RGB Images Driven Recognition of Grapevine Varieties using Densely Connected Convolutional Network

Pavel Škrabánek[1][0000−0001−6194−0467], Petr Doležel[2][0000−0002−7359−0764], and Radomil Matoušek[1][0000−0002−3142−0900]

[1] Brno University of Technology, Brno, Czech Republic,
pavel.skrabanek@vut.cz
[2] University of Pardubice, Pardubice, Czech Republic

Abstract. We present a grapevine variety recognition system based on a densely connected convolutional network. The proposed solution is aimed, as a data processing part of an affordable sensor, at selective harvesters. The system classifies size normalized RGB images according to varieties of grapes captured in the images. We train and evaluate the system on in-field images of ripe grapes captured without any artificial lighting, both facing toward the sun and away from the sun. A dataset created for this purpose consists of 7200 images classified into 8 categories. The system distinguishes between seven grapevine varieties and background, where four and three varieties have red and green grapes, respectively. We study the impact of data augmentation on classification performance of the system. We show that it is possible to achieve average per-class classification accuracies at 97.75 ± 0.25 % and 98.00 ± 0.00 % for red and green grape varieties, respectively. The system also accurately differentiates grapes from background. An overall average per-class accuracy of the system is at 98.00 ± 0.13 %. The results show that conventional cameras in combination with the proposed system allow construction of affordable automatic selective harvesters.

Keywords: Recognition of grapevine varieties · Densely connected convolutional network · Data augmentation · In-field images · Agriculture mechanization.

1 Introduction

Over the past few years, we have observed unprecedented progress of agriculture mechanization towards its full automation. The rapid development in areas such as computer vision and machine learning, likewise affordability of powerful hardware and precise manipulators, allows construction of autonomous robotic systems, e.g. for weed control [21], precise spraying [24, 3] and harvesting. Robots capable of cropping greenhouse vegetables, apples, grapes [3], sweet peppers [1] and even strawberries [33] have been presented. One of the directions of their further development is selective harvesting. A good example is harvesting of

grapevines according to their varieties. The basic prerequisite for such a selective harvester is correct recognition of grapevine varieties.

Recognition of grapevine varieties can be carried out different ways. A traditional recognition method is ampelometry [5]. As the method is visual, it is non-destructive. However, it requires involvement of an expert with extensive training, even when using a specialized software [23]. Accuracy of this method is strongly dependent on skills and experience of the expert. DNA analysis [17] is an example of a more objective approach. However, this method, as well as other wet chemistry techniques, is destructive, time-consuming, labour-demanding and it also requires the involvement of an expert.

Current development of computer vision methods and availability of advanced image sensors, has enabled automation of grapevine variety identification. Methods, which process data provided by a spectrometer [6, 4] or a hyperspectral camera [7], are automatic, non-destructive and time-efficient. Measurements of an interaction of electromagnetic radiation with matter, at many different spectral bands, allow accurate recognition of grapevine varieties. The main disadvantage of this approach is the high purchase price of a spectrometer or a hyperspectral camera. Implementation of such sensors into a selective harvester would significantly increase its price.

Traditional methods aimed at recognition of grapevine varieties are limited by human senses. For example, ampelometry uses eyesight for the grapevine variety recognition. Despite the fact that the human perceives only visible light, mostly in three bands, experts are capable of recognizing tens of varieties. Conventional cameras provide images of comparable attributes. We believe that the images keep information, that allows the classification of grapes according to their varieties, at the same level of accuracy.

An image-based classification of grapevines according to their varieties is a complex task which requires extraction of many discriminative features. An extensive diversity of an outdoor environment further increases the complexity of the feature extraction. The overall complexity of this task requires employment of a state-of-the-art image categorization system.

State-of-the-art image categorization systems are based on deep convolutional networks (deep ConvNets) [14]. Deep ConvNets allow creation of self-contained image categorization systems which ensure both feature extraction and classification of object images. Key factors influencing performance of such a system are a learning capacity of a deep ConvNet and the quality of a training set. The capacity of the network is given by its topology. Modern topologies control the capacity by varying width or depth of networks [13]. Enlarging a deep ConvNet capacity through increasing its width is used e.g. in GoogLeNet [27]. The second approach is to increase the number of network layers (the network depth), while retaining the data processing linearity. Topologies, such as Highway Networks [25], Residual Networks [9], Deep Pyramidal Residual Networks [8], Densely Connected Convolutional Networks (DenseNets) [11] and Cross-Layer Neurons Networks [34] can have tens to hundreds of layers.

Factors such as selection of training samples, their correct categorization, proportional representation of samples with respect to their categories, as well as the total number of samples in a training set predetermine its quality [15, 30]. In the case of deep ConvNets, the amount of training data is of great importance. Unfortunately, the relationship between classification performance and number of samples is logarithmic [26]. Collection and categorization of a sufficiently large set of images is thus often very time consuming, if not impossible [31, 19]. This issue can be overcome by means of data augmentation.

Data augmentation techniques extend datasets by generating synthetic samples. In image classification tasks, the synthetic samples originate as transformations of available images. The augmentation techniques that use randomly located crops, geometric transformations or photometric transformations are intuitive and easy to implement. The simplest example of synthetic samples are randomly located crops from original images [10, 28, 9] that are appropriate for images with objects of interest surrounded by a background. When approximate positions of objects are known, the transformation can be implemented as label-preserving (existing images are transformed such that their labels are preserved) [19]. Geometric transformation-based augmentation techniques include random reflections, random translations, random shearing and random rotations [32, 26, 13, 31]. When properly set up, they show high likelihood of preserving labels. Augmentation techniques based on photometric transformations adjust image lighting and colour, and leave the geometry unchanged [29, 10, 19]. However, for some tasks, colour is a very important distinctive feature, and the colour transformations may discard important information. Due to the putative loss of information, photometric transformations are not always label-preserving [19].

With respect to the aforementioned facts, we expect that automatic recognition of grapevine varieties can be built on RGB images which capture grape clusters and leaves. We propose a grapevine variety recognition system based on a DenseNet topology. A dense connectivity pattern used in DenseNets alleviates a vanishing-gradient problem, and allows creation of very deep networks with high learning capacity [11]. However, the main advantage of the DenseNet topology is proper classification performance of DesneNet based image categorization systems trained on small datasets [31]. For training and evaluation of the system, we form a dataset based on in-field photos captured under various lighting conditions. To ensure high classification accuracy of the system, we propose utilization of data augmentation within a training phase. We study the effect of the augmentation on the classification performance of the system.

2 Materials and Methods

2.1 Data Collection

We captured colour photos (individual images) of common grapevine during a harvest. The data collection was carried out within two days in the morning and in the afternoon in August 2015. We selected various locations in vineyards in Čejkovice, Czech Republic. We used no artificial lighting and we captured

the photos in a direction of sunshine likewise in the opposite direction (both days was partly sunny). The resulting collection of photos includes Welschriesling (WR), Saint Laurent (SL), Gewürztraminer (GT), Pinot noir (PN), Riesling Weiss (RW), Pinot gris (PG), and Veltliner Grün (VG) varieties (names of varieties according to Vitis International Variety Catalogue [16]).

We used camera bodies CANON EOS 1000D and CANON EOS 1100D with CANON ZOOM lenses EF-S 18-55 mm f/3.5-5.6 II and IS II, respectively. Resolutions of the photos are 1936×1288 pixel (px) and 4272×2848 px, respectively. The photos use RGB colour model with 24 bits bit depth. We placed the cameras perpendicular to vineyard rows (in terms of an axis of a lens), in a distance about 1.4 m from the rows, at an altitude of 1.25 m from the ground. A focal length varied between 18 mm and 24 mm.

2.2 Dataset

Photos acquired within the data collection include common grapevine plants with fruit and background. Around ten grape clusters are captured in each photo and they cover less than five percent of the image area. As we expect structures of grape clusters and leaves to be sufficiently discriminative features for the recognition of the varieties, we transform the photos into a dataset of RGB images, where at least 70 % of each image is covered by a grape cluster. The images are 120×120 px crops from the photos. For this purpose, we randomly select between 12 and 14 photos (depending on a density of grape clusters in photos) of each variety. From each photo, we create dozens of unique images (crops are carried out at different locations). For each variety, we create 900 images. Further, we create 900 images capturing a background (leaves without grapes, sky, grass, etc.), i.e. the final dataset consists of 7200 images classified into 8 categories (Fig. 1). In Table 1, we summarize information about the number of selected photos with respect to grapevine varieties (first column), camera bodies (first row) and focal lengths (second row).

Table 1: **Number of photos selected for forming of the dataset.** For each variety (first column), number of used photos is stated with respect to the focal length (second row) and the camera body (first row).

camera body	EOS 1000 D			EOS 1100 D	
focal length	18	21	24	18	23
Gewürztraminer	8	-	4	-	-
Veltliner Grün	8	4	-	-	-
Pinot gris	8	2	-	-	4
Pinot noir	8	-	4	-	-
Riesling Weiss	6	-	4	4	-
Saint Laurent	12	-	0	-	-
Welschriesling	10	-	4	-	-

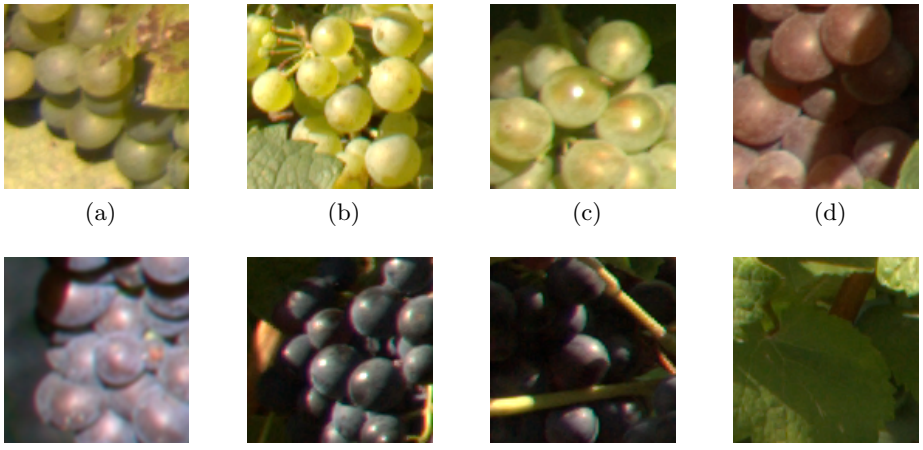

Fig. 1: **Example images for the categories in the dataset:** (a) Veltliner Grün (VG), (b) Riesling Weiss (RW), (c) Welschriesling (WR), (d) Gewürztraminer (GT), (e) Pinot gris (PG), (f) Pinot noir (PN), (g) Saint Laurent (SL), and (h) Background (BG).

2.3 Densely Connected Convolutional Networks

As in other deep ConvNets [14], convolutional, pooling and fully connected layers are arranged in a feed-forward manner to form a DenseNet. Regular patterns occurring in DenseNets allow us to simplify description of their topologies. Let us define two composite building elements which will be used to describe a topology of the presented variety recognition system: a *dense blocks* (DB) and a *transition layer* (TL).

Dense Block Let us consider a n-th DB of d_n layers that is built in a DenseNet of L layers. The input and the output of the n-th DB are placed at i_n-th and o_n-th levels of the network, respectively, i.e. $d_n = o_n - i_n + 1$. Feature maps produced at the ℓ-th level of the network, where $\ell \in [i_n, o_n]$, are given as

$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_{i_n-1}, \ldots, \mathbf{x}_{\ell-1}]), \tag{1}$$

where $H_\ell(\cdot)$ is a non-linear transformation performed at the ℓ-th level, $\mathbf{x}_{i_n-1}$ are feature maps at the input of the n-th DB, $\mathbf{x}_i$ for $i \in [i_n, \ell-1]$ are feature maps produced at i-th to $\ell-1$-th levels of the n-th DB, and $[\mathbf{x}_{i_n-1}, \ldots, \mathbf{x}_{\ell-1}]$ denotes concatenation of the feature maps.

Two variants of the non-linear transformation $H(\cdot)$ can be used in DBs: a basic and a bottleneck version [11]. The basic version is a composite function which consists of a batch normalization (BN) [12], a rectified linear unit (ReLU), and a convolution (Conv) [14], respectively. Using a short notation, the basic

version of $H(\cdot)$ can be written as BN-ReLU-Conv$(h \times w, f, s)$, where s is stride of convolutional filters, f is number of the filters, and h and w are their height and width, respectively. The bottleneck version of $H(\cdot)$ is defined as BN-ReLU-Conv$(1\times1, 4f, 1)$-BN-ReLU-Conv$(h\times w, f, s)$. If necessary, convolutions are zero-padded to keep the feature-map size fixed. For both versions of the composite function $H(\cdot)$, the parameters h, w, s, f are identical for all layers within a DB. We use abbreviations DBa and DBb for DBs with the basic and the bottleneck version of $H(\cdot)$, respectively.

Transition Layer Let us consider a TL connected at the output of the n-th DB (i.e. the TL is placed at the (o_n+1)-th level of the network). The (o_n+1)-th TL produces feature maps

$$\mathbf{x}_{o_n+1} = H_{o_n+1}([\mathbf{x}_{i_n-1}, \mathbf{x}_{i_n}, \ldots, \mathbf{x}_{o_n}]), \tag{2}$$

where $[\mathbf{x}_{i_n-1}, \mathbf{x}_{i_n}, \ldots, \mathbf{x}_{o_n}]$ denotes the concatenation of all feature maps that appear in the n-th DB. H_{o_n+1} is a composite function BN-ReLU-Conv$(1 \times 1, f, 1)$-AP$(2 \times 2, 2)$, where AP$(2 \times 2, 2)$ denotes an average pooling with pools 2×2 and stride 2 [11].

Compactness of the network is controlled by the number of the 1×1 convolutional filters f incorporated in TLs. The number of feature maps produced by the $(o_n + 1)$-th TL is given as $f_{o_n+1} = \lfloor \theta m_n \rfloor$, where θ is a compression factor, $\theta \in [m_n^{-1}, 1]$ and m_n is the number of feature maps produced by the n-th DB.

2.4 Variety Recognition System

Topology The presented variety recognition system is a DenseNet. The network classifies RGB images of dimensions 120×120 px according to varieties of grapes captured in the images. We control number of filters f in DBs by a variable k, where $k = 20$. The network is opened by one DBa which consists of one layer $(d = 1)$ with $2k$ convolutional filters $(f = 2k)$ with kernels of size 7×7 px $(h = w = 7)$, stride by 2 px $(s = 2)$. The following layer is a max pooling layer (MPL) with pools 3×3 px $(h = w = 3)$ stride by 2 $(s = 2)$. The inner parts of the network consist of two DBbs with 6 and 9 layers, respectively. At each layer of a DBb, k filters with kernels of size 3×3 px stride by 1 px ensure the feature extraction. Each DBb in the network is followed by one TL. We set up the compression factor θ at 0.5 for both TLs. The network is closed by a global average pooling (GAP) and a classifier, respectively. The classifier consists of one fully connected layer of eight neurons followed by a softmax function. The topology of the network is summarized in Table 2.

Training and evaluation To train and evaluate the system, we randomly split the dataset into training and validation subsets at the ratio 5:1 while keeping balanced distribution of categories, i.e. the training and the validation subsets consist of 750 and 150 samples of each category, respectively. On the training

Table 2: **Topology of the variety recognition system.** Building elements which form the system are listed with respect to their placement in the network in the first row of the table (the first block is the leftmost one), where DBa and DBb denote the basic and the bottleneck versions of the dense block; MPL is a max pooling layer; TL is a transition layer, GAP denotes a global average pooling, and C is used for a classifier that consists of one fully connected layer followed by a softmax function. The parameters h and w are a height and a width of a filter kernel or of a pool; s is stride of the kernel or of the pool; f is the number of filters at one convolution in a dense block; and d is the number of layers in the dense block.

	DBa	MPL	DBb	TL	DBb	TL	GAP	C
h	7	3	3	-	3	-	5	-
w	7	3	3	-	3	-	5	-
s	2	2	1	-	1	-	5	-
f	$2k$	-	k	-	k	-	-	-
d	1	-	6	-	9	-	-	-

subset, we train the system using the ADAM optimizer [18] for 500 epochs with mini batches of 400 samples, minimizing a cross-entropy error function for the multiclass classification problem [2]. We set up a learning rate, and an exponential decay rate for first and second moment estimates at 10^{-3}, 0.95 and 0.999, respectively. We shuffle samples in the training subset every epoch.

We evaluate classification performance of the trained system on the validation subset. We observe numbers of correctly and incorrectly classified samples. For each category, we calculate classification accuracy of the system. For the i-th category, it is given as

$$\mathrm{acc}_i = \frac{|\mathrm{TP}_i| + |\mathrm{TN}_i|}{n}, \tag{3}$$

where $|\mathrm{TP}_i|$ is the number of correctly classified samples of the i-th category, $|\mathrm{TN}_i|$ is the total number of correctly classified samples of complementary categories to the category i, and n is the total number of samples in the validation subset. To evaluate performance of the system on a subset of categories I, we calculate average per-class accuracy according to

$$\overline{\mathrm{acc}} = \frac{\sum_{i \in I} \mathrm{acc}_i}{|I|}, \tag{4}$$

where $|I|$ is the number of categories in the subset I [22].

As the number of samples in the training subset might be insufficient, we consider training on synthetic samples to be a possible way to improve classification performance of the system. To assess importance of the augmentation, we train the system both with and without utilization of data augmentation. Due to a stochastic character of the training process [20], we repeat the training-validation process thirty times for each training approach. We carry out the experiments in MATLAB R2020b using Deep Learning Toolbox.

Data Augmentation In computer vision, the commonly used data augmentation techniques are randomly located crops, techniques based on photometric transformations, and techniques based on geometric transformations. In our case, the augmentation by randomly located crops cannot be used due to the dimensions of images in our dataset (dimensions of the images match exactly with expected dimensions of input data). As the colour of grapes is an important distinctive feature in the recognition of grapevine varieties, photometric transformations may be counterproductive. When properly adjusted, the geometric transformations such as reflections, translations, shearing, and rotations can ensure label-preserving data augmentation. To ensure a high likelihood of preserving labels, selection of the transformations as well as their settings must be based on an analysis of the dataset.

The images in the dataset capture the background, or grapes and leaves of one of the seven recognized varieties. Leaves in the images with grapes cover only a small part of image areas and they are always located near to image edges. As the leaves can be an important discriminative feature, we consider translations up to ± 3 px in both directions to be label-preserving. The grapes in clusters are arranged to form a typical triangular shape approximately symmetric with respect to the y-axis. With respect to this fact, we consider reflections over the y-axis to be label-preserving while reflections over the x-axis not. Further, we estimate rotations and shears up to ± 20 degree to be label-preserving.

With respect to the limit values, we use for the augmentation of the training subset, random horizontal and vertical translations (range of a translation distance: ± 3 px), random y-axis reflections, random rotations (range of a rotation angle: ± 20 degree), and random horizontal and vertical shears (range of a shear angle: ± 20 degree). Probability of the reflections is 50 %. The translation distance, the rotation and the shear angles are picked randomly from a continuous uniform distribution within the specified intervals. We implement the data augmentation via a MATLAB function `imageDataAugmenter`.

3 Results and Discussion

We obtain 30 sets of results for each performance measure and each training paradigm. To provide a synoptic summary of the results, we calculate arithmetic mean and standard deviation for each measure. As the system recognizes eight categories, we obtain 8 and 8×7 statistics for correctly and incorrectly classified samples, respectively. We summarize the statistics for the training with and without the augmentation in Table 3 and Table 4, respectively. We arrange the data in a confusion matrix manner, i.e. rows and columns represent instances in actual and predicted categories, respectively. In the last columns of the tables, we present statistics of classification accuracies for each category (3).

The overall ($I = \{\mathrm{GT}, \mathrm{VG}, \mathrm{PG}, \mathrm{PN}, \mathrm{BG}, \mathrm{RW}, \mathrm{SL}, \mathrm{WR}\}$) average per-class accuracy (4) of the system trained with the augmentation (Table 3) is at 98.00 ± 0.13 %. When considering only the red ($I = \{\mathrm{GT}, \mathrm{PG}, \mathrm{PN}, \mathrm{SL}\}$) and the green grape varieties ($I = \{\mathrm{VG}, \mathrm{RW}, \mathrm{WR}\}$), we get average per-class accuracies at

Table 3: **Evaluation results for the system trained with the data augmentation.** The table shows arithmetic means and standard deviations of the performance measures. The statistics are organized in the confusion matrix manner. Rows and columns represent instances in actual and predicted categories, respectively. Average per-class accuracies of the categories are summarized in the last column. Distinguished categories are Gewürztraminer (GT), Veltliner Grün (VG), Pinot gris (PG), Pinot noir (PN), background (BG), Riesling Weiss (RW), Saint Laurent (SL), and Welschriesling (WR).

i	GT	VG	PG	PN	BG	RW	SL	WR	acc_i
GT	145.63 ± 3.21	0.10 ± 0.31	2.50 ± 2.56	0.80 ± 1.00	0.53 ± 0.90	0.10 ± 0.31	0.23 ± 0.63	0.10 ± 0.31	0.99 ± 0.00
VG	0.03 ± 0.18	137.57 ± 5.81	0.00 ± 0.00	0.00 ± 0.00	0.57 ± 0.73	4.60 ± 3.46	0.00 ± 0.00	7.23 ± 3.22	0.98 ± 0.00
PG	2.33 ± 1.83	0.00 ± 0.00	141.23 ± 3.37	3.30 ± 2.01	0.10 ± 0.40	0.07 ± 0.25	2.97 ± 1.83	0.00 ± 0.00	0.98 ± 0.00
PN	1.23 ± 1.36	0.07 ± 0.25	3.63 ± 1.90	128.20 ± 6.69	0.57 ± 0.63	0.00 ± 0.00	16.30 ± 5.71	0.00 ± 0.00	0.97 ± 0.00
BG	0.83 ± 0.95	1.10 ± 0.99	1.70 ± 1.78	1.13 ± 1.17	141.63 ± 3.61	1.70 ± 1.37	1.27 ± 1.46	0.63 ± 1.03	0.99 ± 0.00
RW	0.00 ± 0.00	6.50 ± 2.58	0.00 ± 0.00	0.00 ± 0.00	2.17 ± 1.76	136.00 ± 3.81	0.00 ± 0.00	5.33 ± 2.47	0.98 ± 0.00
SL	0.37 ± 0.61	0.00 ± 0.00	2.37 ± 1.27	11.80 ± 3.46	0.63 ± 0.72	0.10 ± 0.40	134.73 ± 4.06	0.00 ± 0.00	0.97 ± 0.01
WR	0.23 ± 0.57	9.33 ± 5.01	0.00 ± 0.00	0.00 ± 0.00	0.37 ± 0.72	6.07 ± 3.12	0.03 ± 0.18	133.97 ± 6.37	0.98 ± 0.00

Table 4: **Evaluation results for the system trained without the data augmentation.**

i	GT	VG	PG	PN	BG	RW	SL	WR	acc_i
GT	137.20 ± 4.69	0.17 ± 0.38	5.70 ± 3.51	2.43 ± 2.24	2.23 ± 1.63	0.07 ± 0.25	1.73 ± 1.31	0.47 ± 0.78	0.98 ± 0.00
VG	0.03 ± 0.18	123.27 ± 8.79	0.23 ± 0.56	0.07 ± 0.25	1.63 ± 1.19	7.47 ± 3.42	0.20 ± 0.41	17.10 ± 6.29	0.96 ± 0.01
PG	5.47 ± 2.75	0.23 ± 0.43	121.67 ± 6.58	9.37 ± 4.44	2.27 ± 1.80	0.17 ± 0.46	10.10 ± 4.26	0.73 ± 1.11	0.95 ± 0.01
PN	3.23 ± 2.14	0.17 ± 0.46	9.40 ± 2.71	108.33 ± 7.61	2.43 ± 2.14	0.10 ± 0.31	26.00 ± 7.74	0.33 ± 0.55	0.93 ± 0.01
BG	1.90 ± 1.56	2.97 ± 1.83	1.87 ± 1.53	2.43 ± 1.61	132.67 ± 5.79	3.97 ± 3.30	2.33 ± 1.88	1.87 ± 1.17	0.97 ± 0.01
RW	0.10 ± 0.40	10.23 ± 4.25	0.20 ± 0.41	0.10 ± 0.31	3.00 ± 2.05	126.97 ± 6.52	0.10 ± 0.31	9.30 ± 4.39	0.96 ± 0.01
SL	1.93 ± 1.55	0.27 ± 0.45	9.40 ± 4.28	23.13 ± 6.39	2.93 ± 1.78	0.13 ± 0.35	111.47 ± 9.26	0.73 ± 1.08	0.93 ± 0.01
WR	0.13 ± 0.35	12.20 ± 4.50	0.83 ± 1.12	0.40 ± 0.56	1.50 ± 1.17	9.23 ± 3.87	0.27 ± 0.53	125.43 ± 5.92	0.95 ± 0.01

97.75 ± 0.25 % and 98.00 ± 0.00 %, respectively. The system accurately distinguishes between grapes and background (8.37 ± 3.61 from 150 samples of background miss classified as grapes, and 4.93 ± 2.72 from 1050 samples of grapes miss classified as background). We also observe that confusions occur almost exclusively among grapes of the same colour (0.53 ± 0.68 samples of red grape varieties classified as a green variety, and 0.30 ± 0.70 samples of green grape varieties classified as a red variety). The system best recognizes Gewürztraminer (145.63 ± 3.21 from 150 samples of Gewürztraminer classified correctly, and only 5.03 ± 2.92 from 1050 samples of other categories miss classified as Gewürztraminer). With only 128.20 ± 6.69 correctly classified samples from 150, we consider Pinot noir to be the most difficult variety for the system trained with the data augmentation. False positive classification of this variety is 17.03 ± 4.73 from 1050 samples of other categories.

For the system trained without the data augmentation (Table 4), we obtain overall average per-class accuracy at 95.38 ± 0.88 %. For the red and the green grape varieties, we get average per-class accuracies at 94.75 ± 0.75 % and 95.67 ± 1.00 %, respectively. We can say that the system, trained without the data augmentation, relatively accurately distinguishes between grapes and background (17.33 ± 5.79 from 150 samples of background miss classified as grapes, and 16.00 ± 5.51 from 1050 samples of grapes miss classified as background). Also, in this case, confusions occur almost exclusively among grapes of the same colour (3.57 ± 2.86 samples of red grape varieties classified as a green variety, and

2.67 ± 1.56 samples of green grape varieties classified as a red one). The system trained without the augmentation best recognizes Gewürztraminer (137.20 ± 4.69 samples from 150 classified correctly, and 12.80 ± 4.66 from 1050 samples of other categories miss classified as Gewürztraminer). The most difficult variety is again Pinot noir (108.33 ± 7.61 samples from 150 classified correctly, and 37.93 ± 9.83 from 1050 samples of other categories miss classified as Pinot noir).

All the results presented so far indicate that training of the system with the data augmentation, results in better classification performance. The significance of the augmentation is best apparent when comparing the numbers of correctly classified samples obtained for the system trained with and without the data augmentation. Let the absolute difference between the numbers of correctly classified samples is for the i-th category given as

$$\Delta|\mathrm{TP}_i| = |\mathrm{TP}_i^+| - |\mathrm{TP}_i^-|, \tag{5}$$

where $|\mathrm{TP}_i^+|$ and $|\mathrm{TP}_i^-|$ are numbers of correctly classified samples of the i-th category for the system trained with and without the data augmentation, respectively. The relative difference is for the i-th category given as

$$\delta_i = \frac{\Delta|\mathrm{TP}_i|}{|\mathrm{P}_i|}, \tag{6}$$

where $|\mathrm{P}_i|$ is the number of samples of the i-th category in the training subset. In Table 5, we state for each category (first row) arithmetic means and standard deviations of the absolute (5) and relative (6) differences (second and third row, respectively).

Table 5: **Differences between the numbers of correctly classified samples resulting from utilization of different training paradigm**. The absolute (second row) and the relative (third row) differences between correct classifications of the system trained with and without the augmentation with respect to the categories (first row).

i	GT	VG	PG	PN	BG	RW	SL	WR
$\Delta\|\mathrm{TP}_i\|$	8.43 ± 6.22	14.30 ± 10.13	19.57 ± 7.79	19.87 ± 11.39	8.97 ± 7.39	9.03 ± 7.10	23.26 ± 10.19	8.53 ± 9.07
δ_i	5.62 ± 4.15	9.53 ± 6.75	13.04 ± 5.19	13.24 ± 7.59	5.98 ± 4.93	6.02 ± 4.74	15.51 ± 6.80	5.69 ± 6.04

As shown by the results in Table 5, the data augmentation improves the correct classification for all categories by up to 15.51 ± 6.80 % (Saint Lauren). Thanks to the data augmentation, the average number of correctly classified samples (the main diagonal of the confusion matrix in Table 3) is, for some categories, close to the maximum (150 samples per category). Considering all these facts, we conclude that precise automatic recognition of grapevine varieties based on RGB images is possible; however, training with the data augmentation is essential to develop an accurate grapevine variety recognition system.

4 Conclusion

We show that in-field colour images of ripe grapes acquired by a conventional camera can be used for classification of grapevines according to their varieties. The presented variety recognition system is capable of distinguishing between seven grapevine varieties, where four and three varieties have red and green grapes, respectively. The system also accurately differentiates grapes from their background. When trained with the data augmentation, the overall average per-class accuracy of the system is at 98.00 ± 0.13 % on images captured without any artificial lighting. Considering all these facts, we conclude that the proposed solution allows construction of affordable automatic selective harvesters.

Acknowledgments

The work was supported from ERDF/ESF "Cooperation in Applied Research between the University of Pardubice and companies, in the Field of Positioning, Detection and Simulation Technology for Transport Systems (PosiTrans)" (No. CZ.02.1.01/0.0/0.0/17_049/0008394). We would like to thank Petr Junek for creating the dataset.

References

1. Bac, C.W., Hemming, J., van Tuijl, B., Barth, R., Wais, E., van Henten, E.J.: Performance evaluation of a harvesting robot for sweet pepper. Journal of Field Robotics **34**(6), 1123–1139 (2017). https://doi.org/10.1002/rob.21709, https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21709
2. Bishop, C.: Pattern Recognition and Machine Learning. Springer-Verlag New York, 1 edn. (2006)
3. Bontsema, J., Hemming, J., Pekkeriet, E., Saeys, W., Edan, Y., Shapiro, A., Hočevar, M., Oberti, R., Armada, M., Ulbrich, H., Baur, J., Debilde, B., Best, S., Evain, S., Gauchel, W., Hellström, T., Ringdahl, O.: CROPS: clever robots for crops. Engineering & Technology Reference **1**(1) (2015). https://doi.org/10.1049/etr.2015.0015
4. Fernandes, A., Utkin, A., Eiras-Dias, J., Silvestre, J., Cunha, J., Melo-Pinto, P.: Assessment of grapevine variety discrimination using stem hyperspectral data and adaboost of random weight neural networks. Applied Soft Computing **72**, 140 – 155 (2018). https://doi.org/https://doi.org/10.1016/j.asoc.2018.07.059
5. Galet, P.: A Practical Ampelography: Grapevine Identification. Comstock Pub. Associates, Ithaca, N.Y, 1 edn. (1979)
6. Gutiérrez, S., Tardaguila, J., Fernández-Novales, J., Diago, M.: Data mining and nir spectroscopy in viticulture: Applications for plant phenotyping under field conditions. Sensors (Switzerland) **16**(2) (2016). https://doi.org/10.3390/s16020236
7. Gutiérrez, S., Fernández-Novales, J., Diago, M.P., Tardaguila, J.: On-the-go hyperspectral imaging under field conditions and machine learning for the classification of grapevine varieties. Frontiers in Plant Science **9**, 1102 (2018). https://doi.org/10.3389/fpls.2018.01102

8. Han, D., Kim, J., Kim, J.: Deep pyramidal residual networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6307–6315 (July 2017). https://doi.org/10.1109/CVPR.2017.668
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016). https://doi.org/10.1109/CVPR.2016.90
10. Howard, A.: Some improvements on deep convolutional neural network based image classification. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2014)
11. Huang, G., Liu, Z., v. d. Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (July 2017). https://doi.org/10.1109/CVPR.2017.243
12. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015. Proceedings of Machine Learning Research, vol. 37, pp. 448–456. PMLR (2015)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84 – 90 (2017)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (May 2015). https://doi.org/10.1038/nature14539
15. Lemnaru, C., Potolea, R.: Imbalanced classification problems: Systematic study, issues and best practices. In: Enterprise Information Systems. pp. 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
16. Maul et al.: Vitis international variety catalogue (2020), www.vivc.de
17. Pelsy, F., Hocquigny, S., Moncada, X., Barbeau, G., Forget, D., Hinrichsen, P., Merdinoglu, D.: An extensive study of the genetic diversity within seven french wine grape variety collections. Theoretical and Applied Genetics **120**(6), 1219–1231 (2010). https://doi.org/10.1007/s00122-009-1250-8
18. Ruder, S.: An overview of gradient descent optimization algorithms. CoRR **abs/1609.04747** (2016)
19. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 60 (Jul 2019). https://doi.org/10.1186/s40537-019-0197-0, https://doi.org/10.1186/s40537-019-0197-0
20. Skrabanek, P., Dolezel, P., Nemec, Z., Stursa, D.: Person detection for an orthogonally placed monocular camera. Journal of Advanced Transportation **2020**, 1–13 (2020). https://doi.org/10.1155/2020/8843113
21. Slaughter, D., Giles, D., Downey, D.: Autonomous robotic weed control systems: A review. Computers and Electronics in Agriculture **61**(1), 63 – 78 (2008). https://doi.org/10.1016/j.compag.2007.05.008
22. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427 – 437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002
23. Soldavini, C., Schneider, A., Stefanini, M., Dallaserra, M., Policarpo, M.: Super ampelo, a software for ampelometric and ampelographic descriptions in vitis. Acta Horticulturae **827**, 253–258 (2009). https://doi.org/10.17660/ActaHortic.2009.827.43
24. de Soto, M.G., Emmi, L., Perez-Ruiz, M., Aguera, J., de Santos, P.G.: Autonomous systems for precise spraying – evaluation of a robotised patch sprayer. Biosystems Engineering **146**, 165 – 182 (2016). https://doi.org/10.1016/j.biosystemseng.2015.12.018

25. Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. pp. 2377–2385. NIPS'15, MIT Press, Cambridge, MA, USA (2015)
26. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 843–852 (2017). https://doi.org/10.1109/ICCV.2017.97
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (June 2016). https://doi.org/10.1109/CVPR.2016.308
28. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
29. Taylor, L., Nitschke, G.: Improving deep learning with generic data augmentation. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1542–1547 (2018). https://doi.org/10.1109/SSCI.2018.8628742
30. Škrabánek, P., Doležel, P.: On reporting performance of binary classifiers. Scientific Papers of the University of Pardubice, Series D **XXIV**, 181–192 (2017)
31. Škrabánek, P., Zahradníková, jr., A.: Automatic assessment of the cardiomyocyte development stages from confocal microscopy images using deep convolutional networks. PLOS ONE **14**(5), 1–18 (05 2019). https://doi.org/10.1371/journal.pone.0216720, https://doi.org/10.1371/journal.pone.0216720
32. Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D.: Understanding data augmentation for classification: When to warp? In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–6 (2016). https://doi.org/10.1109/DICTA.2016.7797091
33. Xiong, Y., Peng, C., Grimstad, L., From, P.J., Isler, V.: Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. Computers and Electronics in Agriculture **157**, 392 – 402 (2019). https://doi.org/https://doi.org/10.1016/j.compag.2019.01.009, http://www.sciencedirect.com/science/article/pii/S0168169918312456
34. Yu, Z., Li, T., Luo, G., Fujita, H., Yu, N., Pan, Y.: Convolutional networks with cross-layer neurons for image recognition. Information Sciences **433-434**, 241 – 254 (2018). https://doi.org/https://doi.org/10.1016/j.ins.2017.12.045