Deep Slap Fingerprint Segmentation for Juveniles and Adults

M. G. Sarwar Murshed [*], Robert Kline[*], Keivan Bahmani[*], Faraz Hussain, Stephanie Schuckers

Electrical and Computer Engineering
Clarkson University, Potsdam, NY
{murshem, kliner, bahmank, fhussain, sschucke}@clarkson.edu

***Abstract*—Many fingerprint recognition systems capture four fingerprints in one image. In such systems, the fingerprint processing pipeline must first segment each four-fingerprint slap into individual fingerprints. Note that most of the current fingerprint segmentation algorithms have been designed and evaluated using only adult fingerprint datasets. In this work, we have developed a human-annotated in-house dataset of 15790 slaps of which 9084 are adult samples and 6706 are samples drawn from children from ages 4 to 12. Subsequently, the dataset is used to evaluate the matching performance of the NFSEG, a slap fingerprint segmentation system developed by NIST, on slaps from adults and juvenile subjects. Our results reveal the lower performance of NFSEG on slaps from juvenile subjects. Finally, we utilized our novel dataset to develop the Mask-RCNN based Clarkson Fingerprint Segmentation (CFSEG). Our matching results using the Verifinger fingerprint matcher indicate that CFSEG outperforms NFSEG for both adults and juvenile slaps. The CFSEG model is publicly available at https://github.com/keivanB/Clarkson_Finger_Segment**

***Index Terms*—juvenile and adult fingerprints, deep slap Segmentation, Mask-RCNN**

1. Introduction

Due to their high accuracy and convenience, fingerprint-based recognition systems are now in widespread use, e.g. in border crossings, cell-phone authentication, and health care. Many fingerprint recognition systems use high-end multi-finger scanners instead of single finger scanners in order to achieve more accurate identification [1]. In such systems, the fingerprint processing pipeline relies on a fingerprint segmentation model to localize individual fingerprints [2] within each four-finger slap. As a result, slap fingerprint segmentation models are an integral part of fingerprint recognition systems as the loss of fingerprint ridge structure due to improper segmentation can significantly degrade the overall matching performance [3].

To the best of our knowledge, all current slap fingerprint segmentation models have been developed using adult datasets [2], [4], [5]. However, since aging is known to affect the size and ridge-valley structure of the fingerprints, lower the quality and diminish the matching performance [6], [7], careful consideration and modeling are required to accurately segment and match slaps from juvenile subjects. However, the lack of publicly available juvenile datasets, as well as privacy concerns, have led to a dearth of work on this topic.

In this work, we have developed a human-annotated in-house dataset of 15790 slaps (Children: 6706, Adult: 9084). Our new dataset allows us to accurately evaluate the performance of the NFSEG model on both adult and juvenile subjects. NFSEG is one of the most popular fingerprint segmentation system published by NIST [4]. This segmentation system is used as a benchmark algorithm to evaluate the performance of a newly developed slap segmentation algorithm. Additionally, we utilized our human-annotated dataset to develop Clarkson Fingerprint SEGmentation (CFSEG) – a novel deep learning-based fingerprint segmentation model designed to accurately segment both adult and juvenile slap fingerprints. Our test results using the *Verifinger* fingerprint matcher (version 10, compliant with NIST MINEX [8]), show that CFSEG outperforms NFSEG in both adult and juvenile cohorts of our dataset.

2. Collection and Annotation of Slap Fingerprints

We combined several in-house adult and juvenile datasets to create a balanced and representative dataset for training and evaluating CFSEG. Our dataset contains a total of 15790 slap fingerprints (adults: 9084, juvenile: 6706) from 203 adults and 242 juvenile subjects. All slaps are captured at 500 ppi using the FBI certified *Crossmatch L Scan Guardian (9000251)* fingerprint scanner [9].

Initially, all slaps are segmented using the NFSEG model. Subsequently, we developed a Graphic User Interface (GUI) based on the open-source `labelImg` [10] platform for fingerprint annotation. Our GUI allows us to utilize NFSEG segmentations as a starting point for each capture. This process significantly reduced our annotation time to a practical time frame. Additionally, this pipeline allows us to utilize the estimated angle of rotation from the NFSEG and rotate each slap to the upright position. In the first stage, human annotators evaluated each slap for the correct angle of rotation. In the second stage, annotators cycle through slaps and correct the fingerprints that are incorrectly

[*] Authors contributed equally to this work.

segmented by NFSEG while the rest of the bounding boxes remain untouched. This process resulted in a cleaned and annotated dataset with a total of 52674 localized fingerprints (adult: 30304, juvenile: 22370). Finally, we used a rigorous two stage 10-fold cross-validation process based on the unique identities in the dataset to derive our training, validation, and test sets. At each fold, 80% of the adult and 80% of the juvenile identities were used for training while the remaining 20% of identities of both types were used for validation (10%) and testing (10%).

3. Clarkson Fingerprint Segmentation (CFSEG)

Our CFSEG model for slap segmentation is based on the Mask-RCNN architecture [11]. In this section, we discuss the overall Mask R-CNN model, the loss functions, and the training scheme used for training the CFSEG.

A. Architecture of Mask R-CNN

Mask-RCNN is a simple and flexible two-stage deep architecture developed to perform semantic segmentation, object localization, and object instance segmentation of natural images [11]. As a result, we adopt Mask-RCNN architecture for CFSEG.

The first stage of this architecture contains the Region Proposal Network (RPN) which proposes candidate object bounding boxes. It consists of two networks, a Convolution Neural Network (CNN) and a Region Proposal Network (RPN). The CNN is the backbone of the Mask R-CNN architecture, and it is responsible for extracting feature maps from the input images. Any CNN model designed for image classification tasks (such as ResNet, MobileNet, or VGG) could be used as the backbone network [12]. Previous research has shown that the ResNet-FPN backbone provides better performance in terms of accuracy and speed on object detection tasks [11]. Therefore, we have used ResNet-101 [13] as the backbone network in our experiments. On the other hand, the RPN is responsible for generating a set of 'Region Proposals', which signify regions in the feature map that have a high probability of containing objects. In the second stage, a small, Fully Connected Neural Network (FCNN) takes the proposed regions from the first stage and predicts bounding boxes and object classes for each of them. Note that the proposed regions generated by the RPN can be of different sizes. However, fully connected layers in the networks only take a fixed size vector to make predictions. The ROIAlign [11], which is a modified version of Max-Pooling, is used to fix the size of these proposed regions. ROIAlign also fixes the misalignment which preserves the exact spatial location of an object and helps to improve overall accuracy.

A fully convolutions network is added (parallel with the existing branch for classification and bounding box regression) which is responsible for predicting segmentation masks in a pixel-to-pixel manner on each Region Of Interest (ROI).

B. Loss function

For training RPNs, a class label indicating whether it is an object or not is assigned to each anchor. Anchor boxes are reference boxes placed at different positions in the input image. A positive label is assigned to an anchor if that anchor has an Intersection over Union (IoU) greater than 0.7 with any ground truth objects. A negative label is assigned to an anchor if that anchor has an IoU less than 0.3 with any ground truth objects. The anchors that have an IoU value between 0.3 and 0.7 are considered neutral and excluded from the training set. Shift and resizing operations are performed before training the RPN which helps make the anchor cover the ground truth object completely. With this definition, a multi-task loss function is used to train the Mask R-CNN. This loss function is divided into three parts which combine the loss of classification, localization, and segmentation as follows:

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

The class, bounding box, and mask losses are represented by L_{cls}, L_{box} and L_{mask} respectively. L_{cls} is the log loss function over two classes carried out as a binary classification by predicting an object being a target object or not. The smooth L1 loss is used for bounding box loss L_{box}. Mask loss L_{mask} is the average binary cross-entropy loss, and it is calculated for each class separately. These calculations prevent competition among the classes when generating masks. The details of different loss functions are as follows:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \frac{\lambda}{N_{box}} \sum_i p_i^* L_{box}^{smooth}(t_i - t_i^*) + (-1 * \frac{1}{m^2}) \sum_{1\le i,j\le m} [y_{ij}\log y_{ij}^k + (1-y_{ij})\log(1-y_{ij}^k)] \tag{2}$$

Here, i represents the index number of an anchor and p_i is the predicted probability of an anchor being an object. The ground-truth binary label is represented by p_i^*. t_i and t_i^* represent the vector with four coordinates of the predicted and ground-truth bounding box respectively. m is the size of the mask output; y_{ij} is the label of a cell (i,j) in the true mask for the region of size $m \times m$; y_{ij}^k is the k-th mask predicted value of the same cell in the mask learned for the ground-truth class k.

C. Training

An image-centric training approached [14] is utilized in this work and only positive anchors are used to determine the loss of the network. Our training approach is as follows:

- N ROIs are generated from each image with the ratio of positive to negative ROIs of 1:3, where N is selected as 64 and 512 for backbone and FPN stages respectively.
- When training RPN, the anchors aspect ratio is 3, and the span is 5 scales. The RPN is trained individually

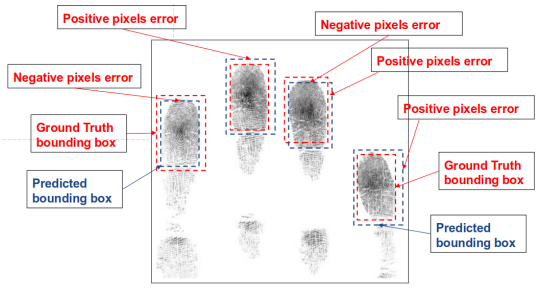

Fig. 1: Example for calculating positive and negative errors between the predicted and ground truth bounding box.

and does not share features with other parts of the Mask-RCNN network.

- The Stochastic Gradient Descent (SGD) with a learning rate of 0.001, momentum of 0.9, and decay of 0.0001 are used to train the network.

During inference, the region proposal number increases to 300 and 1000 for the backbone and FPN stages, respectively. This increment helps the network not miss any potential regions during inference time. The bounding box prediction branch considers each proposal proposed by the backbone and FPN network. Then the non-maximum suppression is performed to remove the low-scoring ROIs. To keep computational overhead to a minimum during training and inference, the mask branch only runs over the 50 highest scoring ROIs.

4. Evaluation and Results

In this work, we utilized both the Mean Absolute Error (MAE) of the predicted bounding boxes and fingerprint matching performance as our evaluation metrics. In this section, we describe our evaluation process and results.

A. Mean Absolute Error (MAE)

The MAE measures the distance between the detected bounding box and the annotated ground-truth bounding box in terms of pixels. Slap fingerprint segmentation models have to find the balance between over-extending the bounding boxes (over-segmentation) vs. excessively reducing the predicted bounding box (under-segmentation). Over-segmentation can lead to the ridge-valley structure of other fingerprints leaking into the bounding box of other fingerprints. This extra noise can potentially degrade the matching performance. On the other hand, under-segmentation can cause the loss of the valuable parts of the fingerprints.

Previous work has identified that under-segmentation beyond 32 pixels from the sides and 64 pixels from the top or bottom of the fingerprint can negatively affect the matching performance [3]. As a result, we calculate and report MAE for *each side* of the bounding box individually, i.e. the error is calculated by measuring how far those four sides are from the four sides of the ground truth bounding box. In Figure 1, the red dotted rectangle represents the ground-truth bounding box generated by a human-annotator and the blue rectangle represents the detected bounding box by a fingerprint detection model. If any side of a detected bounding box captures more information than the side of the ground-truth bounding box, we consider it as a positive error. On the other hand, if any side of a detected bounding box captures less information than the ground-truth bounding box, we consider it as a negative error. Finally, the MAE for each side is calculated using the equation 3, Where, n is the total number of fingerprints in the test dataset. X represents any side such as left, right, top, bottom, of the bounding box.

$$MAE = \frac{\sum_{i=0}^{n} abs(X\ error_i)}{Total\_fingerprints\_in\_the\_dataset} \quad (3)$$

Table 1 shows the mean and standard deviation of the MAE for NFSEG and CFSEG models evaluated using our dataset. We can observe that, compared to NSFEG, CFSEG has lower MAE in almost all directions and produces more precise bounding boxes for both adults and juvenile subjects. Additionally, our results confirm that compared to adults, NFSEG suffers from higher MAE in all directions in the juvenile cohort of our dataset. Furthermore, we observe that in both adult and juvenile subjects, NFSEG has difficulty in accurately localizing the lower boundary of fingerprints. Figure 2 illustrates the histogram of the bottom errors for adult and juvenile subjects. We can observe that while CFSEG maintains the performance in Juvenile subjects, NFSEG is particularly susceptible to this type of error in juvenile subjects. Our failure analysis on such cases revealed that the high error observed in this direction is because of the over-segmentation at the distal interphalangeal joint. Figure 3 depicts an example of this problem, where the red bounding box indicates the ground truth and the blue bounding box presents the bounding box predicted by NFSEG. Additionally, the long tail of the error histogram of the NFSEG model for the juvenile subjects suggests that NFSEG over-segmented many juvenile samples below the Minimum Tolerance Limit (MTL) of -64 pixels suggested by Slapseg-II [3]. This can potentially reduce the matching performance of the system.

TABLE 1: Mean Absolute Error (MAE) [Mean (Std.)] for NFSEG and CFSEG

Age Group	Adults		Children	
Model	NFSEG	CFSEG	NFSEG	CFSEG
Left	07.13(28.59)	08.21(14.66)	12.09(38.19)	08.36(16.25)
Top	13.18(44.70)	13.05(20.26)	28.47(92.51)	13.77(21.13)
Right	09.46(31.52)	07.73(13.78)	13.60(34.55)	07.21(14.16)
Bottom	35.23(84.73)	16.60(24.41)	102.87(159.00)	15.00(22.92)

B. Fingerprint Matching

In the second stage of the evaluation process, we utilized the *Verifinger* fingerprint matcher (version 10, compliant with NIST MINEX [8]) to compare the matching accuracy using the fingerprints segmented with NFSEG and CFSEG. This helps us better analyze the effects of segmentation accuracy on the overall performance of the fingerprint recognition system. In order to evaluate the matching accuracy, we used the annotated ground truth, NFSEG,

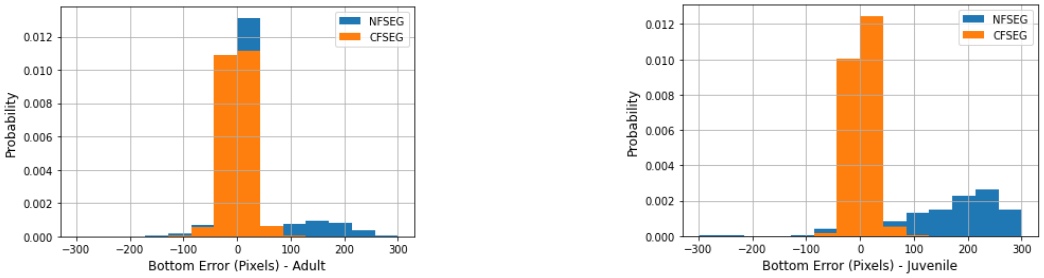

Fig. 2: The histogram of the bottom error (pixels) from predicted bounding boxes by the NFSEG and CFSEG models for adult (Left) and Juvenile (Right) subjects.

Fig. 3: Example of over-segmentation error in NFSEG. The red rectangle shows the ground-truth bounding box while the blue rectangle shows the detected bounding box by the NFSEG.

and CFSEG bounding boxes. For every fingerprint in the dataset, we evaluated all the mated comparisons for our genuine distribution (172348), while randomly selecting 20 non-mated fingerprints to construct an imposter distribution (441322 imposter comparisons). In total, we performed 600,000 comparisons using all 10 fingers.

Our results indicate that for both adults and juvenile cohorts the fingerprints segmented with CFSEG provide higher accuracy across operation points. Additionally, we observe lower overall matching performance in juvenile subjects compared to adults. Finally, Table 2 depicts the True Positive Rate (TPR) at False Positive Rate (FPR) of 0.1% for fingerprints segmented using NFSEG, CFSEG, and human-annotators.

TABLE 2: TPR [Mean (Std.)] at FPR of 0.001 for NFSEG, CFSEG, and Ground Truth.

Model	Adults	Children
NFSEG	0.9972 (0.0027)	0.9675 (0.0135)
CFSEG	0.9977 (0.0026)	0.9687 (0.0135)
Ground-Truth	0.9991 (0.0011)	0.9716 (0.0137)

5. Conclusion and Limitations

The experimental results indicate that our novel CFSEG model provides more precise bounding box predictions with respect to NFSEG for both adult and juvenile subjects. However, CFSEG currently depends on the estimated rotation angle of the slap acquired from NFSEG. This will be addressed in future work by the introduction of a regression sub-network for angle prediction and artificially augmenting the training dataset with rotated slaps. Additionally, our fingerprint matching experiment is conducted using the human-annotated class labels (i.e. index, middle, or ring fingers) for all algorithms to produce a meaningful comparison between the predicted bounding boxes. Future work can focus on designing a class prediction sub-network to integrate this task into the CFSEG and further improve the overall performance. Finally, the trained CFSEG model is publicly available to other researchers at https://github.com/keivanB/Clarkson_Finger_Segment.

References

[1] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[2] P. Feulner, "Slap fingerprint segmentation evaluationII (slapsegII)," https://www.nist.gov/itl/iad/image-group/slap-fingerprint-segmentation-evaluation-ii-slapsegii, Feb. 2010.

[3] C. I. Watson, "SlapSegII Analysis: Matching Segmented Fingerprint Images," pp. 1–53, Nov. 2010.

[4] Kenneth Ko, Wayne J. Salamon, "NIST Biometric Image Software (NBIS)," https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis, February 2010.

[5] P. Gupta and P. Gupta, "A slap fingerprint based verification system invariant to halo and sweat artifacts," *Applied Mathematical Modelling*, vol. 54, pp. 413–428, 2018.

[6] J. Galbally, R. Haraksim, and L. Beslay, "A study of age and ageing in fingerprint biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 1351–1365, 2018.

[7] R. Haraksim, J. Galbally, and L. Beslay, "Fingerprint growth model for mitigating the ageing effect on children's fingerprints matching," *Pattern Recognition*, vol. 88, pp. 614–628, 2019.

[8] C. Watson, G. Fiumara, E. Tabassi, S. Cheng, P. Flanagan, and W. Salamon, "Fingerprint vendor technology evaluation, nist interagency/internal report 8034: 2015," 2014.

[9] "Cross Match L SCAN Patrol and Patrol ID fingerprint scanners," https://neurotechnology.com/fingerprint-scanner-cross-match-l-scan-patrol.html.

[10] Tzutalin, "Labelimg," https://github.com/tzutalin/labelImg, 2015.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[12] S. Gonzalez, C. Arellano, and J. Tapia Farias, "Deepblueberry: Quantification of blueberries in the wild using instance segmentation," *IEEE Access*, vol. PP, pp. 1–1, 08 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.