A transformer-based wood surface defect detection

Miha Ožbot
Faculty of Electrical Engineering,
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
miha.ozbot@fe.uni-lj.si

Janez Perš
Faculty of Electrical Engineering,
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
janez.pers@fe.uni-lj.si

Abstract

In this paper, an approach of image inspection for wood surface defect detection based on the attention mechanism is presented. We address the possibility of applying the Swin Transformer architecture to the problem of wood surface defect detection. We show that transformers present a reliable way to detect complex defects, principally blue-stain. The wood surface visual database used in this study was acquired in association with the local wood manufacturing industry. It consists of 297 images of wood planks from the trees Picea abies and Pinus sylvestris, containing 18 categories of defects. Our method achieves an overall defect detection accuracy rate of 95.3% on a benchmark dataset and 87.8% on our extended defect types dataset.

1. Introduction

In the last decades, a large variety of automated surface inspection (ASI) methods have been proposed, to achieve a reliable identification of wood surface defects. Lumber quality grading was proven to be a tedious task for humans, resulting in a less than optimal detection accuracy rate [2, 18, 20, 11]. Automated detection based on computer vision, promises to solve this problem while speeding up manufacturing significantly.

Wood surface defects lower the value of the whole wood plank as they might cause structural weakness or limit their use in furniture manufacturing and veneers. In the industrial manufacture plant, the segmentation procedure is followed by the cutting of planks to remove the inadequate parts. Therefore, false-positive detection of defects is a considerable problem, as healthy units might be discarded, lowering the profit margins. One of the main goals of our algorithm is to minimize the ratio of false-positive detections, and consequently the discard ratio of healthy units.

Importantly, the production process may cause some abrasions or scratches on the surface of the planks. An additional requirement is the classification of the wood general texture as radial, semi radial, or tangential. Wood texture can affect the wood value and strength, so it should be sorted accordingly. It is important to note, that multiple textures can be present on a single plank. Incidentally, any wobble of the passing plank as it moves by the fixed camera may cause an abrupt shift in the intensity of the measured color, further complicating the detection problem.

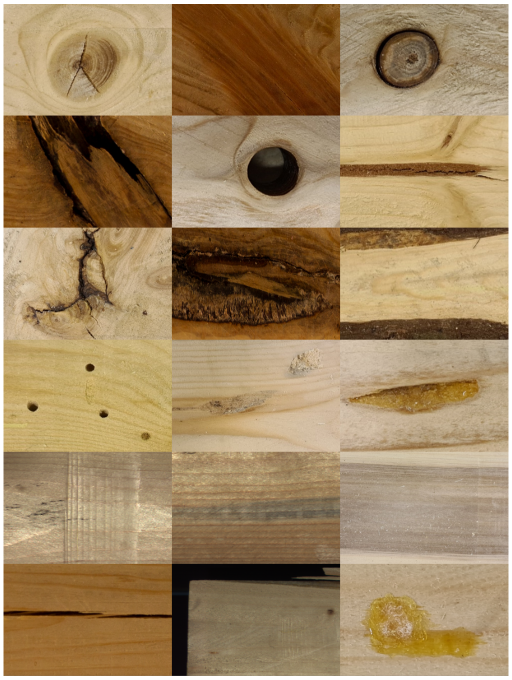

Figure 1. In order of appearance (top left to bottom right), the 18 defect types are: healthy knot (1), healthy knot spike (2), black knot (3), black knot spike (4), fallen knot (5), pith (6), bark pocket (7), rot/decay (8), bark/wain (9), worm holes (10), worm gallery (11), resin pocket (12), fibre damage (13), blue stain (14), brown stain (15), cracks/splits (16), lack of material (17), and resin (18).

A persistently challenging task in wood surface segmentation is the detection of a discoloration called Blue stain, which is caused by microscopic fungi, and may occur as a blue, gray, or brown variant. Supposedly, Blue stain does not cause structural weakness of the wood but can lower the aesthetic value of the planks used as building, or furniture material. Other stains are also possible on the wood surface, which are sometimes confused for blue stain, i.e. chemical brown stain that is caused by drying, or lubricant oil from machinery. This defect is mainly visible on the blue channel, thus it requires the use of techniques that take into consideration color images. Consequently, this is one of the most challenging tasks in wood surface defect detection. The blue stain defect detection presents a significant challenge even to a human observer. It does not have a typical shape, it can envelop other defects, it can stretch across a large portion of the observed wood plank, it can be mistaken for other defects, and it requires color information. To address these challenges we use the Swin Transformer [15] to tackle the problem of wood surface feature extraction, focussing mainly on the blue stain.

2. Related work

An overview of the classically used method for wood surface quality inspection is presented in [11]. In recent years, a variety of deep convolution neural networks managed to achieve surface defect detection accuracy that meets the industry's 95% accuracy standards. In [20], a faster region-based convolutional network (Faster R-CNN) was used for wood defect detection. A Transfer learning method was used for the initial training, based on the pretrained neural network models: AlexNet, VGG-16, ResNet, and GoogleLeNet. Transfer learning based on ResNet was also done in [8]. Transfer learning consistently outperformed training from scratch in all examples. The wood texture was also classified with a different dataset. In another study, the problem of complex non-uniform plank background segmentation was addressed by image binarization and several local thresholding methods [16]. A compact CNN for surface defect detection, classification, and segmentation was presented in [9]. It uses a depthwise separable convolution to reduce the number of calculations and weights. A fully convolutional neural network (CNN) was used in [6]. Notably, they used polar transformation to obtain different wood surface defect shapes. A pair of inception convolutional modules were used to capture features from multiple scales. This study was followed up with a similar 16-layer deep CNN [7]. The authors emphasized the need for an automatic wood feature learning method, in contrast with complex extractions of features. In [22] a CNN feature extractor with an Extreme learning machine (ELM) classifier was used for wood defect box semantic segmentation.

A different approach is to detect anomalies in a predominantly non-anomalous dataset. A surface anomaly detection with mixed supervised learning was presented in [1], producing segmentation and classification outputs. Unfortunately, the wood type that this study focuses on is heavily anomalous and such methods would result in suboptimal detection.

The attention-based Transformers, originally meant for language translation and text generation [21], have been getting top scores on the classification, detection, and semantic segmentation challenges, but have not yet been applied on wood surface datasets. Transformers are used to address the problem of different data sizes and global spatial connections, which is crucial for the detection of blue-stain defects, that can stretch over large portions of a wood plank.

Our paper is most similar to the Swin Transformer [15]. It uses hierarchical architecture and shifting input windows to address the problem of multiple object scales. Its successor, the Swin Transformer V2 [14], used logarithmic space coordinate transformation with relative position bias, to address the problem of different scales. The attention dot product was replaced by a scaled cosine calculation with learnable scaling to limit any single pixel dominance. CoAtNet [3] combines transformers and convolution, by the summation of a kernel to the self-attention calculation, and by adding convolution layers before the transformer layer to lower dimensionality.

3. Methods

3.1. Datasets

We introduce a new dataset that consists of 297 RGB color images of wood planks from the spruce (Picea abies) and pine (Pinus Sylvestris) trees. Although this is not a large number of images, these wood types have a large density of defects, and especially prevalent are knot defects. Conversely, blue stain defects are relatively uncommon, which present a challenge. Images of size $W \times H = 4144 \times 384$ pixels were obtained with the use of a SICK Ranger-E55434 camera that enables a linear scan of the passing plank, which had a length of 1500, and width of 130 millimeters. The image detection system was shielded from the outside light disturbances and illuminated with a fixed light source but was still subject to random disturbances of the wood plank movement and approach angle. The images were preprocessed to remove the background and roughly realigned to address the mentioned plank movement, although some distortion artifacts remained present in the data. A rough annotation of defects was done with box-shaped labels by experts from the wood industry.

Additionally, a large publicly available benchmark dataset [17] was used to compare results with previous work. This dataset contains 43000 labeled wood surface defects divided into ten defect types, which mostly match ours, and importantly include the blue stain defect. Furthermore, this dataset was not acquired in a controlled environment but, similarly to our dataset, was collected in an industrial setting.

3.2. Transformers

We experiment with a Transformer feature extraction backbone for wood surface defect detection. It uses a mechanism called attention to calculating relations between elements in the input data sequence. For use on images, an image window is divided and flattened into a sequence of patches. The order of the patches in the sequence is important and is taken into consideration with a position embedding. An attention mechanism (1) calculates a weighted sum of value V vectors based on a scalar product of key K are query Q vectors, which are a linear transformation of the input signal X with learnable parameters W.

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^T}{\sqrt{d_k}} + B\Big)V, \quad (1)$$

where $A \in \mathbb{R}^{M^2 \times d_k}$, $Q \in \mathbb{R}^{M^2 \times d_k}$, $K \in \mathbb{R}^{M^2 \times d_k}$, $V \in \mathbb{R}^{M^2 \times d_k}$ are attention, query, key, and value matrices, respectively; $B \in \mathbb{R}^{M^2 \times M^2}$; M^2 is the number of patches in a windows; and d_k is the key and query dimensionality.

Every Transformer block is comprised of multiple parallel attention calculations collectively called multi-head self-attention (MSA), with their respective keys, queries and values. Incidentally, self-attention indicates that all the keys, queries, and values vectors used in the attention calculation came from the same block.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_i)W^O \quad (2)$$

$$\text{head}_i = \text{Attention}(QW_i{}^Q, KW_i{}^K, VW_i{}^V) \quad (3)$$

where $W_i^Q \in \mathbb{R}^{d_k \times d_m}, W_i^K \in \mathbb{R}^{d_k \times d_m}, W_i^V \in \mathbb{R}^{d_k \times d_m}$ are learnable parameter matrices; i is the head index; and d_m is a model dimension parameter. Feed-forward layers are used before every MSA calculation to compute non-linear hierarchical features by combining the outputs of all attention heads into one output. This also allows the network to have fixed internal layer output sizes, even when modified for larger fine-tuning input images. A LayerNorm (LN) is implemented before or after every MSA to increase computation stability and enable the comparison of feature values between stages, while residual connections solve the problem of exploding/vanishing gradients during training.

Although transformers are effective at storing information in their parameters, they have one innate disadvantage when compared to CNN, that is, they are not invariant to translations in the input data, which is overcome by training transformers with a huge number of examples [4]. Training this type of model architecture end-to-end is an arduous task, requiring an enormous hand-labeled dataset and GPU processing power. Therefore, a pretrained Transformer model is commonly leveraged as a backbone feature extractor. In the case of object detection, it can be fine-tuned in conjunction with a classifier layer and a region proposal algorithm.

The aforementioned Swin Transformer [15] has four stages of multiple successive twin transformer blocks with patch merging modules in between each stage. It can be used as a backbone structure for a variety of image recognition tasks. It divides the input image into windows and each window into patches, while also shifting the windows locations in between consecutive transformer layers. This scaling and window crossing enables the detection of differently sized defects and is also circumvents the issue of quadratic complexity scaling with image size since this operation scales linearly. The smallest version of the Swin transformer is stylized as Swin-T and is the one we used in this study.

The object classification average precision (AP) used in the evaluation is calculated as the mean area under the precision-recall curve for all classes, and for a range of Intersection over Union (IoU) values as proposed for the COCO dataset evaluation [13].

4. Experiments

We perform experiments to compare the attention mechanism with the best convolutional neural networks results for the general wood defect detection task, and we examine the question of how effective are Transformers at detecting blue stain defects. First, we perform object detection with the classes that match both datasets, i.e. healthy knots, black knots, cracks, missing knots, resins, and blue stain. In the second test, we use all of the available defect classes. The specifications of the hardware and software used for this experimentation are presented in table 1.

Hardware	Software
NVIDIA GeForce RTX T400 2GB	Python 3.7.5
Intel Core i7-10700 2.90GHz	CUDA 10.1 + cudnn7
16GB RAM	PyTorch 1.7.1
	torchvision 0.8.2
	timm 0.3.2

Table 1. Experimental hardware and software environments

Our novel spruce and pine wood surface dataset images were split and reshaped into images of size 265×265 pixels, to meet the Transformer backbone requirements. The images were then augmented by random cropping, horizontal reflections, scaling, rotation, contrast change, and saturation change. The original dataset was extended to achieve a better balance of defect types, which is important since this study focuses on an uncommon defect type. For easier handling, it was reformed into a COCO dataset format, and divided into subsets with a ratio of 3:1:1, for training, validation, and testing, respectively. Additionally, we performed

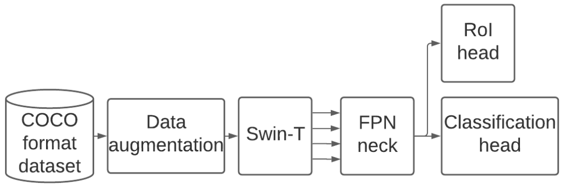

Figure 2. The Swin Transformer backbone architecture with a Feature Pyramid network neck, an object detection head, region proposal algorithm head.

Figure 3. Loss function during training of the Mask R-CNN classifier for our dataset and the benchmark dataset

Classifier	Dataset	Evaluation	
		$\mathrm{AP}^{\mathrm{box}}$	$\mathrm{AP}^{\mathrm{box}}_{75}$
Mask R-CNN [5]	our		
Faster R-CNN [19]	our		
Mask R-CNN [5]	benchmark [17]		
Faster R-CNN [19]	benchmark [17]		

Table 2. Comparison of the evaluated methods with the Swin-T backbone. $\mathrm{AP}^{\mathrm{box}}_{75}$ denotes the average precision for IoU over 0.75.

data augmentation based on the AutoAugment [23] policy during the training procedure. Each of the two datasets was used separately for both fine-tunning and inference.

The Swin-T backbone [15] was trained on the COCO object detection and classification dataset [13]. The two datasets were compared in the inference task with the trained classifier heads, i.e a Mask R-CNN [5], and a Faster R-CNN [19]. A Feature Pyramid Network (FPN)[12] was used with the Swin Transformer hierarchical feature maps from each of the four stages to improve feature quality. For optimization, the Adam optimization technique [10] was used, which is a stochastic gradient descent method that the first-order and the second-order moments for the optimization. Multiple learning rates, beta, and dropout values were tested.

5. Results

6. Conclusion

7. References

[1] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 8 2021.

[2] Ibrahim Cetiner, Ahmet Ali Var, and Halit Cetiner. Classification of knot defect types using wavelets and knn. *Elektronika ir Elektrotechnika*, 22:67–72, 12 2016.

[3] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. 6 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 10 2020.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] Ting He, Ying Liu, Chengyi Xu, Xiaolin Zhou, Zhongkang Hu, and Jianan Fan. A fully convolutional neural network for wood defect location and identification. *IEEE Access*, 7:123453–123462, 2019.

[7] Ting He, Ying Liu, Yabin Yu, Qian Zhao, and Zhongkang Hu. Application of deep convolutional neural network on feature extraction and detection of wood defects. *Measurement: Journal of the International Measurement Confederation*, 152, 2 2020.

[8] Junfeng Hu, Wenlong Song, Wei Zhang, Yafeng Zhao, and Alper Yilmaz. Deep learning for use in lumber classification tasks. *Wood Science and Technology*, 53:505–517, 3 2019.

[9] Yibin Huang, Congying Qiu, Xiaonan Wang, Shijun Wang, and Kui Yuan. A compact convolutional neural network for surface defect inspection. *Sensors 2020, Vol. 20, Page 1974*, 20:1974, 4 2020.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014.

[11] Martin Kryl, Lukas Danys, Rene Jaros, Radek Martinek, Pavel Kodytek, and Petr Bilik. Wood recognition and quality imaging inspection systems, 2020.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. 12 2016.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. 5 2014.

[14] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. 11 2021.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 3 2021.

[16] Wei Luo and Liping Sun. An improved binarization algorithm of wood image defect segmentation based on non-uniform background. *Journal of Forestry Research*, 30:1527–1533, 8 2019.

[17] Kodytek Pavel, Bodzas Alexandra, and Bilik Petr. Supporting data for deep learning and machine vision based approaches for automated wood defect detection and quality control., Apr 2021.

[18] Vincenzo Piuri and Fabio Scotti. Design of an automatic wood types classification system by using fluorescence spectra. *undefined*, 40:358–366, 5 2010.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 6 2015.

[20] Augustas Urbonas, Vidas Raudonis, Rytis Maskeliunas, and Robertas Damaševičius. Automated identification of wood veneer surface defects using faster region-based convolutional neural network with data augmentation and transfer learning. *Applied Sciences (Switzerland)*, 9, 11 2019.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 2017-December, 2017.

[22] Yutu Yang, Xiaolin Zhou, Ying Liu, Zhongkang Hu, and Fenglong Ding. Wood defect detection based on depth extreme learning machine. *Applied Sciences (Switzerland)*, 10:1–14, 11 2020.

[23] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. Learning data augmentation strategies for object detection. 6 2019.