

Mobile based Human Identification using Forehead Creases: Application and Assessment under COVID-19 Masked Face Scenarios

Rohit Bharadwaj^a, Gaurav Jaswal^b, Aditya Nigam^c, and Kamlesh Tiwari^a

^a Dept. of CSIS, Birla Institute of Technology and Science Pilani, Jhunjhunu-333031, Rajasthan, INDIA

^b Dept. of EE, Indian Institute of Technology Delhi, Hauz Khas-110016, New Delhi, INDIA

^c SCEE, Indian Institute of Technology Mandi, Mandi-175005, Himachal Pradesh, INDIA

{f20170633,kt}@pilani.bits-pilani.ac.in, gauravjaswal@ee.iitd.ac.in, aditya@iitmandi.ac.in

Abstract

In the COVID-19 situation, face masks have become an essential part of our daily life. As mask occludes most prominent facial characteristics, it brings new challenges to the existing facial recognition systems. This paper presents an idea to consider forehead creases (under surprise facial expression) as a new biometric modality to authenticate mask-wearing faces. The forehead biometrics utilizes the creases and textural skin patterns appearing due to voluntary contraction of the forehead region as features. The proposed framework is an efficient and generalizable deep learning framework for forehead recognition. Face-selfie images are collected using smartphone's frontal camera in an unconstrained environment with various indoor/outdoor realistic environments. Acquired forehead images are first subjected to a segmentation model that results in rectangular Region Of Interest (ROI's). A set of convolutional feature maps are subsequently obtained using a backbone network. The primary embeddings are enriched using a dual attention network (DANet) to induce discriminative feature learning. The attention-empowered embeddings are then optimized using Large Margin Cosine Loss (LMCL) followed by Focal Loss to update weights for inducing robust training and better feature discriminating capabilities. Our system is end-to-end and few-shot; thus, it is very efficient in memory requirements and recognition rate. Besides, we present a forehead image dataset (BITS-IITMandi-ForeheadCreases Images Database¹) that has been recorded in two sessions from 247 subjects containing a total of 4,964 selfie-face mask images. To the best of our knowledge, this is the first to date mobile-based forehead dataset and is being made available along with the mobile application in the public domain. The proposed system has achieved high performance results in both closed-

¹Database can be accessed from <http://ktiwari.in/projects/foreheadcreases/>.

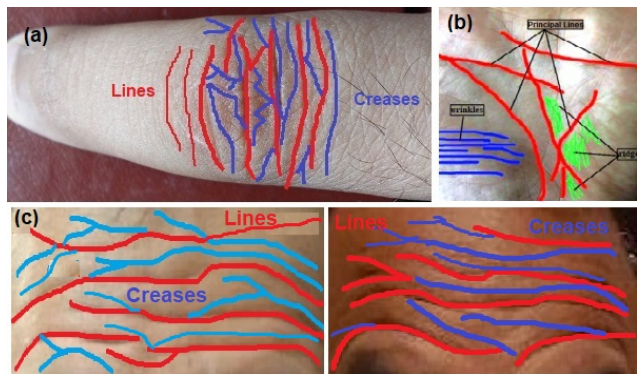


Figure 1: Examples images of (a) finger knuckle print, and (b) palmprint are showing similar line and crease-patterns to (c) forehead creases that indicate the addition of forehead patterns for biometric recognition could address the challenges in masked face scenarios.

set, i.e., CRR of 99.08% and EER of 0.44% and open-set matching, i.e., CRR: 97.84%, EER: 12.40% which justifies the significance of using forehead as a biometric modality.

1. Introduction

The most successfully adopted biometrics to unlock smartphones are face selfie (introduced with Android 4.0), iris scanner (introduced with Galaxy Note 7), and fingerprint touch (introduced with Apple's iPhone 5s) mechanisms [2]. Additionally, commercial solutions for mobile biometric recognition based on inbuilt smartphone sensors are already available. Since COVID-19 can spread through contact and contaminated surfaces, face, iris, voice, and fingerprint-based classical biometric methods have undergone serious challenges. These are 1) accessing face authentication when a mask hides a major part of the face can prevent us from unlocking our smartphone, 2) touching fingerprint scanner might increase the possibility of contamination, thereby triggering infectious diseases. 3) iris au-

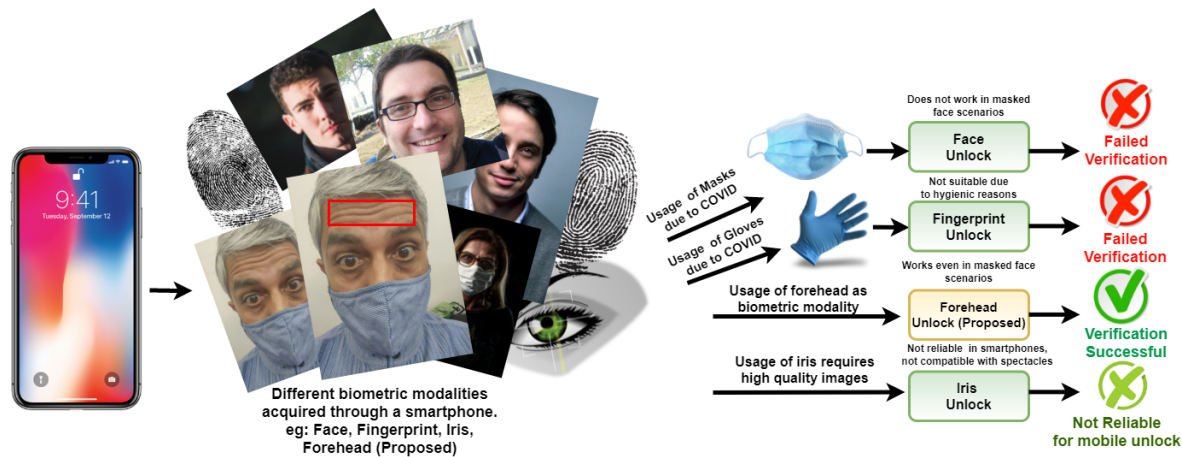


Figure 2: Overview of challenges for reliable biometric recognition and suitability of forehead crease patterns for smartphone authentication under COVID-19 masked face scenarios.

thentication seems more dependent on high-quality images. Unlike face recognition [17], periocular [1] and voice [16] may be utilized as alternate recognition systems, however, they are not very reliable when substantial portions of the face are occluded. Some recent studies have shown consequences of face coverings upon the quality of speech [5]. Face mask and low-quality images might present an obstruction to reliable segmentation of periocular region [18]. Due to the problems mentioned above, there is a reinforced requirement to look for alternate biometric identifiers that could be utilized in the COVID-19 era, particularly regarding wearing face masks. Aiming towards this, our work investigates forehead creases (under facial expressions) as a biometric trait that can give competitive results compared to the face and periocular in the COVID-19 era when masks must be worn by everyone and where touchable interactions must preferably be avoided.

Related Work and Challenges Before turning to the technical description of our approach, we first provide background on prior studies in forehead recognition. To the best of our knowledge, forehead crease patterns are a significantly less explored biometric modality in the literature so far. The first-ever work in forehead recognition was presented by [11], but the purpose is not helpful for smartphone-based applications. Their system acquired forehead images using a complex imaging system based on near-infrared laser scanning. Also, their setup is coped up with hygienic concerns, sensor surface noise, and user inconvenience as one has to maintain a head posture and then put the head on a surface for data collection, thus not suitable in COVID-19 applications. Therefore, the approach detailed in this reference is a baseline work without many optimizations of forehead patterns for smartphone-based user recognition. It is worth mentioning that acquiring forehead images using smartphones offers higher user convenience

and contactless imaging aspects. In such forehead images, skin creases and lines are the significant sources of information which are observed to be distinct and permanent for each individual due to the differences in frontal bone morphology and thickness of the skin tissues [12].

We assume and hypothesize that forehead creases have similar image properties to finger knuckle print [7, 14] and palmprint [13, 19] which are well-established traits in establishing human identities. Figure 1 indicates the anatomical view of forehead patterns that has full potential to be considered as biometric modality as similar to palmprint and knuckle print. However, the uniqueness of such wavy and horizontal lines of varying thickness is insufficient to establish the personal identity in non-cooperative conditions or recognition at a distance. Moreover, contactless imaging via smartphone often introduces deformation of forehead patterns due to pose, scale, and illumination; thus, a full potential from the forehead biometrics is yet to be realized. Figure 2 illustrates the need for forehead-based smartphone recognition systems when other well-known biometrics are causing severe challenges due to facial masks and urgent hygienic concerns.

1.1. Our Work and Contribution

To the best of the authors' knowledge, this is the first work motivated to advance the smartphone-based forehead recognition capabilities in masked face scenarios, especially in the context of the COVID-19 pandemic. Our mobile application requires bare minimum hardware and memory requirements, i.e., an android smartphone device and 5 MB of space; thus, there is increased scope of reproducibility. Our dataset comprises 4,964 image examples taken in two different sessions where we have roughly 20 image poses per 247 subjects. However, we assume that the mobile app user is cooperative during contactless forehead imag-

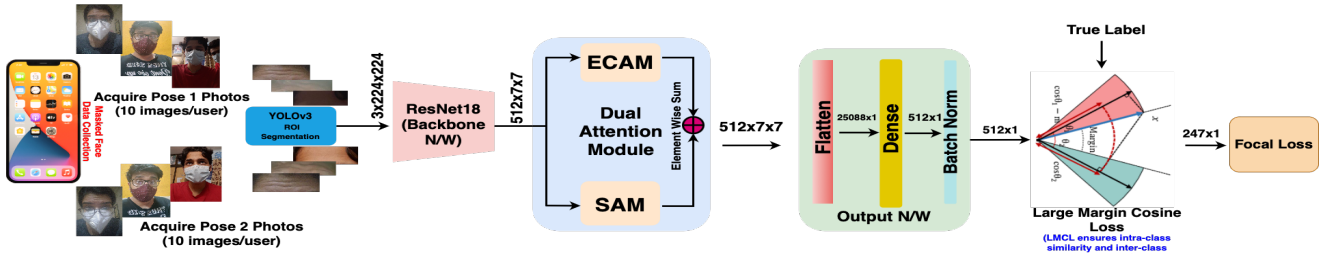


Figure 3: Block diagram of the proposed forehead recognition network assisted with dual attention network. The network performs discriminative feature learning in cosine space that has richer embedding representation than state-of-art networks.

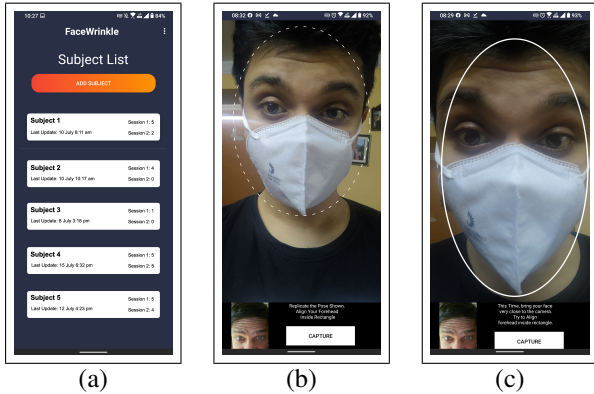


Figure 4: Data collection app. (a) interface, (b) Pose-1 from a fixed distance, (c) Pose-2 with camera close to face.

ing, and such cooperation is often expected from fingerprint, iris, or other contactless biometrics systems.

Contributions: Our work² describes a complete forehead segmentation and recognition pipeline using an end-to-end deep metric learning approach whose detailed architecture is depicted in Figure 3. To this end, the following main aspects are considered: 1) how to achieve a consistent forehead localization under multiple spatial scales, varying poses, occlusion, and illumination? 2) how to obtain a computationally efficient yet effective baseline CNN model for Mobile based forehead recognition? 3) how to take advantage of the metric learning domain to improve the recognition performance in data-scarce conditions?

Contactless data collection with varying scale, pose, occlusion, and discrete crease patterns raise several forehead RoI segmentation and recognition challenges. This paper introduces a challenging masked face dataset and attempts to work on the challenges above. The key contributions of this paper are summarised in the following points:

1. To the best of our knowledge, our work is the first smartphone-based biometric recognition approach with forehead images taken under masked face scenarios.
2. We present deep metric learning using Large Margin Cosine Loss that is highly optimized for extracting dis-

criminative features for forehead recognition.

3. Incorporating a dual attention mechanism that considers spatial attention to learn semantic regularities and channel attention to compute the correlation between all the channels independently.
4. Devised an intelligent data acquisition strategy using the front camera of mobile devices. The samples are acquired remotely in two sessions, and our android application facilitates the process by providing a screen GUI.
5. Developed a forehead-image database consisting of 4,964 selfie masked faces and corresponding cropped ROI's captured from 247 subjects.
6. Similar to face recognition methods, the proposed framework has been extended to provide open-set recognition performance.

We demonstrate forehead-image recognition performance using standard evaluation measures: correct recognition rate (CRR), equal error rate (EER), decidability index (DI), and the Receiving Operating Characteristic (ROC) curve. Besides, we used the Jaccard index (Intersection Over Union), precision, and recall to measure the performance of RoI segmentation.

The rest of the paper is structured into four main sections. Section II describes the image acquisition setup via mobile app and detail about the dataset. Section III presents the methodology of our approach and includes the details of the network and training considered in this work. Section IV details experiments, matching protocol employed for performance evaluation. Key findings and discussions are summarized in the last section.

2. Forehead Image Acquisition

We use a non-contact and remote imaging setup to acquire masked face samples in non-uniform lighting and changing background conditions. Keeping in mind the significance of isolation/social distancing, a special android application is developed to ease the smartphone-based data-acquisition process remotely. The data is taken in two sessions, wherein a subject must provide five photographs of the two poses in each session. The above two poses are

²Code is available at:

<https://github.com/rohit901/ForeheadCreases>

Subjects	Accuracy (%)	Precision	Recall	IOU
15	72.80	0.76	0.85	0.78
30	73.11	0.77	0.85	0.79
45	73.84	0.77	0.86	0.79

Table 1: Segmentation performance under different training datasets.

taken into consideration to account for the variance due to distance. The Android app GUI first allows users to bring their faces into a guided bounding box to keep consistency in face postures across the dataset, as shown in Fig 4 (b) and Fig 4 (c). Pose 1 requires a fixed distance between the camera and the phone, while pose 2 requests the user to close the gap between the camera and the phone. The user has to make a surprise reaction and lift their eyebrows exactly like shown in Fig 4 (b) & (c), for making the forehead wrinkle patterns to be visible. Most importantly, session two photographs are taken after a minimum time gap of 1 day to ensure the temporal stability of forehead patterns. In this way, each subject has provided 20 ($5samples \times 2poses \times 2sessions$) photographs. The mobile application is designed as a client-server model; thus, all the data is collected remotely in an unconstrained environment.

3. Proposed Approach

In this section, we first summarize the whole pipeline of a forehead recognition system and then elaborate on the details of the proposed approach, including segmentation and backbone network with dual attention mechanism assisted and simple yet effective large margin cosine loss [21]. We argue that rather than designing and training new CNNs for forehead recognition, using CNNs whose architectures have already been proven successful in large-scale datasets can yield good performance without tedious architecture steps. The pipeline of a complete forehead recognition system is illustrated in Figure 3, including forehead RoI segmentation, feature extraction while focusing on discriminative patterns and matching. Without loss of generality, we use Yolo v3 for segmentation task, ResNet18 is chosen as our backbone network driven by dual attention (DANet) ([22], [4]) and LMCL [21] is used as loss function. However, there is a need to use attention mechanisms further to enhance feature discrimination in both spatial and channel dimensions. The dual attention mechanism is incorporated into the backbone network to discover more discriminative feature representation for forehead images than the standard feature learning.

3.1. Extracting the Region of Interest

Our non-ideal forehead dataset’s inherent problems, which we have mainly focused on, are changing face postures, blurred boundaries, complex backgrounds, illumina-

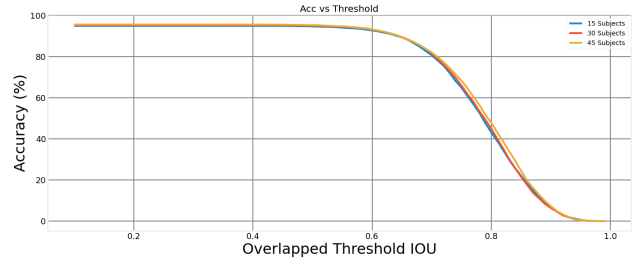


Figure 5: Forehead segmentation accuracy on different training datasets with respect to threshold on IOU

tion variation, and occlusions by hairs, glasses, and dots worn on foreheads. In Fig 6, some of the challenging images containing occlusion, illumination, etc., are highlighted.

Since the data was taken remotely due to the lockdown restrictions, many users could not provide good quality face data. After locating the raw forehead regions via app GUI, one can use any state-of-the-art detection networks for RoI segmentation. We deployed a simple yet effective pre-trained YOLOv3 model [15] (which initially was trained on Imagenet) on around 600 raw forehead images from 30 subjects. Using the trained weights from the YOLOv3 model, we obtained bounding boxes over the rest of the 4, 340 forehead images from 217 subjects. Each subject has a maximum of 20 images, and a few have less than 20 images because all the detections whose confidence score is less than 0.6 were dropped. The state-of-the-art evaluation parameters like Jaccard index, accuracy, precision, and recall are computed for all the studied cases and are shown in Fig 5 & Fig 7. The mean values of these metrics on IOU thresholds ranging from 0.1 to 0.99 with a step size of 0.01 are displayed in Table 1. Figure 6 (a) and (b) show a few challenging masked face images and corresponding forehead ROI’s where our network performs well, while Figure 6 (c) and (d) shows examples of forehead images suffered from various type of noise factors (occlusion, illumination, glasses, etc.) and corresponding poor quality segmentation results. Figure 7 shows the precision-recall graph for different training data sets, where bars and dashed lines represent precision and recall, respectively. Up to $IOU \leq 0.7$, the precision and recall of the segmentation network over all the training data sets remains very high and decreases significantly after $IOU = 0.8$.

3.2. Backbone Network

Given a cropped forehead image $x^{(i)} \in \mathbb{R}^{C \times H \times W}$ as input, which is obtained from the segmentation network as described in the previous section, the backbone network outputs the feature map $E^{(i)} \in \mathbb{R}^{C' \times H' \times W'}$. Mathematically,

$$E^{(i)} = f(x^{(i)}; \theta^{(1)}) \quad (1)$$



Figure 6: Masked face and corresponding forehead ROI using the segmentation network.

where $\theta^{(1)}$ denotes all the parameters of the backbone network, C denotes the number of channels of the input image ($C = 3$ for colored images), H denotes the height of the image, W denotes the width of the image. Similarly, C' , H' , W' denotes the feature maps' number of channels, height, and width, respectively. Since our in-house dataset is relatively small, we decided to use an efficient DNN with comparatively fewer parameters to prevent overfitting on training data. Also, to address the vanishing gradients problem, without any loss of generality, we have considered ResNet-18 [6] as our backbone network f , which is the simplest yet effective DNN architecture.

The number of parameters or the cardinality of the set $\theta^{(1)}$ comes out to be around 22 million. The feature map $E^{(i)}$ is extracted from the conv5_x layer of ResNet-18, and it is further passed as input to our dual attention submodule, which is described in the next section.

3.3. Dual Attention Module (DAM)

In order to increase the discriminatory power and select leading forehead features, an attention mechanism has been utilized to guide the network to automatically select regions that can most contribute to an accurate matching. Since forehead ROI's have a lot of background in the form of eyebrows, acne, hair, and sometimes even makeup, it leads to distorted features and can reduce the discriminatory power. To ensure that the backbone network is not confused by such unimportant features and only learn

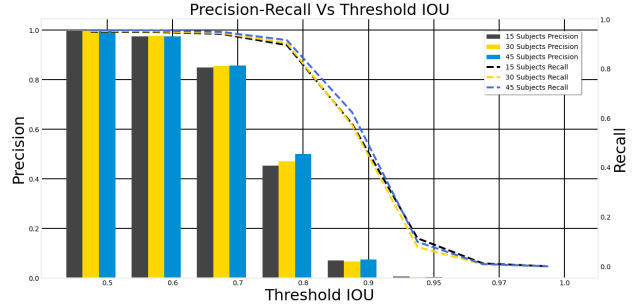


Figure 7: Precision/recall for different data sets w.r.t. IOU

domain-specific forehead patterns, we employ a dual attention module (DAM) mechanism to jointly model the interdependencies of spatial and channel level forehead features. DAM applies spatial attention module (SAM) and channel attention module (CAM) independently over the outputs of the last convolution layer of the backbone network.

The goal of the spatial attention module is to aggregate the semantically similar pixels in the spatial domain of the input feature map. Though the spatial attention mechanism attends to the entire input feature map based on content (pixel values), it does not consider the spatial positions of the pixel.

Spatial Attention Module: Forehead images show plenty of spatial variations of lines and wrinkles that extends from one end of the image to the other. The spatial attention module (SAM) [4] allows us to capture the spatial semantic regularities present in the forehead patterns. Given the output feature map $E^{(i)}$ obtained from the backbone network, it is then passed through SAM to emphasize the spatial dependencies. This module can be represented as $A_{SAM}(E^{(i)}; \theta^{(2)}) \in \mathbb{R}^{512 \times 7 \times 7}$, where $\theta^{(2)}$ denotes the weights of SAM module. Mathematically,

$$F_1^{(i)} = E^{(i)} \otimes A_{SAM}(E^{(i)}; \theta^{(2)}) + E^{(i)} \quad (2)$$

where $F_1^{(i)}$ is the final spatial highlighted features.

Channel Attention Module: To capture the channel-wise relationship of features obtained from the backbone network, it is important to give higher significance to essential feature maps and thereby reduce the prominence of redundant or unnecessary information. This further enables more stable training. To solve the same purpose, we deployed ECANet [22] as our channel attention module. This module can be implemented easily by 1D convolution of size F ($F = 5$) and only introduce a total of 6 training parameters. This module can be represented as $A_{CAM}(E^{(i)}; \theta^{(3)}) \in \mathbb{R}^{512 \times 7 \times 7}$, where $\theta^{(3)}$ denotes the weights of CAM module. The output from this module is multiplied with the extracted feature map $E^{(i)}$ to get channel-wise highlighted features $F_2^{(i)}$ denoted as:

$$F_2^{(i)} = E^{(i)} \otimes A_{CAM}(E^{(i)}; \theta^{(3)}) \quad (3)$$

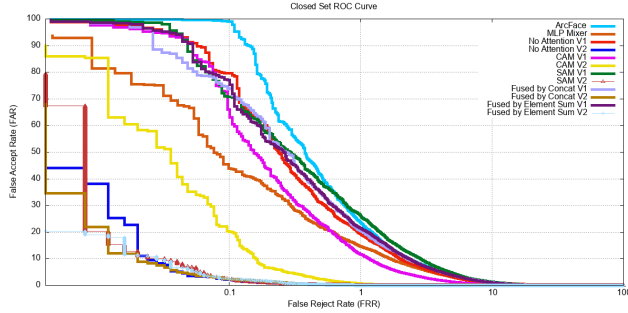


Figure 8: ROC curves for the closed set experiments. The x-axis is shown in log scale.

3.4. Feature Fusion

To effectively fuse the attention outputs obtained from CAM and SAM, we have considered the following two scenarios.

First, we decided to concatenate the outputs obtained from the two attention modules. Mathematically,

$$O^{(i)} = F_1^{(i)} \parallel F_2^{(i)} \quad (4)$$

where $O^{(i)} \in \mathbb{R}^{2C' \times H' \times W'}$ is the feature map obtained after concatenating the two attention module outputs, and \parallel denotes the concatenation operator.

Second, we performed an element wise addition between the two attention outputs. Mathematically,

$$O_j^{(i)} = F_{1j}^{(i)} + F_{2j}^{(i)} \quad (5)$$

where, $F_{1j}^{(i)}$ is the j -th element of $F_1^{(i)}$, $F_{2j}^{(i)}$ is the j -th element of $F_2^{(i)}$, $O_j^{(i)} \in \mathbb{R}^{C' \times H' \times W'}$ is the j -th element of output feature map. Thus, fusing the attention outputs from SAM and CAM, we can aggregate similar but less noticeable forehead wrinkle patterns to highlight their feature representation and reduce the influence of salient features like forehead skin, background, etc. The fused output O is then flattened to 1-dimension, passed into a fully-connected layer followed by a batch-normalization layer to obtain the final output feature vector $I^{(i)} \in \mathbb{R}^{512}$, which represents the embedding of the input forehead RoI image $x^{(i)}$. In this way, we have jointly modeled the spatial and channel-wise feature relationships, and this process can be observed from Figure 3.

3.5. Loss Function and Network Training

Large Margin Cosine Loss (LMCL) [21] is utilized for training the entire network in an end-to-end fashion. Training our model using this loss function results in learning features that are distinguishable in the cosine space. To have less intra-class variation (variation within same subject due to different poses, illumination, etc.) and more inter-class variations (variation between two different subjects),

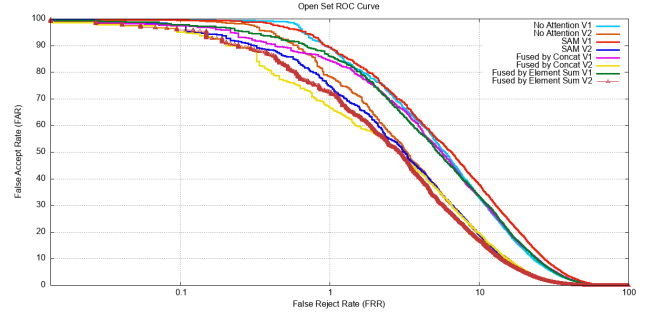


Figure 9: ROC curves for the open set experiments. The x-axis is shown in log scale.

$m \geq 0$ is set as enforced margin so that features from different subjects are correctly matched. Given a set of two feature embeddings, the margin constrained can be defined as:

$$\cos \theta_1 - m > \cos \theta_2 \quad (6)$$

$$\cos \theta_2 - m > \cos \theta_1 \quad (7)$$

As our classification is binary, i.e., genuine and imposter, Eq 6 is the condition for matching within class 1 (genuine) and Eq 7 is the condition for matching within class 2 (imposter). Thus, based on margin constrained, the LMCL loss function can be formulated as:

$$L = \frac{1}{N} \sum_i -\log \frac{e^{s(\cos(\theta_{y_i}, i) - m)}}{e^{s(\cos(\theta_{y_i}, i) - m)} + \sum_{j \neq y_i} e^{s \cos(\theta_j, i)}} \quad (8)$$

where, $\cos(\theta_j, i) = W_j^T O^{(i)}$ and W_j is the weight vector of the j -th class.

Network training: Our dataset of 247 subjects was divided roughly equally into train and test sets. We have set LMCL parameters $m = 0.35$, $s = 300$, and used feature embedding dimension as 512. The output obtained from LMCL is a vector of class probabilities. To facilitate back-propagation, output from LMCL is further passed to the focal loss [10] module with $\gamma = 2$. Additionally, the model weights are updated using adam optimizer [9], initial learning rate is set to $lr = 3 \times 10^{-4}$ and it is decayed by $\gamma = 0.1$ at every 20 epochs. We trained the network for 100 epochs with $L2$ weight penalty of $\lambda = 5 \times 10^{-4}$.

4. Experimental Results

This section describes matching protocols, ablation study, and comparative methods used for performance evaluation.

4.1. Matching Protocol

We used our trained model to evaluate various performance parameters to validate our approach. The image

	Subject Count	Matching Strategy	CRR (%)	EER (%)	DI
Model V1	75	Open Set	96.08	17.50	1.44
Model V2	75	Open Set	97.84	12.40	1.66
Model V1	247	Closed Set	97.22	4.46	1.91
Model V2	247	Closed Set	99.08	0.44	2.79

Table 2: Closed Set and Open Set System performance on forehead dataset.

present in our train set is used as a gallery, while the image in the test set is used as a query to perform matching. Each query image is matched with each of the gallery images. The trained model is used during inference to obtain the embeddings from the gallery and query image set and match those embeddings. The embedding dimension is set to 512. Basically, a matching score is computed between a query and a gallery embedding by using L^2 (or the Euclidean) distance. A score between two embeddings is termed as **genuine matching** if they are obtained from the same subjects; otherwise, it is an **imposter matching**. We expect the matching score for genuine matching to be very low, while the matching score for imposter matching to be higher because the learned metric distance is a measure of dissimilarity. In this task, we perform the following two kinds of matching experiments: closed set matching and a hard matching strategy, i.e., open set matching as described below.

Closed Set Matching: In this matching strategy, we consider all 247 subjects, and we divide the entire dataset almost equally into training (gallery) and testing set (query). So, we have 2462 images in the gallery and 2502 images in the query (one subject has more than 20 images). In this manner, we get 22768 genuine matchings and 5565663 imposter matchings. We evaluate matching performance on two kinds of models. First, **Model V1**, which is our proposed network trained on the non-augmented dataset. Second, **Model V2**, which is our proposed network trained on the augmented dataset.

Data Augmentation. For data augmentation, we sampled 5 points randomly from each of 45×45 , 47×47 , 49×49 , and 51×51 neighborhoods around the center of the original bounding box of each ROI. We generated new ROI’s using these sampled points as the center of the new bounding boxes while keeping the height and width of the boxes unchanged. Thus, the dataset is augmented by 20 times, which is $2462 \times 20 = 49,240$ images. However, it is important to note that the data augmentation process is done only for the training set, and **Model V2** is then trained on augmented data.

Using the trained model’s weights, we evaluated matching performance on our standard non-augmented dataset as described above. All the details about performance parameters such as correct recognition rate (CRR), equal error rate (EER), decidability index (DI) for the closed set matching can be observed from Table 2. The respective ROC plots

Model	Scenario	Subjects/ Poses	CRR %	EER %
V1 (No attention)	closed set	247	97.64	4.18
V2 (No attention)	closed set	247	99.36	0.38
V1 (No attention)	open set	75	96.92	17.06
V2 (No attention)	open set	75	97.44	12.98
V1 (Fused by Element Wise Sum)	open set	75	96.08	17.50
V2 (Fused by Element Wise Sum)	open set	75	97.84	12.40
V1 Fused by Concatenation	open set	75	95.52	17.45
V2 Fused by Concatenation	open set	75	97.71	13.22
V1 (SAM)	open set	75	93.43	19.19
V2 (SAM)	open set	75	96.90	12.95
V1 (SAM)	closed set	247	97.00	4.88
V2 (SAM)	closed set	247	99.32	0.39
V1 (CAM)	closed set	247	98.10	3.02
V2 (CAM)	closed set	247	99.48	0.90
V1 (Fused by Concatenation)	closed set	247	96.96	4.47
V2 (Fused by Concatenation)	closed set	247	99.24	0.48
V1 (Fused by Element Wise Sum)	closed set	247	97.22	4.46
V2 (Fused by Element Wise Sum)	closed set	247	99.08	0.44

Table 3: Ablation study of network and its components using different matching protocols.

for this experiment are shown in Figure 8. On account of results, the following observations can be drawn: 1) Model V2 achieves the best individual performance in terms of both CRR (99.08%) and EER (0.44%). This is an excellent performance, and it is getting close to iris [23], finger knuckle [8] biometrics systems. 2) Considering data augmentation, the network is allowed to see different variations of forehead patterns which drastically increases the generalisability of the network.

Open Set Matching: In this experiment, we consider a challenging matching strategy where 30% subjects are excluded from training. More specifically, our proposed model is first trained after removing 75 subjects from our dataset. Thus, the model is trained only on 172 subjects by splitting the dataset equally into training and testing sets. This trained model is then used to generate matching scores on the remaining unseen 75 subjects. In particular, all the images from the remaining unseen 75 subjects are split equally into gallery and query. Therefore, we consider 744 images in the gallery and 742 images in the query set. In this manner, a total of 7386 genuine matchings and 544662 imposter matchings are obtained. We again evaluate matching performance on two kinds of models: **Model V1** (trained on the non-augmented dataset), and **Model V2** (trained on the augmented dataset). The Data augmentation process remains the same as described earlier. Only the training set images from 172 subjects are augmented, and the model V2 is trained on this augmented dataset. The trained model’s weights are used to generate matching scores on the remaining unseen 75 subjects to evaluate the matching performance. This unseen dataset is the same as used for Model V1. The performance parameters for the open set matching can be observed from the Table 2, and respective ROC plot can be seen from Figure. 9 On account of results, the following necessary observations can be realized: 1) Model V2 has better performance than Model V1 in both CRR(97.84% vs. 96.08%) and EER (12.40% vs. 17.50%), which justifies the importance of doing data aug-

Approach	Dataset Statistics	Testing Protocol	CRR %	EER %
ResNet18 (Work by [11])	30 Subjects, 270 images	Validation Accuracy	94.47	NA
ResNet101 (Work by [11])	30 Subjects, 270 images	Validation Accuracy	98.55	NA
MLP Mixer [20] (Our Work)	247 Subjects, 4964 images	Standard Matching	92.31	4.08
ArcFace Loss [3] (Our Work)	247 Subjects, 4964 images	Standard Matching	97.47	4.54
Proposed (element sum fusion)	247 Subjects, 4964 images	Standard Matching	97.22	4.46
Proposed (concat fusion)	247 Subjects, 4964 images	Standard Matching	96.96	4.47
Proposed (SAM only)	247 Subjects, 4964 images	Standard Matching	97.00	4.88
Proposed (CAM only)	247 Subjects, 4964 images	Standard Matching	98.10	3.02
Proposed (element sum fusion) *	247 Subjects, 4964 images	Standard Matching	99.08	0.44
Proposed (concat fusion) *	247 Subjects, 4964 images	Standard Matching	99.24	0.48
Proposed (SAM only) *	247 Subjects, 4964 images	Standard Matching	99.32	0.39
Proposed (CAM only) *	247 Subjects, 4964 images	Standard Matching	99.48	0.90

Table 4: Comparison with other methods and settings. (* denotes model trained using augmented dataset.)

mentation. 2) Despite the challenging nature of open set matching, our model V2 still improves EER on augmented data compared to model V1, which signifies the strength of our network.

Overall, it is evident from Table 2 that performance on the open set testing protocol gets deteriorated compared to the closed set results. However, the CRR is very good, and it further increases our confidence in the generalisability of our models on unseen datasets.

4.2. Ablation Study

This section considers individual modules for our model architectures’ performance evaluation to better understand bottleneck network and attention mechanisms. The model architectures, i.e., Model V1 (non-augmented dataset) and Model V2 (augmented dataset) under various scenarios, are considered, and their matching results are shown in Table 3.

We first investigate how the attention criterion in terms of SAM and CAM independently affects the performance. Considering that forehead images are not that complex, a single attention mechanism might be sufficient to achieve good performance and generalisability. Experiment results on open set matching, as shown in 6th and 10th rows of Table 3 describe that our model with dual attention achieves better performance in comparison to any of the individual attention blocks. However, in the case of closed set matching, SAM shows comparable performance w.r.t. dual attention. From the above two observations, it can be stated that the dual attention mechanism provides more generalisability. The element-wise sum operation in our DAM does the better joint learning of spatial and channel-wise features, which justifies its higher accuracy than the concatenation operation in both closed and open set matching protocols. On the other side, CAM reports best individual results (CRR: 98.10, EER: 3.02) on non-augmented data compared to SAM (CRR: 97.00, EER: 4.88). Conversely, we see the reverse trend when there is sufficient training data available or when we are using an augmented dataset. The scenario SAM (CRR: 99.32, EER: 0.39) has the better performance because of its low EER as compared to CAM (CRR: 99.48, EER: 0.90). The *No Attention* scenario gives the best performance (CRR: 99.36, EER: 0.38) for augmented data, as can be seen from 2nd row in the case of closed set matching.

However, the same scenario performs poorly (CRR: 97.44, EER: 12.98) in open set matching as shown in 4th row of Table 3. Finally, one can conclude that to ensure model generalisability, as well as higher performance, the use of dual attention-based SAM and CAM and combining the attention outputs according to Eq 5, gives us better results.

4.3. Comparative Analysis

To the best of our knowledge, this is the first work reporting forehead crease patterns for Mobile based person recognition. The other paper [11], which is tangentially related to the scope of the current study (already been discussed in previous sections), cannot be directly compared due to the following reasons: 1) authors have reported their results using k-fold cross-validation; since they have not followed the standard matching protocol, there is no EER to report from their work. 2) authors have not reported any results on open set matching, and also, the dataset used by them is not publicly available. Moreover, due to the significantly fewer samples used in their experiments, the generalisability of their model cannot be inferred. To make comparative analysis more significant, we have computed experimental results on our dataset by using other well-known network architectures: MLP Mixer[20]; ArcFace loss[3]. From Table 4, one can observe that our models (Model V1, Model V2) trained using LMCL loss are performing better in all studied cases than the models trained using MLP Mixer and ArcFace.

5. Conclusion

This work investigates the usefulness of the forehead creases, under surprised facial expressions, as a biometric modality for smartphone-based user recognition. In the COVID-19 situation, when a person is mandated to wear a face mask and not touch surfaces, other biometric modalities such as face and fingerprint cause serious challenges. We device touch-less image acquisition using a mobile application and present a generalized deep learning framework with an attention-guided mechanism that further regularizes metric learning for better inter-class variability. The system is evaluated on a masked face dataset acquired from 247 subjects that contain 4,964 selfie images. Our proposed network reports high performance results in the open and closed set matching protocols: CRR: 99.08%, EER: 0.44% on the closed set, and CRR: 97.84%, EER: 12.40% on open set experiments. These outperforming results are comparable to the performance of iris, finger knuckle, and palmprint biometrics. Thus, we validate our assumption for considering forehead creases as a biometric modality to improve masked face scenarios.

References

- [1] Tiago de Freitas Pereira and Séastien Marcel. Periocular biometrics in mobile environment. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2015.
- [2] Alexander De Luca, Alina Hang, Emanuel Von Zezschwitz, and Heinrich Hussmann. I feel like i’m taking selfies all day! towards understanding biometric authentication on smartphones. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1411–1414, 2015.
- [3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [4] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.
- [5] David Haws and Xiaodong Cui. Cyclegan bandwidth extension acoustic modeling for automatic speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6780–6784. IEEE, 2019.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [7] Gaurav Jaswal, Amit Kaul, and Ravinder Nath. Knuckle print biometrics and fusion schemes—overview, challenges, and solutions. *ACM Computing Surveys (CSUR)*, 49(2):1–46, 2016.
- [8] Gaurav Jaswal, Aditya Nigam, and Ravinder Nath. Finger knuckle image based personal authentication using deep-matching. In *2017 IEEE international conference on identity, security and behavior analysis (ISBA)*, pages 1–8. IEEE, 2017.
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [11] Jirapong Manit, Luise Preuße, Achim Schweikard, and Floris Ernst. Human forehead recognition: a novel biometric modality based on near-infrared laser backscattering feature image using deep transfer learning. *IET Biometrics*, 9(1):31–37, 2019.
- [12] Jirapong Manit, Achim Schweikard, and Floris Ernst. Deep convolutional neural network approach for forehead tissue thickness estimation. *Current Directions in Biomedical Engineering*, 3(2):103–107, 2017.
- [13] Aditya Nigam and Phalguni Gupta. Designing an accurate hand biometric based authentication system fusing finger knuckleprint and palmprint. *Neurocomputing*, 151:1120–1132, 2015.
- [14] Aditya Nigam, Kamlesh Tiwari, and Phalguni Gupta. Multiple texture information fusion for finger-knuckle-print authentication system. *Neurocomputing*, 188:190–205, 2016.
- [15] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [16] Rahim Saeidi, Ilkka Huhtakallio, and Paavo Alku. Analysis of face mask effect on speaker recognition. In *Interspeech*, pages 1800–1804, 2016.
- [17] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 773–782, 2019.
- [18] Juan Tapia, Marta Gomez-Barrero, Rodrigo Lara, Andres Valenzuela, and Christoph Busch. Selfie periocular verification using an efficient super-resolution approach. *arXiv preprint arXiv:2102.08449*, 2021.
- [19] Kamlesh Tiwari, Devendra K Arya, GS Badrinath, and Phalguni Gupta. Designing palmprint based recognition system using local structure tensor and force field transformation for human identification. *Neurocomputing*, 116:222–230, 2013.
- [20] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.
- [21] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
- [22] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: efficient channel attention for deep convolutional neural networks, 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.
- [23] Zijiang Zhao and Ajay Kumar. Towards more accurate iris recognition using deeply learned spatially corresponding features. In *Proceedings of the IEEE international conference on computer vision*, pages 3809–3818, 2017.