Proceedings of the Federated Conference on Computer Science and Information Systems pp. 101–105
DOI: 10.15439/2018F286
ISSN 2300-5963 ACSIS, Vol. 15

Barley Variety Recognition with Viewpoint-aware Double-stream Convolutional Neural Networks

Przemysław Dolata
Wrocław University of Science and Technology
ul. Wyb. Wyspiańskiego 27,
50-370 Wrocław, Poland
Email: przemyslaw.dolata@pwr.edu.pl

Jacek Reiner
Wrocław University of Science and Technology
ul. Wyb. Wyspiańskiego 27,
50-370 Wrocław, Poland
Email: jacek.reiner@pwr.edu.pl

***Abstract*—Varietal homogeneity is an important factor in quality of malting barley, but its inspection is difficult. Biochemical methods are expensive and inefficient, while machine vision suffers due to high variability of the grains' features. In our previous work, we have shown a convolutional neural network for simultaneous feature extraction and classification of image data basing on multiple views. It was suggested that for machine vision inspection, the observed side of a grain should be taken into account – dorsal and ventral sides of each kernel exhibit different features. In this study we present a viewpoint-aware convolutional neural network, which learns to extract specialized features from images of dorsal and ventral sides of barley grains. We show that it increases the average classification accuracy by 0.6% and sensitivity by 2.3% with respect to the viewpoint-ignorant architecture on our dataset.**

I. Introduction

In the case of food products, the quality of ingredients which they are produced from plays a significant role, hence their comprehensive inspection is necessary. The sooner the potential fault can be detected, the lower is the actual production cost including wasted materials. Therefore, effective and quick quality assessment requires automation. Among non-invasive methods, machine vision has a special significance. However, the difficulty in identifying quantitative features and their considerable variability recently draws attention to artificial intelligence methods, especially deep learning.

An example of such a natural product is barley, especially its malting varieties, which is a key ingredient in the production of beer and whiskey. Any deficiencies in its quality immediately affect quality and therefore value of the finished product. Hence, the examination of purchased barley includes detection of impurities or damage as well as moisture content and protein content measurement. Typically, such assessment is performed visually by sampled statistical process control (SPC) as it is technically infeasible to inspect all individual grains in a shipment weighing several tons. This process is however very tedious and, due to the difficulty of the task as well as the human's fatigue, error-prone.

This work was supported from the internal budget of Mechanical Faculty of WUST. The source material (barley grain) for the dataset preparation was supplied from a project financed by National Centre for Research and Development - Project PBS3/A8/38/2015.

The subtle flavors of beer and whiskey are determined, inter alia, by enzymes associated with varieties of barley. However, control of varietal homogeneity without expensive bio-chemical tests is still an unresolved problem. One possible approach, using machine vision methods, seems particularly promising due to its potential speed and no need for direct interaction with the grains. The aforementioned difficulty of feature identification becomes even more challenging in the case of barley grains, because they exhibit different features on dorsal and ventral sides.

In this study, we present a machine vision approach to recognition of barley varieties using convolutional neural networks. We propose a neural network architecture with two feature extraction streams, each specialized to process images of a specific side of the grain. This architecture is complemented with a preprocessing recognition step, in order to identify the dorsoventral orientation of each grain. The paper is arranged as follows: in section II we reintroduce a double-stream convolutional neural network from our previous work, in section III we describe the novel architecture, as well as the dataset and training methods, and in section IV we experimentally evaluate performance of the models and compare them.

II. Related works

There are several known approaches to barley grain varietal recognition. All of them rely on digital image processing and feature extraction. The features are usually hand-engineered, e.g. edges, texture descriptors, and low dimensional or reduced to a low dimension. Most works also employ some form of machine learning to perform recognition. Zapotoczny *et al.* [1] explore possibilities of classifying images of barley kernels using principal component analysis (PCA), and linear or non-linear discriminant analysis (LDA/NDA). Nowakowski *et al.* [2] use extracted features as learning vectors for an artificial neural network (a multilayer perceptron). Hailu and Meshesha [3] present a classifying ensemble of k nearest neighbors and an artificial neural network. Those approaches yield promising results, but they are only tested on very small datasets (up to several hundred images in up to 5 classes). The scope of work by Szczypiński *et al.* [4] is significantly larger, their dataset comprising over 13,000 images of 11 varieties.

In all of those works, feature extraction and classification are considered separate parts of a system, where the features remain fixed and the classifier is designed using machine learning (ML). Recent advancements in ML made it possible to learn the feature extraction function and classification as a single system. Convolutional neural networks (CNNs) can be trained on raw images, without the difficulty of manually designing the feature extractor. Despite their applications to other agriculture-related problems (e.g. [5]), there have been no attempts to classify barley varieties with CNNs.

In our previous paper [6] we have presented a CNN for detection of defects and impurities in barley grain. Our approach made use of the double-sided imaging capacity of the acquisition system presented by [7]. The double-stream CNN was able to extract features from images of both sides of the grain. Then it fused the feature vectors together, creating a single representation from both images. This enabled it to utilize the information contained within both views of the object to predict its class.

However, due to the unpredictable nature of the imaging process, it was never known which side of each grain was actually visible on which image. Therefore the network had to be robust to this unpredictability, effectively discarding the information about dorsoventral orientation of the grains.

III. Experiment Setup

A. Reference neural network architecture

As a reference model we reintroduce the double-stream convolutional neural network from our previous work [6]. This architecture consists of two streams – that is, two separate CNNs – each assigned to a specific camera in order to process one image of each grain. At some point, depending on the setup details, representations produced by those streams are merged into a single stream. Classification is performed by a feed-forward fully-connected (FC) neural network, whose input is this merged representation.

Dorsal and ventral sides of a grain may exhibit different features. However, during the imaging process the dorsoventral orientation of the grains cannot be constrained. Therefore it is not known which side of the grained is imaged by which camera - the cameras cannot be assigned to any particular viewpoint (i.e. dorsal or ventral). In order to provide robustness against this unpredictability, the two feature extraction networks have shared parameters. That means that even though these are two separate networks with different data flowing through them, their parameters are shared, i.e. constrained to always be equal (fig. 1). Since the camera viewpoints are irrelevant in this setting, we term this architecture viewpoint-ignorant.

The actual CNN implementation used in this study is derived from AlexNet [8], comprising 5 convolutional layers of decreasing kernel size (respectively, 11x11, 5x5, 3x3, 3x3 and 3x3) with 3 overlapping pooling operations between them, and 3 fully-connected layers: first two followed by dropout layers [9], the last one by a softmax operation. After each convolutional and FC layer, a ReLU nonlinearity is applied [8]. The FC layers originally consist of 4096 neurons each except the last one, which is scaled depending on the number of classes. In order to reduce overfitting we limit the capacity of the network by replacing those layers with significantly smaller ones: 64 and 16 neurons each.

A double-stream network is constructed by instantiating two copies of each convolutional layer, although the

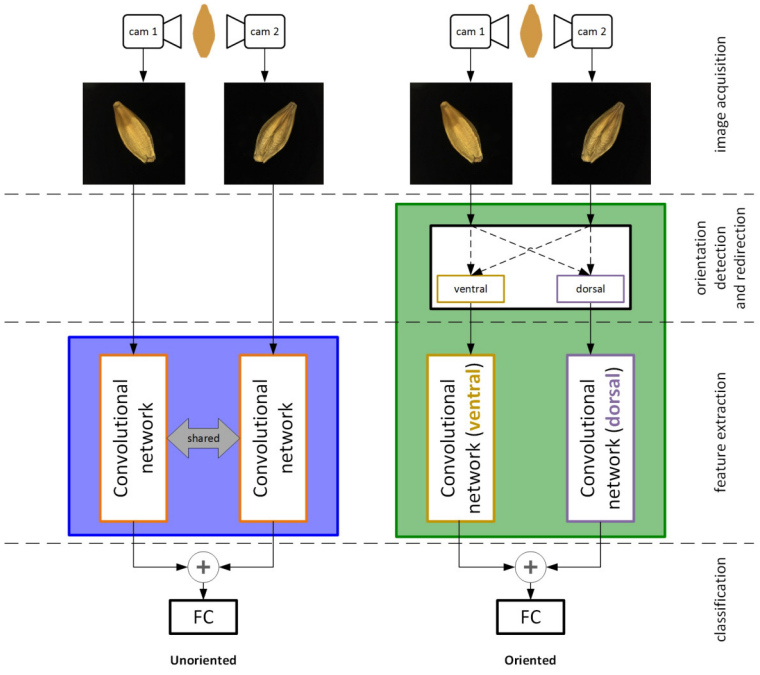

Fig 1. Double-stream CNN architectures: viewpoint-ignorant (on the left) and viewpoint-aware (on the right).

parameters (and gradients thereof) of each pair are constrained to be equal. Each stream will process an image of a different side of the grain, and during training their gradients will be summed. The streams will be merged after the last pooling operation by simple tensor concatenation. The following FC layers will proceed normally, although the input of the first FC layer will be larger to accommodate the concatenated outputs of the streams.

B. Proposed architecture

We propose a network architecture in which the streams are not robust to random dorsoventral orientation of the grains, but instead each learns to extract features specific to each side. In this setting it is assumed that each image the network receives is taken from a particular viewpoint – one of the ventral, another of the dorsal side. We term this network viewpoint-aware.

The novel architecture consists of two streams constructed exactly like in the reference architecture, except for the parameter and gradient equality constraints, which are lifted. This results in two entirely separate feature extraction networks, each associated with a specific view of the object. The rest of the architecture, particularly fusion of the streams by concatenation, and FC layers with softmax, remain structurally identical with the reference architecture.

The assumption that each stream receives a specific view of the object, as opposed to image from a specific camera, requires that the dorsoventral orientation of the object be identified before any processing. In order to achieve this, we introduce a preprocessing step in which the side of a grain is recognized, and the two images are redirected to their associated feature extraction streams according to the result. We solve the task of viewpoint recognition by reducing it to binary classification.

Both the CNN streams and the imaging cameras are ordered, so there is a natural correspondence between them – assume stream 1 is associated with ventral view, then camera 1 can capture either ventral or dorsal side. Therefore, viewpoint recognition becomes a binary classification task: either the images are acquired in the right alignment or not, in which case they need to be switched. For this, we use a CNN of the same structure as the reference architecture with shared streams, except for the final FC layer, which only has 2 outputs (correct alignment or switching needed). The complete system is shown in (fig. 1).

C. Dataset

The data used throughout this research was acquired using a prototype imaging system. The device captured RGB images of individual grains using two cameras located coaxially, opposing to each other, allowing for acquisition of top and bottom views of each grain (details in [7]). Due to the nature of the grain partitioning and transportation subsystem, the dorsoventral orientation of the objects was not predictable. For this reason, no orientation labels could be assigned to the images at the data preparation stage.

Barley was acquired from a research supply. The grains came already separated into 8 varieties, which could be grouped into spring (S) or winter (W), as well as malting (M) and fodder (F) varieties. There were exactly 2 varieties in each of the group combinations (SM, SF, WM, WF).

A total of 3169 pairs of top/bottom images were acquired, ranging between approximately 200 and 500 pairs per variety. During preprocessing, they were cropped so that the grains were visible in the center of the images, and then resized to 256x256. For the purpose of training and cross-validation, the dataset was split into 3 disjoint subsets. For every cross-validation bin, one of those subsets was used as a training set, while the two remaining ones were merged into a validation set. We applied data augmentation on each training set, appending copies of each image rotated 16 times. Table I shows the dataset composition (pre-augmentation).

TABLE I.
DATASET COMPOSITION

Variety	No. training samples	No. validation samples	Percent of total
SM Bordo	140	278	13.2%
SM Kormoran	163	327	15.5%
SF Mercada	133	266	12.6%
SF Skarb	129	257	12.2%
WM Vanessa	116	232	11.0%
WM Vincenta	166	334	15.8%
WF Kobuz	138	274	13.0%
WF Zenek	72	144	6.8%
Total	1057	2112	100%

D. Training procedure

Neural network training was performed using the Caffe framework [10], with Nvidia DIGITS front-end for task management, using a Nvidia GTX TITAN Z graphics processing unit (GPU) with 2 banks of 6 GB VRAM and 2880 CUDA cores each. To reduce the possibility of overfitting the data, the transfer learning technique was used: each of the convolutional layers was initialized from an AlexNet model pre-trained on ImageNet, a dataset of 1.5 million natural images of various origin in 1000 classes (pre-training, performed independently by Jeff Donahue, BVLC, was not a part of this study). The remaining layers' weights were initialized with Gaussian noise of mean 0 and standard deviation of 0.01, while biases were initialized with a constant of 0.1 each.

Networks were trained using multinomial logistic loss function and Nesterov Accelerated Gradient (NAG) optimization method [11], which is a variation of stochastic gradient descent with momentum. Major training hyper-parameters were: momentum $\mu = 0.9$, batch size 128 (as large as could fit in the GPU memory), initial learning rate $\alpha = 0.01$ (as high as the training could still converge at).

For variety recognition, the reference double-stream viewpoint-ignorant network was compared with the proposed viewpoint-aware network. Both networks were trained for 25 epochs: the first 2 epochs at learning rate 0.01, then until epoch 20 at rate 0.001 and for the remaining time at 0.0001. Each process was repeated 3 times for each of the cross-validation folds. Results (F1 measure and confusion

matrices) are reported by averaging of each 3 cross-validation models.

A naïve setup for the viewpoint-aware network would consist of a preprocessing network embedded into the architecture. However, since this sub-network is not being trained at this stage, its presence would only increase the memory and computation power requirement of the entire system, making the training process significantly slower. For this reason, the dorsoventral orientation recognition network was trained separately.

First, a subset of 500 images was selected from the main dataset and annotated manually (with another 500 images selected and annotated for the purpose of validation). The network was trained on this dataset for 20 epochs. After the first 8 epochs learning rate was reduced to 0.001, and after another 8 to 0.0001.

Then, the preprocessing network was queried once over the entire dataset to generate the auxiliary labels containing the information about dorsoventral orientation of each grain. The double-stream viewpoint-aware variety recognition network only read those labels at training time, reducing the preprocessing step only to redirecting images to the feature extraction streams as needed. When using such network in production environment, both the preprocessing sub-network and the variety recognition network would have to be instantiated at the same time.

IV. Results

A. Viewpoint recognition

The dorsoventral orientation recognition network reached 99.8% accuracy after 20 epochs of training. There was little to no overfitting, as both training and validation loss were equal to about -4 in logarithm. The 0.2% error rate was caused by a single image, in which the grain was not imaged from neither the dorsal nor ventral side, but from the sides (fig. 2). This is a rare occurrence which the imaging system allows, but due to an insignificant fraction of such images in the dataset, we decided to ignore their influence – those cases were not handled in any particular way.

B. Variety recognition

Training of a single variety recognition network took approximately 80 minutes (compared to less than 2 minutes for the preprocessing network). All of the models displayed a satisfactory fit – validation loss was actually lower than training, and accuracy was higher (fig. 3). Explanation for this counter-intuitive phenomenon is in dropout. During training, 50% of the fully-connected neurons are randomly deactivated, artificially increasing prediction difficulty, when during validation, no neurons are disabled (their activations are scaled down by a factor of 0.5 to preserve the total magnitude of the activation). This has a significant anti-overfitting effect.

Classification results comparing the viewpoint-ignorant and viewpoint-aware networks, averaged over cross-validation folds, are shown in Table II. Sensitivity and specificity are defined for binary classification, so the table

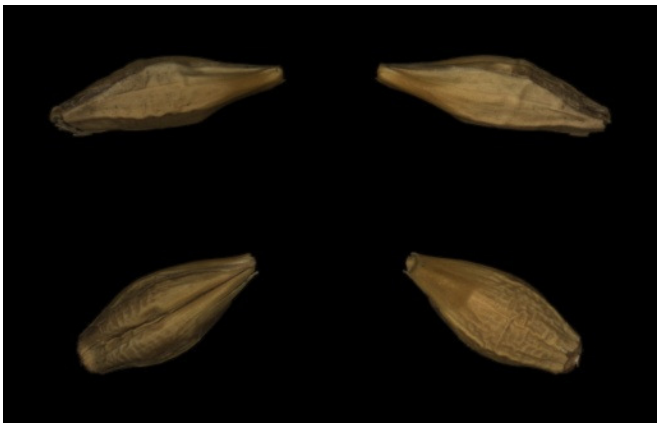

Fig 3. Difficult case of a sidewise grain orientation (top) versus normal grain exhibiting its ventral (bottom left) and dorsal sides (bottom right)

contains averages of values obtained for each class as a one-versus-all classification.

In fig. 4 we compare confusion matrices for viewpoint-ignorant and viewpoint-aware models. The matrices are normalized row-wise, so a percentage on each tile corresponds to a fraction of images from a given row that were recognized as belonging to the given column (true positive ratio on diagonal). In most cases, the TPR is higher for the viewpoint-aware network – most notably for SF Zenek, an increase from 51.9% to 69.7%. With two classes (SM Bordo, WM Vanessa) the viewpoint-aware network performed worse in terms of TPR. However, in those cases the classification precision (ratio of correct predictions to all predictions as this class, interpreted as probability that a prediction is correct) was significantly higher: 92.80% vs 91.92% for Bordo and 81.43% vs 77.60% for Vanessa.

This confirms that the viewpoint-aware approach is more powerful on average, but the scale of the difference would depend on weights assigned to errors of each kind.

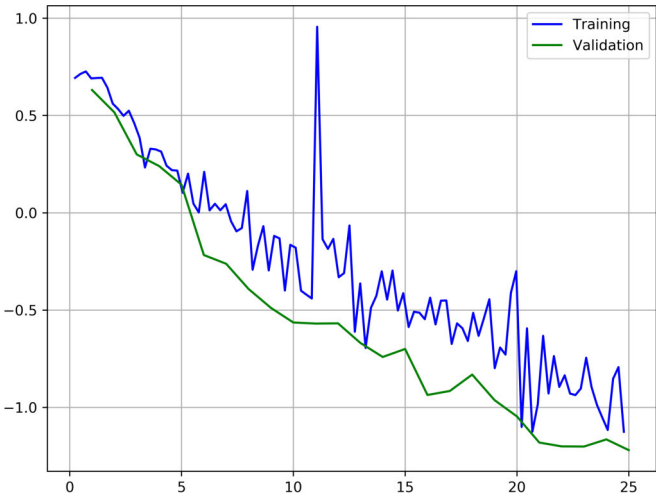

Fig 2. Loss function logarithm on training and validation sets throughout training

TABLE II.
VARIETY CLASSIFICATION RESULTS

Measure	Viewpoint-ignorant	Viewpoint-aware
Accuracy	96.65%	97.24%
Sensitivity	86.63%	88.97%
Specificity	98.09%	98.42%

V. CONCLUSIONS

We have presented a viewpoint-aware double-stream convolutional neural network and proved its superior performance at classification of barley grain varieties. The system performed better on average than a previously shown viewpoint-ignorant network, with slight variations in performance at particular classes.

Those differences might depend on the actual properties of the grains themselves. A detailed study into grain classification would have to account for many such factors, for example phenotypic variability of barley across vegetation seasons.

The system could in principle be trained in an end-to-end setup, if only the dorsal/ventral annotations were available. Due to the image acquisition technique as well as the nature of the imaged objects, obtaining those annotations during data acquisition is not trivial. This is however a limitation of the data imaging system, not our proposed machine learning system.

ACKNOWLEDGMENT

We wish to thank Piotr Lampa and Krzysztof Wall for performing data acquisition.

REFERENCES

[1] P. Zapotoczny, M. Zielinska, and Z. Nita, "Application of image analysis for the varietal classification of barley:," *Journal of Cereal Science*, vol. 48, no. 1, pp. 104–110, Jul. 2008. doi: 10.1016/j.jcs.2007.08.006

[2] K. Nowakowski, P. Boniecki, R. J. Tomczak, S. Kujawa, and B. Raba, "Identification of malting barley varieties using computer image analysis and artificial neural networks," presented at the *Fourth International Conference on Digital Image Processing (ICDIP 2012)*, 2012, vol. 8334, p. 833425. doi: 10.1117/12.954155

[3] B. Hailu and M. Meshesha, "Applying Image Processing for Malt-barley Seed Identification," presented at the *Conference: Ethiopian the 9th ICT Annual Conference 2016 (EICTAC 2016)*, Addis Ababa, 2016.

[4] P. M. Szczypiński, A. Klepaczko, and P. Zapotoczny, "Identifying barley varieties by computer vision," *Computers and Electronics in Agriculture*, vol. 110, pp. 1–8, Jan. 2015. doi: 10.1016/j.compag.2014.09.016

[5] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, "Deep-plant: Plant identification with convolutional neural networks," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 452–456. doi: 10.1109/ICIP.2015.7350839

[6] P. Dolata, M. Mrzygłód, and J. Reiner, "Double-stream Convolutional Neural Networks for Machine Vision Inspection of Natural Products," *Applied Artificial Intelligence*, vol. 31, no. 7–8, pp. 643–659, Sep. 2017. doi: 10.1080/08839514.2018.1428491

[7] P. Lampa, M. Mrzygłód, and J. Reiner, "Methods of manipulation and image acquisition of natural products on the example of cereal grains.," *Control & Cybernetics*, vol. 45, no. 3, 2016.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[10] Y. Jia et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," 2014, pp. 675–678. doi: 10.1145/2647868.2654889

[11] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the Importance of Initialization and Momentum in Deep Learning," in Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, Atlanta, GA, USA, 2013, pp. III–1139–III–1147.

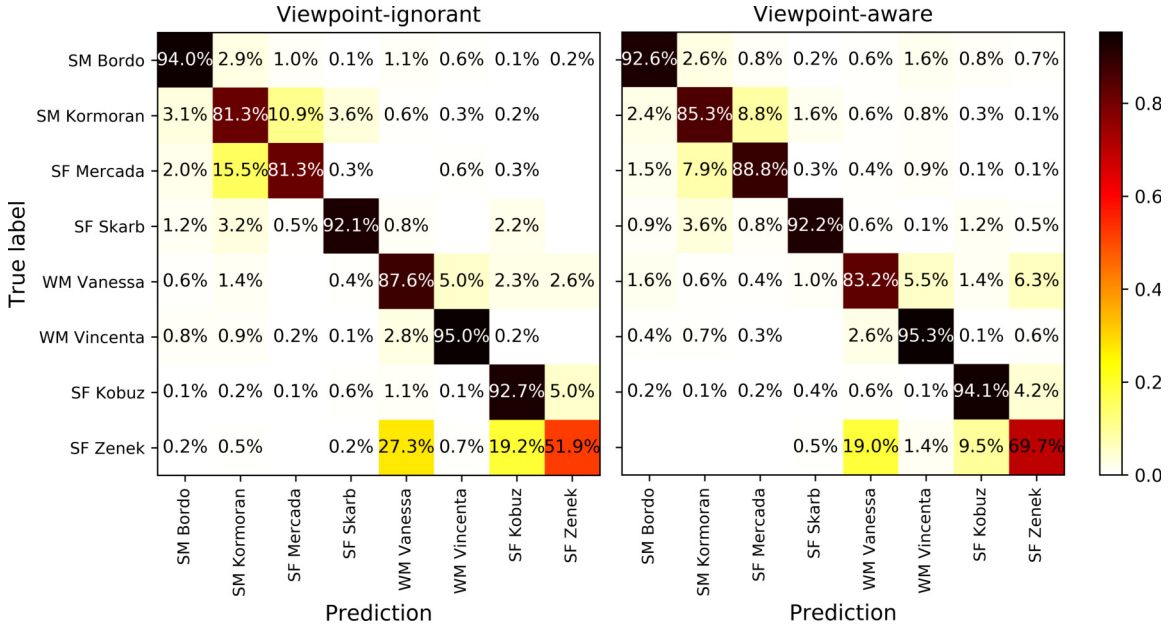

Fig 4. Confusion matrices for viewpoint-ignorant and viewpoint-aware models