F. Boutros, N. Damer, M. Fang, M. Gomez-Barrero, P. Neto,
K. Raja, C. Rathgeb, A. Sequeira, and M. Todisco: (Hrsg.): BIOSIG 2025,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2025

Deep Learning-Based Approaches for Contactless Fingerprints Segmentation and Extraction

M.G. Sarwar Murshed [1], Syed Konain Abbas [2], Sandip Purnapatra [3], Daqing Hou [3], and Faraz Hussain [3]

Abstract:

Fingerprints are widely recognized as one of the most unique and reliable characteristics of human identity, making them a preferred choice for biometric-based authentication systems. The use of contactless fingerprints has emerged as an alternative. This paper focuses on the development of a deep learning-based segmentation tool for contactless fingerprint localization and segmentation. Our system leverages deep learning techniques to achieve high segmentation accuracy and reliable extraction of fingerprints from contactless fingerprint images. In our evaluation, our segmentation method demonstrated an average mean absolute error (MAE) of 30 pixels, an error in angle prediction (EAP) of 5.92 degrees, and a labeling accuracy of 97.46%. These results demonstrate the effectiveness of our novel contactless fingerprint segmentation and extraction tools.

Keywords: Contactless, Deep learning, Segmentation, Fingerprints, Biometrics

1 introduction

Biometric recognition systems play important roles in different domains due to their accuracy and quick processing time. One of the most popular and extensively utilized biometric characteristics for authentication is fingerprints. Each fingerprint is unique and remains consistent over time. To enhance security and accuracy of a fingerprint-based authentication system, multiple fingerprints, such as the fingerprint slap, are used for a user instead of a single fingerprint. The first step in using a slap fingerprint to authenticate a user is to separate or segment each finger in the slap [Ca15]. There are several publicly and commercially available slap fingerprints segmenters, such as NIST NFSEG [Ko07] and Neurotechnology Verifinger Segmenter [GJ23].

The term fingerphoto refers to a contactless fingerprint image that may be taken with any camera, including a simple mobile phone camera [MV21]. In Figure 1, a sample fingerphoto is shown.

[1] University of Wisconsin–Green Bay, Dept. of Computer Science, WI, USA,
murshedm@uwgb.edu, https://orcid.org/0000-0002-0059-0632

[2] Clarkson University, Dept. of Computer Science, Potsdam, NY, USA,
abbas@clarkson.edu, https://orcid.org/0009-0003-1645-3792

[3] Clarkson University, Dept. of Electrical and Computer Engineering, Potsdam, NY, USA,
purnaps@clarkson.edu, https://orcid.org/0009-0004-0464-3481;
dhou@clarkson.edu, https://orcid.org/0000-0001-8401-7157;
fhussain@clarkson.edu, https://orcid.org/0000-0001-8971-1850

Most contactless fingerprint authentication systems utilize fingerphotos, which are contactless fingerprint images that are captured from multiple fingers. Segmenting all fingertips from fingerphotos is an active area of research in contactless biometric authentication systems. Segmentation plays a crucial role in fingerprint matching, as observed in the existing literature [Ma22].

In this paper, we describe the proposed contactless fingerphoto segmentation system developed by enhancing existing contact-based fingerprint segmentation model, CRFSEG [Mu24], by updating its architecture and training it on a contactless dataset. The proposed contactless segmentation model demonstrates higher accuracy when evaluated on our in-house contactless fingerprint dataset consisting of 23,650 slaps, which were annotated by human experts. Special attention was given to optimizing the deep learning architecture for this purpose. The paper presents the following novel contributions:

- Built an in-house contactless dataset of 23,650 fingerphotos (94,600 single fingers). Annotated all fingerphoto images[4] manually to establish a ground-truth baseline for the accuracy assessment of fingerprint segmentation systems.
- Trained a deep learning-based slap segmentation model for contactless fingerphotos that can handle arbitrarily oriented fingerprints.
- Assessed the performance of the proposed model using the following metrics: MAE, EAP, and accuracy in fingerprint matching.

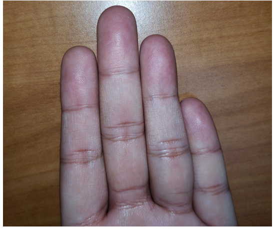


Fig. 1: A sample fingerphoto image taken with a standard mobile phone camera. Fingerphotos are typically acquired by capturing an image of human fingers using a smartphone camera and often include multiple fingers within the frame.

2 Related Work

Researchers have utilized various segmentation methods to achieve precise contactless fingerprint segmentation. Wild et al. introduced a Skin-Mask finger segmentation technique to segment contactless finger photos [Wi19]. They employed a filtering approach with the veriFinger fingerprint matcher and the NFIQ 2.0 quality metric on a dataset of 2582 contact-based and 1728 contactless images from 108 fingers, resulting in TAR values ranging from 95.5% to 98.6% at FAR=0.1%. However, the technique assumes relatively uniform skin color, which may not hold true for dirty or have been exposed to sunlight. It is sensitive to environmental factors like lighting conditions and skin color variations, and background elements resembling skin tones may affect segmentation accuracy.

[4] The total number of images is 23,650. Out of 23,650, 2150 fingerphotos were collected from 29 subjects and manually annotated. The remaining 21,500 were generated by the augmentation method.

Malhotra et al. introduced a method for segmenting the distal phalange in contactless fingerprint images by combining a saliency map and a skin-color map [Ma20]. They employed a random decision forest matcher and feature extraction with a deep scattering network on a dataset of 1216 contact-based and 8512 contactless images from 152 fingers, achieving an EER ranging from 2.11% to 5.23%. Despite impressive results, the algorithm requires extensive hyperparameter tuning and struggles with accurately distinguishing fingerprints in noisy backgrounds or excessively bright lighting conditions.

Grosz et al. proposed an auto-encoder-based segmentation approach [Gr21]. They utilized a spatial transformer for 500 PPI deformation and scale correction on contactless fingerprints. Their dataset comprised three parts: one with 8,512 contactless and 1,216 contact-based fingerprints from 152 fingers, another with 2,000 contactless and 4,000 contact-based fingerprints from 1,000 fingers, and a ZJU dataset with 9,888 contactless and 9,888 contact-based fingerprints from 824 fingers, serving as the basis for evaluation. Their approach achieved impressive EERs of 1.20%, 0.72%, 0.30%, and 0.62% on these datasets, respectively. Many existing fingerprint segmentation methods, particularly those developed for contact-based slap images [Mu24], are not well-suited for segmenting contactless fingerphoto due to several inherent challenges. These challenges include distinct visual characteristics/patterns, arbitrary finger orientations, variable lighting conditions, and complex backgrounds, which can significantly impact the accuracy of traditional detection models. While some studies have proposed contactless segmentation methods [Wi19, Ma20, Gr21], they often rely on restrictive conditions such as consistent skin tone or controlled capture settings, and often lack robustness across diverse environments. Our work addresses this gap by introducing a deep learning-based segmentation model for contactless fingerphotos. Unlike prior approaches, we incorporate orientation-aware anchors using an Oriented Region Proposal Network (O-RPN) to facilitate robust detection of rotated fingertips. We further validate our model on a large, manually annotated dataset of contactless slap images, demonstrating significant improvements in segmentation performance.

3 Research Methods

In this section, we explain how we collected fingerphotos from adult subjects, data annotation, data augmentation, and ground truth creation. Then, we describe the proposed neural network architecture. Finally, we discuss the metrics used to evaluate different slap segmentation algorithms.

3.1 Slap Dataset

In this section, we discuss contactless datasets including fingerphoto collection, annotation, augmentation, and ground truth creation.

3.1.1 Contactless Fingerphoto Collection

Contactless fingerphoto collection is a crucial component of contactless fingerprint biometrics research. The collection process involved acquiring fingerprint images without any physical contact with a camera on a mobile device. This dataset comprises a total of 2150 fingerphotos. To introduce such variation into the dataset used in this study, we employed data augmentation techniques. Through augmentation, we obtained an additional 21,500 augmented images, thereby expanding the dataset to a total 23,650 images. The details of the dataset are shown in the Table 1.

Tab. 1: Our dataset has 23,650 fingerphotos. Out of these, 2150 fingerphotos were collected from 29 subjects at Clarkson University. For image collection, we used different types of mobile phones such as Samsung S20, iPhone 7, iPhone X, and Google Pixel. The images were manually annotated by experts. To create rotating slap images, we utilized the 2,150 finger photos that were annotated by humans to create more images using augmentation techniques. This resulted in the generation of 21,500 more augmented images by rotating all the finger photos at different angles (-90 to 90 degrees). The original dataset contains real-world variations, including differences in lighting, background, skin tone, and hand positioning. Consequently, our augmentation strategy was primarily focused on rotation to address orientation variability, which represents the most significant challenge in contactless fingerprint segmentation. This augmentation is intended to help the model become invariant to different types of rotations of finger photos.

Dataset	Total fingerphotos	Lefthand fingerphotos	Righthand fingerphotos
Bonafide	2150	1118	1032
Augmented	21500	11180	10320
Total	23650	12298	11352

Data annotation plays a crucial role in developing a well-structured and representative dataset, which is essential for building a reliable deep learning-based image segmentation model. The annotation process involves drawing a bounding box around each fingertip in a fingerphoto and assigning a corresponding label to each fingertip. The fingertip labels used include Left-Index, Left-Middle, Left-Ring, Left-Little, Left-Thumb, Right-Index, Right-Middle, Right-Ring, Right-Little, and Right-Thumb. All of the fingerphotos in the dataset used in this study were annotated and tested by humans.

3.2 Deep Learning Architecture for Contactless fingerphoto Segmentation

A previous study [Mu24] developed a two-stage Faster R-CNN architecture for segmenting contact-based slap fingerprints. We utilized a similar two-stage faster architecture for segmenting contactless slap fingerphotos. The contactless fingerphoto deep learning architecture, see Figure 2, comprises three key structural components: the box head, the oriented region proposal network, and the backbone network.

3.2.1 Backbone Network

our backbone network, featuring the ResNet-101 with FPN (Feature Pyramid Network) architecture, is designed to extract feature maps of different sizes from contactless finger-

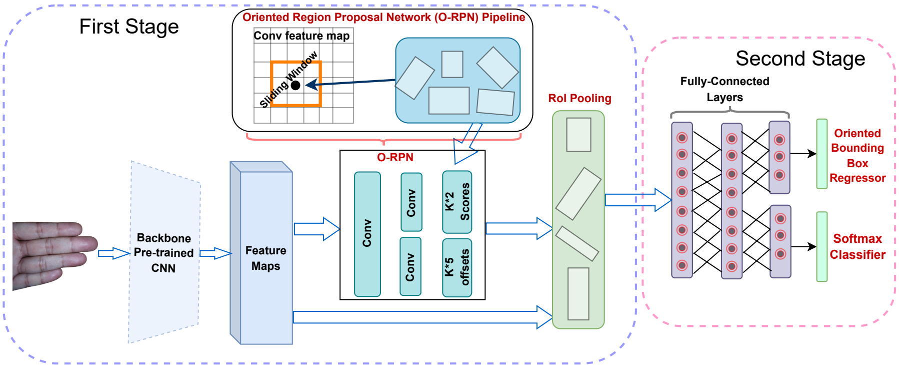


Fig. 2: The complete architecture for the proposed model includes several processing stages and is intended for precise fingerprint segmentation. For the feature maps, we need to provide input contactless fingerprint image to a pre-trained CNN model on the ImageNet dataset. To generate oriented anchors, O-RPN needs to run on all levels of feature maps. By selecting spatial features from O-RPN and from the output of the backbone network, ROI pooling layers generate fixed-length feature vectors. These fixed-length feature vectors are then passed through the fully connected layers. There are two parallel branches that receive the output of the fully connected layers, referred to as the Softmax Classifier and Oriented Bounding Box Regressor. The Softmax Classifier branch contains softmax layers for multiclass classification, while the Oriented Bounding Box Regressor branch contains regressors for bounding box regression.

photos. This network incorporates both stem blocks and bottleneck blocks. To optimize calculations, extract key features, and expand channel capabilities, bottleneck blocks with three convolution layers of varying kernel sizes (1×1, 3×3, and 1×1) are employed [He16]. Within the stem block, the focus is on reducing the input image size, achieved through Convolution 2D layers, ReLU activation, and max-pooling layers, to minimize computational cost while retaining all essential information. The oriented region proposal network (O-RPN) then takes in the multiscale semantic-rich feature maps generated by the backbone network.

3.2.2 Oriented Region Proposal Network (O-RPN)

The proposal of regions of interest (ROIs) by the O-RPN is a crucial step as it helps in generating accurate rotated bounding boxes for both axis-aligned and rotated objects. This is in contrast to the conventional regional proposal network (RPN) used in the Faster R-CNN architecture, where only axis-aligned regions are proposed. The proposal of oriented regions is achieved by using a sliding window approach with a 3×3 kernel on the feature maps. Consequently, the system generates anchor boxes with diverse aspect ratios, scales, and orientations. If we have k_a different orientations, k_s different scales, and k_r different aspect rations, then $K = k_a \times k_s \times k_r$ anchor boxes are generated for each position in the feature map. Most of these anchor boxes might not have targeted objects in them. Subsequently, convolution layers and parallel output layers called localizing and classifying layers are employed to differentiate anchor boxes with target objects from others. The classifying layer assigns foreground or background labels according to the intersection-over-union (IoU) score with ground truth boxes while learning offsets (x,y,w,h, θ) is done by the localization layer for the foreground boxes. The term *foreground* is used to denote regions of interest (proposals) that exhibit a substantial overlap with the bounding boxes of the ground truth objects in the image. In simpler terms, these are areas most likely to contain a prominent object. The classifying layer assigns a foreground label to these regions. Conversely, regions of interest that lack a significant overlap with any ground truth bounding box are designated as the *background*. The probability of an object being present in these areas is lower. The classifying layer assigns a background label to these regions. Regarding the regression offset, the localizing layer generates (K×5) parameterized encodings, while the classifying layer produces (K×5) parameterized scores for region classification. The anchor generation strategy is illustrated in Figure 2. For training O-RPN, seven different orientations (-$\pi/4$, -π/6, -π/12, 0, π/12, π/6, π/4), three aspect ratios (1:1, 1:2, 2:1) and three scales (128, 256 and 512) are used to generate anchors. In conclusion, the selection of seven orientations, three aspect ratios, and three scales involves a trade-off between maintaining computational efficiency during training and inference while ensuring the model's capacity to capture the diversity of objects in images [Ji19]. All anchors are then categorized as positive, negative, and neutral: positive anchors have an IoU overlap greater than 0.7 with ground-truth boxes, negative anchors have an IoU less than 0.3, and neutral anchors fall between 0.3 and 0.7 and are discarded during training. In contrast to Faster R-CNN, where horizontal anchor boxes are used, our approach utilizes oriented anchor boxes and oriented ground-truth boxes. The loss functions employed to train the O-RPN are defined by the following equations:

$$L_{o-rpn} = L_{cls}(p, u) + \lambda u L_{reg}(t, t^*) \tag{1}$$

Here, L_{cls} is the classification loss, p is the predicted probability across the foreground and background classes by the softmax function, u represents class label for anchors, where u = 1 for foreground containing fingerprint and u = 0 for background; $t = (t_x, t_y, t_h, t_w, t_\theta)$ denotes the predicted regression offset value of an anchor calculated by the network, and $t^* = (t_x^*, t_y^*, t_h^*, t_w^*, t_\theta^*)$ represents ground truth. λ is a balancing parameter that manages the balance between class loss and regression loss. Only the regression loss is enabled if u = 1 for the foreground and there is no regression for the background. The classification loss function is defined as the cross-entropy loss between the ground-truth label u and the predicted probability p:

$$L_{cls}(p, u) = -u \cdot \log(p) - (1 - u) \cdot \log(1 - p) \tag{2}$$

Tuple t and t^* are calculated like this:

$$\begin{aligned} t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a, \\ t_w = \log(w/w_a), t_h = \log(h/h_a), \\ t_\theta = \theta - \theta_a \end{aligned} \tag{3}$$

$$
\begin{aligned}
t_x^* &= (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a, \\
t_w^* &= \log(w^*/w_a), t_h^* = \log(h^*/h_a), \\
t_\theta^* &= \theta^* - \theta_a
\end{aligned}
\tag{4}
$$

Where x, x_a and x^* denote predicted box, anchor and ground truth box, respectively; similar for y, h, w and θ. The smooth-L1 loss is adopted for bounding box regression as follows:

$$L_{reg}(t, t*) = \sum_{i \in x,y,w,h,\theta} u.\text{smooth}_{L1}(t_i^* - t_i) \tag{5}$$

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{6}$$

3.2.3 Box Head

The O-RPN network, as part of the architecture, generates 1000 proposal boxes and objectless logits by default. These proposal boxes are then projected into the feature space using an ROI pooling layer. The output of the ROI pooling layer is reshaped and fed into the fully connected (FC) layers. The FC layers generate a RoI vector, which is then passed through a predictor with two branches: the rotated bounding box regressor and the classifier.

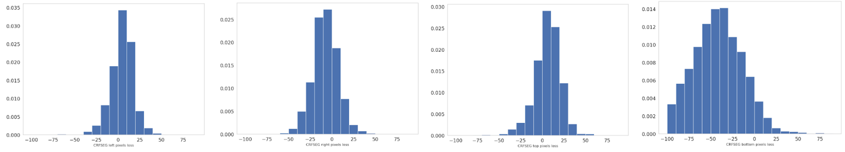
3.3 Training

The proposed model was developed using Detectron2, a framework created by Facebook AI Research (FAIR) [Wu19], which supports advanced deep learning-based object detection algorithms. We utilized the Faster R-CNN algorithm and customized the Detectron2 code to implement the oriented regional proposal network (ORPN), added new layers for handling rotated bounding boxes, to meet our specific requirements for accurate slap fingerprint segmentation. We performed a total of 40,000 training iterations for the fingerprint segmentation model. The learning rates started at 10^{-4} and were decreased by a ratio of 0.1 at specific intervals (4000, 8000, 12000, 18000, and 25000, 32000 iterations). The weight decay was set to 0.0005, and the momentum was set to 0.7. Throughout the experiments, we employed multi-scale training, eliminating the need for scaling the input before feeding it into the neural network. The contactless fingerprint dataset, consisting of 23,650 finger photos, is divided using an 80:10:10 train/validate/test split ratio. A 10-fold cross-validation technique is employed to construct and assess the model, and the outcomes are presented in the results section.

4 Results

This section presents a comprehensive analysis of our findings. We employed four distinct metrics widely used in evaluating fingerprint segmentation models: Mean Absolute Error,

Fig. 3: The Mean Absolute Error (MAE) assesses the accuracy of an object detection system by measuring the distance of each side of the detected bounding boxes with the ground truth bounding boxes. This figure illustrates the MAE in pixels for the left, right, top, and bottom sides of the bounding boxes predicted by the contactless segmentation system.



Left Right Top Bottom

Error in Angle Prediction, Fingerprint Labeling Accuracy, and Fingerprint Matching Accuracy.

4.1 Mean Absolute Error

MAE measures the pixel-wise distance between the predicted and actual boundaries of the bounding boxes, indicating segmentation precision [Mu24]. The proposed model achieved MAEs of 26.09 (left), 27.33 (right), 20.23 (top), and 52.92 (bottom), all below the NIST tolerance threshold of 64 pixels. Further details are illustrated in Figure 3 and summarized in Table 2.

Tab. 2: The Mean Absolute Error (MAE) and its standard deviation were computed to assess the performance of the contactless segmentation system on contactless fingerphoto dataset. The MAE was determined by averaging the absolute differences between each side of the detected bounding box and the corresponding side of the ground-truth bounding box, measured in pixels. A lower MAE value indicates better performance in accurately segmenting the fingerphotos.

Dataset	Side	MAE (Std. dev.)
Contactless	Left	26.09 (65.36)
	Right	27.33 (64.29)
	Top	20.23 (52.97)
	Bottom	52.92 (90.93)

4.1.1 Error in Angle Prediction (EAP)

The EAP metric quantifies the deviation between the predicted angles of bounding boxes and their corresponding ground truth, reflecting the rotational accuracy [Mu24]. The proposed model demonstrated an average EAP of 5.92°, with a standard deviation of 11.98°, indicating strong rotational robustness. The performance is depicted in Figure 4.

4.1.2 Label Prediction Accuracy of the contactless segmentation model

This metric evaluates how accurately the model assigns the correct finger labels using the Hamming Loss measure [KD15] [Mu24]. The model achieved 97.46% labeling accuracy on our contactless dataset, demonstrating its superior accuracy in predicting fingerprint labels.

Fig. 4: The histogram of the error in fingerphoto angle prediction by the contactless segmentation model on the fingerphoto dataset. This error is calculated by subtracting the angles predicted by the models from the ground-truth angles of the fingerphoto images. Low standard deviation values indicate better results.

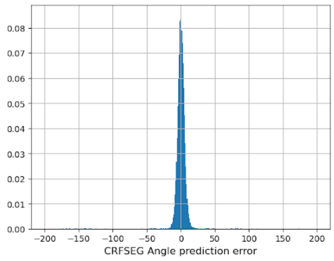
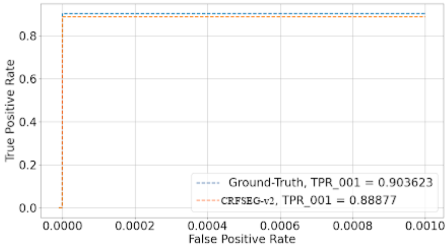


Fig. 5: The Receiver Operating Characteristics (ROC) for the fingerprint matching performance of Ground Truth, a newly developed fingerphoto segmentation model in the contactless dataset.



Tab. 3: The True Positive Rate (TPR) at a False Positive Rate (FPR) of 0.001 is evaluated for both ground-truth and contactless segmentation model segmented fingerprints within the dataset. The results indicate the proposed model performed close to the ground-truth level in terms of fingerprint matching.

Model	Accuracy
Ground-truth	90.36%
Contactless Model	88.88%

4.1.3 Fingerprint Matching

To calculate the matching accuracy, we generated two sets of segmented fingerprint images from the contactless fingerphoto dataset. One set of segmented fingerprints is generated using the information of human-annotated (ground truth) bounding boxes and another set is generated using the information of the proposed model (segmentation model) generated bounding box information. After segmenting contactless finger photos, we applied the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique to enhance the contactless segmented fingerprint images, as mentioned in [Ch23]. After that we evaluated all possible genuine comparisons for each fingerprint in the segmented fingerprint set, while randomly selecting 100 non-mated fingerprints to create an imposter distribution. The segmentation model achieved 88.88% matching accuracy, compared to 90.36% with human-labeled boxes. Performance is shown in Table 3 and the ROC curve in Figure 5.

5 Comparison with the Contact-Based Fingerprint Segmentation Model

For the contact-based fingerprint segmentation model, the mean absolute errors (MAEs) for the left, right, top, and bottom sides were 18.48, 17.77, 18.31, and 18.25, respectively, as reported in Murshed et al. [Mu24]. In comparison, the contactless fingerprint segmentation model yielded higher MAEs of 26.09, 27.33, 20.23, and 52.92 for the same sides. A lower MAE indicates better performance. Regarding the error in angle prediction (EAP), the contactless segmentation model achieved an average EAP of 5.92° with a standard deviation of 11.98°, whereas the contact-based dataset achieved a comparable average EAP of 5.86° with a lower standard deviation of 8.05°. The smaller EAP and lower standard deviation means better localization results. The label prediction accuracy of the contact-based segmentation model was 98.80%, while the contactless segmentation model achieved a slightly lower accuracy of 97.46%. Similarly, the fingerprint matching accuracy for the contact-based segmentation method was 97.17%, compared to 88.88% for the contactless segmentation method.

Overall, the MAE, EAP, and labeling accuracy of our contactless segmentation model remain competitive with high-precision segmentation systems trained on contact-based slap images. However, the performance of the contactless model is lower than that of the contact-based model. This is primarily because contact-based fingerprints are captured with specialized sensors in controlled environments, while contactless fingerprints are acquired with mobile phone cameras under natural, unconstrained conditions. The resulting variability in lighting, background, and finger orientation introduces noise and reduces segmentation and matching performance.

6 Conclusion

In this paper, we have examined the potential of contactless fingerprint authentication systems as a promising alternative to traditional contact-based authentication methods. By employing deep learning techniques and leveraging convolutional networks trained with an end-to-end approach, we have created a highly accurate segmentation system tailored for the precise segmentation and extraction of contactless fingerphoto images. Our evaluations of real-world datasets reveal the effectiveness and reliability of the proposed method, highlighting its capability to overcome common challenges such as rotation and noise. A significant advantage of the proposed system is its flexibility to be fine-tuned with additional datasets, improving its performance and applicability across a broad and diverse range of finger photo images. To the best of our knowledge, this level of adaptability and accuracy in segmentation is something that is missing in current commercial slap segmentation systems, which typically do not offer easy fine-tuning capabilities for diverse finger photo datasets.

7 Acknowledgements

This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant Number 1650503.

References

[Ca15] Cadd, S. et al.: Fingerprint Composition and Aging: A Literature Review, 2015.

[Ch23] Chakraverti, S. et al.: De-noising the image using DBST-LCM-CLAHE: A deep learning approach, 2023.

[GJ23] Grosz, S. A.; Jain, A. K.: AFR-Net: Attention-driven fingerprint recognition network, 2023.

[Gr21] Grosz, S. A. et al.: C2cl: Contact to contactless fingerprint matching. IEEE Transactions on Information Forensics and Security 17, pp. 196–210, 2021.

[He16] He, K. et al.: Deep Residual Learning for Image Recognition, 2016.

[Ji19] Jia, J. et al.: EMBDN: An efficient multiclass barcode detection network for complicated environments, 2019.

[KD15] Khamis, S.; Davis, L. S.: Walking and talking: A bilinear approach to multi-label action recognition, 2015.

[Ko07] Ko, K.: User's guide to nist biometric image software (nbis), 2007.

[Ma20] Malhotra, A. et al.: On Matching Finger-Selfies Using Deep Scattering Networks, 2020.

[Ma22] Maltoni, D. et al.: Fingerprint sensing, 2022.

[Mu24] Murshed, M. S. et al.: Deep Age-Invariant Fingerprint Segmentation System, 2024.

[MV21] Marasco, E.; Vurity, A.: Fingerphoto Presentation Attack Detection: Generalization in Smartphones, 2021.

[Wi19] Wild, P. et al.: Comparative Test of Smartphone Finger Photo vs. Touch-based Cross-sensor Fingerprint Recognition, 2019.

[Wu19] Wu, Y. et al.: Detectron2, https://github.com/facebookresearch/detectron2, 2019.