Training Deep Neural Network Embeddings for Contactless Palmprint Recognition

Lukas Wirkestrand

Master's thesis
2024:E38

Lund Institute of Technology
Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Centrum Scientiarum Mathematicarum

***Abstract*—This study delves into the potential of contactless palmprints within large-scale biometric frameworks, focusing on improving candidate narrowing through an encoder-based approach. Utilizing deep neural networks and trained via semi-hard triplet learning, the research transforms palm images into distinctive feature vectors for precise identification and candidate selection. Comprehensive analysis involving various architectures, datasets, and preprocessing techniques achieved a closed-set rank 10 retrieval rate of 99.4% on the HandID and Tongji datasets. Additionally, the Average Number of Hands (ANH) metric was introduced for model comparison, revealing that Model 62 outperformed others across multiple tests. Although the models are not yet sufficient as standalone end-to-end classifiers, they exhibit strong potential when combined with additional classifiers. Comparisons to previous studies underscore the promising performance of palmprint biometrics, highlighting their potential in specific domains like access security and payment.**

***Index Terms*—Contactless Palmprint Recognition, Machine Learning, Deep Neural Network, Triplet Learning**

I. Introduction

The field of biometrics harnesses the unique and universal characteristics of individuals to verify identities and plays a pivotal role in modern security, access, and law enforcement. There are a multitude of modalities available, each possessing benefits and drawbacks. Among the most popular modalities is fingerprint identification, often depicted in detective narratives and now integrated into nearly every mobile device. Fingerprints have several advantages, such as small sensor requirements, high information density, and the ability to differentiate even amongst identical twins [1]. However, the recent COVID-19 pandemic has shown some limitations for fingerprint biometrics, such as reliance on contact and the potential for distortions [2]. Another popular modality is face recognition, which performs well due to the large amount of data available but suffers criticism on the basis of privacy concerns.

Similarly to fingerprints, palmprints are also rich with biometric information such as the principal lines, wrinkles, and ridges [3]. Palmprints offer distinct advantages over fingerprints, including the potential for application in lower resolutions [4] and containing additional identifying details due to their increased size. They also possess advantages over facial recognition in terms of privacy concerns and controlled data acquisition. As recently as May 2023, Beijing Metro launched a pilot program for fare payement via palmprint, citing hygiene and accessibility as the main benefits [5].

Despite these advantages, research on the usage of palmprint biometrics remains relatively limited compared to other modalities. This thesis addresses this gap by exploring the effectiveness of deep neural network embeddings for contactless palmprint recognition, providing a comprehensive analysis of various methods and their performances.

In modern biometrics, machine learning has emerged as the dominant approach. Originally, models based on handcrafted features dominated the field, but the last decade has seen significant progress in deep neural networks trained on large datasets [4]. Fingerprint and face recognition have greatly benefited from this new approach, bolstered by decades of data collection. Recently, an increasing number of palmprint datasets have become available, prompting the question of whether similar success can be achieved for palmprints using these advanced machine learning techniques. This thesis aims to determine the potential of these techniques, focusing on the implementation and efficacy of deep neural network embeddings for contactless palmprint recognition.

A. My Work and Contribution

This study investigates developing machine learning models for palmprints to be used in large-scale biometrics systems. Specifically, it aims to address the many-to-few classification problem, which involves narrowing down a large number of potential candidates to a few likely matches. Doing so would save computational time by not applying more complicated and time consuming models on a large number of candidates. The goal is to create a model that can achieve a high rank 10 retrieval rate, meaning that the correct individual is among the top ten candidates retrieved by the model. By addressing this research question, I seek to advance palmprint biometrics and facilitate their integration into new products.

The objective of the model is to accurately ascertain the identity of individuals within a known database based on an unseen palm image. My hypothesis posits that a model that embeds each input palm image into a distinct feature vector should sufficiently differentiate between individuals. Functioning essentially as an encoder, the model processes palm images and outputs corresponding feature vectors. The primary aim is to devise a training pipeline that ensures the encoder generates a unique feature vector for each input image, with images of the same palm yielding highly similar feature vectors. These unique encodings could then be used as part of a larger pipeline for classification.

B. Preface

This master's thesis was conducted at Precise Biometrics, located in Lund, Sweden, based on topics suggested by the company. While Precise Biometrics provided data, a laptop, and guidance, all research and writing was undertaken by me.

II. Previous Work

The extensive research conducted in fingerprint and facial biometrics has great value and applicability to palmprint recognition. Fingerprints and palmprints share the characteristic of having distinct ridges, while facial recognition encounters similar challenges to palmprints, such as background noise, region of interest (ROI) framing, and varying lighting and pose conditions. This section reviews previous work in these biometric areas and discusses its relevance to palmprint recognition. Additionally, studies specifically on palmprints are presented, analyzing their applicability and limitations.

In 2015, Schroff et al. [6] introduced FaceNet, a model that improved facial biometrics in large-scale systems by using compact embeddings to map face images to feature vectors. FaceNet utilized a deep neural network (DNN) trained with triplet learning, detailed in Sec. III-A, to generate these

embeddings. FaceNet encountered two major problems: the pose and the illumination of the subject were highly varied. These are also problems within palmprints, as the hand poses and lighting of the images complicate model precision. Hand pose affects the principal line position and ridge counts, while illumination can significantly alter pixel intensities. FaceNet achieved generalization towards these by using a large dataset of labeled faces, a luxury not as available for palmprints, leaving it uncertain if existing datasets can provide palmprint models with similar generalization.

In 2019, Afifi at Google [7] presented the 11k Hands dataset, featuring images of the palm and back of the hand, which was used to train a gender classification model. Afifi highlighted the advantage of palmprints in the controlled conditions in which images are taken. Like Schroff et al., Afifi employed a deep convolutional neural network but added a dual-stream approach, meaning they add a separate pipeline trained on a detailed version of the images. Their preprocessing applies a smoothing filter to create a low-frequency image that the original image is then divided by to obtain a high-frequency 'detailed' image.

Also in 2019, Thapar et al. [8] introduced PVSNet, which used the palm-vein structures underneath the palm surface for identification. They claimed that palm-vein images offer advantages in preventing spoofing and do not require physical sensor contact. Using these images, they trained an encoder-decoder using a deep convolutional neural network with triplet learning, explained in Sec. III-A, inspired by U-net. An alternative to the palm-vein images are 'surface' images, which are taken with a wavelength that highlights the surface ridges. Furthermore, 'subdermal' images of the palm's capillary beds could be used; an alternative not explored by Thapar et al. The question of whether to employ surface, subdermal, or normal images will be a subject of analysis in this report.

Most recently, Grosz et al. [9] presented a new state-of-the-art model for palmprint identification that combined a 50-layer deep neural network and a visual transformer to generate image embeddings. Classification was performed using the Euclidean distance between embeddings. Grosz et al. attribute their improvements to a large, newly captured dataset, soon to be publicly available. Grosz et al.'s paper was found after the results were already in hand, and since it has similar approaches to mine, their results are a valuable benchmark. However, the lack of overlap between their test datasets and mine limits direct comparability.

III. Theory & Methods

This section presents the necessary theory and approach used in this study to develop a deep neural network (DNN) for palmprint analysis. Essentially, the network processes a 200x200-pixel image of a palm and generates a 128-dimensional embedding vector. Identification is performed by embedding an unseen palm and measuring the L2 norm to nearby enrolled embeddings, as illustrated in Fig. 2. Enrolling involves encoding palm images from an individual and calculating the average embedding vector. Each palm was treated as a different class, meaning each person has two classes. For stability, each 128-dimensional embedding is constrained to the 128-dimensional hypersphere. The many-to-few problem, meaning reducing the number of candidates, is addressed by returning a list of the 10 most similar candidates based on euclidean distance in the 128-dimensional feature space. The proposed model's training strategy relies on triplet learning, explained in Sec. III-A, inspired by the principles presented by FaceNet [6] and PVSNet [8]. The training pipeline is illustrated in Fig. 1.

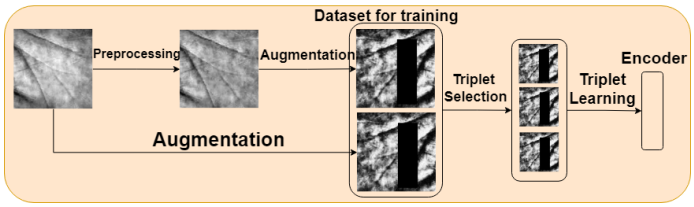


Figure 1: Training pipeline including preprocessing, data augmentation and training using triplet learning.

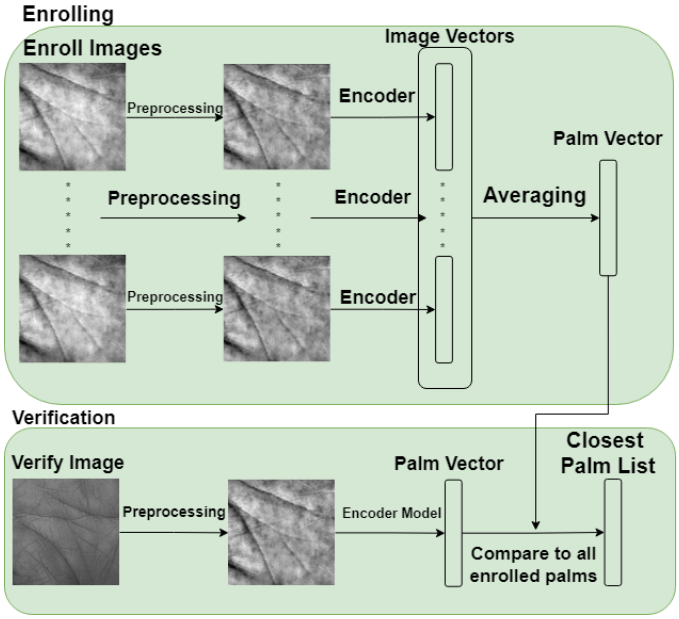


Figure 2: Testing Pipeline: enrolling a set of images and then comparing verify images to the enroll images. The trained encoder creates embeddings of the images, from which an average is taken to get a palm embedding.

A. Triplet Loss

Triplet learning consists of training on three images: an 'anchor' serving as a reference point, a 'positive' image depicting the same palm as the anchor, and a 'negative' image portraying another palm. The objective of triplet learning is to iteratively train the model to embed the anchor and positive images similarly while creating a clear separation from the negative image. Conceptually, triplet learning aims to minimize the L2 norm between all embeddings of the same

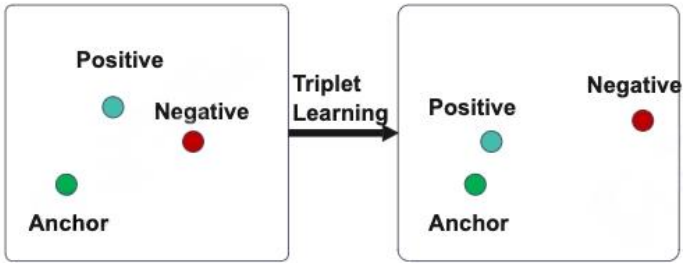
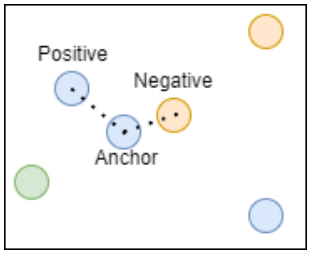
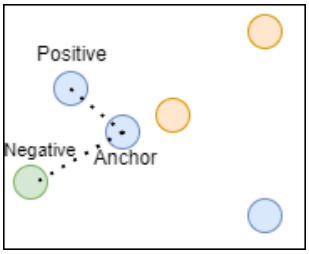


Figure 3: Visualization of triplet learning. After learning, the distance between the anchor and positive embeddings is smaller, while the distance between the anchor and negative embeddings is greater.



(a) Hard triplet selection : The positive is randomly selected from the blue class and the negative is the closest different class.



(b) Semi-hard selection: The positive is randomly selected from the blue class and the negative is the closest different class farther from the positive.

Figure 4: Comparison of triplet selection techniques. Each color indicates a different class.

palm (anchor-positive pairs) while maximizing the norm between embeddings of different palms (anchor-negative pairs), as shown in Fig. 3. The pytorch function *TripletMarginLoss* is used, which is mathematically defined as:

$$L(a,p,n) = \max[||f(a)-f(p)||_2 - ||f(a)-f(n)||_2 + \text{margin}, 0],$$

where a, p, and n represent the anchor, positive, and negative images, respectively, and $f(\cdot)$ is the mapping of the neural network. The margin hyperparameter defines the 'closeness' criterion between the anchor-positive pair and the anchor-negative pair. If the negative image lies closer to the anchor than the positive image does anchor, the margin constitutes the minimum separation of the two pairs. This means that the model learns to discriminate between positive and negative images effectively. When the margin is not violated, i.e., when

$$||f(a) - f(p)||_2 - ||f(a) - f(n)||_2 + \text{margin} < 0,$$

the resulting loss function is 0, and the model learns nothing. Therefore, it is vital to properly tune the margin and select triplets effectively.

B. Triplet Selection

The naïve approach to selecting triplets is to randomly determine a positive image from the anchor's class alongside a negative image from a different class. However, this selection scheme often results in 0 loss since the margin is not violated for a negative image far from the anchor, resulting in wasted computational time without learning.

For triplet learning, hard and semi-hard triplet selection, also called hard or semi-hard negative mining, are the most popular methods. Hard triplet selection involves selecting triplets such that the negative image is as close to the anchor as possible, while the positive image is randomly selected from the anchor class. This forces the model to train on triplets it finds difficult to differentiate, improving performance but potentially leading to local minima issues, as outlined in [6]. Semi-hard triplet selection means selecting negative images whose embeddings are farther away from the anchor embedding than the positive image's, but still as close as possible, ensuring non-zero loss during training. See Fig. 4 for a visualization. Although techniques exist to avoid the local minima of hard triplet selection, as described in [10], I initially chose the simpler approach of semi-hard triplet selection with the intention of revisiting the choice later. However, complications with implementation resulted in only trying semi-hard triplet selection throughout the entire study.

To practically implement semi-hard negative triplet selection, a distance matrix between all training images is precalculated at the beginning of each epoch using the model, which is then used as a reference throughout the entire epoch. This approach is less precise than caluclating new distances for each batch but significantly reduces computational complexity. Compared to the naïve approach, my approach has a higher computational cost but yields superior results by allowing the proper implementation of semi-hard triplet selection. As the model learns to differentiate images of different classes, they will begin spreading out due to the margin enforcing the minimum distance, thus increasing the risk of finding no suitable triplets for learning in later epochs. To compensate for this, two strategies are employed. Firstly, the embeddings are constrained to the 128-dimensional hypersphere. Secondly, inspired by Thapar et al. [8], the margin is dynamically increased over the course of the model training. To ensure the model trains sufficiently on the final margin, most models were trained using a 'flat' margin, which means the final half of epochs all utilized the final margin, see Fig. 5 for illustrations of the margins used.

C. Dataset Analysis

In this section, I discuss the various datasets used to train and test the model. Multiple previous palmprint studies have graciously provided datasets for academic use, and internal collections at Precise Biometrics have also supplemented the datasets. During data collection, images of entire hands are taken, but only the central part of the palm is used for the model. Fig. 6 shows example images from the Tongji dataset, with Fig. 6d used for preprocessing. Most datasets had 200x200-pixel ROI images available; when they did not, Precise Biometrics provided tools for ROI extraction. A list of the datasets used is provided in Tab. I.

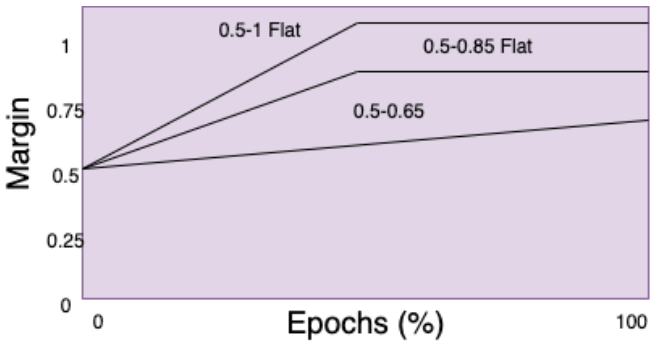


Figure 5: Visualization of margins used by the 3 models. Flat margin implies training half the epochs on the final margin.

TABLE I: Datasets used including number of unique hands and total images. Casia and Google11k contains images of both sides of the hand; only those of the palm were used.

Dataset	# of Unique Hands	Total Images	Source
Training			
Casia (Palmar)	617	5286	[11]
Google11k (Palmar)	190	5380	[7]
IITD	455	2512	[12]
MPD	200	16000	[13]
Precise Biometrics 1	224	1321	Internal
Precise Biometrics 2	42	334	Internal
Testing			
HandID 1 (H. 1)	77	365	Internal
HandID 2 (H. 2)	80	795	Internal
Tongji	600	12000	[14]

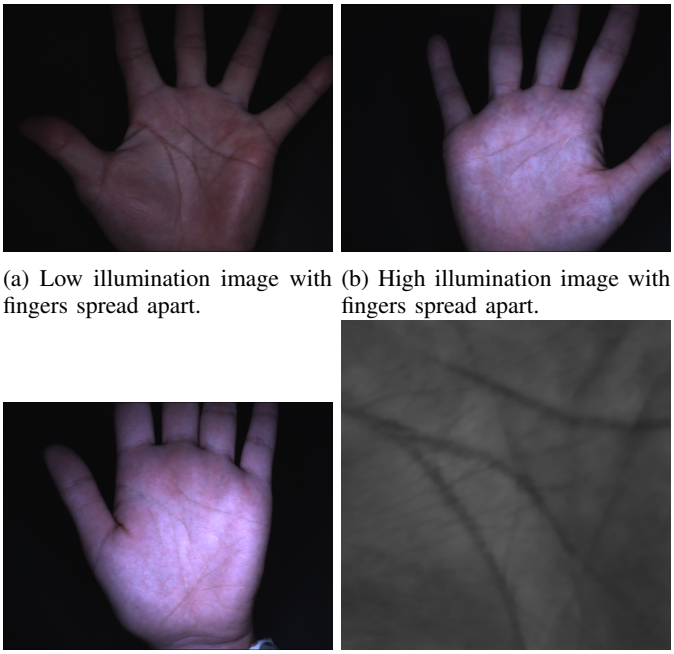
My models were trained using the datasets labeled 'Training' in Tab. I, with 20% of the data removed and used as a validation set. Special care was taken to perform the validation split based on individuals in order to avoid having the same person in both the training and validation sets. Additional models were also trained on individual datasets to gauge if any datasets were adversely affecting the models. In particular, MPD was removed and used as a test set in one ablation study since it appeared very uniform. Two collections were provided by Precise Biometrics for testing: HandID 1 and HandID 2, consisting of 77 and 80 palms, respectively. The Tongji dataset was obtained after initial results were obtained and thus was not used for training, but it served as a sizable test set due to its considerable 12000 images.

There are two major factors that vary in palmprint images: illumination and pose, as visualized in Fig. 6. Both Google11k and IITD include pose variation by asking the subject to open and close their hand while using uniform light. Casia does not specify pose variation and uses uniform light, while MPD has open hands with varied light. The internal collections, Precise Biometrics 1 and Precise Biometrics 2, have variations in lighting but not in poses. HandID 1 and HandID 2 contain no variation in pose, with HandID 1 exhibiting significantly more variation in illumination than HandID 2. Tongji has a slight variation in pose and a large variation in illumination. An overview of the dataset variation is shown in Tab. II.

In conclusion, there is some variation within both pose and illumination in the training datasets, but it may not be sufficient to ensure that models will generalize well to these variations. Differences in pose are difficult to simulate through preprocessing, but illumination differences could be replicated.

TABLE II: Pose and light variation in the datasets.

Dataset	Pose Variation	Light Variation
Casia (Palmar)	?	×
Google11k	✓	×
IITD	✓	×
MPD	×	×
Precise Biometrics 1 & 2	×	✓
HandID 1	×	✓
HandID 2	×	×
Tongji	✓	✓



(a) Low illumination image with fingers spread apart.

(b) High illumination image with fingers spread apart.

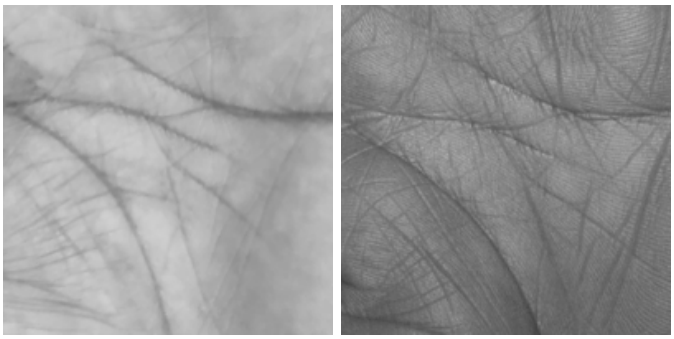
(c) High illumination image with fingers held together.

(d) Grayscale region of interest of Fig. 6a.

Figure 6: Various illuminations and poses of original images of different hands from the Tongji dataset.

Some of the datasets, specifically HandID 2 and Tongji, contain both 'enroll' and 'verify' images. For Tongji, these are images taken in different sessions six months apart. For HandID 2, each person has two sets of verify images: surface and subdermal. The surface image is taken with a wavelength that results in a heightened surface resolution of the image, as illustrated in Fig. 7. Fig. 7a shows an enroll image taken with a regular phone camera, and Fig. 7b shows the surface image. The subdermal image is taken with a wavelength that highlights the palm's capillary beds but is not available for illustration. These vary slightly from the vein-structure images used by PVSNet [8]. Either surface or subdermal could feasibly be used as the verification image in the model; however, both suffer from the issue of being a different modality than both the enroll images and the training set, which are regular

light images.



(a) HandID 2 enroll image, taken with a phone camera using regular light.

(b) HandID 2 'surface' verify image, captured at a wavelength emphasizing the palm surface.

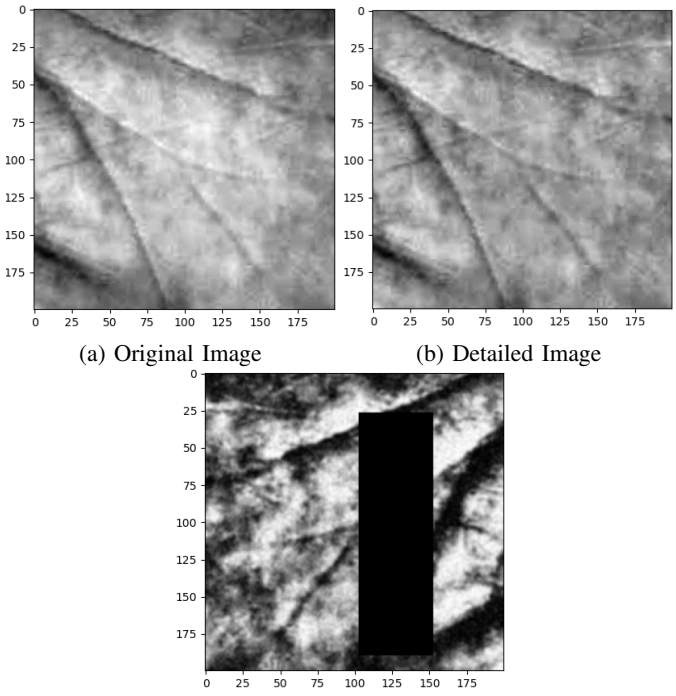
Figure 7: Normal and surface images of the same hand from HandID 2.

D. Preprocessing and Data Augmentation

Palmprints contain many ridges and lines, which have a high contrast compared to the surrounding area. To enhance these lines, high-frequency 'detailed' images were created by elementwise dividing the original images by a smoothed version of the image, similar to the approach by Afifi [7]. Fig. 8 illustrates this preprocessing step. Specifically, Fig. 8a shows the original image from the dataset, while Fig. 8b depicts the detailed preprocessing of that image. Models were trained on the original dataset, the detailed dataset, and both datasets combined. When training on both datasets, the original and detailed images of the same hand were treated as the same class to prevent the model from learning to differentiate based on camera type. The test datasets, including the surface and subdermal images of the HandID 2 dataset, were also preprocessed into detailed versions.

Properly training deep neural networks (DNNs) requires vast amounts of data. The 30000 training images are only about 10% of the data used in similar studies, such as that by Grosz et al. [9]. To bolster the dataset, various data augmentation techniques were employed, as presented in Tab. III. The data augmentation involves adding noise, performing gaussian blur, flipping the images horizontally, and erasing a randomly sized rectangle from the image. Since the model utilizes triplet learning, each augmentation operation involving rectangle removal or horizontal flipping was applied equally to all three images to avoid training a right/left hand invariance. Conversely, noise and blur were applied only to the anchor image to increase the model's robustness against sensors that might provide noisy or blurry images. The augmentation parameters were obtained through a brief hyperparameter optimization on the validation set. Fig. 8c shows an image with data augmentation.

In summary, preprocessing and data augmentation were crucial steps in preparing the palmprint images for model training. By creating detailed versions of the images and applying various augmentation techniques, I aim to enhance the model's ability to generalize to different lighting and pose conditions, as well as increase its robustness to noise and blur. These steps ensured that the model was trained on a diverse set of images, improving its performance and reliability in real-world applications.



(a) Original Image

(b) Detailed Image

(c) Augmented Image

Figure 8: Preprocessing and data augmentation used. Images of the same hand.

TABLE III: Data augmentation applied during training. Noise added had 0 mean and 0.1 variance. Various torchvision packages were utilized. Gaussian blur was implemented using *transforms.functional.gaussian_blur* with either a 50/50 kernel size of 3 or 1. Horizontal flipping was implemented using *transforms.functional.hflip*. Random erasing was implemented using *transforms.RandomErasing*.

Augmentation	Batch Chance	Parameter
Noise	70%	N(0,0.1)
Gaussian Blur	70%	Kernel: 3/1 (50/50)
Horizontal Flipping	50%	-
Random Erasing	40%	Rectangle 2-21 % of the image

E. Network Architecture

The architecture of the neural network is based on the Zeiler& Fergus model [15], featuring convolution layers with 5 **x** 5 or 3 **x** 3 kernels and ReLU activation. Down-sampling is performed using 2 **x** 2 max-pooling layers, and optimization is done using the ADAM optimizer. The detailed architecture is outlined in Tab. IV.

Different model iterations were tested, as described in more detail in the ablation study in Sec. V-C. Throughout the project, three models emerged as the best performing on

TABLE IV: Architecture of neural network used.

Layer	Size in	Size out	Kernel
Conv2d	200x200x1	200x200x32	5x5
MaxPool2d	200x200x32	100x100x32	2x2
Conv2d	100x100x32	100x100x64	3x3
MaxPool2d	100x100x64	50x50x64	2x2
Conv2d	50x50x64	50x50x128	3x3
MaxPool2d	50x50x128	25x25x128	2x2
Conv2d	25x25x128	25x25x256	3x3
MaxPool2d	25x25x256	12x12x256	2x2
Conv2d	12x12x256	12x12x512	3x3
MaxPool2d	12x12x512	6x6x512	2x2
Conv2d	6x6x512	6x6x1024	3x3
MaxPool2d	6x6x1024	3x3x1024	2x2
Flatten	3x3x1024	9216x1x1	-
Linear 1	9216x1x1	512	-
Linear 2	512	256	-
Linear 3	256	128	-

different metrics. These models, referred to as models 52, 62, and 74, will be presented in the subsequent sections.

F. Training the Model

The models were trained using an NVIDIA GeForce RTX 3090 GPU. A validation set containing 20% of the images, distributed similarly to the training set as shown in Tab. I, was used for hyperparameter tuning. The training process involved using different datasets, a varying number of epochs, batch sizes, margin adjustments, and learning rates. Each model's specific configuration is shown in Tab. V. The different training strategies for the models aimed to optimize the models for both accuracy and generalization, leveraging detailed preprocessing and data augmentation to enhance performance.

TABLE V: Hyperparameters of trained models. Margins are illustrated in Fig. 5.

Model	Datasets	Epochs	Batch	Margin	Learning Rate
52	Detailed	30	64	0.5-1 Flat	0.00002
62	Og+Detailed	15	24	0.5-0.65	0.00002
74	Og+Detailed	50	24	0.6-0.8 Flat	0.000015

G. Evaluation Metrics

To evaluate the quality of the models, several evaluation metrics were utilized on the three test datasets. Previous works, such as [9], utilize the true accept rate (TAR) at 0.01% false accept rate (FAR) as a benchmark for model performance. This metric measures the end-to-end model capability and incorporates the potential for misclassifying a person, making it a very useful industry standard. However, I do not use TAR at 0.01% FAR because my model's goal is to narrow down to the 10 most similar candidates. These candidates could then be fed into a classifier, such as a support vector machine, to create an end-to-end classifier where TAR at 0.01% FAR would be a suitable evaluation metric. If the person in question is not enrolled in the model, it would be significantly quicker computationally to identify them at this later stage. Consequently, all tests are conducted on a closed-set, meaning the correct person is assumed to be enrolled in the database already. While this approach does not allow for direct comparison with previous studies using TAR at 0.01% FAR, I can still compare using the closed-set rank 1 retrieval rate, as reported by Grosz et al. [9].

Evaluation metrics were calculated by encoding all images in the test dataset and then iteratively removing one image to verify against the rest. For the HandID 2 and Tongji datasets, an additional test was performed by encoding all enroll images and iteratively encoding the verify images.

1) Average Number of Hands: Since the goal of the model is to narrow down the number of candidates for identification and not necessarily provide an end-to-end identification pipeline, I introduce the Average Number of Hands (ANH). This metric measures how many unique enrolled palms lie closer to the palm being verified than the correct person does; an ANH of 1 means that the most similarly encoded palm comes from the same person.

To calculate ANH, all enroll images are encoded before iteratively encoding the verify images and checking which enrolled images lie closest, as shown in Fig. 2. The score of an image is determined by the number of people checked, including the image itself, before finding an enrolled image with the same class as the verify image. The average of all verify images is the ANH for that dataset. Since the encodings are limited to a 128-dimensional hypersphere, a larger dataset should result in a larger ANH, as more data makes the closest hand less likely to be of the same class.

ANH was also used as the evaluation metric for the impact of different embedding dimensionalities. For these tests, model 62 was used as a baseline, changing only the size of the final output layer, as it appeared to be the best-performing model at the time.

2) Closed-set Rank 1/10 Retrival Rates: Two additional metrics useful for assessing the model's identification performance are the closed-set rank 1 and rank 10 retrival rates. Rank N retrieval rates can be thought of as returning a list of the N most similar enrolled hands, and closed-set means that the hand tested is guaranteed to be enrolled. Closed-set rank 1 retrieval rate effectively shows how well the model performs as an end-to-end classifier on a given dataset. It calculates the percentage of times the most similar enroll image to a varify image is of the same class. Closed-set rank 10 retrival rate provides an overview of the model's performance on the many-to-few problem of narrowing down options. It calculates the percentage of times an image of the verify class is present within the 10 most similar enroll images returned for each verify image.

IV. Results

This section presents the experimental results obtained from the models, including the Average Number of Hands (ANH), closed-set rank 1 and 10 retrieval rates, and the effects of varying embedding dimensionalities. Each model's results were obtained by training three identical models with different seeds, and the mean and standard deviation were calculated.

A. Average Number of Hands

The ANH results for the three models are shown in Tab. VI. I observe that model 62 outperforms the other models across all datasets, closely followed by model 74, and model 52 performs significantly worse. The results on HandID 2 are significantly better than on HandID 1, while HandID 2 Surface and HandID 2 Subdermal perform poorly. The larger Tongji dataset performs about the same as HandID 1.

TABLE VI: ANH of my models on the test datasets. HandID 2 (H. 2) contains three versions of the images: enroll, surface (Sur.) and subdermal (Sub.). For Tongji, the enroll images are used.

Model	H. 1	H. 2	H. 2 Sur.	H. 2 Sub.	Tongji
52	1.42 ±0.04	1.19 ±0.01	5.29 ±0.64	5.60 ±1.21	1.38 ±0.06
62	**1.29 ±0.06**	**1.10 ±0.03**	**3.78 ±0.38**	**3.55 ±0.49**	**1.32 ±0**
74	1.37 ±0.04	1.20 ±0.03	3.92 ±0.46	3.88 ±0.48	1.44 ±0.12

B. Closed-set Retrieval Rates

The closed-set retrieval rates using only enroll images are presented in Tab. VII. Here we also observe that model 62 narrowly outperforms model 74 and significantly outperforms model 52 across all datasets. Model 62 achieves a rank 10 retrieval rate of around 99.36% on both the HandID 1 and HandID 2 datasets. The rank 1 retrieval rates for model 62 are more varied, at around 92-95%. The results on Tongji are the best, reaching a rank 10 retrieval rate of 99.54%.

TABLE VII: Closed-set rank 1 (R. 1) and rank 10 (R. 10) retrieval rates for the three models (M.) on the enroll images of test sets HandID 1 (H.1), HandID 2 (H.2) and Tongji (T.).

M.	H.1 R.1	H.1R.10	H.2 R.1	H.2R.10	T. R.1	T. R.10
52	92.42 ±0.72	98.81 ±0.13	85.45 ±1.51	98.91 ±0.15	97.17 ±0.34	99.42 ±0.0623
62	**94.88 ±0.47**	**99.36 ±0.26**	**91.65 ±2.50**	**99.37 ±0.36**	**98.24 ±0.050**	**99.54 ±0.019**
74	91.87 ±0.85	99.36 ±0.52	88.09 ±1.93	99.33 ±0.21	97.16 ±0.49	99.34 ±0.15

After testing using the enroll images, the performance of the models on the verify images of HandID and Tongji are presented in Tab. VIII. These perform substantially worse than only using the enroll images for all models, especially on the surface and subdermal images.

C. Embedding Dimensionality

The ANH for models with varying embedding dimensionalities is shown in Tab. IX. The results show minimal differences between the embedding dimensionalities.

V. Discussion

A. Semi-Hard vs. Hard Triplet Selection Schemes

One of the initial decisions was whether to apply hard or semi-hard triplet selection, as described in Sec. III. Hard

TABLE VIII: Closed-set rank 1 (R. 1) and rank 10 (R. 10) retrieval rates for the three models (M.) on the HandID 2 (H. 2) and Tongji (T.) dataset using the verify images. Both surface (Sur.) and subdermal (Sub.) images are included.

M.	R.1 Sur.	R.10 Sur.	R.1 Sub.	R.10 Sub.	T. R.1	T.R.10
52	36.87 ±2.78	82.05 ±1.61	36.96 ±2.94	81.26 ±2.60	60.96 ±5.51	87.35 ±2.35
62	**45.45 ±0.56**	**88.42 ±1.97**	**47.51 ±1.62**	88.85 ±1.56	**74.72 ±1.59**	**93.90 ±0.80**
74	43.13 ±1.25	88.24 ±1.35	42.16 ±1.61	**88.90 ±1.02**	66.57 ±0.87	90.79 ±0.87

TABLE IX: ANH of various embedding dimensionalities. Model 62 was used as a baseline for all the models.

Embedding Dim.	HandID 1	HandID 2	Tongji
64	1.43 ± 0.02	1.12 ± 0.03	1.33 ± 0.06
128	1.29 ± 0.06	1.10 ± 0.03	1.32 ± 0.000028
256	$\mathbf{1.29 \pm 0.023}$	1.12 ± 0.0069	$\mathbf{1.27 \pm 0.059}$
512	1.29 ± 0.076	$\mathbf{1.09 \pm 0.022}$	1.31 ± 0.047

triplet selection poses the risk of getting stuck in local minima, but previous studies, such as Xuan et al. [10], present modifications to the loss function that mitigate these risks. However, integrating these modifications with the existing *MarginLossFunction* proved challenging, and other questions were prioritized. Consequently, semi-hard triplet selection was used initially and continued throughout the study due to its acceptable performance. Further research into hard triplet selection could yield interesting results, and a proper implementation would likely match the efficacy of semi-hard triplet selection.

B. Performance and Comparison

The results are best understood in context with each other and with previous studies. This section provides a comparative perspective and outlines the rationale behind the findings.

1) Average Number of Hands: The Average Number of Hands (ANH) performance, presented in Tab. VI, is a novel metric used here to compare the models. Unsurprisingly, model 52 consistently performed the worst due to its simpler data augmentation strategy. When only one version per model was trained, this was not certain, but with the larger sample size presented in Tab. VI, I can dismiss model 52 as inferior to my other models. This result highlights the importance of data quantity in deep neural networks. Models 62 and 74 were trained on double the amount of data using the heavy data augmentation outlined in Sec. III-D. The results indicate that model 62 outperforms model 74 by a narrow margin across all datasets.

Interestingly, the models performed better on HandID 2 than on HandID 1, likely due to the controlled lighting conditions during HandID 2's collection, as displayed in Tab. II. This implies that my models have not achieved the desired generalization to light conditions. One possible explanation for this is that not enough images in the training dataset have lighting variations; only Precise Biometrics 1 and Precise Biometrics 2 contain light variations but account for only around 5.4%

of the data. If the Tongji dataset had been available during training, which contains both variations in pose and light, the models are likely to generalize better to these variations.

In Sec. III-G1, I hypothesized that a larger dataset would have a worse ANH due to the increased candidates for missclassification. However, the similar ANH results between HandID 1 and the much larger Tongji dataset disprove this hypothesis and suggest that the 128-dimensional feature vector space remains manageable even with a larger dataset.

For HandID 2, the performance significantly dropped when classifying subdermal and surface images compared to enroll images. This drop is attributed to the different modalities of these images, and preprocessing was insufficient to bridge this gap. An attempt to combine surface and subdermal embeddings by averaging proved ineffective, resulting in worse performance. Using surface or subdermal images for the training of a palmprint classifier remains possible but is severely limited by the amount of data available.

2) Closed-Set Rank 1/10 Retrieval Rate: The closed-set rank 1 retrieval rate shown in Tab. VII for enroll images is around 92-95%, insufficient for end-to-end classification. In comparison, Grosz et al. [9] achieved a rank 1 retrieval rate of 99.71% on the Casia dataset, highlighting the gap between the datasets and methodologies. Specifically, Grosz et al. utilized around 10 times the data and trained a multi-streamed model utilizing both a deep neural network and a visual transformer. My models, while not intended as final classifiers, could serve as a strong pre-classifier for a support vector machine, given the 99.4% rank 10 retrival rate, which would be an interesting subject for future research. The results shown by [9] indicate that improvements remain possible, and I believe that further model refinement could improve the rank 10 retrieval rate to near 100%.

Model 62 outperforms the others in rank 1 retrieval rates, although rank 10 retrieval rates showed no significant difference between models 62 and 74. This indicates a slight superiority of model 62 over model 74.

Verification when enrolling with one subset of images and then verifying with another yielded significantly worse results, as seen in Tab. VIII, reinforcing the challenge of generalizing across different image modalities and conditions. The Tongji dataset, with verify images collected 6 months later, performed significantly better than HandID 2's, whose verify images were surface and subdermal images. There is minimal difference between using surface and subdermal images, both resulting in a rank 1 retrieval rate of around 44 % and a rank 10 retrieval rate of around 88%. From these results, it is evident that the model does not properly generalize to the surface and subdermal images. In order to do so, I believe adding images of surface and subdermal modality to the training dataset would be necessary.

Likewise, comparing the Tongji verify results in Tab. VIII to the results of using only the enroll images for Tongji in Tab. VII, we see that using the enroll images performs significantly better. This is more surprising than the case for HandID 2, since Tongji's verify images are of the same modality as its enroll images. Using the Tongji verify images, model 62 achieves a rank 10 retrieval rate of 93.9% versus 99.54% when using only enroll images. This result likely stems from a difference in the enroll and verify Tongji datasets, performed in collections 6 months apart.

C. Ablation Study

This section discusses the various models and hyperparameters tested during the study, aiming to justify the final model choices.

1) Training on Different Datasets: Apart from training models on all datasets presented in Tab. I, models were also trained on individual datasets. These models unilaterally performed worse than the models trained on all data. MPD is a dataset of specific note that contains a large quantity of images that visually appear very similar to one another. An untrained, newly initiated model trained on this dataset performed surprisingly well; this led me to believe that MPD might be the cause of significant overfitting, and a model was trained without this dataset. However, this model performed worse than the previous best models. No other datasets were singled out in this way, thus it remains a possibility that other combinations of datasets could lead to improved results.

2) Embedding Dimensionality: Like Schroff et al. present in [6], I also investigate the impact of embedding vector length on model performance. In their study, they find that embedding dimensionalities of 64, 128, 256, and 512 all have similar results. They posit that this is because the larger embedding vectors need more training to achieve the same accuracy. Grosz et al. [9] also try similar dimensionalities around this range and conclude that the difference is minimal. Similarly, I found very limited differences in such models, as shown in Tab. IX. Furthermore, the results in Tab. VI bolster this since the models achieve similar ANH on HandID 1 and Tongji datasets, which have significantly different amounts of images. I suspect that this is due to the lack of complexity of palmprints when compared to the dimensionality required for embedding linguistic meaning, such as in large language models; the quality of the images is not sufficiently high that enough details can be picked up to warrant a higher embedding dimensionality. An interesting future study would be to try more extreme embedding dimensionalities to observe when the invariance breaks and ascertain why.

3) Comparison to Simple Hand Crafted Feature Vector: In order to ascertain how my models perform compared to a baseline, simple handcrafted features were created on the original + detailed dataset by taking a single averaging convolution of the 200x200-pixel image. I attempt to emulate the 128-dimensionality by using an 18x18 kernel with stride 18, resulting in a feature vector of 121 dimensions. After obtaining the handcrafted feature vectors for each image, the same steps are taken as with the model testing: calculating a distance matrix for each test point and then comparing the points to see whether they are closest to another point of the same class or not. This resulted in an ANH of 5.34, compared to 1.13 for model 62. Conversely, using an untrained, newly initialized

model with random parameters resulted in an ANH of 5.51. These results thus appear very similar and, unsurprisingly, much worse than those of my trained models.

4) Palmprint as a Competitive Biometric Modality: My results, as well as those of [9], indicate that palmprints, as a biometric modality, exhibit promising potential for security access applications. However, it is essential to recognize that while palmprint models show competence in distinguishing individuals, they do not yet surpass the comprehensive development and application of fingerprint and face recognition technologies. In scenarios like mobile phones, where fingerprint sensors and facial recognition are predominant and efficient, palmprints are unlikely to supplant them. Yet, in contexts such as office access or personal payments, palmprints offer advantages, including not storing sensitive facial images and enabling user-initiated interactions, like raising one's hand for authentication, unlike facial recognition systems that lack such user prompts. It is possible that the growing development and increased research into palmprints will bring them to the forefront of biometric modalities in the coming decade.

VI. Conclusion

In conclusion, my models confirm the hypothesis that employing an encoder adequately addresses the many-to-few issue in narrowing down candidates for palmprint recognition. This study introduces the Average Number of Hands (ANH) metric, providing a new way to compare model performance. While my proposed model, model 62, is not flawless and would be insufficient as a standalone end-to-end classifier, its 99.4% closed-set rank 10 retrieval rate indicates potential when combined with another classifier. Although there are challenges with different image modalities and the need for more robust pre-processing techniques, recent advancements in palmprint biometrics underscore its emergence as a formidable contender alongside fingerprint and face biometrics. Future research could explore the implementation of hard triplet selection, test different dataset combinations, and investigate extreme embedding dimensionalities to further optimize performance. With ongoing development, palmprint technology could become a preferred biometric modality in specific domains such as access security and payment, complementing existing systems and offering unique advantages.

Use of AI

ChatGPT 3.5 and 4o were used in a limited amount throughout this work. Most questions asked were regarding the syntax of relatively simple Python functions. The code supplied was always tested and usually worked after resolving some compiling errors through further communication with ChatGPT. ChatGPT was also used to refine paragraph structures for academic writing. The AI spellchecker QuillBot was also used thoroughly, mainly adding commas and changing prepositions.

Acknowledgement

Portions of the research in this paper use the CASIA Palmprint Database collected by the Chinese Academy of Sciences' Institute of Automation (CASIA).

Thanks to Precise Biometrics for their support and collaboration throughout this research endeavor, which significantly enriched the study's scope and feasibility. My sincere thanks to my supervisors at Precise Biometrics, Johan Windmark and Ellen Åström, for their invaluable insight and expertise, which provided a solid foundation for me to build upon. I also appreciate the contributions of Axel Kärrholm and Diego Figueroa Llamosas, whose help in developing ideas and creating a vibrant working environment made my time at Precise Biometrics truly enjoyable.

Special thanks to my supervisor at LTH, Johanna Engman, for her knowledge, guidance, and availability for direction and feedback. I always left the supervisory meetings reinvigorated and with a clear path forward.

References

[1] X. Tao, X. Chen, X. Yang, and J. Tian, "Fingerprint recognition with identical twin fingerprints," *PLoS One*, vol. 7, p. e35704, Apr. 2012.

[2] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, "Deep contactless fingerprint unwarping," 04 2019.

[3] P. P. Sarangi, M. Panda, S. Mishra, and B. S. P. Mishra, "Chapter 3 - multimodal biometric recognition using human ear and profile face: An improved approach," in *Machine Learning for Biometrics* (P. P. Sarangi, M. Panda, S. Mishra, B. S. P. Mishra, and B. Majhi, eds.), Cognitive Data Science in Sustainable Computing, pp. 47–63, Academic Press, 2022.

[4] D. K. Sharma, B. Tokas, and L. Adlakha, "Chapter 2 - deep learning in big data and data mining," in *Trends in Deep Learning Methodologies* (V. Piuri, S. Raj, A. Genovese, and R. Srivastava, eds.), Hybrid Computational Intelligence for Pattern Analysis, pp. 37–61, Academic Press, 2021.

[5] D. Juan, "Beijing introduces palm-print access on subway line," *China Daily*, 2023.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[7] M. Afifi, "11k hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, 2019.

[8] D. Thapar, G. Jaswal, A. Nigam, and V. Kanhangad, "Pvsnet: Palm vein authentication siamese network trained using triplet loss and adaptive hard mining by learning enforced domain specific features," in *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pp. 1–8, 2019.

[9] S. A. Grosz, A. Godbole, and A. K. Jain, "Mobile contactless palmprint recognition: Use of multiscale, multimodel embeddings," *arXiv preprint arXiv:2401.08111*, 2024.

[10] H. Xuan, A. Stylianou, X. Liu, and R. Pless, "Hard negative examples are hard, but useful," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 126–142, Springer, 2020.

[11] P. Poonia and P. K. Ajmera, "Palm-print recognition based on image quality and texture features with neural network," in *2021 Sixth International Conference on Image Information Processing (ICIIP)*, vol. 6, pp. 41–46, 2021.

[12] A. Kumar, "Incorporating cohort information for reliable palmprint authentication," in *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 583–590, IEEE, 2008.

[13] Y. Zhang, L. Zhang, R. Zhang, S. Li, J. Li, and F. Huang, "Towards palmprint verification on smartphones," *CoRR*, vol. abs/2003.13266, 2020.

[14] L. Zhang, L. Li, A. Yang, Y. Shen, and M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach," *Pattern Recognition*, vol. 69, pp. 199–212, 2017.

[15] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013.

Master's Theses in Mathematical Sciences 2024:E38
ISSN 1404-6342

LUTFMA-3542-2024

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/