Two-stage CNN-based wood log recognition *

Georg Wimmer[1][0000−0001−5529−0154], Rudolf Schraml[1][0000−0002−4441−0461], Heinz Hofbauer[1], Alexander Petutschnigg[2], and Andreas Uhl[1][0000−0002−5921−8755]

[1] University of Salzburg, Jakob Haringer Str. 2, 5020 Salzburg, Austria
{gwimmer,uhl}@cs.sbg.ac.at, rudolf.schraml@sbg.ac.at
[2] University of Applied Sciences Salzburg, Markt 136a, 5431 Kuchl, Austria

Abstract. The proof of origin of logs is becoming increasingly important. In the context of Industry 4.0 and to combat illegal logging there is an increasing motivation to track each individual log. This work presents a two-stage convolutional neural network (CNN) based approach for wood log tracing based on digital log end images. First, the log cross section is segmented from the background by applying a CNN-based segmentation method using the Mask R-CNN framework. In the second step, wood log recognition is applied using CNNs that are trained on the segmented wood log images using the triplet loss function. Our proposed two-stage CNN-based approach achieves Equal Error Rates between 0.6 and 3.4% on the six employed wood log image data sets and clearly outperforms previous approaches for image based wood log recognition.

Keywords: wood log tracking · deep learning · segmentation.

1 Introduction

Methods for wood log tracking are an essential component in solving a wide variety of problems and requirements of an ecological, legal and social nature. Currently, this mainly relates to proof the origin of wood products, e.g. by certification companies like the Forest Stewardship Council (FSC). However, efforts towards traceability down to the individual log have been intensified by a variety of stakeholders. The motivation for this is that, on the one hand, illegal logging can be better combated and, on the other hand, the identification of each individual log forms a basis for steps towards forest-based industry 4.0. In the context of Industry 4.0, Radio Frequency Identification (RFID) is the state-of-the-art for object recognition/tracking. However, like a set of other tracking technologies for wood logs (e.g. punching, coloring or barcoding log ends [14]), RFID requires physical marking of each tree which suffers costs. An alternative to physical marking is to use biometric characteristics to recognize each individual log. A short summary on biometric log tracking using various characteristics is presented in [7]. In a series of works between 2014–2016 we investigated wood

* This work is partially funded by the Austrian Science Fund (FWF) under Project No. I 3653

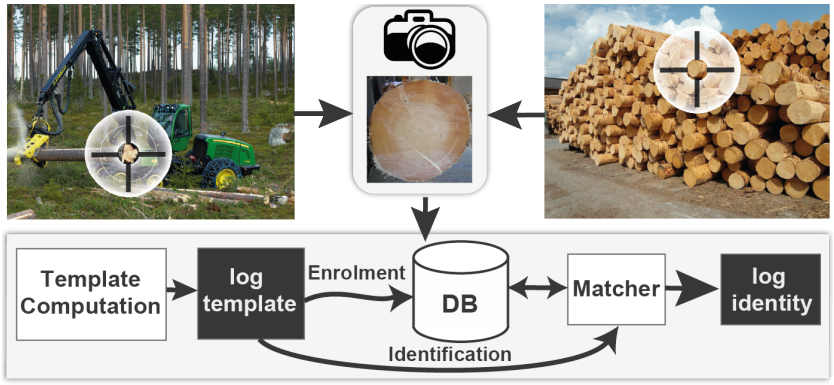


Fig. 1. Exemplary scheme of enrolment and identification for wood log tracking

log tracking based on digital log end images in regard to the distinctiveness and robustness of the annual ring pattern. For a literature review we refer to [6]. Figure 1 presents the scheme of such a log tracking system. Significant for this work is that the utilized approaches were inspired by human fingerprint and iris-recognition methods. Those rely on traditional feature extraction methods (e.g. Gabor filterbanks) and moreover require a sophisticated pre-processing (segmentation, pith estimation, rotational pre-alignment) of each log end image prior to feature extraction. Comparison of the extracted features is also complex. Furthermore, it has to be noted that for our previous works manually segmented log end images and determined pith positions were utilized. Time has passed and deep learning based approaches have become state-of-the-art. Not surprisingly, deep learning-based methods have also been investigated in many application areas in the forestry and timber industry in recent years. Exemplary applications are wood species identification using cross-section (CS) images [13, 5], remote sensing-based tree species classification [1] or lumber grading [3] using wood board surface images.

In this work we apply convolutional neural networks (CNNs) for two-stage wood log recognition. CNNs are used for segmentation of the CS in the log end image as well as for feature extraction that offers advantages in many ways. The experimental evaluation is based on a database (DB) which was utilized in [10] and a new DB denoted as 100 logs DB (HLDB). This work significantly contributes to biometric log end recognition by showing that a CNN-based approach does not require to determine the pith position and moreover no rotational pre-alignment is required. Results show that CNN-based segmentation and feature extraction shows a similar performance as the results presented in [10] which are based on groundtruth data and traditional feature extraction methods. The experimental evaluation on the new HLDB, for which no groundtruth data is available, underpins this statement and shows the weaknesses of the traditional methods that are mainly caused by inaccurate segmentation and pith estimation results.

(a) $CSLD_{DB1}$ (b) $CSLD_{DB2}$ (c) $CSLD_{DB3}$

Fig. 2. Example image of each of the three sub-databases of CSLD

Section 2 introduces the DBs and describes the CNN-based segmentation of the CS as well as the CNN-based log recognition and the experimental setup. Section 3 presents the results and Section 4 concludes this work.

2 Material and Methods

2.1 Data Bases

Two different DBs are utilized: (i) the Cross Section Log Database (CSLD) which was already utilized in [9, 10] and (ii) a new database referred to as 100 Logs Database (HLDB).

The CSLD is a combination of images from three sub-databases with 50, 120 and 109 different logs, respectively. The images are recorded with a Canon EOS DMark with a 35mm lens. In total, this database comprises 2270 CS images from 279 different logs. For this database, a manual segmentation of the log cross section is available. For a detailed description of CSLD we refer to [10]. The CSLD is utilized (i) to compare the CNN-based results to previous results which were based on annotated groundtruth data and (ii) to train a segmentation CNN in order to segment the images of HLDB, for which no manual segmentation is available. Figure 2, shows one example image of each of the three sub-databases of the CSLD dataset.

HLDB comprises different datasets which were all taken from the same 100 logs. CS-Images were acquired from both ends of each log. The first two datasets HLDB_{FH} and HLDB_{FL} were taken in the forest (see Fig. 3(a)) using a Lumix camera (Panasonic FZ45 Lumix camera) and a Huawei smartphone (Huawei P8 Lite 2017), respectively. Both datasets consist of 4 images for each log end. After two images the camera was rotated by approximately 45 degrees and two more images were captured. The next dataset, denoted as Sawmill dataset (HLDB_{SM}), was captured after cutting off a thin disc from each log end (see Fig. 3(b)). Three images with different rotations for each fresh cross-cut log end were acquired using the Huawei smartphone.

The CS-Images of $\text{HLDB}_{FH,FL,SM}$ were taken without tripod which causes different rotations, slightly varying perspectives toward the CS and slightly different positions of the CS in the image. Finally, one side of the 200 discs was acquired using a , once raw (HLDB_R) and once after they were sanded (HLDB_S).

(a) Forest site - HLDB$_{FH,FL}$

(b) Sawmill yard - HLDB$_{SM}$ - discs were cut for HLDB$_{R,S}$

Fig. 3. 100 Logs Image Database (HLDB): Fig. 3(a) shows log piles close to the forest at which the forest datatsets were acquired. Fig. 3(b) shows the data acquisition at the sawmill yard where discs of each log end were cut off.

The captured CSs are mirrored versions of the CSs in the Sawmill dataset. For HLDB$_R$ four and for HLDB$_S$ six CS-Images with different rotations were captured, respectively. Fig. 4 shows exemplary images for all HLDB datasets from the bottom end of log labelled #E001. It can be observed that the CS-Images of the two forest datasets HLDB$_{FH,FL}$ look quite similar since the images were taken with the same surrounding, the same log cut pattern and hardly any time shift between the taking of the images of the two datasets. The CS-Images captured at the sawmill yard HLDB$_{SM}$ look completely different because of the fresh cut that results in a totally different saw cut pattern and wood coloration. The disc CS-Images HLDB$_{R,S}$ are captured under idealistic conditions and serve as a reference in the experiments, especially the sanded CS-Images in HLDB$_S$ which show a undisturbed annual ring pattern.

2.2 CS-Segmentation

Prior to any feature extraction, the CS area in the CS-Image needs to be localized and segmented from the background. We apply the Mask R-CNN framework [2] to get a segmentation mask. As net architecture we employ the ResNet-50

(a) HLDB$_{FH}$ (b) HLDB$_{FL}$ (c) HLDB$_{SM}$ (d) HLDB$_{R}$ (e) HLDB$_{S}$

Fig. 4. HLDB: Exemplary images for both log ends of log #E001 from all datasets HLDB$_{FH,FL,SM,R,S}$

architecture using a model pretrained on the COCO dataset. The segmentation net is then trained on CSLD for which groundtruth segmentation masks are available. The segmentation net is trained for 30 epochs in order to differentiate the CS from the background. Then, the trained segmentation net is applied to the HLDB datasets to segment the CS from the background. To also get CNN-based segmentation masks for CSLD, we apply a 4-fold cross validation, where one fold consists of a fourth of the 279 logs (the images of one log are all in the same fold). 3 folds are used to train the segmentation net and the trained net is applied to segment the images of the remaining fold.

The obtained segmentation mask of a CS-Image, which consists of probability values between 0 and 1 for each pixel of the image, is binarized. All values of the CNN segmentation mask that are below the threshold value t ($t = 0.5$ for HLDB$_{SM,R,S}$ and CSLD and $t = 0.25$ for HLDB$_{FH,FL}$) are set to zero and the remaining values are set to one. The binarized segmentation mask of the CS is further used to set the background (all image positions with a zero in the segmentation mask) to black. Finally, each CS-Image is reduced to the smallest possible square shaped image section so that the CS (all image positions with a '1' in the segmentation mask) is still completely included in the image together with a five pixel thick black border on each side of the image. A schematic representation of the segmentation including the extraction of the square shaped image patch containing the CS is displayed in Fig. 5. For CSLD we can quantitatively assess the outcome of the segmentation. Averaged over all CS-Images in CSLD, 99.26% of the pixels per image were correctly segmented. In Figure 6, we present exemplar outcomes of the segmentation and patch extraction process for HLDB$_{FL,FH,SM}$. The segmentation outcomes on HLDB$_{SM,R,S}$ all look perfectly fine based on the authors' visual impression. For the two forest datasets HLDB$_{FH,FL}$, most images were well segmented, but on some images, parts of the log CS were predicted as background which was the reason why we used a smaller threshold ($t = 0.25$ instead of 0.5) to binarize the segmentation masks. This reduced the risk to predict parts of the CS as background (as can be seen in Figure 6(d)) but also led to the problem that for some images a bit of background surrounding the log CS was predicted as being part of the log CS (see Fig. 6(f)).

The advantage of our proposed segmentation and square image patch extraction approach for log recognition is that the background of a log CS image does not influence the log recognition. For CNN based recognition systems, where the

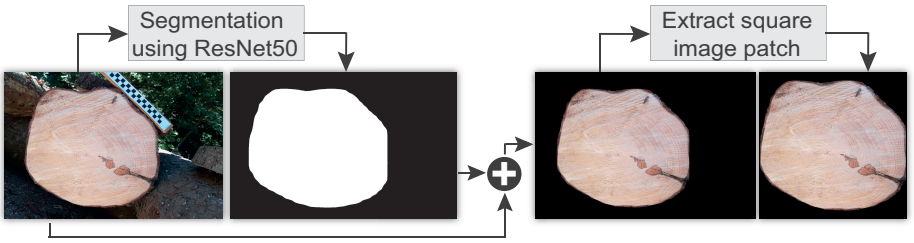
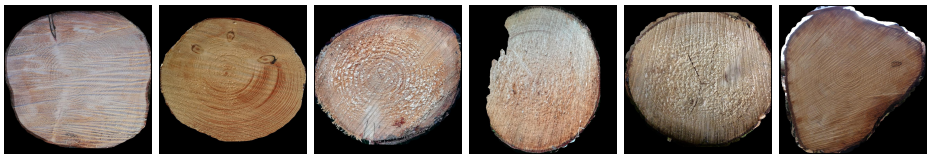


Fig. 5. Segmentation & patch extraction of the CS-Image



(a) HLDB_{SM} (b) HLDB_{SM} (c) HLDB_{FH} (d) HLDB_{FH} (e) HLDB_{FL} (f) HLDB_{FL}

Fig. 6. Exemplary outcomes of the segmentation in combination with square image patch extraction

images usually have to be resized to a fixed size before feeding them through the network, an additional advantage is the reduced loss of image quality. The segmented square shaped image patches are clearly smaller than the original CS-Image and so less information on the log is lost by reducing the image resolution to fit the required CNN input size.

2.3 Wood log recognition using CNN triplet loss

In biometric applications, the problem with common CNN loss functions (e.g. the SoftMax loss) is that CNNs are only able to identify those subjects which have been used for the training of the neural network. If new subjects are added in a biometric application system, then the nets need to be trained again or else a new subject can only be classified as one of the subjects that were used for training (the one that is most similar to the newly added subject with respect to the CNN). This of course renders the practical application of common CNN loss functions impossible for biometric applications.

Contrary to more common loss functions like the Soft-Max loss, the triplet loss [12] does not directly learn the CNN to classify images to their respective classes. The triplet loss requires three input images at once (a so called triplet), where two images belong to the same class (the so called Anchor image and a sample image from the same class, further denoted as Positive) and the third image belongs to a different class (further denoted as Negative). The triplet loss learns the network to minimize the distance between the Anchor and the Positive and maximize the distance between the Anchor and the Negative. The triplet

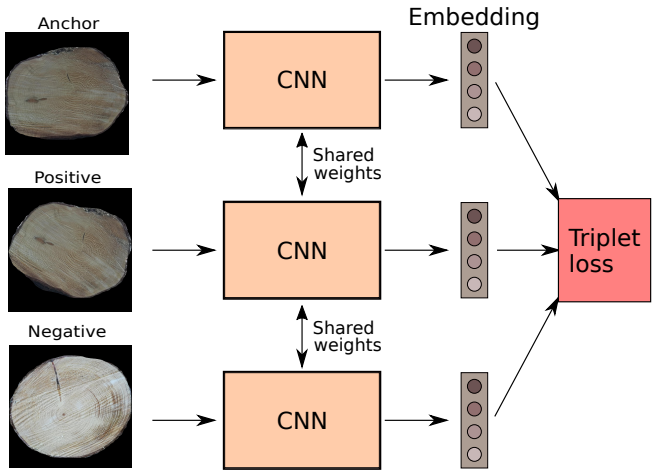


Fig. 7. CNN training using the triplet loss

loss using the squared Euclidean distance is defined as follows:

$$L(A, P, N) = \max(||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \alpha, 0), \qquad (1)$$

where A is the Anchor, P the Positive and N the Negative. α is a margin that is enforced between positive and negative pairs and is set to $\alpha = 1$. $f(x)$ is an embedding (the CNN output) of an input image x. Figure 7 shows the scheme of learning a CNN using the triplet loss. A triplet of training images (Anchor, Positive and Negative) is fed through the CNN resulting in an embedding for each of the three images. The embeddings of the three images are then used to compute the triplet loss to update the CNN.

For our application this means that the CNN is trained so that the Euclidean distances between the CNN feature vectors of all log CS-Images of the same class (log) is small, whereas the Euclidean distance between any pairs of CS-Images from different logs is large. We employ hard triplet selection [12] (only those triplets are chosen for training that actively contribute to improving the model) and the Squeeze-Net (SqNet) architecture [4]. SqNet is a small neural networks that is specifically created to have few parameters and only small memory requirements.

The size of the CNN's last layer convolutional filter is adapted so that a 256-dimensional output vector (embedding) is produced. Training is performed on batches of 128 images. To make the CNN more invariant to shifts and rotations and increase the amount of training data, we employ data augmentation for CNN training. The images are randomly rotated in the range of 0-360 ° and random shifts in horizontal and vertical directions are applied by first resizing the input images to a size of 234×234 and then extracting a patch of size 224×224 (the best working input size using the SqNet for log recognition) at a random position of the resized image (± 10 pixels in each direction). The CNN is trained for 400

epochs, starting with a learning rate of 0.001 that is divided by 10 every 120 epochs.

2.4 Experimental Setup

In this work, a 4-fold cross validation is employed. For each dataset, the CNN is trained four times, each time using three of the folds for training and evaluation is applied on the remaining fold. Each fold consists of a fourth of the logs of a dataset, where all images of a log are in the same fold. We further denote these experiments as "SqNet". In a second experiment, we additionally use training data from logs of other datasets. For the HLDB datasets, we additionally use the CSLD dataset for training (the CSLD dataset is added to the three training folds). In case of the CSLD dataset, we additionally employ all the images of HLDB_{FH} for training. We further denote these experiments with additional training data from another dataset as "SqNet+". For performance evaluation we have decided to present verification results, i.e. we compute the Equal Error Rates (EERs) for the different datasets achieved with SqNet and SqNet+.

The EER is well suited to compare the CNN-based results to results achieved with traditional approaches and with results achieved in prior works (e.g. in [10]). We have to consider that each of the four trained CNNs per dataset (one per fold) has a different mapping of the images to the CNN output feature space. Thus, feature vectors of different folds cannot be compared in the evaluation and the EER has to be computed separately for each fold. We report the mean EER over the four folds.

As already mentioned before, the HLDB datasets consist of images from both log ends which show no obvious visible similarities. To employ the maximum number of images for CNN training, both ends are considered as different classes thus resulting in 200 classes in total. To avoid any bias by assigning different classes to the two sides of a log, we exclude those triplets during training where the Anchor and the Negative are from the same log but different sides. The same is applied for EER computation, where those comparison scores (scores between images from different sides of the same log) are ignored.

Comparison Methods: In order to assess the performance of CNN-based wood log recognition we compute EERs using the fingerprint- and iris-based approaches proposed in [10]. For rotational pre-alignment, the CM (center of mass to pith estimate vector) strategy is applied.

The iris-based results IRIS_V and IRIS_H are computed in the exact same way as described in [10]. Features are computed with the Log Gabor configuration LG (64/08) and for matching the shifting is done in the range of -21 to 21 feature vector positions.

For the fingerprint-based approach we utilize a modified approach, based on a circular grid, as introduced in [8], which does not require to compute feature vectors for rotated versions of the registrated CS-Image. Identical as the template comparison procedure for the iris-based approach, rotation compensation in the matching stage is performed by shifting the feature vectors of each band. This

circular grid fingerprint-based approach is referred to as FP_{CG}. Contrasting to our previous work, we do not utilize manually extracted groundtruth data and instead use the CNN-based CS-Segmentation results and the pith position determined using the approach described in [11].

3 Results and Discussion

In Table 1, we present the EERs of the CNN-based approaches and the comparison methods for each dataset. The best result for each dataset is marked in bold letters. The main finding in Table 1 is that the CNN-based approaches are clearly superior to the traditional ones. SqNet+ performs slightly better than SqNet. So additional training data improves the results, although the additional data is from another database with different image acquisition conditions and a different camera.

Interestingly, the CNN results are quite similar across the different HLDB datasets, despite the differing image acquisition conditions and cameras. For example, the images of the HLDB_S dataset offer a perfectly visible annual ring pattern (sanded surface) and the scale and viewpoint of the images is constant, whereas the two datasets acquired at the forest ($\text{HLDB}_{FH,FL}$) offer a poor visibility of the annual ring pattern due to the saw cut pattern and the images were taken under varying scales and viewpoints. Despite that, the CNN results of the forest datasets are slightly better than those of the HLDB_S. So this together with the fact that the CNN results on HLDB_R (raw CS) are slightly better than those on HLDB_S (sanded CS) indicates that the saw cut pattern is an important feature for CNNs that could be even more important for wood log recognition than the annual ring pattern. Varying viewpoints and scales of the recorded images does not seem to be a problem for the CNNs.

Considering the CSLD results, an EER of 0.9% achieved by IRIS_V with manually segmented images was the best result presented in [10]. Using the automated CS-segmentation instead of the manually segmented image data, the results of the comparison approaches deteriorate greatly ($3.4 - 5.8\%$ EER). The results for the CNN-based approaches using segmentation groundtruth data (SqNet $= 0.7\%$, SqNet+ $= 0.6\%$) outperform our previous results and the EERs achieved with the automated CS-Segmentation (SqNet/+ $= 1\%$) are only slightly worse than those with manually segmented image data.

Thus, another main advantage of our proposed approach is that the proposed CS-Segmentation is well suited to be used with the CNN-based recognition. This statement is confirmed by the EERs presented for the HLDB datasets, which are all below 3.4% for our proposed approach. The EERs presented for the comparison approaches are not even close to this performance. By comparing the EERs for the HLDB computed with the traditional approaches, it is obvious that the EERs computed for $\text{HLDB}_{S,R}$ are better than those computed for $\text{HLDB}_{FL,FH,SM}$. The reason is that for $\text{HLDB}_{S,R}$ rotational pre-alignment is more accurate than for the other datasets because of the more accurate CNN segmentation for $\text{HLDB}_{S,R}$. However, these observations highlight the main advan-

	Methods	CSLD	$HLDB_{FH}$	$HLDB_{FL}$	$HLDB_{SM}$	$HLDB_{S}$	$HLDB_{R}$
CNN	SqNet	0.7 / 1.0	3.2	2.4	3.1	**3.4**	2.8
	SqNet+	**0.6 /1.0**	**2.8**	**1.7**	**2.6**	**3.4**	**2.6**
Comp.	$Iris_{H}$	2.12 [10] /5.8	26.8	27.3	22.4	10.6	12.5
	$Iris_{V}$	0.9 [10] / 3.4	20.5	21.0	15.1	5.3	6.1
	FP_{CG}	2.1/3.9	16.9	19.5	20.3	8.0	8.3

Table 1. Recognition performance (Mean EER in [%]) on the 6 log CS datasets for the proposed CNN-based method using the two different training strategies, as well as three comparison approaches (Comp.) using classical hand-crafted biometric features. On the CSLD dataset, recognition is applied on the manually segmented images (value before the slash) and the CNN segmented images (value after slash).

tage of the CNN-based approaches: They work in combination with a fully automated CNN-based segmentation and do not require any rotational pre-alignment prior to feature extraction.

4 Conclusion

Recently, there has been an increasing interest in methods for tracking roundwood on the basis of each individual log. In prior works we proposed a physical free approach using log end images and methods inspired by fingerprint and iris recognition-based approaches. Results were promising and showed good performances when using groundtruth data for segmentation of the log end in each image. However, in a real world application a fully automated system is required. In order to close this gap, we employ a CNN-based segmentation approach combined with a CNN-based log recognition approach which is compared to results achieved with the traditional log recognition approaches when using automatically segmented images. Results showed, that the CNN-based wood log recognition works well in combination with the CNN-based segmentation. On the contrary, the traditional approaches suffer from the inaccuracies of the CNN-based segmentation which affects the required rotational pre-alignment strategy. It can be concluded that the proposed two-stage CNN-based wood log recognition approach is well suited for individual wood log tracking. What remains is to prove that this two-stage approach also works in a realistic scenario, i.e. if logs can be tracked when using imagery captured at various stages of the log tracking chain.

References

1. Fricker, G.A., Ventura, J.D., Wolf, J.A., North, M.P., Davis, F.W., Franklin, J.: A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. Remote Sensing **11**(19), 2326 (oct 2019). https://doi.org/10.3390/rs11192326

2. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2017). https://doi.org/10.1109/iccv.2017.322
3. Hu, J., Song, W., Zhang, W., Zhao, Y., Yilmaz, A.: Deep learning for use in lumber classification tasks. Wood Science and Technology **53**(2), 505–517 (feb 2019). https://doi.org/10.1007/s00226-019-01086-z
4. Iandola, F.N., Moskewicz, M.W., Ashraf, K., Han, S., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. CoRR **abs/1602.07360** (2016), `http://arxiv.org/abs/1602.07360`
5. Olschofsky, K., Köhl, M.: Rapid field identification of cites timber species by deep learning. Trees, Forests and People **2**, 100016 (dec 2020). https://doi.org/10.1016/j.tfp.2020.100016
6. Schraml, R., Charwat-Pessler, J., Entacher, K., Petutschnigg, A., Uhl, A.: Round-wood tracking using log end biometrics. In: Proceedings of the Annual GIL Meeting (GIL'2016). pp. 189–192. LNI, Gesellschaft für Informatik (2016)
7. Schraml, R., Charwat-Pessler, J., Petutschnigg, A., Uhl, A.: Towards the applicability of biometric wood log traceability using digital log end images. Computers and Electronics in Agriculture **119**, 112–122 (2015). https://doi.org/10.1016/j.compag.2015.10.003
8. Schraml, R., Entacher, K., Petutschnigg, A., Young, T., Uhl, A.: Matching score models for hyperspectral range analysis to improve wood log traceability by fingerprint methods. Mathematics **8**(7), 10 (2020). https://doi.org/10.3390/math8071071, `https://www.mdpi.com/2227-7390/8/7/1071`
9. Schraml, R., Hofbauer, H., Petutschnigg, A., Uhl, A.: Tree log identification based on digital cross-section images of log ends using fingerprint and iris recognition methods. In: Proceedings of the 16th International Conference on Computer Analysis of Images and Patterns (CAIP'15). pp. 752–765. LNCS, Springer Verlag, Valetta, MLT (2015). https://doi.org/10.1109/ICIP.2015.7351488
10. Schraml, R., Hofbauer, H., Petutschnigg, A., Uhl, A.: On rotational pre-alignment for tree log end identification using methods inspired by fingerprint and iris recognition. Machine Vision and Applications **27**(8), 1289–1298 (2016). https://doi.org/10.1007/s00138-016-0814-2
11. Schraml, R., Uhl, A.: Pith estimation on rough log end images using local fourier spectrum analysis. In: Proceedings of the 14th Conference on Computer Graphics and Imaging (CGIM'13). Innsbruck, AUT (Feb 2013). https://doi.org/10.2316/P.2013.797-012
12. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (June 2015). https://doi.org/10.1109/CVPR.2015.7298682
13. Tang, X.J., Tay, Y.H., Siam, N.A., Lim, S.C.: MyWood-ID. In: Proceedings of the 2018 International Conference on Computational Intelligence and Intelligent Systems - CIIS 2018. ACM Press (2018). https://doi.org/10.1145/3293475.3293493
14. Tzoulis, I., Andreopoulou, Z.: Emerging traceability technologies as a tool for quality wood trade. Procedia Technology **8**(0), 606–611 (2013)