Timber Tracing with Multimodal Encoder-Decoder Networks

Zolotarev Fedor, Eerola Tuomas, Lensu Lasse, Kälviäinen Heikki, Haario Heikki, Heikkinen Jere, Kauppi Tomi

DOI: 10.1007/978-3-030-29891-3_30

Timber Tracing with Multimodal Encoder-Decoder Networks

Fedor Zolotarev[1], Tuomas Eerola[1], Lasse Lensu[1], Heikki Kälviäinen[1], Heikki Haario[1], Jere Heikkinen[2], and Tomi Kauppi[2]

[1] Lappeenranta-Lahti University of Technology LUT, School of Engineering Science, Department of Computational and Process Engineering, Machine Vision and Pattern Recognition Laboratory, P.O. Box 20, FI-53851 Lappeenranta, Finland
firstname.lastname@lut.fi

[2] Finnos Oy, Tukkikatu 5, FI-53900 Lappeenranta, Finland
firstname.lastname@finnos.fi

Abstract. Tracking timber in the sawmill environment from the raw material (logs) to the end product (boards) provides various benefits including efficient process control, the optimization of sawing, and the prediction of end-product quality. In practice, the tracking of timber through the sawmilling process requires a methodology for tracing the source of materials after each production step. The tracing is especially difficult through the actual sawing step where a method is needed for identifying from which log each board comes from. In this paper, we propose an automatic method for board identification (board-to-log matching) using the existing sensors in sawmills and multimodal encoder-decoder networks. The method utilizes point clouds from laser scans of log surfaces and grayscale images of boards. First, log surface heightmaps are generated from the point clouds. Then both the heightmaps and board images are converted into "barcode" images using convolutional encoder-decoder networks. Finally, the "barcode" images are utilized to find matching logs for the boards. In the experimental part of the work, different encoder-decoder architectures were evaluated and the effectiveness of the proposed method was demonstrated using challenging data collected from a real sawmill.

Keywords: convolutional neural networks · encoder-decoder networks · multimodal translation · machine vision · sawmilling.

1 Introduction

Sawmilling is a complex process with a large number of different stages and variables. A typical sawmill contains various measuring devices, such as laser scanners and RGB cameras. These devices measure both the raw material and end products during the different process stages. These measurements are not connected with each other which prevents the full utilization of the measured information. The measurements are typically used to sort the logs or the boards to various categories, but the information about an individual log or board is

lost while the material moves to the next process stage. However, tracing the material through the whole process is of great interest as it would allow various improvements for the sawmill process, such as better process control, the optimization of sawing, the prediction of the end-product quality in early stages of the process, and more accurate grading of the final product using sensor fusion. Other benefits include ensuring the legality and sustainability of the product origin [3] and avoiding faulty products at an earlier production step [5].

Tracking material is challenging in a typical sawmill environment where the raw material does not go through a straightforward process pipeline. In the multiple process stages, the material is stored in warehouses or storage areas without record keeping of individual items. This happens, for example, after the log measurements and during drying. Tracking of the material becomes even more difficult due to the various transformations the raw material goes through during debarking of the logs and sawing. Invasive tracking techniques such as labels or tags have been proposed [3]. However, they are too expensive to be used in the large scale, are easily damaged or corrupted during processing, and do not typically provide the means to track every board obtained from a single labeled log. This calls for non-invasive tracking techniques that utilize the measurable properties of the material itself to implement the matching.

In this paper, we propose a method to match grayscale images of boards to the laser scans of the surfaces (point clouds) of the corresponding logs. The method starts with the construction of heightmaps from the point clouds and utilizes a multimodal encoder-decoder network to translate the images and the heightmaps into matchable representations. The encoder-decoder network is trained using "barcode" images representing the starting and end points of knot clusters obtained from discrete X-ray tomography. The resulting network is able to generate similar "barcodes" from both the board images and the log surface heightmaps. Using these "barcodes", simple cross-correlation can be applied to find the best match for each board, i.e., the log where the board was sawn from. The proposed approach is visualized in Fig. 1.

It should be noted that the discrete X-ray tomography data is required only for the training phase. Once the networks have been trained, only the laser scanners and the cameras are required for the production deployment. We demonstrate the effectiveness of the proposed method with a challenging data set collected from a real sawmill.

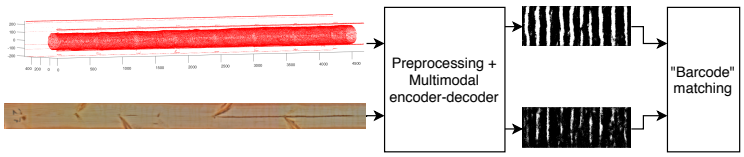


Fig. 1: Board-to-log matching. Given the point clouds of log surfaces and an image of a board, the task is to find the matching log for the board.

2 Related work

2.1 Timber tracing

Various solutions for tracing timber in sawmills have been proposed. Most of the solutions focus either on the re-identification of logs before sawing (log-to-log matching) or on matching boards before and after drying (board-to-board matching). The methods include cross-section images of the log ends [11], laser scans and RFID-tags for logs [2], and image-based matching for boards [8,9]. While the existing solutions provide a high tracing accuracy for the beginning and the end of the sawmill process, extending the tracing to cover the actual sawing (board-to-log matching) remains a challenge due to the following obvious reasons: 1) the material goes through a remarkable transformation and 2) the imaging modalities are different.

In [4], a method was proposed where knot clusters visible in X-ray scans of logs were matched with knots detected from boards. The presented preliminary results were promising. This approach, however, relies heavily on the performance of the knot detectors. Moreover, it requires an expensive X-ray scanner in front of the saw. To find a remedy for these restrictions, we utilize an end-to-end multimodal translation approach to translate the log measurements and board images into similar representations that can be matched to each other.

2.2 Multimodal translation

Fully convolutional networks The concept of fully convolutional networks was introduced in [7]. Its core idea is to remove the fully connected layers at the end of the network and instead make the input and output of the network to be of the same size. The fully convolutional networks are widely used for a variety of tasks, including multimodal translation, although a large proportion of them were initially created for the purpose of semantic segmentation. U-Net [10] is an example of such a network. It consists of the sequentially connected contracting and expansive paths, meaning that an image is first downsampled, and then upsampled back to the original size. The first part can be considered as an encoder and the second part as a decoder. It also utilizes skip connections that transfer features directly from the first half of the network to the second. This approach allows to capture smaller details and avoids the loss of resolution in the resulting image. The somewhat similar approach is used in SegNet [1] where the idea is to perform upsampling according to the indices that were selected during the max pooling operations. Compared to U-Net, only the indices are transferred using skip connections, as opposed to a large number of intermediate features.

Multimodal encoder-decoder networks In order to match data with different modalities, a method is needed to convert data from each modality to similar representations. A common approach for this is to use the encoder-decoder network. For example, in [6], U-Net is used as the encoder-decoder to translate data to different modalities. The latent representations, i.e., the outputs of the

encoders, are concatenated for all the modalities and then decoded. Instead of concatenating different latent representations, the authors of [13] enforce the alignment of the latent space, meaning that any object should have one latent representation, independent of the modality. This common latent representation is used to align the encoder-decoder pairs, allowing for the direct translation between the modalities that do not have explicit pairs of examples.

3 Proposed method

The proposed method to determine the corresponding log for each board consists of the following three steps:

1. Generate a heightmap from the laser point cloud.
2. Use encoders and decoders to transform the heightmap and the board image into the knot "barcodes".
3. Transform the "barcodes" into one-dimensional signals and perform cross-correlation to find the matching log for the board.

3.1 Generating heightmaps

Information about the log surface is obtained in the form of the 3D point clouds. In the beginning it is necessary to remove data outliers generated due to the noise and obstructions during the data retrieval process. The point cloud is processed in layers that correspond to the circular cross sections of the log. First, the log center is calculated for the layer by using the median of all points. The points further than $2.5 \cdot \mathrm{MAD}$ are removed, where MAD is the median absolute deviation of distances from the center. After that, the circle is fitted to the points using the least squares method. If the residual error of the fit is larger than a given threshold t ($t = 200$ has been used in this work), the layer is considered too noisy and removed. For the fitted circle with radius r, points with distance greater than $1.1r$ or less than $0.9r$ from the center are removed next. An example of the resulting point cloud can be seen in Fig. 2.

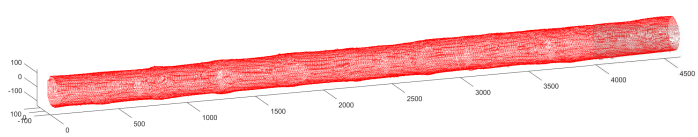
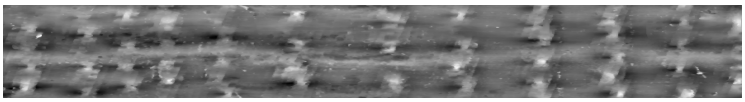


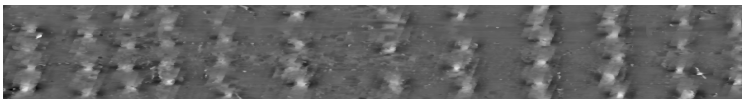
Fig. 2: Log point cloud after the removal of outliers.

A straightforward way to generate the heightmap is to fit a circle to each layer and to map the distance between each point and the circle. However, it was noticed that the resulting heightmaps contained large elevated areas making this approach unsuitable for our task. Those elevations appeared as a result of the

log shape not being perfectly round but instead being the shape of a deformed circle that smoothly changes along the length of the trunk. The solution was to average each layer to generate the approximate smooth surface of the log. This was done by using moving average on the distances to the center of the fitted circle. First, averaging with a window of size 15 is done on each layer, and then averaging is performed along the length of the log with a window of size 35. The final heightmap is calculated by using the distance from the measured point to the smoothed surface. This way the large elevations are eliminated and only small bumps that correspond to the locations of branches remain. Horizontal dimension corresponds to the length of a log, vertical corresponds to an angle to the log center. A visual comparison of the two approaches can be seen in Fig. 3.



(a) Heightmap of distances to the fitted circles.



(b) Heightmap of distances to the average surface.

Fig. 3: Different approaches for generating a heightmap. (a) uses distances to the fitted circles, while (b) uses distance to the averaged surface. Large areas of elevation in (a) arise from the variance of the general shape of the log. (b) highlights only small bumps that roughly correspond to the locations of branches.

3.2 Multimodal encoder-decoder networks

Multimodal encoder-decoder networks provide a tool to translate from one image modality to another. The following three modalities are considered in this work: log heightmaps, board images, and "barcode" images. The log heightmaps and the board images are 2D representations of the logs and the boards, respectively, and the "barcodes" correspond to the approximate locations of the knot clusters in the log. The information in the "barcodes" is essentially a one-dimensional list of the start and end positions of the knot clusters in the longitudinal direction of the log presented as a 2D binary image. This information is extracted from the logs using an X-ray scanner before they are sawn. The log heightmap and the board image are the source modalities (input) and the "barcode" image is the target modality (output of the translation process). Therefore, the X-ray generated "barcodes" form the ground truth to train the encoder-decoder networks. However, it should be noted that the "barcodes", generated from discrete X-ray

tomography, are only used during the training phase and, therefore, an X-ray scanner is not required in order to deploy this method into production.

The schematic of the translation process of heightmaps and images to barcodes is presented in Fig. 4. By using only a single decoder on the latent representations of any other modality, we aim to align them. Skip connections from the encoders are also used along with the latent representation. The following three architectures were chosen for the encoder-decoder networks: U-Net [10], Segnet [1], and All Convolutional U-Net. The main difference between U-net and Segnet is the use of the skip connections. Before any downsampling operations, U-net transfers all features directly to the decoder, allowing for the preservation of the small details. Segnet, on the other hand, transfers only max pooling indices that are used to upsample the image. The advantage of this approach is the invariance to the modality-specific features. All Convolutional U-Net is a modification of the classic U-Net that was inspired by [12]. Instead of using the pooling and upsampling operations, the network learns the down- and upsampling operations by utilizing strided convolutions and strided transposed convolutions, respectively. It is done by completely removing max pooling and adding strides to the preceding convolution layers. The same operation is performed for the upsampling layers, except that the convolution layers are changed to transposed convolutions in order to perform the upsampling. Horizontal alignment of knots along the length of the log is used as a ground truth, which is essentially one-dimensional. Vertical locations of individual knots can differ even for matching logs and boards, owing to the fact that they depend on the measuring angle (in case of a heightmap) and the sawing angle (in case of a board image). Consequently, the loss function has been constructed to specifically optimize the horizontal intensity distribution in the resulting images. The loss function has the following form

$$\bar{\boldsymbol{p}} = \sum_{y=0}^{h-1} P[y,x], \qquad \bar{\boldsymbol{t}} = \sum_{y=0}^{h-1} T[y,x], \qquad loss(P,T) = \frac{\bar{\boldsymbol{p}} \cdot \bar{\boldsymbol{t}}}{||\bar{\boldsymbol{p}}||\ ||\bar{\boldsymbol{t}}||}, \tag{1}$$

where P, T are predicted and true images respectively, h is the height of an image. It is essentially a cosine of an angle between vectors representing horizontal alignment of knots. The value range is $[0,1]$ due to the fact that each vector element is nonnegative (vector elements are sums of pixel values, which are normalized to $[0,1]$ range). The training is performed on pairs of matching log and board: loss from Eq.1 is calculated using the log and board "barcodes" as predicted images and x-ray cluster locations as true images. Both losses are then summed up to get the final loss for the iteration. This is done in order to train decoder on both modalities simultaneously.

3.3 "Barcodes" cross-correlation

After the heightmap and the board image are translated into the "barcode" images it is possible to measure their similarity. This is done by summing up barcodes along their height dimension which results in 1D signals representing

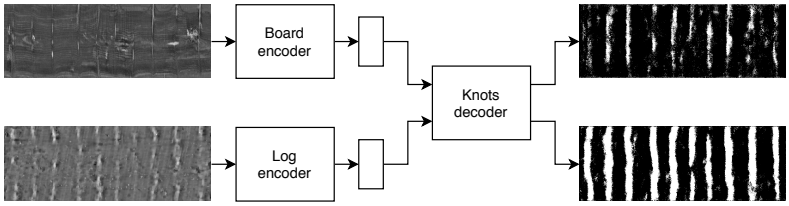


Fig. 4: Schematic of the multimodal translation process. It should be noted that there is only a single decoder.

the locations of knots along the length of logs or boards. Before the matching, the signals are normalized to be unit vectors. The similarity of the normalized signals is computed by using cross-correlation. The log that contains the highest similarity with the given board is selected as a match.

The conversion to 1D signals, instead of using the "barcodes" directly, is done in order to emphasize the similar properties of logs and boards. As was previously mentioned, vertical locations of knots differ even in the matching pairs of logs and boards. The proposed approach is invariant to the vertical locations of knots as it does the matching based on the horizontal knot distribution.

4 Experiments and results

4.1 Data

The data for the experiments consisted of 50 logs and 274 boards sawn from them. The logs were selected from 5 different categories with varying diameters and quality, 10 logs from each category. The logs in Category 1 were numbered from 1 to 10, the logs in Category 2 from 11 to 20 and so on. The logs in Categories 1 to 3 (log numbers 1-30) were low quality logs with a larger amount of knots enabling easier matching while the logs in Categories 4 to 5 (log numbers 41-50) were high quality logs with a small amount of knots. A laser scanner was used to generate point clouds of the log surfaces, and an X-ray scanner to extract the locations of the knot clusters. Each log was color coded manually so that it was possible to track them through the sawing process providing the ground truth for the board-to-log matching.

After the logs were sawn, the resulting boards were imaged with an RGB camera system. The boards representing the main yield were imaged from both upper and lower sides while the side boards were imaged only from one side resulting in 393 board images in total. The different sides of the same board were treated as separate samples in the experiments. The images were converted to grayscale, rescaled to 448×160, and the colors were inverted. Inverting the colors was done to make the knots appear as white instead of black. Examples of the data used for training are presented in Fig. 5.

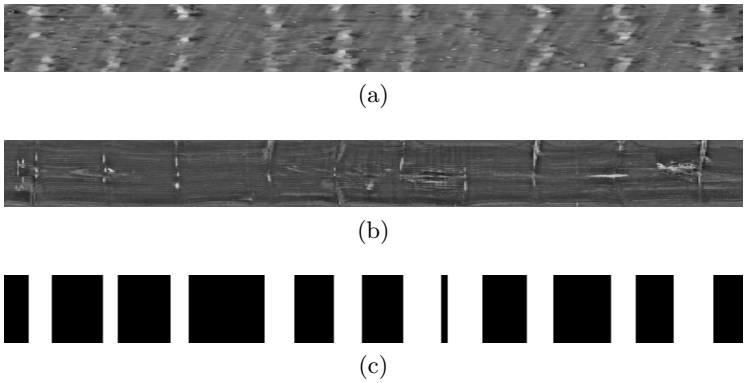


Fig. 5: Examples of preprocessed data: (a) A log heightmap; (b) A board image; (c) A "barcode" image of the knot clusters.

4.2 Experimental arrangements

The encoders and the decoders were trained simultaneously on pairs of log heightmaps and board images from the same log as shown in Fig. 4. After the conversion of the heightmap and the board image the loss function was computed as described in Sec. 3.2 using the ground truth "barcode" images, generated using discrete X-ray tomography. During the training, data was augmented by random flips, small gaussian noise and vertical cyclical shift of the heightmaps. The vertical shift emulates different log rotations during laser measurements. Networks were trained by using Adam optimizer with the default parameters and learning rate of 0.001 and batches of size 7. Precision, recall and F-score for each log were calculated by treating the logs as classes for the board identification. The quality of the board-to-log matching was assessed using the weighted average F-score for all the logs. First, F-score is calculated for each class, then average is calculated by weighting each class score according to the support size of that class.

Due to the heterogeneous nature of the available data (5 different log categories), it is necessary to evaluate the effects of different categories on the training. In order to do that, one log category was completely excluded from the training and used as a test set. After that, 10 randomly selected logs from the rest were used for validation, and remaining logs were used to train the network, repeating this procedure 10 times for each test category. In order to compare the selected architectures of encoder-decoder networks, each network was trained 20 times, and results on the whole dataset and validation subsets was compared. Out of the 50 logs, 15 were used for validation and were excluded from the training set. Boards sawn from those logs were also excluded, meaning that approximately 30% of data was used for the validation. Finally, usage of additional information was studied. Similarly to [4], the lengths of the logs and the boards were used in the matching process. The results with and without the augmentation were compared. The augmentation was done simply by subtracting

the difference of lengths from the cross-correlation value. Before subtracting, the difference was multiplied by 0.001. Boards are sawn along the length of the log, therefore they have approximately the same length. Negative length difference is used as a similarity measure along with a correlation coefficient.

4.3 Results

Effects of log category on the training The effect of the log quality on matching (category) was tested by completely excluding each category from training and testing the network on that set after training. The results are presented in Fig. 6. F-scores are rather high for low quality logs with large amounts of knots (1-30), but when high quality logs (31-50) are excluded from the training, the network is unable to generalize the learned features to them. This could mean that the higher the quality of the log, the more representative it is and possesses more subtle features. Whereas on the low quality logs the network quickly overfits on the few more prominent features that are inherent for the low quality wood. Therefore, it might be a good rule of thumb to include more high quality logs into the training procedure.

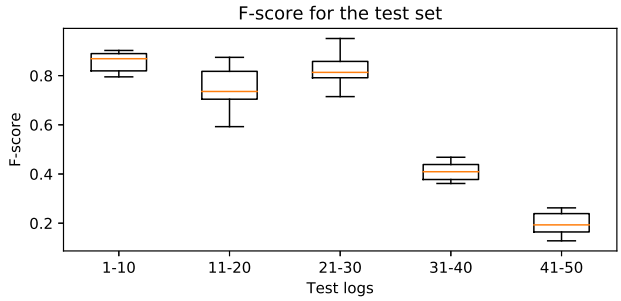
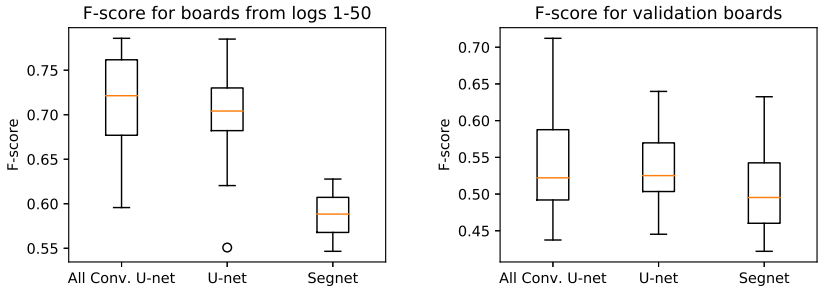


Fig. 6: The distribution of average F-scores on the test dataset. The box spans from lower to higher quartiles with a line at the median value. Lines extended from the box show the range of data with outliers depicted as white circles.

Architecture comparison The results of training 20 models of each encoder-decoder architecture are presented in Fig. 7. Validation set used comprised of 15 logs in total, 3 logs were taken from each category of quality. This was done in order to make the training set more representative. U-net and Convolutional U-net have similar scores on the validation set, both outperforming Segnet. All Convolutional U-net also has the highest score ceiling. When tested on the whole dataset, Convolutional U-net has the highest median score. Thus, All Convolutional U-net can be considered as the best architecture overall. The large spread

of the scores can be explained by the nature of this particular translation task. In contrast to an ordinary classification task, there is an additional step of computing 1D signals and matching them together. This makes the result dependent not only on the sample itself, but on a large number of other samples that are used in the cross-correlation. Moreover, the sample size is rather small, and deep neural networks are notoriously dependent on the amount of the available training data. Keeping in mind results of the experiment from Section 4.3, the training could be made more robust by including more high quality logs into the training set.



(a) (b)

Fig. 7: The distribution of average F-scores for the selected architectures of the encoder-decoder networks: (a) All logs and boards used as input; (b) Only boards in the validation set used as input and matched to all logs.

Utilizing log and board lengths Table 1 shows results of matching for the All Convolutional U-Net network. The results presented are mean values of scores calculated with each of 20 networks trained in Section 4.3. Each 10 logs represent a category of increasing quality. It can be seen that as the quality increases the matching accuracy drops. This is expected as the main features that are used during the matching, i.e., knots, are less prevalent or completely absent in the high quality wood. The overall quality of matching is still reasonably high, reaching almost perfect scores for the logs of lower quality. The overall F-score of 0.72 is already reasonably high, but it can be improved even further by utilizing the lengths of boards and logs. It is evident from Table 2 that the size information itself is not enough to perform precise matching. However, as it can be seen from Table 3, augmenting the "barcode" matching with the size information shows a significant improvement in the matching accuracy, especially for the high quality logs. This demonstrates the flexibility of the approach, allowing for the utilization of the additional information that is available in a sawmill.

Table 1: Average board matching scores over all runs without size augmentation. Maximum scores over all runs are presented in parentheses. P is precision, R is recall and F is F-score.

Log IDs	1-10	11-20	21-30	31-40	41-50	1-30	31-50	1-50	Validation
P (%)	99 (100)	98 (100)	96 (98)	82 (96)	74 (87)	95 (97)	74 (84)	77 (83)	65 (80)
R (%)	95 (98)	92 (99)	85 (93)	61 (75)	56 (70)	91 (96)	58 (73)	74 (81)	52 (66)
F (%)	97 (99)	93 (99)	87 (95)	67 (79)	60 (73)	91 (96)	62 (75)	72 (79)	54 (71)

Table 2: Board matching results using only lengths.

Log IDs	1-10	11-20	21-30	31-40	41-50	1-30	31-50	1-50
P (%)	68	50	41	51	82	38	54	34
R (%)	42	34	26	25	52	34	38	36
F (%)	49	37	31	33	60	35	41	33

Table 3: Average and maximum board matching scores with size augmentation.

Log IDs	1-10	11-20	21-30	31-40	41-50	1-30	31-50	1-50	Validation
P (%)	99 (100)	99 (100)	97 (98)	93 (99)	91 (100)	98 (100)	89 (96)	87 (92)	87 (94)
R (%)	98 (100)	96 (100)	93 (98)	72 (80)	72 (83)	96 (99)	72 (81)	83 (88)	68 (78)
F (%)	98 (100)	97 (100)	94 (98)	78 (87)	77 (88)	96 (99)	76 (85)	81 (87)	73 (81)

5 Conclusion

A multimodal translation-based method for matching sawn timber to the logs from which they were sawn was proposed. The method uses the board images and the laser scanner data of the log surfaces as the input and utilizes the discrete X-ray tomography data to train a multimodal encoder-detector network translating the board and log data to a matchable representations. It should be noted that the X-ray data about the knot clusters is required only in the training phase. Once the network has been trained only the laser scanner and camera systems are needed providing a cost-effective method for timber tracing. The effects of different log categories on the training were studied and it can be concluded that the proposed method generalizes best when trained on higher quality logs. Three different architectures of the encoder-decoder networks were evaluated and a modified version of U-Net was chosen based on the evaluation. Furthermore, it was shown that the method performance can be improved by using additional information, such as lengths of boards and logs. The future work includes the evaluation of the method with larger datasets. In addition, modifying this method to utilize features not limited to the knots could greatly improve its effectiveness when working with high quality wood.

Acknowledgements

The research was carried out in the DigiSaw project (No. 2894/31/2017) funded by Business Finland and the participating companies. The authors would like to thank Finnos Oy, FinScan Oy, and Stora Enso Wood Products Oy Ltd for providing the data for the experiments.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 2481–2495 (2017)
2. Chiorescu, S., Grönlund, A.: The fingerprint method: Using over-bark and under-bark log measurement data generated by three-dimensional log scanners in combination with radiofrequency identification tags to achieve traceability in the log yard at the sawmill. Scandinavian journal of forest research **19**(4), 374–383 (2004)
3. Dykstra, D.P., Kuru, G., Taylor, R., Nussbaum, R., Magrath, W.B., Story, J., et al.: Technologies for wood tracking: verifying and monitoring the chain of custody and legal compliance in the timber industry. World Bank, Washington, DC (2002)
4. Flodin, J., Oja, J., Grönlund, A.: Fingerprint traceability of sawn products using industrial measurement systems for x-ray log scanning and sawn timber surface scanning. Forest products journal **58**(11), 100–105 (2008)
5. Kozak, R.A., Maness, T.C.: A system for continuous process improvement in wood products manufacturing. Holz als Roh- und Werkstoff **61**(2), 95–102 (2003)
6. Kuga, R., Kanezaki, A., Samejima, M., Sugano, Y., Matsushita, Y.: Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In: ICCV Workshops (2017)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
8. Pahlberg, T., Hagman, O., Thurley, M.: Recognition of boards using wood fingerprints based on a fusion of feature detection methods. Computers and Electronics in Agriculture **111**, 164–173 (2015)
9. Põlder, A., Juurma, M., Tamre, M.: Wood products automatic identification based on fingerprint method. Journal of Vibroengineering **14**(2) (2012)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
11. 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: International Conference on Computer Analysis of Images and Patterns. pp. 752–765 (2015)
12. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
13. Wang, Y., van de Weijer, J., Herranz, L.: Mix and match networks: Encoder-decoder alignment for zero-pair image translation. In: CVPR (2018)