Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 1299 – 1316
© TÜBİTAK
doi:10.55730/1300-0632.3850

Classification and phenological staging of crops from in situ image sequences by deep learning

Uluğ BAYAZIT[1,*], **Turgay ALTILAR**[1], **Nilgün GÜLER BAYAZIT**[2]
[1]Computer Engineering Department, Faculty of Computer and Informatics Engineering,
İstanbul Technical University, İstanbul, Turkey
[2]Mathematical Engineering Department, Faculty of Chemical and Metallurgical Engineering,
Yıldız Technical University, İstanbul, Turkey

Received: 15.10.2021 • **Accepted/Published Online:** 15.04.2022 • **Final Version:** 31.05.2022

Abstract: Accurate knowledge of crop type information is not only valuable for verifying the declaration of farmers to obtain subsidy or insurance for the grown crop, but also for generating crop type maps that serve a variety of purposes in land monitoring and policy. On the other hand, accurate knowledge of crop phenological stage can help farm personnel apply fertilization and irrigation regimes on a timely basis. Although deep learning based networks have been applied in the past to classify the type and predict the phenological stage of crops from in situ images of fields, more advanced deep learning based networks, that learn and make such inferences from temporal windows of sequences of field images taken by cameras at stationary coordinates and looking directions, have not been reported to date. This work proposes a conceivable architecture for learning and making inferences from such data. Specifically, the feature vectors of the images in a temporal window of the image sequence for a crop cycle are extracted by a first stage deep convolutional neural network and their temporal dependencies are exploited by a second stage recurrent neural network. Experiments on a dataset of image sequences from 63 fields of 5 different types of crops reveal that the proposed system can achieve over 80% accuracy in crop type classification and under 0.5 mean absolute error in phenological stage number estimation. The learning performances improve with the size of the temporal window and the fine-tuning of the deep convolutional neural network used for feature extraction. The performances achieved with the proposed system are superior to those obtained by applying classical machine learning methods to handcrafted texture and color features.

Key words: Image sequence, crop type, classification, phenology, prediction, convolutional neural network, recurrent neural network

1. Introduction

The advent of Industry 4.0 has significantly increased production quantity and quality in agronomy due to advanced decision-making by interconnected computerized systems on vast amounts of data. Classification and prediction type problems in agronomy have mainly benefited from the recent advances in deep learning.

In deep learning based computer vision, the recently popularized convolutional neural network (CNN) has effectively reduced the number of trainable parameters so that the trained system generalizes well to the visual data outside of the training set. This aspect along with the widespread availability of low cost parallel processing hardware (graphical processing units and clusters of computing nodes) and open source software development frameworks have facilitated the widespread application of deep learning based computer vision to

[*]Correspondence: ulugbayazit@itu.edu.tr

many research areas, including agronomy and field crops.

In agronomy and field crops research, computer vision problems of crop/weed detection, fruit counting, plant recognition, land cover classification, disease detection, phenotype/genotype classification, yield prediction, crop type clas6sification, phenological stage recognition, health and/or quality monitoring have been predominantly addressed by employing CNN, and, to a lesser extent, long short-term memory (LSTM), as the deep learning network architectures.

Accurate classification of crop type and prediction of phenological stage by an intelligent system facilitates the field imagery to be annotated, archived and associated with other types of data (satellite, atmospheric measurement) without human supervision. Verification of farmer's declaration of the planted crop type by correct and automatic identification of crop type makes it possible for the state to give correct subsidy, or for the insurance firms to offer the right insurance against adverse atmospheric conditions, flooding, wildfires, landslides or damage by birds or boars. Mosaicing the determined crop type information can facilitate the generation of crop type maps over a farmland which serve various purposes in land monitoring and policy, including, but not limited to, the improvement of yield models, the development of hydrological models and the modeling of flood damage estimation.

On the other hand, automatic and correct identification of phenological stage of a crop helps the farmers to timely administer irrigation, fertilization (split application of nitrogen), disinfection regimes to maximize yield. Timely grazing livestock to benefit them with quality nutrients also depends on accurate knowledge of phenological stages. Prior works apply deep learning to plant/crop classification and crop phenological stage prediction by processing mostly still images, or sequences of aerial/satellite images.

The work of [14] applies a CNN to learn unsupervised features for 44 plant species. Similarly, but more expansively, the work of [20] applies transfer learning for classifying 1000 species of trees, herbs, ferns using a pretrained CNN. [26] employs a CNN having three convolutional layers to classify images of field image sequences of 16 plant species obtained from the Agricultural Monitoring and Information System (TARBIL) in Turkey and reports an accuracy of 97.47%. For plant phenotyping, root and shoot feature identification and trait loci discovery in root architecture datasets have been performed by deep CNNs in [17]. Three different legumes are differentiated with CNNs by using leaf vein patterns in [9]. [22] demonstrates the performance of a deep CNN network for learning 100 ornamental plant species on an image dataset collected by a mobile phone in natural scenery. Application of deep residual CNNs to large scale plant identification is addressed in [3, 10].

In crop type identification from aerial imagery, [19] combines local histograms with convolutional units to improve performance. By employing recurrent neural networks (RNN), [21] performs the same task on satellite image sequences. High resolution pixelwise classification of oil radish, barley, weed, stump, soil, equipment and other types is performed by adding a deconvolution layer to a VGG-16 based deep network by [15]. Similarly, for crop classification, [4] shows the superiority of CNNs and autoencoders to a baseline image stacking approach on multitemporal optical and synthetic aperture radar image sequences. For crop classification on multispectral multitemporal remote sensing data, [12] shows the superiority of 3D CNN to 2D CNN and other conventional methods.

In [25], the crop phenology recognition problem is approached by applying a 2D CNN on image data (both [25] and the current work use the field image data provided by TARBIL). In [5], the local extrema in the rate of change in curvature of generalized sigmoid based model of green chromatic coordinate vegetation index are employed along with 2D CNN, 3D fully convolutional network and Siamese network in order to predict the phenological transition dates of vegetation in the PhenoCam image dataset. Recently, [8] has applied a CNN to single images of citrus crop to predict its phenological stage for aiding deficit irrigation techniques and [18] has applied CNNs on 138,000 images to identify the growth stages of wheat and barley up to stem elongation.

The works of [6, 24] determine the onset date of the heading stage for rice and wheat, respectively, by first employing the ResNet-50 network to detect flowering regions and regions with spikes and then taking the date of the image for which 50% of the regions are labelled as positive. A logistic function is fitted to the count data in [24].

The primary contribution of the current work, that sets it apart from the above works, is the application of recurrent networks to in situ images captured in a time window. This way, the temporal dependencies among feature vectors extracted from these consecutive images can be exploited to achieve a learning performance that exceeds the performance obtained by deep learning on individual images. Even though the current work addresses the use of stationary cameras for capturing field imagery over entire crop cycles, the techniques developed here can be readily extended to sequences of images collected at precise GPS coordinates and looking directions by mobile cameras (like those mounted on a UAV) in a more practical application scenario.

In order to learn the type or the phenological stage of the crop, the current approach employs CNNs to extract feature vectors from images. Rather than exploiting the knowledge imparted by the feature vector of a single image, the proposed approach exploits the joint knowledge of the feature vectors of the currently evaluated image and its most recent precedents in a temporal window. Use of multiple consecutive feature vectors (images) makes the learning method more robust against bad weather conditions or shaky cameras that result in occasional hazy or blurry shots, or against the appearance of foreign objects or unwanted impressions on the field surface.

Additionally, the phenological stage prediction process for an image can use the knowledge of the prior images to its advantage. For example, in the early leaf development stages, past images contain less green component than the past images in the leaf development stage. Therefore, in this work, an LSTM network is employed to exploit temporal dependencies among feature vectors extracted from individual images in the causal temporal window and the relation between the temporal window size and the learning performance is analyzed. Similar LSTM-CNN architectures have been investigated for video classification and human action/activity recognition in [1, 7, 16], and more relevantly, for phenotyping from plant image sequences [23], but have not been applied to type and phenological stage learning problems in agricultural computer vision to date.

Since crop plant bodies can exhibit irregular displacements due to wind or occlusions due to self-weight between consecutive images, the temporal correlation among corresponding pixels in consecutive images is rather small. Consequently, in order to exploit the temporal dependencies of successive images, applying an LSTM network on the global features extracted from them by a CNN network is preferred to directly feeding the images to a ConvLSTM network.

Secondarily, the current work shows that the performance of multiple LSTM units operating on multiple feature vectors of multiple dimensionalities (both low and high) can exceed the performance of a single LSTM unit operating on a single, high dimensional feature vector. It also shows that smoothing stage prediction results can further improve performance.

Fine-tuning of the CNN network used to extract features from images and preprocessing of the images by contrast stretching are the other design issues studied in the current work. Finally, the proposed deep learning approaches are compared against classical machine learning methods operating on handcrafted color and texture features.

2. Materials and methods

2.1. The dataset

The ITUTAR-FIS field image sequence dataset consists of 63 sequences (each covering one crop cycle) of 6216 images, that were acquired from TARBIL and captured by 63 ground stations cameras for six different crop types (8 sunflower, 19 winter wheat, 11 barley, 9 cotton, 1 chickpeas, 15 maize) between 2015 and 2017. Figure 1 shows a segment of a sequence of barley field images. For each crop type, the statistics of the stage lengths (in days) and the number of images in each stage are presented in Tables 1 and 2, respectively.

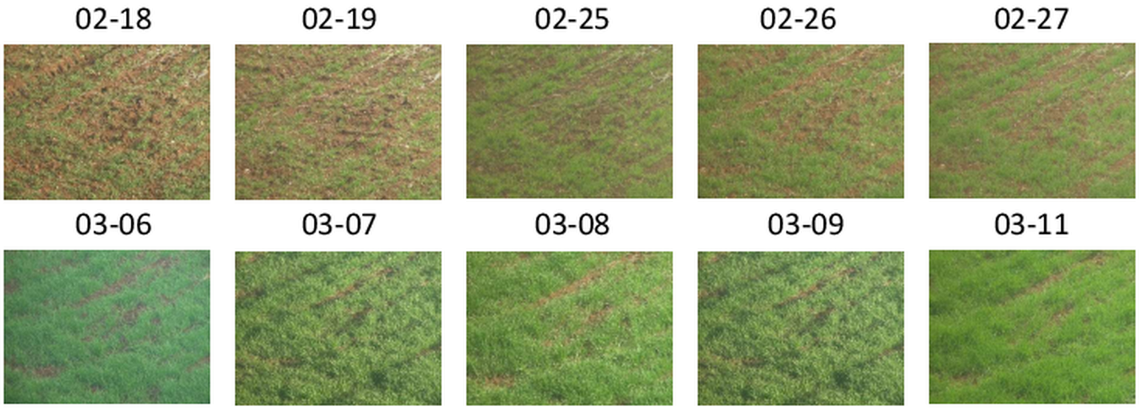

Figure 1. A number of consecutive images of barley collected in 2016 at the station labelled as Adıyaman 02.03.

Table 1. For each crop type, the average (and the standard deviation) of the number of days of each stage.

Stage	Barley	Sunflower	Wheat	Maize	Cotton	Chickpeas
0	21.36 (8.5)	14.38 (9.1)	29.37 (18.9)	11.00 (2.2)	8.33 (1.6)	34.00 (0.)
1	29.18 (11.2)	23.38 (4.7)	25.42 (17.6)	26.47 (4.4)	15.67 (2.2)	21.00 (0.)
2	51.82 (26.8)	21.38 (3.1)	33.47 (11.)	21.47 (3.1)	24.22 (3.6)	38.00 (0.)
3	37.09 (9.9)	14.75 (3.3)	33.16 (10.1)	17.87 (6.5)	22.56 (3.8)	21.00 (0.)
4	31.18 (7.8)	26.38 (7.5)	34.63 (10.4)	10.33 (2.6)	19.78 (7.4)	17.00 (0.)
5	15.73 (3.4)	34.38 (13.1)	19.74 (7.5)	17.40 (6.0)	32.11 (10.4)	24.00 (0.)
6	16.73 (3.7)		20.11 (6.3)	35.27 (20.7)	44.11 (11.9)	
7	19.55 (3.8)		23.79 (11.5)			

Table 2. For each crop type, the average (and the standard deviation) of the number of images of each stage.

Stage	Barley	Sunflower	Wheat	Maize	Cotton	Chickpeas
0	9.18 (5.5)	7.88 (6.4)	7.58 (4.3)	4.40 (2.2)	3.22 (1.6)	23.0 (0.)
1	9.64 (2.9)	15.75 (3.3)	8.47 (2.6)	14.73 (4.6)	7.89 (2.9)	17.0 (0.)
2	14.27 (4.7)	15.13 (4.3)	13.32 (5.9)	15.60 (3.3)	13.00 (3.5)	23.0 (0.)
3	15.45 (4.0)	11.38 (3.5)	13.84 (4.3)	14.27 (5.7)	13.11 (5.1)	7.0 (0.)
4	15.82 (4.1)	20.13 (5.9)	16.74 (6.9)	7.87 (2.8)	11.44 (4.8)	15.0 (0.)
5	10.55 (2.5)	28.63 (11.4)	11.63 (5.9)	11.87 (5.1)	20.44 (6.5)	14.0 (0.)
6	11.09 (3.5)	1.13 (0.4)	11.32 (4.4)	24.33 (11.7)	29.89 (11.3)	1.0 (0.)
7	12.91 (6.1)		15.95 (6.0)	1.33 (1.3)	1.00 (0.)	
8	1.09 (0.3)		1.16 (1.2)			

The associated metadata is a list of the phenological stage onset dates of each station-cycle. The onset dates have been determined by field crop experts of TARBIL who have viewed the image sequences.

Table 3 presents the phenological stages of the six crops used in the current study. Figure 2 shows images of maize captured at onset dates of its 8 phenological stages during the crop cycle of one station. Integers are assigned as phenological stage labels (from 0 at sowing to maximum value at harvest) to images at stage onset dates in each sequence. The images having dates between two onset dates are assigned real labels that are linearly interpolated from the integer labels of these onset dates. Figure 3 depicts the resulting piecewise linear curve for the crop cycle of Figure 2. In order to properly apply linear interpolation, the dates of all images are converted to integer values that represent the number of days counted from January 1st of the sowing year.

Table 3. Phenological stage names and their BBCH scale [13] ranges.

Stage	Barley	Sunflower	Wheat	Maize	Chickpeas	Cotton
0	Seeding 00-07	Seeding 00-08	Seeding 00-07	Seeding 00-07	Seeding 00-08	Seeding 00-08
1	Emergence 09-10	Emergence 09-10	Emergence 09-10	Emergence 09-10	Emergence 09-10	Emergence 09-10
2	Leaf development 11-20	Leaf development 11-39	Leaf development 11-20	Leaf development 11-20	Leaf development 11-49	True leaves - booting 11-49
3	Tillering 21-29	Bud formation 51-59	Tillering 21-29	Cob development 30-39	Flowering 50-69	Bud formation 51-59
4	Stem elogation 30-39	Flower opening 61-69	Stem elogation 30-39	Tasseling, silking 51-69	Pod filling 70-79	Flower opening 60-69
5	Booting, ear emergence 41-59	Seed development, ripening 71-97	Booting, heading 41-59	Kernel development 71-79	Maturity 80-97	Yield formation 70-79
6	Flowering 61-69	Harvest 99	Flowering 61-69	Grain filling, maturity 83-97	Harvest 99	Ripening 80-97
7	Grain filling, ripening 71-97		Grain filling, ripening 71-97	Harvest 99		Harvest 99
8	Harvest 99		Harvest 99			

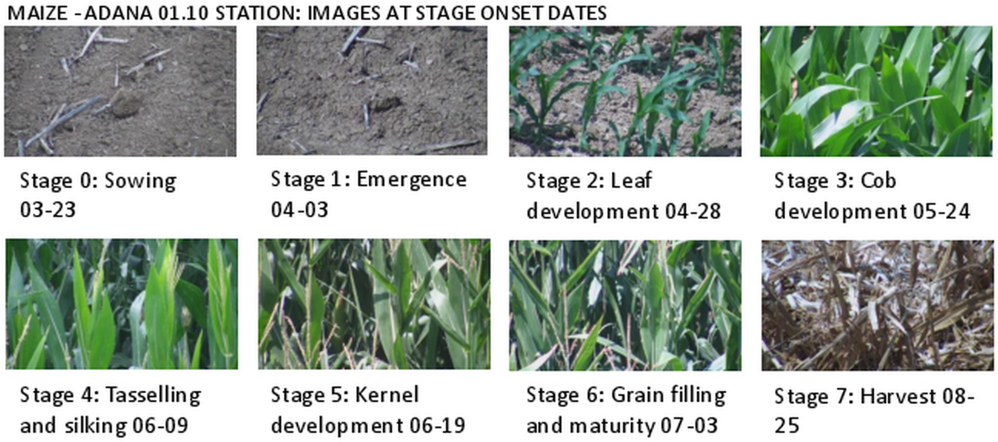
MAIZE -ADANA 01.10 STATION: IMAGES AT STAGE ONSET DATES
Stage 0: Sowing 03-23
Stage 1: Emergence 04-03
Stage 2: Leaf development 04-28
Stage 3: Cob development 05-24
Stage 4: Tasselling and silking 06-09
Stage 5: Kernel development 06-19
Stage 6: Grain filling and maturity 07-03
Stage 7: Harvest 08-25

Figure 2. Phenological stages of maize and their onset dates observed at station Adana 01.10 in 2016

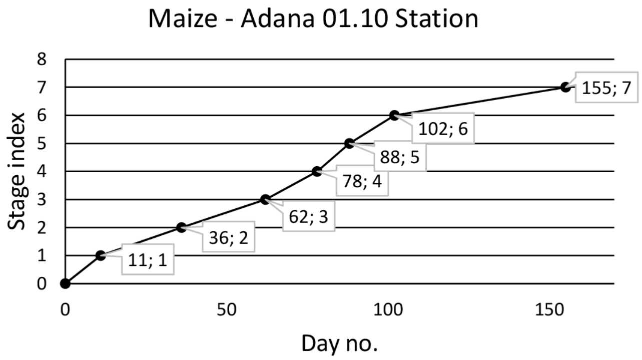

Figure 3. Piecewise linear interpolation of stage indices for dates between onset dates.

2.2. The overall system architecture

The proposed system shown in Figure 4 consists of VGG-16 (a deep 2D-CNN) and LSTM networks. The VGG-16 network extracts a spatial feature vector from each of the most recent w images in a causal time window. One or more LSTM units, operating on feature vectors of one or more dimensionalities, exploit the temporal dependencies among these feature vectors to classify the type or predict the phenological stage.

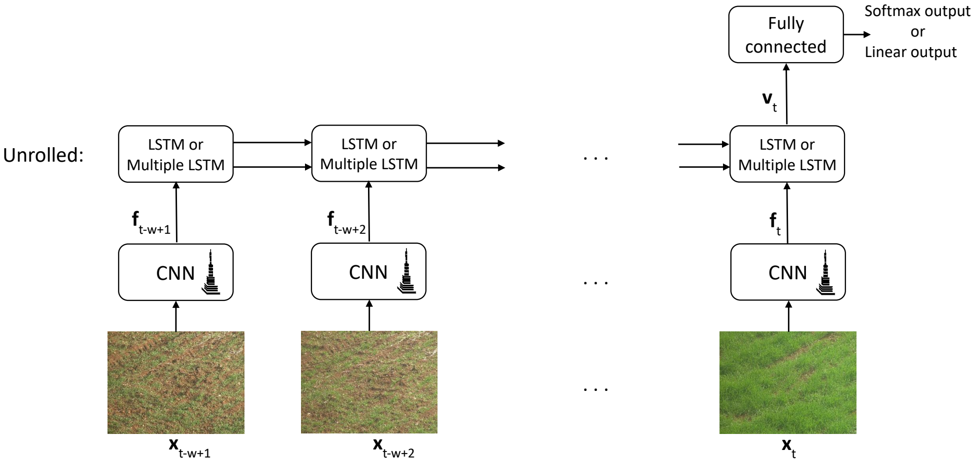

Figure 4. Block diagram of the proposed crop type classification/phenological stage prediction system.

2.3. The VGG-16 network for feature extraction

The VGG-16 network consists of a stack of convolutional (convolution followed by nonlinear activation) and pooling layers. Each convolutional layer extracts local, translation invariant features at its input, and each pooling layer combines nearby local features, decimates the result and feeds the following convolutional layer with lower resolution feature maps. Although each convolution kernel has small local support such as 3 × 3, a succession of such convolutions and poolings results in a much larger effective receptive field at the higher layers of the network whereby globally significant features in the input image can be exposed. In this work, the fully connected top layers of the VGG-16 network are only used to fine-tune the convolutional layers by using individual images and are not used in the actual decision-making process.

2.4. The LSTM unit

The LSTM [11] network unit is a popular variant of the recurrent neural networks that uses state information. The unit inputs the state vector, $\mathbf{s}_{t-1}$, and the output vector, $\mathbf{v}_{t-1}$, from time step $t-1$ as well as the input feature vector, $\mathbf{f}_t$, at time step t to yield the state vector, $\mathbf{s}_t$, and the output vector, $\mathbf{v}_t$, at time step t. Let σ be the sigmoid function used for activation nonlinearity and $\tanh$ be the hyperbolic tangent function used for normalization nonlinearity. These single variable functions are applied to each component of their input vectors. Let $\mathbf{b}_g$ stand for the bias vector of gate g and $\circ$ denote the Hadamard product. Based on $\mathbf{f}_t$ and $\mathbf{v}_{t-1}$, the part of the previous state that needs to be removed (kept) is determined with the remove gate activation

$$\mathbf{r}_t = \sigma(V_r\mathbf{f}_t + Q_r\mathbf{v}_{t-1} + \mathbf{b}_r).$$

The new information in normalized form and input gate activation which determines its strength are

$$\begin{aligned}\tilde{\mathbf{s}}_t &= \tanh(V_m\mathbf{f}_t + Q_m\mathbf{v}_{t-1} + \mathbf{b}_m), \\ \mathbf{i}_t &= \sigma(V_i\mathbf{f}_t + Q_i\mathbf{v}_{t-1} + \mathbf{b}_i),\end{aligned} \tag{1}$$

respectively. The retained part of the previous state is enhanced by the new information to yield the new state

$$\mathbf{s}_t = \mathbf{r}_t \circ \mathbf{s}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{s}}_t$$

and is normalized and modulated by the output gate activation $\mathbf{o}_t$ to yield the bounded output as $\mathbf{v}_t$ where

$$\begin{aligned}\mathbf{o}_t &= \sigma(V_o\mathbf{f}_t + Q_o\mathbf{v}_{t-1} + \mathbf{b}_o), \\ \mathbf{v}_t &= \mathbf{o}_t \circ \tanh(\mathbf{s}_t).\end{aligned} \tag{2}$$

2.5. System training approach

Jointly training the VGG-16 and LSTM networks to attain the best learning performance has high complexity since not only all training images need to be read from disk or kept in memory, but also the intermediary feature vectors extracted by VGG-16 need to be kept and updated. Hence, a memory efficient, two stage approach is proposed for training.

In the first stage, the pretrained VGG-16 network is fine-tuned for predicting crop type/phenological stage from the feature vector of a single image. A small number of training epochs is typically sufficient for the VGG-16 network weights to converge. After fine-tuning, the frozen VGG-16 network without the fully-connected output layers is used to extract the feature vector from each image in the training and test datasets. In the second stage, the LSTM network is trained by inputting temporal windows of training feature vectors.

In crop type classification, the one hot encoded label of the window is the type label of any image in the window. The LSTM network output is processed by a fully connected layer with softmax activation so that the a posteriori crop type probability given LSTM output can be maximized in decision-making. The loss minimized in training is categorical crossentropy.

In crop phenological stage prediction, the label of each window is set to the last date of the images in the window. The LSTM network output is processed by a fully connected layer with linear activation so that a real value is estimated. The loss minimized in training is the mean absolute error.

After LSTM training, a window of feature vectors of a test sequence may be fed to the LSTM network to obtain the actual prediction for the window.

2.6. Multiple LSTM units applied on feature vectors of multiple dimensionalities

Better phenological stage prediction performance is attainable with the use of two LSTM units operating on two feature vectors of different dimensionalities. It is envisioned that while the LSTM operation on the original high dimensional feature vector, $\mathbf{f}_t$, can exploit temporal dependencies among local characteristics like regional color or texture content, the LSTM operation on the low dimensional feature vector, $\mathbf{f}_t^l$, can exploit temporal dependencies among global characteristics like overall color or texture content. The low dimensional (32) feature vector, $\mathbf{f}_t^l$, is obtained from the high dimensional ($7 \times 7 \times 512$) feature vector output of the VGG-16 network by means of a two-layer fully connected neural network. This two-layer network is trained in the second stage mentioned in Section 2.5. The fully connected layer that outputs the real value for the phenological stage number is concurrently trained on the concatenation of the low and high dimensional feature vectors output by the two LSTM units. The module of multiple LSTM units that can substitute for the single LSTM unit is shown in Figure 5.

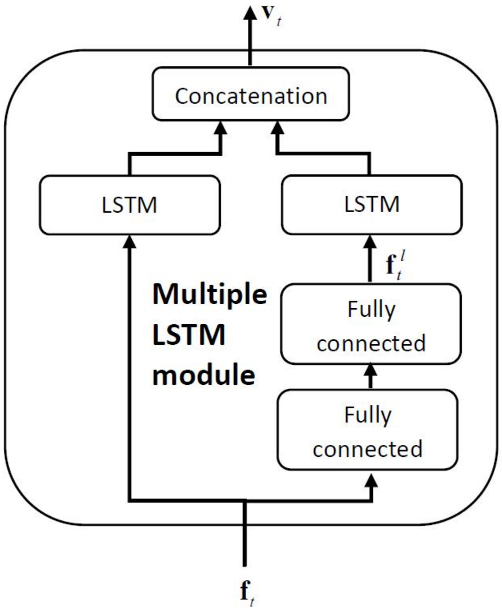

Figure 5. Two LSTM units operating on both low and high dimensional feature vectors.

2.7. Smoothing of stage predictions

The phenological stage predictions can be smoothed to remove the inherent noise by applying a moving average (MA) filter. The three filter taps weigh $\hat{p}_{n-1}$, $\hat{p}_n$, $\hat{p}_{n+1}$, the predictions for the nearest past, current and nearest future images, respectively, to yield the filtered prediction for the current image as

$$\hat{p}_{MA} = \alpha\hat{p}_n + (1-\alpha)(\frac{d_n - d_{n-1}}{d_{n+1} - d_{n-1}}\hat{p}_{n+1} + \frac{d_{n+1} - d_n}{d_{n+1} - d_{n-1}}\hat{p}_{n-1}), \tag{3}$$

where d_n is the nth observation day (image) and α determines the strength of smoothing. The past and future predictions get weights in proportion to the relative proximity of their dates to the date of the current prediction. The proposed noncausal filter is suitable for applications allowing delayed decision-making. Each pass of the filter requires the prediction to be delayed by one observation.

2.8. Implementation details

The proposed system has been implemented in Python and Keras with Tensorflow 1.12 backend accelerated by CUDA on a GTX 1080Ti graphical processing unit. In addition to stagewise training, training the networks with batches of samples helps reduce the memory consumption. For training the VGG-16 network, batches have been generated by partitioning the randomly shuffled images of the training image sequences. For training the LSTM network, batches have been generated by partitioning the randomly shuffled training samples. In this case, each sample corresponds to a subsequence (a temporal window) of feature vectors of consecutive images from one image sequence. Batch size has been set to 32 for training both networks. Network weights have been optimized by Adam optimizer having the learning rate of 0.001.

Before being fed to the VGG-16 network, the high resolution (2K or 4K) images have been resized to 224 × 224 resolution and their pixel color intensities have been scaled to the range [0,1]. The VGG-16 network has been fine-tuned on individual images for 4 epochs for type classification and for 6 epochs for phenological stage prediction. After fine-tuning, its fully connected top layers have been removed and the 1-D feature vectors obtained by flattening the 7 × 7 × 512 feature tensors at the output of this truncated network have been processed by the LSTM network.

3. Results and discussion

The data, excluding the single chickpeas sequence, has been partitioned into 10 approximately equal size folds where all batches of each sequence reside in one fold only. Then, group 10 fold cross validation has been performed by including the chickpeas image sequence in the training sets.

Validation accuracy and validation mean absolute error (MAE) in units of stage no., averaged over 5 runs, are reported for crop type classification and phenological stage prediction, respectively. We mostly report MAE in units of no. stages rather than no. days, since in the latter case, the significance of error varies largely with the type of crop and its phenological stage. In other words, error expressed in units of no. stages is a meaningful normalization for representing performance over the entire crop cycle. Unless otherwise stated, the results reported are for the single LSTM unit.

Validation and test sets are the same in the current study since hyperparameter (optimizer choice, learning rate, epoch number, batch size etc.) optimization has not been pursued. Rather, we focus on the influence of each of the investigated design issues on performance.

The nonparametric McNemar's test has been used to confirm whether a statistically significant difference exists between the performances of two classification approaches. The squared difference of the off-diagonal elements of the 2 × 2 contingency matrix is normalized by the sum of these elements to yield the test statistic. The continuity corrected statistic is expressed as

$$\chi^2 = \frac{(\|\beta - \gamma\| - 1)^2}{\beta + \gamma}, \tag{4}$$

where β and γ are the two joint frequencies of only one classifier deciding correctly. The statistic is a chi-square distributed random variable with one degree of freedom. The p-value is the probability of the statistic exceeding its realization.

The nonparametric Wilcoxon signed-rank test has been used to confirm whether the prediction errors arising from two prediction approaches are statistically different. When the sum of the ranks of the positive paired sample deviations, W^+, is taken as the test statistic, a normal distribution with a mean value of

$$\mu_W = \frac{n(n+1)}{4} \tag{5}$$

and a standard deviation of

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{6}$$

is used to approximate the distribution of the statistic since we deal with a large number of images. The ties are ranked as the average of the ranks were they broken with infinitesimal adjustments. The variance is reduced by $\frac{t^3-t}{48}$ for each group of t tied ranks. The continuity corrected statistic

$$Z = \frac{W^+ - \mu_W - 0.5}{\sqrt{{\sigma_W}^2 - \sum_i \frac{t_i^3 - t_i}{48}}} \tag{7}$$

is standard normal distributed. In our case, the p-value is the probability of Z exceeding its realization.

3.1. Ablation studies on design choices

3.1.1. Contrast stretching of input images

Firstly, we considered the effect of contrast stretching of color components of input images on learning performances. In Figure 6, where the first and second stages were trained for 4 and 25 epochs, respectively, for crop type classification, contrast stretching is seen to yield a negative effect on accuracy. The difference in the accuracies with and without contrast stretching varies between 0.05 and 0.07 over the 25 epochs of training the LSTM network. Null hypothesis of classification with and without contrast stretching having equal accuracy is rejected by McNemar's test with the p-value of 0.00281.

In Figure 6, the first and second stages were trained for 4 and 100 epochs, respectively, for crop phenological stage prediction. Here, the MAE with contrast stretching is larger than the MAE without contrast stretching at early epochs of training the LSTM, and only slightly lower at late epochs when the VGG-16 feature extractor network is not fine-tuned to fit the data well. However, when VGG-16 network is fine-tuned, contrast stretching yields a marked performance advantage possibly because reduced lighting variation (due to weather conditions, time of day) enables the learning system to generalize better to test data. The null hypothesis of the equivalence of the MAEs of the two cases is rejected by the Wilcoxon test for two paired groups with a p-value of 0.0357. Based on these findings, contrast stretching has been applied only for crop phenological stage prediction for the rest of the experimental work presented here.

3.1.2. Fine-tuning VGG-16 network

In Figure 7, the average accuracy of crop type classification is plotted as a function of LSTM training epoch no. (1–25) for 1–4 epochs of fine-tuning the VGG-16 network. As expected, no fine-tuning is distinctly inferior to one epoch of fine-tuning since the null hypothesis of the two cases yielding same accuracy is rejected by McNemar's test with a p-value of 0.00064 (for 25 epochs LSTM training). On the other hand, the null hypothesis of one epoch fine-tuning and 4 epochs fine-tuning VGG-16 yielding the same accuracy is rejected by McNemar's test p-value of 0.0182 (for 25 epochs LSTM training).

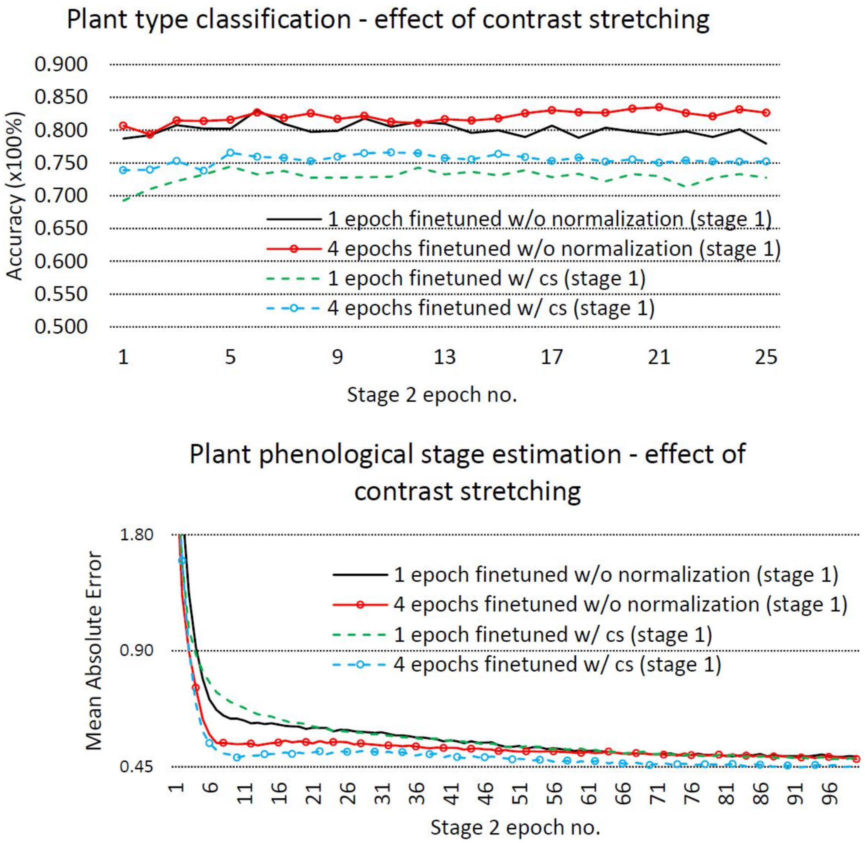

Figure 6. Effect of contrast stretching preprocessing operation on performances. Left: crop type classification, Right: crop phenological stage prediction $(w = 32)$.

In Figure 7, the dependence of MAE of crop phenological stage prediction on the number of epochs of fine-tuning the VGG-16 network is more explicit. However, there is little performance improvement observed after 4 epochs of fine-tuning. The difference in MAE between 1 epoch and 4 epochs of VGG-16 fine-tuning is statistically significant with a p-value of 0.0100 by the Wilcoxon signed rank test (for 100 epochs of LSTM training). Generally speaking, fine-tuning VGG-16 improves performance since the crop field images are quite different with regards to the structural components (color, texture, edges) than the ImageNet training images used in pretraining to set the initial weights of VGG-16.

3.1.3. VGG-16 vs. other deep networks for feature extraction

We have also compared the use of VGG-16 network against the more recent, deeper ResNet101 and DenseNet201 networks (implemented in Tensorflow 2) in feature extraction. Using no contrast stretching, 4 epochs fine-tuning, 25 epochs of LSTM training and a window size of 32, type classification accuracy turned out to be 32.9% and 55.2% (49.8% and 27.5% lower than VGG-16) with ResNet101 and DenseNet201, respectively. Using contrast stretching, 6 epochs fine-tuning 6 epochs finetuning, 100 epochs of LSTM training and a window size of 32, stage prediction MAE with ResNet101 or DenseNet201 turned out to be 2.747 and 2.762 ($\approx$ 6 times VGG-16). The poor generalization performances of ResNet101 or DenseNet201 can be attributed to the limited size of our data. For this data size, VGG-16 network offers a good model complexity.

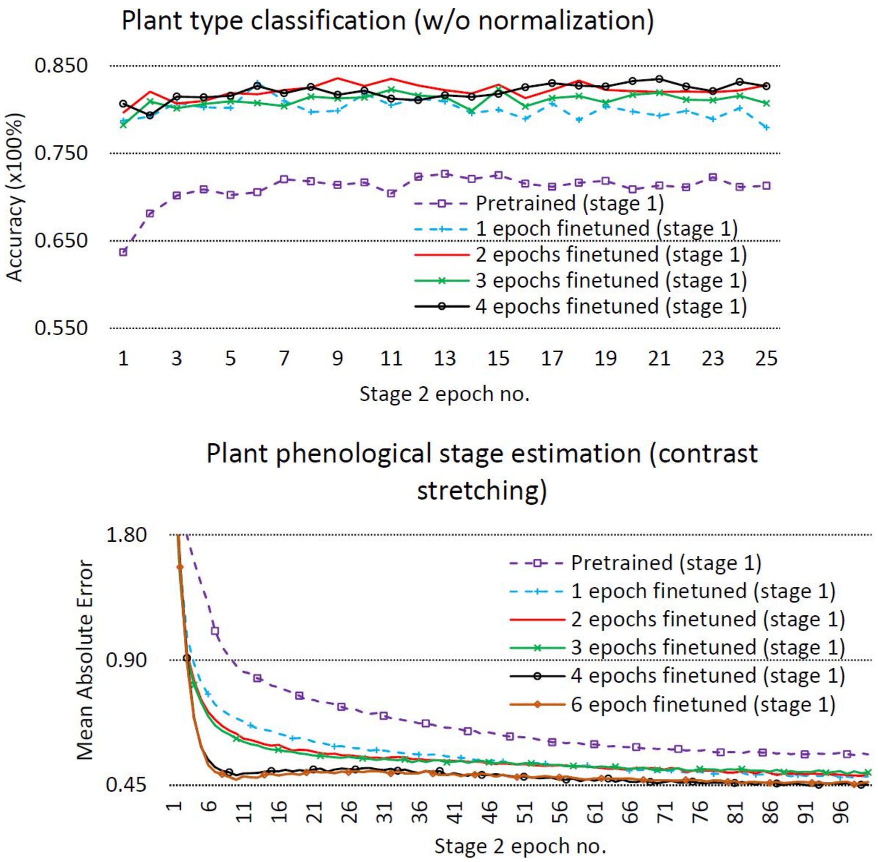

Figure 7. Effect of fine-tuning stage 1 feature extractor network on performances. Left: crop type classification accuracy without color component normalization (contrast stretching), Right: crop phenological stage prediction MAE with contrast stretching $(w = 32)$.

3.1.4. LSTM window size

From Figure 8, it is seen that increasing the LSTM temporal window size from $w = 1$ to $w = 32$ increases the learning performances. Here, the case of $w = 1$ is implemented as a dense layer that outputs class/prediction information at the same time instant as the input. The null hypotheses of the type classification accuracies and phenological stage prediction MAEs being equal for $w = 1$ and $w = 32$ are rejected by McNemar's and Wilcoxon signed rank tests with p-values of 0.00106 and 0.00041, respectively. This supports the main claim of the paper that, for learning type and stage information with deep networks, processing sequences of images yields an advantage over processing individual images.

3.2. Type classification performance with finalized design choices

Table 4 presents the confusion matrix for crop type classification where the optimal design choices determined in the previous section were used. The average accuracy, Kohen's Kappa score, F1 score are 0.826, 0.776, 0.823, respectively. The discrimination between barley and wheat is the weakest. Barley is more often (33.0%) mistaken for wheat than wheat is mistaken for barley (15.8%). Average accuracy of 0.826 is not low since images at the early stages of the crop cycle impart little information about the crop planted.

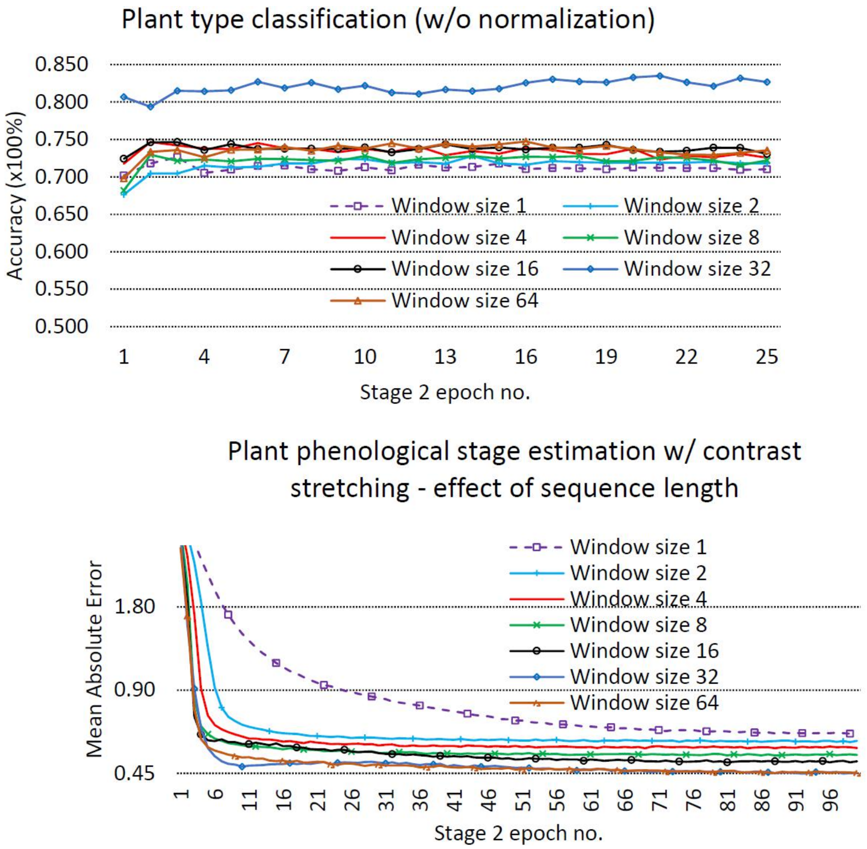

Figure 8. Effect of window size, w, in 2nd stage LSTM network on performances. Left: crop type classification accuracy with no color component normalization (contrast stretching), Right: crop phenological stage prediction MAE with contrast stretching.

Table 4. Confusion matrix for crop type classification (rows: actual class, columns: predicted class). VGG-16 feature extractor network was fine-tuned for 4 epochs, window size = 32.

	Barley	Sunflower	Wheat	Maize	Chickpeas	Cotton
Barley	710	3	350	26	4	7
Sunflower	1	612	42	93	0	52
Wheat	293	19	1562	22	0	4
Maize	14	43	2	1341	1	15
Chickpeas	0	0	0	0	0	0
Cotton	0	71	0	1	0	828

3.3. Stage prediction performance with finalized design choices and investigated enhancements

3.3.1. Single and multiple LSTM units

Table 5 presents the MAEs for the predictions of the phenological stage number where contrast stretching was employed. For all crops, the single LSTM unit yields an average MAE of 0.415 (12 days) whereas the module of two LSTM units yields an average MAE of 0.380 (10.9 days). The null hypothesis of the equivalence of the MAEs of predictions by the two systems is rejected by the Wilcoxon signed rank test with the p-value of 0.0291 which confirms the advantage of the module of two LSTM units over a single unit. The module of two LSTM units is justified for use with barley, wheat and maize for which the most significant advantages in MAE (4% to 9%) are observed.

Table 5. The phenological stage prediction MAE in no. stages (no. days) with a single LSTM unit applied on high dimensional feature vector, and two LSTM units applied on low and high dimensional feature vectors. VGG-16 was fine-tuned for 4 epochs, window size = 32.

	Barley	Sunflower	Wheat	Maize	Cotton	Overall
Single LSTM unit	0.327 (10.0)	0.291 (7.6)	0.341 (10.5)	0.611(15.2)	0.483 (16.3)	0.415 (12.0)
Two LSTM units	0.276 (7.9)	0.313 (8.3)	0.325 (10.4)	0.543 (13.3)	0.428 (13.9)	0.380 (10.9)

Interestingly, limiting the training image data for the phenological stage prediction network to the image data of the crop type used in the test does not yield better performance as might be expected. For example, the average prediction MAE was determined as 0.411 and 0.416, for barley and for wheat, respectively, when the phenological stage prediction network consisting of a single scale LSTM unit was trained solely on image sequences of barley or solely on image sequences of wheat. This points to the limited size of data available for each crop. When all data for all crops is used to train the network, better generalization ability is gained.

Figure 9 shows the error sequences for phenological stage prediction by using the module of the multiple LSTM units on the sequences obtained from the 19 wheat stations. The maximum absolute error is 1.5 (no. stages) and the average over 19 stations is 0.325.

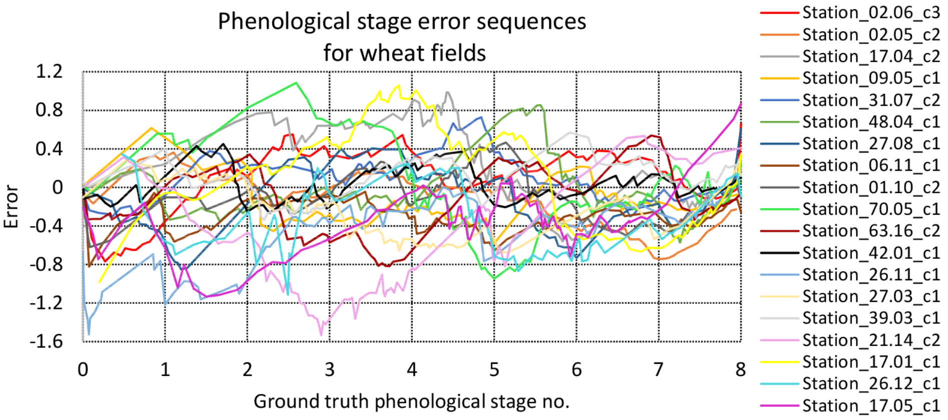

Figure 9. Phenological stage prediction error sequences for wheat fields.

We have investigated, why, the errors for stations 21.14 (pink) and station 70.05 (green) are large at the beginning of the tillering stage, by examining the images for these stations at that time. We observe that the green content for station 21.14 is substantially less than the green content for station 70.05 since the former image shows large lumps of soil whereas the latter, taken on closeup range, shows only leaves. Although human experts can focus on the relevant color information and structure in leaves while labelling these images, the proposed system apparently relies on average color content. The first few images of station 26.11 (light blue) contain many dark regions (shadows) between lumps of soil which may explain the large (1.5) stage prediction errors for these images.

The prediction errors in the estimation of the onset dates of tillering, stem elongation and heading stages of wheat with the module of the multiple LSTM units, calculated by linearly interpolating the prediction errors on the two closest days before and after each onset date are 12.1, 12.3, and 7.1 days, respectively. Although the average absolute error of 0.8 days achieved for the onset date of the heading stage by (Velumani et al., 2020) is comparatively much smaller, one can note that not only the data is different in (Velumani et al., 2020), but also the proposed system is capable of predicting the phenological stage throughout the crop cycle and not just a single onset date.

3.3.2. Smoothing

For smoothing the phenological stage prediction values, one and two pass applications of the MA filter have been tested. The center tap value weighing the present sample has been set at $\alpha = 0.8$ and the tap values weighing the past and future predictions have a sum of $1 - \alpha = 0.2$. For each station, each pass of the MA filter strictly reduces the MAE. For the single LSTM unit, the average MAEs for no MA filtering, one pass MA filtering, two pass MA filtering are 0.341,0.336, and 0.334, respectively. For the module of multiple LSTM units, the corresponding average MAEs are 0.325, 0.321, and 0.319, respectively. For the single and multiple LSTM units, the null hypotheses of the equivalence of the MAEs with no MA filtering and two-pass MA filtering are not supported by the Wilcoxon signed rank test with p-values of 0.0415 and 0.0589, respectively. These figures also suggest that MA filtering shows promise in improving prediction performance.

3.4. Comparisons with classical machine learning methods and handcrafted features

In Table 6, the performances of the classical machine learning methods applied on color level cooccurence matrices (CLCM) features [2] of individual images and the proposed system are compared. CLCM employs a $3 \times 3 \times 3$ cube with the center at the target pixel's target component. Thirteen of the 26 one step neighbors of the center are employed to determine 13 probability matrices. Thirteen Haralick features are derived from each matrix. By placing the three color components at the center of the cube in turn, a total of $13 \times 13 \times 3 = 507$ features are derived for each image.

Table 6. (a) Crop type classification, (b) Crop phenological stage prediction, based on consecutive images in a window of size w with classical machine learning on handcrafted features.

(a)

	kNearestClassifier		SupportVectorClassifier		RandomForestClassifier	
	w/o CS	w/ CS	w/o CS	w/ CS	w/o CS	w/ CS
w=1	0.4962	0.4895	0.5579	0.5680	0.5399	0.5464
w=8	0.5025	0.4968	0.5694	0.5782	0.5535	0.5532
w=32	0.4978	0.5080	0.5632	0.5674	0.5733	0.5689

(b)

	kNearestRegressor		SupportVectorRegressor		RandomForestRegressor	
	w/o CS	w/ CS	w/o CS	w/ CS	w/o CS	w/ CS
w=1	1.3784	1.3829	1.1896	1.1803	1.2376	1.2339
w=8	1.3496	1.3562	1.0978	1.0951	1.2194	1.2136
w=32	1.0740	0.9957	0.8759	0.8646	0.7893	0.7635

The results in these tables are based on a grid search for the best combination of C, the regularization parameter and γ, the kernel coefficient for support vector machines (SVM) ($(C, \gamma) \in \{0.001, 0.01, 0.1, 1, 10\} \times \{0.001, 0.01, 0.1, 1\}$), a grid search for the best combination of n, the number of trees, m, the number of features used for splitting a node, and d, the maximum tree depth for random forests (RF) ($(n, m, d) \in$

$\{25, 50, 100, 150, 200, 300, 500, 800\} \times \{p, \log p, \sqrt{p}\} \times \{5, 8, 10, 15, 25, 40\}$) where p is the number of features, and an exhaustive search for the best k in k-Nearest Neighbors (kNN) ($k \in 1, 3, 5, \ldots, 15$). The folds in cross validation are the same as those used for the proposed deep learning system. Among the three window sizes considered, $w = 1$ implies learning and inference from the feature vector of a single image. For $w = 8$ and $w = 32$, the concatenated feature vectors of the images in the window is input to the classifier/predictor.

Although increasing the window size makes a marked positive impact on the performances of CLCM&SVM, CLCM&kNN and CLCM&RF in phenology prediction, the best performances with these approaches are quite inferior to the performances with the proposed system (compare Table 6 against Figure 8).

4. Conclusion

This work primarily demonstrated by statistical validation that classifying crop type and predicting crop phenological stage by means of a LSTM recurrent neural network, operating on deep CNN features extracted from a nontrivial size temporal window ($w > 1$) of a field image sequence, yields a performance advantage over that achieved by the same deep CNN, operating on individual field images (w=1). The optimum window size is dependent on the frequency of capturing images. Secondly, other than the window size, best design choices related to contrast stretching of the input images and fine-tuning the feature extractor network have been determined for each of the two learning problems. Thirdly, deploying multiple LSTM units on feature vectors with multiple dimensionalities, and smoothing the prediction results have also been shown to yield statistically significant improvements in stage prediction performance. Finally, the proposed deep learning system was shown to exploit temporal dependencies far better than classical machine learning methods employing handcrafted color and texture features.

Although the numerical results are pretty much data dependent, the statistically drawn inferences are likely to hold over similar datasets. In the future, the data of the previous crop cycle(s) of a field can be acquired and used in training to improve the generalization performance of the proposed system to the data of its current crop cycle.

A public repository hosting the source codes and a sample subset of the dataset used in the current work may be found under https://github.com/UlugBayazit/Agricultural-deep-learning-from-in-situ-image-sequences.

Acknowledgement

This work was funded by İstanbul Technical University Scientific Research Project (BAP) MGA-2017-40929.

References

[1] Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A. Sequential deep learning for human action recognition. In: Lepri B, Salah A A (editors). 2nd International Workshop on Human Behavior Understanding (HBU); 2011. pp. 29–39.

[2] Benco M, Hudec R, Kamencay P, Radilova M, Matuska S. An advanced approach to extraction of colour texture features based on GLCM. International Journal of Advanced Robotic Systems 2014; 11 (7): 1-8. doi: 10.5772/58692

[3] Bodhwani V, Acharjya D, Bodhwani U. Deep residual networks for plant identification. Procedia Computer Science 2019; 152 (1): 186-194. doi: 10.1016/j.procs.2019.05.042

[4] Castro J, Feitoza R, Cue La Rosa L, Achanccaray Diaz P, Sanches I. A comparative analysis of deep learning techniques for sub-tropical crop types recognition from multitemporal optical/SAR image sequences. In: 2017 30th Conference on Graphics, Patterns and Images (SIBGRAPI); Niteroi, Brazil; 2017. pp. 382-389.

[5] Cheng Z. Detecting phenological transition dates of vegetation based on multiple deep learning models. MS, TU Delft University of Technology, 2600 AA Delft, The Netherlands, 2018.

[6] Desai SV, Balasubramanian VN, Fukatsu T, Ninomiya S, Guo W. Automatic estimation of heading date of paddy rice using deep learning. Plant Methods 2019; 15 : 76. doi: 10.1186/s13007-019-0457-1

[7] Donahue J, Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S et al. Long-term recurrent convolutional networks for visual recognition and description. In: IEEE 2015 Conference on Computer Vision and Pattern Recognition (CVPR); Boston, MA, USA; 2015. pp. 2625-2634.

[8] Forcén M, Pavón Pulido N, Pérez Noguera D, Berríos Reyes P, Pérez Pastor A et al. Machine Learning-based inference system to detect the phenological stage of a citrus crop for helping deficit irrigation techniques to be automatically applied. In: EGU General Assembly Conference Abstracts; Vienna, Austria; 2020. p. 18284.

[9] Grinblat G, Uzal L, Larese M, Granitto P. Deep learning for plant identification using vein morphological patterns. Computers and Electronics in Agriculture 2016; 127 (7): 418-424. doi: 10.1016/j.compag.2016.07.003

[10] Heredia I. Large-scale plant classification with deep neural networks. In: The Computing Frontiers Conference; Sienna, Italy; 2017. pp. 259-262.

[11] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation 1997; 9 (8): 1735-1780. doi: 10.1162/neco.19997.9.8.1735

[12] Ji S, Zhang C, Xu A, Shi Y, Duan Y. 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. Remote Sensing 2018; 10 (1): 75. doi: 10.3390/rs10010075

[13] Lancashire P, Bleiholder H, Boom T, Langeluddeke P, Stauss R et al. A uniform decimal code for growth stages of crops and weeds. The Annals of Applied Biology 1991; 119 (3): 561-601. doi: 10.1111/j.1744-7348.1991.tb04895.x

[14] Lee SH, Chan CS, Wilkin P, Remagnino P. Deep-plant: Plant identification with convolutional neural networks. In: IEEE International Conference on Image Processing; Quebec City, Canada; 2015. pp. 452-456.

[15] Mortensen AK, Dyrmann M, Karstoft H, Jørgensen RN, Gislum R. Semantic segmentation of mixed crops using deep convolutional neural network. In: International Conference on Agricultural Engineering: Automation, Environment and Food Safety; Aarhus, Denmark; 2016: 1-6.

[16] Ng J-H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R et al. Beyond short snippets: Deep networks for video classification. In: IEEE 2015 Conference on Computer Vision and Pattern Recognition (CVPR); Boston, MA, USA; 2015: 4694-4702.

[17] Pound M, Burgess A, Wilson M, Atkinson J, Griffiths M et al. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. GigaScience 2016; 6 (10): 1-10. doi: 10.193/gigascience/gix083

[18] Rasti S, Bleakley C, Silvestre GCM, Holden NM, Langton D et al. Crop growth stage estimation prior to canopy closure using deep learning algorithms. Neural Computing and Applications 2020; 33: 1733–1743. doi: 10.1007/s00521-020-05064-6

[19] Rebetez J, Satizábal HF, Mota M, Noll D, Büchi L et al. Augmenting a convolutional neural network with local histograms - a case study in crop classification from high-resolution UAV imagery. In: European Symposium on Artificial Neural Networks; Bruges, Belgium; 2016: 515-520.

[20] Reyes AK, Caicedo JC, Camargo JE. Fine-tuning deep convolutional networks for plant recognition. In: Proceedings of the Working Notes of CLEF; Toulouse, France; 2015.

[21] Rußwurm M, Körner M. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In: IEEE 2017 Conference on Computer Vision and Pattern Recognition Workshops; Honolulu, HI, USA; 2017: 1496–1504.

[22] Sun Y, Liu Y, Wang G, Zhang H. Deep learning for plant identification in natural environment. Computational Intelligence and Neuroscience 2017; 2017: 1–6. doi: 10.1155/2017/7361042

[23] Taghavi Namin S, Esmaeilzadeh M, Najafi M, Brown TB, Borevitz JO. Deep phenotyping: deep learning for temporal phenotype/genotype classification. Plant Methods 2018; 14: 66. doi:10.1186/s13007-018-0333-4

[24] Velumani K, Madec S, de Solan B, Lopez-Lozano R, Gillet J et al. An automatic method based on daily in situ images and deep learning to date wheat heading stage. Field Crops Research 2020; 252: 107793. doi: 10.1016/j.fcr.2020.107793

[25] Yalcin H. Plant phenology recognition using deep learning: Deep-Pheno. In: 6th International Conference on Agro-Geoinformatics (Agro-Geoinformatics); Fairfax, VA, USA; 2017: 1–5.

[26] Yalcin H, Razavi S. Plant classification using convolutional neural networks. In: 5th International Conference on Agro-Geoinformatics (Agro-Geoinformatics); Tianjin, China; 2016: 1–5.