CNN based Finger Region Segmentation for Finger Vein Recognition

Bernhard Prommegger, Dominik Söllinger, Georg Wimmer and Andreas Uhl
University of Salzburg
Department of Artificial Intelligence and Human Interfaces
Salzburg, Austria
E-Mail: {**bprommeg**, dsoellinger, gwimmer, uhl}@cs.sbg.ac.at

***Abstract*—Finger region segmentation is an important step in a biometric finger vein recognition toolchain. Its aim is to separate the finger region from background and all other objects of the image. So far, finger region extraction for finger vein recognition systems has mainly used classical image processing based systems. In this work three state-of-the art convolutional neural network (CNN) based architectures for segmentation, namely *Mask R-CNN*, *CCNet* and *HRNet*, are evaluated. A major advantage of the presented CNN-based approach compared to classic image processing approaches is that the images neither have to be preprocessed nor any parameters have to be optimized. All that is required is a sufficient number of already segmented finger vein images for training.**

I. Introduction

Finger region segmentation is the task of generating a binary mask that separates finger region pixels from non-finger region pixels. In general, segmentation tasks are nowadays almost exclusively approached with CNNs (e.g. [1], [2], [3]). This is not quite the case in biometrics. Although CNN based approaches are already used for some modalities (e.g. [1] for face, [4] for iris or [5] for fingerprints), others still rely mainly on classical image processing systems. For example in finger vein biometrics, state-of-the-art finger segmentation systems still utilize classical image processing systems, e.g. [6], [7], [8].

In a CNN based segmentation approach, the segmentation network is trained and evaluated on ground truth data. Since such a ground truth is not available for many segmentation tasks, it has to be created by manual segmentation and annotation. This is a time consuming task. In the course of this work, such a ground truth was created.

The main contributions of this paper are:

1) The analysis of three state-of-the-art segmentation networks on five publicly available data sets in three different training scenarios. The training scenarios differ in the data used for training.
2) The creation of a finger segmentation ground truth for all five sets which will be made publicly available.

The rest of the paper is structured as follows: While the CNN architectures used are presented in section II, section III briefly describes the data sets used. The experimental setup and the results are described in section IV. Section V summarizes the findings and gives a brief outlook on the planned future work.

II. CNN Models

This work evaluates three state-of-the art semantic segmentation networks, namely *Mask R-CNN* [1], *CCNet* [2] and *HRNet* [3], for finger region segmentation. Except of resizing and normalization (which are standard pre-processing steps for CNNs and require hardly any computation time or adaption to different datasets) no pre-processing is needed. For *Mask R-CNN* and *CCNet* the images are resized to 256x192 pixels, for *HRNet* to 1024x768 pixels (*HRNet* rescales by ¼ in its stem, which will again result in a 256x192 pixel output).

A. Mask R-CNN

Mask R-CNN is a two stage network for object instance segmentation. It is based on *Faster R-CNN* [9] which already returns a bounding box and a class label for every candidate object. *Mask R-CNN* now adds a third branch that additionally returns the mask of the object. If more than one candidate is returned, always the one with the largest area (bounding box) is selected.

The model in use is the one provided by Torchvision using a *ResNet-50* backbone pre-trained on COCO[1]. The network is trained for 30 epochs with a batch size of 8 and a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.005, a momentum of 0.9 and a weight decay of 0.0005. The learning rate is decreased by factor 10 every fifth epoch.

B. CCNet

CCNet increases the segmentation results by considering the semantic relation of the pixels with an image. It uses vertical and horizontal criss-cross attention modules to collect the context information. Then a further cycle operation is taken, and each pixel can finally capture the full image correlation.

For the experiments the implementation provided by the authors[2] is used. The network is trained for 30 epochs with a batch size of 16 and a SGD optimizer with an initial learning rate of 0.005, a momentum of 0.9 and a weight decay of 0.0005. The model is trained from scratch.

[1]https://cocodataset.org
[2]https://github.com/speedinghzl/CCNet

978-1-6654-6962-3/22/$31.00 ©2022 IEEE

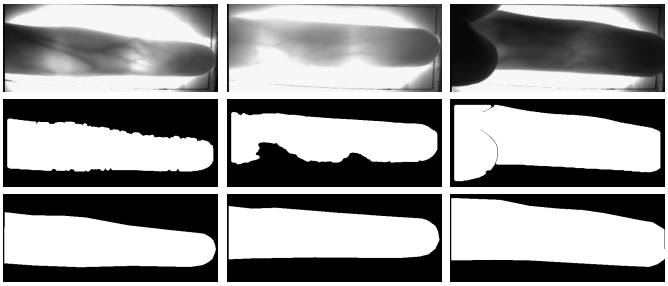

Fig. 1. Sample images from HKPU: From top to bottom: original image, provided mask, manually segmented mask. Left: good quality image, middle: overexposed image, right: image where parts of the finger region is covered by adjacent fingers.

C. HRNet

HRNet is a network that was originally developed for human pose estimation, and later extended to also support semantic segmentation. In contrast to other semantic segmentation networks, it attempts to preserve the high resolution throughout the whole process. This is achieved by maintaining multiple resolutions in parallel with repeated information exchange, the so called multi-resolution fusion, across the different resolutions. As representation head we used the *HRNet*V2.

In this article a re-implementation of the code provided by the authors[3] is used. The model uses the exact same structure as described in section 3 of [3]. The network is trained for 30 epochs with a batch size of 12. The used Adam optimizer was initialized with a learning rate of 0.0001 and never adopted. As loss function the cross entropy loss is used. The model is trained from scratch.

III. Data Sets

The data used in the experiments was taken from several finger vein data sets: *The Hong Kong Polytechnic University Finger Image Database* (HKPU) [10], *PLUSVein-Contactless Finger and Hand Vein Database* (PLUSVein-CL) [11], *PROTECT MultiModal Dataset v2* (PMMDB) [12] and *University of Twente Finger Vascular Pattern Dataset* (UTFVP) [13]. The corresponding ground truth was created from a single person and is available on our website[4].

A. HKPU

The HKPU contains finger vein and finger texture images of 156 subjects from two fingers acquired in two separate sessions. It also provides masks for the finger region. The quality of some of the acquired finger vein images is rather poor. Many images are overexposed. Sometimes, however, other objects (e.g. neighbouring fingers) occlude the finger that is actually to be acquired. As a result of this, finger regions segmentation becomes difficult. This is also reflected in the provided masks. Figure 1 shows some representative samples from the data set. The images in the top row are the finger vein images provided by HKPU, in the middle row there are the provided finger masks and in the bottom row the manually segmented finger regions, respectively.

[3]https://github.com/HRNet

[4]http://wavlab.at/sources/Prommegger22a

TABLE I
Overview of finger vein images used to manually segment the finger region (ground truth) per data set

Dataset	Subjects	Fingers	Images	Comment
HKPU	20	2	480	subjects 1-20
PLUSVein-CL	60	6	612	not published
PMMDB-FR	31	4	613	palmar view (0°) session 1
PMMDB-FV3	16	6	480	LED version palmar view session 2
UTFVP	20	6	480	subjects 1-20

The images in the left column are from an image of good quality. The extracted mask corresponds quite well to the actual finger region, even if it is a bit ragged on its outline. The images in the middle column come from an overexposed image. Some of the overexposed areas are not recognized as finger regions in the supplied mask. In the last image, some parts of the finger region are covered by neighbouring fingers. In the mask provided by the HKPU both, the actual finger as well as the areas from the neighbouring fingers are considered as finger region. It is worth mentioning that the HKPU does not treat such cases consistently throughout the data set: sometimes the covering objects are included in the provides masks (as in this example), sometimes they are excluded. The question now is how to best deal with such areas: Exclude the covered areas from the finger region or define the finger region as if the whole finger is visible? In this work, the authors decided to segment the finger region as if the whole finger would be visible. The mask for sample 3 (bottom row) shows an example how partially covered fingers are handled during manual segmentations.

For the experiments, the finger regions where manually segmented for the first 20 subjects in the data set. This sums up to a total of 480 finger region masks. Table I summarizes the information of the manually segmented images for all five data sets in use.

B. PLUSVein-CL

The data set provides hand and finger vein images together with corresponding ROI images acquired in a contactless acquisition scenario. It is important to mention that the ROIs of the finger vein images do not reflect the finger regions, but a rectangular region that has been cut out with a defined size at a defined position. Therefore, they cannot be used to train a CNN for finger region segmentation. In general, the images in this data set are of uniform quality. The main difference to the other data sets is that they contain also parts of neighbouring fingers. The left column in figure 2 shows an image and the corresponding finger mask. Above and below the middle finger (intended for acquisition) one can also see parts of the index and ring finger.

The published data set contains finger vein images of six fingers from 42 subjects. In addition to the published data, there exist finger vein images of further 60 subjects for which

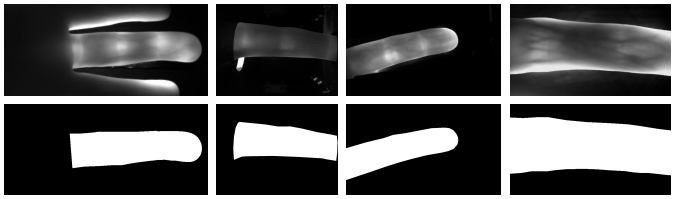

Fig. 2. Sample images (top row) and manually segmented finger mask (bottom row) for PLUSVein-CL, PMMDB-FR, PMMDB-FV3 and UTFVP.

only one or max two samples of each finger were acquired. As this additional data provides more variety (more subjects), the authors decided to use this data to create the ground truth.

C. PMMDB

This data set contains biometric data for several modalities including face, periocular, iris, anthropometrics and hand- and finger veins acquired in two sessions (session 1 indoors, session 2 outdoors). The finger vein images have been acquired using two different sensors: the *PLUS OpenVein finger vein scanner* [14] (the data acquired with this sensor will be called as PMMDB-FV3 in this paper) and the *PLUS multi-perspective finger vein scanner* [15] (data is called PMMDB-FR). For PMMDB-FV3 we use only the palmar images acquired in the indoor session as the outdoor data is of low quality. For PMMDB-FR both sessions have been acquired indoors. Therefore, the authors decided to use data of session 1 as this session contains more samples.

The images acquired for both data sets are of uniform quality. The second (PMMDB-FR) and third column (PMMDB-FV3) of figure 2 shows sample images of this data set. The data set itself does not provide finger masks or ROI images.

D. UTFVP

The UTFVP provides finger vein images from 60 volunteers acquired in two sessions. At each session two samples per finger were captured. The acquired images are of high quality. One sample and its corresponding finger region is visualized in the right columns of figure 2. The data set does not provide any finger region or ROI information. For the ground truth, all images of the first 20 of the 60 subjects where used.

IV. Experiments

In the experiments, the applicability of three state-of-the-art segmentation networks (*Mask R-CNN*, *CCNet* and *HRNet*) for the segmentation of finger regions is evaluated. The experiments are carried out on five different data sets (HKPU, PLUSVein-CL, PMMDB-FR, PMMDB-FV3 and UTFVP). There are three different scenarios how the data for training the CNNs is put together:

1) **Training on a single data set** (`DS only`): Training and evaluation data are taken from the same data set. This means that if HKPU is to be analysed, the CNN will only be trained on data from HKPU.
2) **Training on all data sets** (`All DS`): Regardless of the evaluation data set, the CNN model is trained with data from all five data sets.

Fig. 3. Representative segmentation images for HKPU. Left: manual generated mask, right: mask generated by *CCNet* using `DS only` scenario, left: difference image.

3) **Excluding evaluation data set from training** (`All but DS`): In the last scenario, no data from the dataset to be evaluated is used during training. This means that if HKPU is to be analysed, the CNN is trained on data from the other four datasets (PLUSVein-CL, PMMDB-FR, PMMDB-FV3 and UTFVP).

To ensure the separation of training and evaluation data, the datasets are divided into four folds. Three folds are used for training and one for evaluation. The experiments are repeated four times such that each fold is used exactly once for evaluation and three times for training. This way, results for the whole data set were generated without any overlap in training and evaluation data. The assignment of the images to the fold is based on subject. This means that all images of the same subject are always in the same fold. The assignment of the subjects to the folds is done random.

The results of our experiments along with sample images are shown in the following sections. To assess the quality of the segmentation, the number of misclassified pixels over the entire image (finger region and background) and again individually for the finger region, i.e. number finger region pixels classified as background, and vice versa are used as performance indicators. Due to the different properties of the data sets, e.g. the size of the image and the size of the fingers within the images, providing the differences in absolute pixel values would lead to values that are not comparable between the different data sets. Therefore, all values are given in %, where 100% corresponds to the number of pixels in the finger region of the manually created masks. So that the quality of the segmentation can be correctly assessed, two standard metrics for segmentation, namely the intersection over union (IoU) and the DICE score, are also included in the evaluation. The results are evaluated by training scenario, CNN model and evaluation data set and can be found in table II. For a better overview, the results are discussed per data set.

A. HKPU

The results for the HKPU in table II shows excellent segmentation results for `DS only` and `All DS`. The average number of incorrectly classified pixels is around 2%. The misclassified pixels are shared equally between finger region and background pixels. When looking at the difference image between a generated masks of an representative image (total misclassified pixels = 2.2%) and the corresponding mask of the ground truth in Fig. 3 for these scenarios, one can see that these pixels essentially form a stripe at the border between the finger region and the background. Such errors at the border between finger region and background cannot be prevented and one can therefore speak of an almost perfect segmentation result. The same effect could also be observed for all other data

TABLE II

NUMBER OF MISCLASSIFIED PIXELS IN THE WHOLE IMAGE, FINGER REGION PIXELS CLASSIFIED AS BACKGROUND AND BACKGROUND PIXELS CLASSIFIED AS FINGER REGION. ALL NUMBERS ARE GIVEN IN % RELATIVE TO THE NUMBER OF FINGER REGION PIXEL IN THE MANUALLY SEGMENTED MASK.

Evaluation Data Set	Training Mode	CNN Model	IoU Mean	DICE Mean	Misclassified Pixels Min	Misclassified Pixels Mean	Misclassified Pixels Max	Finger Region as Background Min	Finger Region as Background Mean	Finger Region as Background Max	Background as Finger Region Min	Background as Finger Region Mean	Background as Finger Region Max
HKPU	DS only	M R-CNN	0.95	0.97	1.01%	5.25%	35.00%	0.02%	3.77%	34.58%	0.24%	1.48%	4.95%
		CCNet	0.97	0.99	0.91%	2.76%	20.19%	0.13%	1.54%	18.73%	0.21%	1.23%	6.55%
		HRNet	0.98	0.99	0.58%	1.71%	31.22%	0.08%	1.05%	31.19%	0.00%	0.66%	5.16%
	All DS	M R-CNN	0.97	0.99	0.93%	2.98%	17.08%	0.03%	1.62%	16.26%	0.24%	1.36%	5.77%
		CCNet	0.98	0.99	0.81%	1.99%	15.69%	0.07%	1.01%	15.19%	0.06%	0.97%	3.22%
		HRNet	0.98	0.99	0.57%	2.38%	34.22%	0.04%	1.45%	34.12%	0.03%	0.93%	11.55%
	All but DS	M R-CNN	0.92	0.95	1.68%	9.04%	144.32%	0.00%	3.94%	99.97%	0.10%	5.11%	80.30%
		CCNet	0.67	0.80	19.13%	47.22%	143.36%	0.00%	5.19%	97.12%	0.00%	42.03%	90.11%
		HRNet	0.73	0.84	4.80%	37.37%	94.70%	0.00%	3.94%	93.29%	1.41%	33.43%	74.94%
PLUSVein-CL	DS only	M R-CNN	0.96	0.98	1.75%	3.86%	12.86%	0.07%	1.46%	7.42%	0.50%	2.39%	12.09%
		CCNet	0.94	0.97	2.28%	6.27%	53.32%	0.14%	2.58%	26.42%	0.27%	3.69%	52.72%
		HRNet	0.95	0.98	0.83%	4.81%	48.13%	0.11%	2.75%	31.62%	0.05%	2.07%	48.02%
	All DS	M R-CNN	0.97	0.98	1.43%	3.53%	13.79%	0.01%	1.16%	8.73%	0.45%	2.36%	12.94%
		CCNet	0.96	0.98	1.55%	3.84%	19.12%	0.22%	2.00%	14.33%	0.20%	1.84%	17.58%
		HRNet	0.95	0.97	0.87%	5.17%	65.88%	0.00%	2.42%	25.53%	0.10%	2.75%	65.88%
	All but DS	M R-CNN	0.59	0.72	2.78%	88.23%	379.01%	0.01%	3.26%	51.45%	1.04%	84.97%	364.87%
		CCNet	0.57	0.72	9.06%	79.09%	304.21%	0.35%	4.16%	100.00%	1.51%	74.93%	288.98%
		HRNet	0.75	0.85	3.74%	39.96%	274.20%	0.21%	5.52%	55.43%	0.00%	34.43%	249.02%
PROTECT-FR	DS only	M R-CNN	0.97	0.98	1.26%	3.26%	21.17%	0.07%	1.50%	20.31%	0.40%	1.75%	3.46%
		CCNet	0.97	0.99	1.03%	2.92%	23.08%	0.18%	1.72%	22.73%	0.14%	1.20%	8.67%
		HRNet	0.99	0.99	0.44%	1.30%	23.30%	0.10%	0.90%	22.97%	0.02%	0.40%	3.15%
	All DS	M R-CNN	0.97	0.99	1.33%	2.99%	17.36%	0.04%	1.26%	15.15%	0.28%	1.73%	3.88%
		CCNet	0.98	0.99	0.79%	1.95%	26.30%	0.12%	1.11%	20.25%	0.15%	0.84%	7.89%
		HRNet	0.99	0.99	0.45%	1.36%	19.55%	0.04%	0.70%	19.07%	0.09%	0.66%	4.49%
	All but DS	M R-CNN	0.95	0.98	1.47%	4.72%	29.91%	0.03%	1.67%	26.93%	0.36%	3.05%	8.45%
		CCNet	0.86	0.92	2.11%	14.16%	42.13%	0.18%	12.64%	41.03%	0.00%	1.51%	13.41%
		HRNet	0.87	0.93	0.87%	16.84%	127.19%	0.01%	1.66%	25.70%	0.00%	15.19%	127.14%
PROTECT-FV3	DS only	M R-CNN	0.96	0.98	1.96%	4.46%	15.81%	0.02%	2.43%	11.47%	0.38%	2.03%	13.63%
		CCNet	0.95	0.98	1.76%	4.77%	21.65%	0.04%	2.26%	10.27%	0.13%	2.51%	19.84%
		HRNet	0.97	0.99	0.92%	2.90%	15.45%	0.19%	1.76%	13.61%	0.10%	1.14%	7.29%
	All DS	M R-CNN	0.96	0.98	1.75%	4.15%	18.28%	0.06%	2.11%	8.29%	0.34%	2.04%	16.94%
		CCNet	0.97	0.98	1.27%	3.25%	13.55%	0.08%	1.62%	8.30%	0.17%	1.63%	6.33%
		HRNet	0.97	0.98	0.78%	3.25%	24.33%	0.02%	1.62%	24.02%	0.14%	1.63%	16.28%
	All but DS	M R-CNN	0.92	0.96	2.91%	8.31%	21.53%	0.18%	5.13%	18.34%	0.59%	3.18%	10.78%
		CCNet	0.91	0.95	2.64%	9.89%	81.18%	0.03%	2.26%	17.23%	1.07%	7.63%	78.72%
		HRNet	0.84	0.91	1.39%	16.30%	116.37%	0.28%	14.73%	77.92%	0.04%	1.57%	104.25%
UTFVP	DS only	M R-CNN	0.95	0.97	0.86%	5.53%	67.43%	0.06%	4.44%	66.01%	0.11%	1.09%	4.17%
		CCNet	0.99	1.00	0.27%	0.71%	3.20%	0.04%	0.32%	2.16%	0.11%	0.39%	2.99%
		HRNet	1.00	1.00	0.19%	0.39%	6.24%	0.01%	0.17%	2.14%	0.01%	0.22%	6.13%
	All DS	M R-CNN	0.98	0.99	0.67%	2.32%	6.51%	0.08%	1.49%	5.86%	0.18%	0.83%	2.05%
		CCNet	0.99	1.00	0.25%	0.57%	8.98%	0.03%	0.29%	8.82%	0.03%	0.28%	0.75%
		HRNet	1.00	1.00	0.16%	0.38%	6.13%	0.02%	0.15%	1.59%	0.02%	0.22%	6.04%
	All but DS	M R-CNN	0.96	0.98	0.82%	3.85%	100.00%	0.23%	3.43%	100.00%	0.00%	0.42%	1.41%
		CCNet	0.96	0.98	1.24%	3.80%	26.55%	0.97%	3.65%	26.55%	0.00%	0.15%	1.00%
		HRNet	0.98	0.99	0.23%	2.17%	17.04%	0.10%	1.76%	16.52%	0.00%	0.40%	8.39%
All Data Sets	DS only	M R-CNN	0.96	0.98	0.86%	4.38%	67.43%	0.02%	2.60%	66.01%	0.11%	1.78%	13.63%
		CCNet	0.97	0.98	0.27%	3.60%	53.32%	0.04%	1.73%	26.42%	0.11%	1.87%	52.72%
		HRNet	0.98	0.99	0.19%	2.30%	48.13%	0.01%	1.37%	31.62%	0.00%	0.93%	48.02%
	All DS	M R-CNN	0.97	0.98	0.67%	3.20%	18.28%	0.01%	1.50%	16.26%	0.18%	1.70%	16.94%
		CCNet	0.98	0.99	0.25%	2.38%	26.30%	0.03%	1.24%	20.25%	0.03%	1.14%	17.58%
		HRNet	0.98	0.99	0.16%	2.58%	65.88%	0.00%	1.30%	34.12%	0.02%	1.28%	65.88%
	All but DS	M R-CNN	0.86	0.91	0.82%	25.16%	379.01%	0.00%	3.38%	100.00%	0.00%	21.78%	364.87%
		CCNet	0.79	0.87	1.24%	32.39%	304.21%	0.00%	5.86%	100.00%	0.00%	26.53%	288.98%
		HRNet	0.83	0.90	0.23%	23.11%	274.20%	0.00%	5.33%	93.29%	0.00%	17.78%	249.02%

sets. The results for the `All but DS` scenario are different. As the HKPU images differ vastly from those of the other data sets (overexposure, occlusions), the CNNs did not learn these properties, and therefore, are not capable to correctly segment the finger region. The best results are obtained for *Mask R-CNN* with an average pixel difference of 9%. For *HRNet*, this value is at 37%, for *CCNet* even at 47%. The difference can be explained by the way the CNNs work. *Mask R-CNN* provides bounding boxes for each candidate while the other two returns probabilities for each pixel. For *Mask R-CNN* we select only the candidate with the largest area, which often corresponds to the region of the desired finger.

Fig. 4 shows the masks generated for all three CNNs and training scenarios for a difficult sample image, where parts of the finger are covered by its neighbouring fingers. In the `DS only` scenario, the CNN was able to segment the occluded part of the finger region correctly. In the `All DS` scenario, only *HRNet* managed to somehow segment the

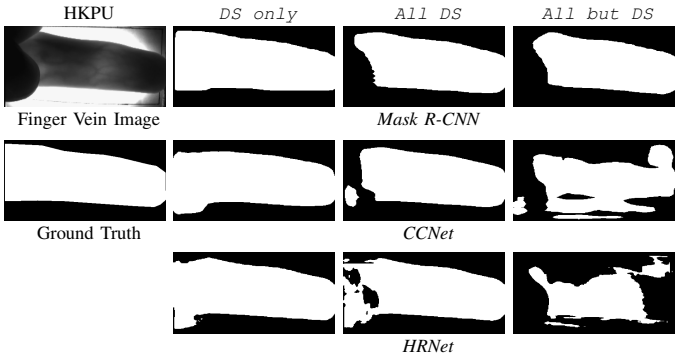

Fig. 4. Segmentation masks generated for HKPU

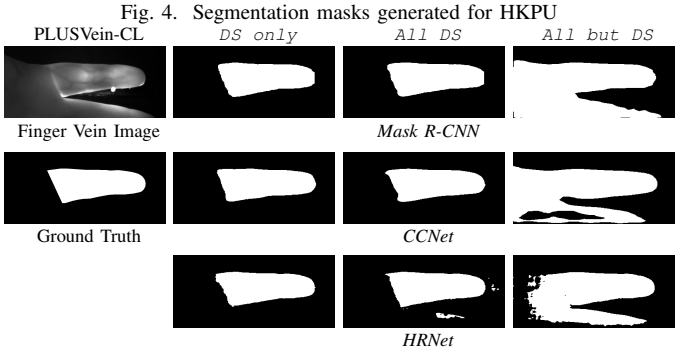

Fig. 5. Segmentation masks generated for PLUSVein-CL

occluded region. The *HRNet* image also shows a typical, easily fixable error: holes. These can be closed with the help of morphological operations. The second typical error can be seen in the *CCNet* image. The segmentation provides two separate components. This can be addressed by using only the largest object in the image. For the `All but DS` scenario only *Mask R-CNN* provides an acceptable result. The mask of the other two CNNs are not usable.

B. PLUSVein-CL

The PLUSVein-CL data set differs from the other data sets by the fact, that the vein images contain also parts of the neighbouring fingers. For the `DS only` and `All DS`, the performance is just slightly inferior to the one of HKPU. The interesting scenario is `All but DS`: Fig. 5 shows the generated masks for a typical image. One can see, that all three CNNs classify not only the region of the intended finger as foreground, but also the region of the neighbouring finger. This behaviour was to be expected since the finger vein images of all other datasets do not contain adjacent fingers. As such, the CNN has not learned to distinguish between intended and adjacent fingers and classifies both as foreground (finger regions). This can be seen by the high error rates of the wrong classified background pixels.

C. PMMDB-FR

As with the previous data sets, very good results are achieved for the `DS only` and `All DS` scenarios. However, the segmentation results in the `All but DS` scenario are notably better than for HKPU and PMMDB-FR. This is mainly due to the fact that the data set does not show any particular

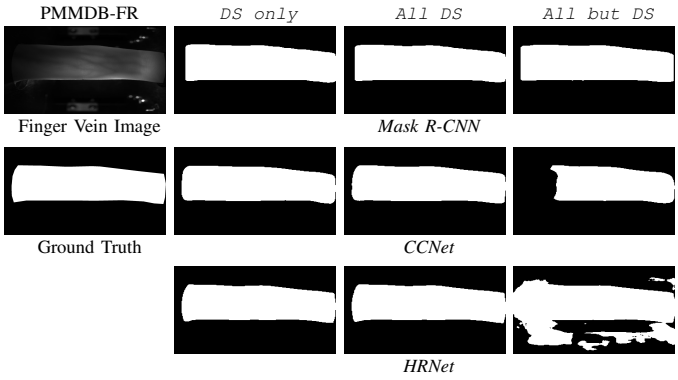

Fig. 6. Segmentation masks generated for PMMDB-FR

difficulties: the finger regions have a relatively even grey value and the background is also uniformly dark. There are no particularly overexposed images, occlusions or other objects (e.g. adjacent fingers) in the images. Since PMMDB-FR shares these properties with PMMDB-FV3 and UTFVP, there are similar images in the training data set and the segmentation of the evaluation data works better. The segmentation with *Mask R-CNN* works particularly well. Error rates similar to the ones of the other scenarios are achieved.

For the example image in Fig. 6 we chose one with a rather bad result for the `All but DS` scenario. For *Mask R-CNN* you can see the very well segmented mask. *CCNet* had problems with the slightly darker root of the finger. In *HRNet*, some background regions were incorrectly classified. These problems are reflected in the error rates of the table II: On average, *CCNet* classifies more finger region pixels as background than background pixels as finger region. With *HRNet* it is exactly the opposite.

D. PMMDB-FV3

The results for PMMDB-FV3 are basically the same as for PMMDB-FR. Only the error rates are 1-2% worse than for PMMDB-FR. Interestingly, the errors for *CCNet* and *HRNet* are exactly the opposite of PMMDB-FR: *CCNet* classifies the finger regions better, *HRNet* the background. Since the results are essentially the same as with PMMDB-FR, we have omitted example images of this database for reasons of space.

E. UTFVP

The overall best results are obtained for UTFVP. Especially *HRNet* with an average of $<0.4\%$ of misclassified pixels and *CCNet* $<0.7\%$, respectively, show an exceptionally good performance. But also the segmentation results for the `All but DS` scenario are surprisingly good. With $<3.9\%$ of misclassified pixels for all three scenarios, the results are better than on PLUSVein-CL in the `DS only` and `All DS` scenarios. These results are reasoned by the high quality of the images within this data set. For *Mask R-CNN*, however, it turned out that the proposed candidates are not always ideal. The proposed bounding boxes are sometimes a bit too small. This leads to slightly higher rates of misclassified foreground pixels and hence higher error rates. As with PMMDB-FV3, we do not present sample masks due to reasons of space.

V. Conclusion and Future Work

In the experiments we evaluated how well different CNN architectures can segment the finger regions of various finger vein data sets. All together, three state-of-the-art CNNs architectures (*Mask R-CNN*, *CCNet* and *HRNet*) were examined on five different databases, (HKPU, PLUSVein-CL, PMMDB-FR, PMMDB-FV3 and UTFVP), in three different scenarios: (1) Training and evaluation on the same data set (`DS only`), (2) training with data from all 5 data sets (`All DS`) and (3) training on 4 data sets and evaluation on the fifth dataset (`All but DS`). The set-up is relatively simple: the CNNs were trained without pre-processing or augmentation of the vein images using manually segmented masks as ground truth. For `DS only` and `All DS` all three CNN architectures delivered excellent results on all data sets. With `All but DS`, good results were also achieved for the three data sets (PMMDB-FR, PMMDB-FV3 and UTFVP). For UTFVP, the results of the `All but DS` scenario were even comparable to the other scenarios. Since HKPU and PLUSVein-CL differed remarkable from the other data sets, no satisfactory segmentation results could be achieved for the `All but DS` scenario.

The results show that a CNN can be trained to segment finger regions from finger vein images acquired with different sensors (`All DS` scenario), even if the data from a sensor is not present in the training data (`All but DS`). Unsurprisingly the latter scenario only works if at least one data set in the training data has similar properties to the excluded data set. If not, the segmentation results are unsatisfactory.

In order to see whether one CNN works in general better than the others, we also evaluated the results of all data sets together. Since our error rates are always calculated with respect to the size of the finger region of the individual image, this is a straight forward task. The results can be seen in table II. For all three scenarios, the results for all three CNN architectures are very similar. With `DS only` and `All DS`, all three CNN architectures provide excellent results, although *Mask R-CNN* performs slightly worse than *CCNet* and *HRNet*. Between the latter two the difference is very small. In the third scenario, `All but DS`, *HRNet* performs best and *CCNet* worst. *HRNet* and *CCNet* seem to work slightly better than *Mask R-CNN*. To really select an architecture as the best, the results are too close together.

In our further work we will try to further improve the results. We will incorporate data augmentation into our training strategy, increase the number of training data, and add more data sets. The focus will be on the `All but DS` scenario.

VI. Acknowledgments

This project has received funding from the FWF project Advanced Methods and Applications for Fingervein Recognition under grant No. P 32201-NBL.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[2] Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang, "Ccnet: Criss-cross attention for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[4] Nianfeng Liu, Haiqing Li, Man Zhang, Jing Liu, Zhenan Sun, and Tieniu Tan, "Accurate iris segmentation in non-cooperative environments using fully convolutional networks," in *2016 International Conference on Biometrics (ICB)*, 2016, pp. 1–8.

[5] Christof Kauba, Dominik Sllinger, Simon Kirchgasser, Axel Weissenfeld, Gustavo Fernndez Domnguez, Bernhard Strobl, and Andreas Uhl, "Towards using police officers' business smartphones for contactless fingerprint acquisition and enabling fingerprint comparison against contact-based datasets," *Sensors (Special Issue Biometric Sensing)*, vol. 21, no. 7, pp. 1–42, 2021.

[6] Lu Yang, Gongping Yang, Lizhen Zhou, and Yilong Yin, "Superpixel based finger vein roi extraction with sensor interoperability," in *2015 International Conference on Biometrics (ICB)*, 2015, pp. 444–451.

[7] Yanan Gao, Jianxin Wang, and Liping Zhang, "Robust roi localization based on image segmentation and outlier detection in finger vein recognition," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 20039–20059, 2020.

[8] Qiong Yao, Dan Song, and Xiang Xu, "Robust finger-vein roi localization based on the 3 criterion dynamic threshold strategy," *Sensors*, vol. 20, no. 14, 2020.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[10] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2228–2244, April 2012.

[11] Christof Kauba, Bernhard Prommegger, and Andreas Uhl, "Combined fully contactless finger and hand vein capturing device with a corresponding dataset," *Sensors (Special Issue Biometric Systems)*, vol. 19(22), no. 5014, pp. 1–25, 2019.

[12] Chiara Galdi, Jonathan Boyle, Lulu Chen, Valeria Chiesa, Luca Debiasi, Jean-Luc Dugelay, James Ferryman, Artur Grudzie, Christof Kauba, Simon Kirchgasser, Marcin Kowalski, Michael Linortner, Patryk Maik, Kacper Micho, Luis Patino, Bernhard Prommegger, Ana F. Sequeira, ukasz Szklarski, and Andreas Uhl, "Protect: Pervasive and user focused biometrics border project a case study," *IET Biometrics*, vol. 9, no. 6, pp. 297–308, 2020.

[13] B. T. Ton and R. N. J. Veldhuis, "A high quality finger vascular pattern dataset collected using a custom designed capturing device," in *2013 International Conference on Biometrics (ICB)*, 2013, pp. 1–5.

[14] Christof Kauba, Bernhard Prommegger, and Andreas Uhl, *OpenVein—An Open-Source Modular Multipurpose Finger Vein Scanner Design*, pp. 77–111, Springer International Publishing, Cham, 2020.

[15] Bernhard Prommegger, Christof Kauba, and Andreas Uhl, "Multi-perspective finger-vein biometrics," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Los Angeles, California, USA, 2018, pp. 1–9.