER with KL divergence ($T = 1$)

Modality	CMC (%)	EER (%)
p2f (ethnic)	95.309	5.754
p2f (ar)	87.640	9.174
f2p (ethnic)	95.400	5.641
f2p (ar)	90.989	9.321

Table 7.6: CMC and EER with KL divergence ($T = 2.5$)

Modality	CMC (%)	EER (%)
p2f (ethnic)	95.309	6.659
p2f (ar)	87.640	9.546
f2p (ethnic)	95.400	6.496
f2p (ar)	90.989	9.759

Chapter 8

Discussion and Conclusion

8.1 Summary of Findings

This study demonstrated that the periocular region is a reliable biometric trait under mask-wearing conditions, where conventional face recognition systems often fail. The proposed Conditional Multimodal Biometrics (CMB) framework successfully integrates periocular and face modalities, achieving improved recognition performance. The newly introduced CMB loss further enhanced representation quality by pulling intra-subject embeddings closer and pushing inter-subject embeddings apart. While KL divergence was tested as an alternative similarity metric, cosine similarity generally produced superior and more efficient results.

8.2 Limitations

Despite promising results, several limitations remain. First, the datasets used (particularly YTF) suffer from low resolution, which negatively affected accuracy. Second, although KL divergence was optimized with broadcasting, its computation remains slower than cosine similarity. Finally, the experiments were limited to 2D image settings, making direct extension to 3D or video-based applications non-trivial.

8.3 Future Work

Future research can expand this work in multiple directions:

- Extending periocular–face recognition to 3D and video-based settings.
- Developing lightweight architectures for deployment on mobile and embedded devices.
- Evaluating robustness under more challenging real-world scenarios, such as extreme pose variation, lighting changes, and occlusions beyond masks (e.g., sunglasses, helmets).
- Exploring advanced metric learning methods, such as ArcFace or CosFace, in combination with CMB loss.

In conclusion, this project shows that periocular recognition combined with end-to-end face recognition is a practical solution to the challenges posed by mask-wearing environments. Conditional Multimodal Biometrics provides a promising direction for developing robust and flexible recognition systems.

Reference

1. Renu Sharma, Arun Rossa, *Periocular Biometrics and its Relevance to Partially Masked Faces.*

2. H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, *The Elements of End-to-end Deep Face Recognition: A Survey of Recent Advances.*

3. Tiong-Sik Ng, Cheng-Yaw Low, Jacky Chen Long Chai, and Andrew Beng Jin Teoh, *Conditional Multimodal Biometrics Embedding Learning for Periocular and Face in the Wild.*

4. Florian Schroff, Dmitry Kalenichenko, James Philbin, *FaceNet: A Unified Embedding for Face Recognition and Clustering.*

5. Hongxia Deng, *MFCosface: A Masked-Face Recognition Algorithm Based on Large Margin Cosine Loss.*

6. Constantino Álvarez Casado, *Real-time face alignment: evaluation methods, training strategies and implementation optimization.*