Segmentation of Bean-Plants Using Clustering Algorithms

Serkan Kartal[1], Sunita Choudhary[2], Michal Stočes[1], Pavel Šimek[1], Tomáš Vokoun[1], Vojtěch Novák[1]

[1] Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences in Prague, Czech Republic

[2] International Crops Research Institute for Semi-arid Tropics (ICRISAT), Patancheru, India

Abstract

In recent years laser scanning platforms have been proven to be a helpful tool for plants traits analysing in agricultural applications. Three-dimensional high throughput plant scanning platforms provide an opportunity to measure phenotypic traits which can be highly useful to plant breeders. But the measurement of phenotypic traits is still carried out with labor-intensive manual observations. Thanks to the computer vision techniques, these observations can be supported with effective and efficient plant phenotyping solutions. However, since the leaves and branches of some plant types overlap with other plants nearby after a certain period of time, it becomes challenging to obtain the phenotypical properties of a single plant. In this study, it is aimed to separate bean plants from each other by using common clustering algorithms and make them suitable for trait extractions. K-means, Hierarchical and Gaussian mixtures clustering algorithms were applied to segment overlapping beans. The experimental results show that K-means clustering is more robust and faster than the others.

Keywords

Introduction

Today, in parallel with the continuous growth of the global population (Kitzes et al., 2008), a continuous increase in food production must be achieved. In order to achieve this increase in food production, agricultural ecosystems need to be better understood by monitoring and analysing continuously. This process requires the analysis of huge agricultural data and the use of new information technologies (Kamilaris et al., 2016). This observation and analysis process can be examined under two headings; large scale observation and small-scale observation. Large-scale observation is performed by remote sensing vehicles such as satellites and drones which provide wide-view snapshots of the agricultural environments (Domínguez et al., 2015; Kumhálová et al., 2014, Kumhálová et al., 2013; Pantazi et al., 2016). When this system is applied to agriculture, it provides many advantages in collecting information on soil properties and cultivation areas. On the other hand, with small-scale observations, more detailed and sensitive information, such as disease detection, plant count and quality control are collected (Khan et al., 2018; Ramos et al., 2017; Suh et al., 2018; Zhang et al., 2018).

Collecting and reporting the data of plants in the growth phase is important factor to ensure agricultural sustainability and increase productivity (Raišienė et al., 2019; Muhammad et al., 2019). This process can help to analyse the relationship between plants and environmental factors so as to provide effective agricultural guidance (Bannayan and Sanjani, 2011; Jannoura et al., 2015; Gebbers and Adamchuk, 2010). In addition, having information about crop counts and growth stage allows farmers to perform field operations such as irrigation and thinning accurately and on time (Cook and Veseth, 1991). However, the collection of growth phase information is currently done through labour-intensive manual observations. This is a time-consuming process since it has to be carried out and reported every day or several times a day. Moreover, the labour-intensive approach is not objective, because observers can understand the same criterion differently, which can lead to errors.

Computer vision-based methods can be effectively used to monitor the growth of crops, as they do not contain the limitations of the labour-intensive process. These methods make a great contribution to the development of precision agriculture by making it easier to observe and measure the effects of different environmental factors on crops. There are many applications of computer vision technology in agricultural automation systems, such as monitoring of crop status (Sakamoto et al., 2012; Vega et al., 2015), weeds identification (Guerrero et. al, 2012) yield estimation (Payne et al., 2013), disease detection (Yuehua et al., 2007; Pourreza et al., 2015), and quality control. Kamilaris and Prenafeta-Boldú (2018) review the studies used Machine Learning Algorithm (MLA) in agricultural and food production challenges emphasizing that those algorithms provide high accuracy and outperforms commonly used image processing techniques.

However, one of the frequently encountered problem in this process is that the plants are getting closer each other as they grow (Figure 1). If the plants come into contact with each other or overlap, this makes the problem even bigger. This problem makes it difficult to extract features belongs single plant such as height and digital biomass. In this study, it is aimed to separate bean plants in the certain region by applying clustering algorithms used in the literature. According to the results of the application, the advantages and disadvantages of the algorithms are discussed. Then, the most suited algorithm is employed and the dataset having separate plants has been prepared.
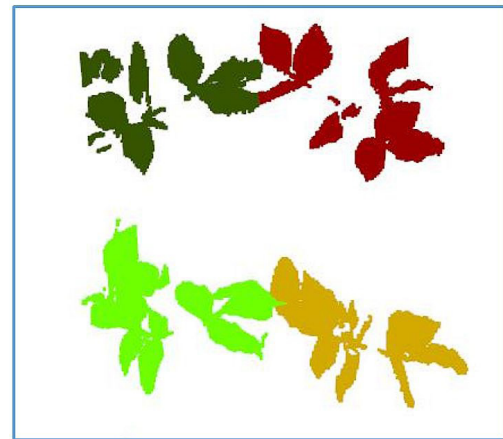
Source: own data processing

Figure 1: An example views of plants starting to overlap each other.

Materials and methods

The bean dataset used in this study was provided by International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in India. A large platform called "LeasyScan" has been developed at ICRISAT (**http://phenospex.com/blog/**) for the plant monitoring. LeasyScan is a laser based continuous plant monitoring system generates 3D point clouds data in every 2 hours on each of the 4800 sectors. Each data file has a time stamp that allows plants to be monitored and analysed depending on the time and environmental conditions. A part of a sample data file created during the scanning process of beans is given in Figure 2.

Pre-processing step is applied to the raw data to segment tray and soil data from the beans. This pre-processing is performed depending on the mathematical calculations achieved by using width, height and depth information of the trays. Later on, beans data that exist in each tray are saved as a separate data file. The final version

Source: own data processing

Figure 2: View of a sample scan file.

of the sample data file obtained after the pre-processing step is given in Figure 3.
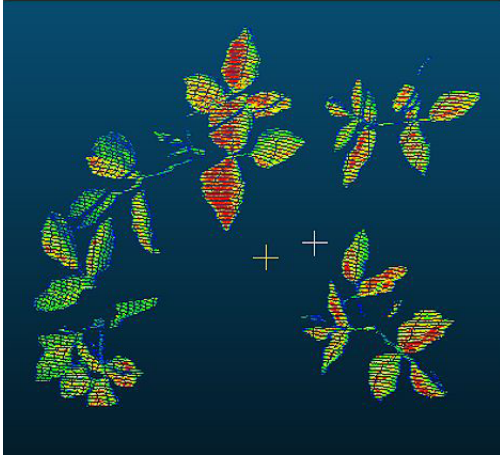

Source: own data processing

Figure 3: Sample data file after pre-processing process.

As seen in Figure 3, each data file consists of more than one bean. In order to observe the development of plants that grow close to each other and to calculate the phenotypic properties separately, firstly, the plants must be separated from each other. Although there are some algorithms developed in the literature to separate large plants, such as trees, an efficient algorithm has not been developed yet to separate small plants (Itakura and Hosoi, 2018; Zhang et al., 2019). In an exemplary segmentation algorithm used to separate trees, the first step is to cut the tree data at a certain height from the ground to create a dataset containing only tree trunks. Then, the points where these trunks are located are detected and accepted as the cluster centre for each tree. In the last step, the upper parts of the tree are clustered according to their closeness to the trunks/cluster centres. Since, there is no woody stem that can be detected and used as cluster centre, alternative approaches are required for the small plants.

In order to overcome this problem and separate the beans from each other, four widely used clustering algorithms have been utilized in this study. These algorithms are K-means, Gaussian mixtures and two different versions of Hierarchical clustering algorithms, respectively. All these algorithms require the number of clusters to be specified. Figure 4. illustrates the characteristics of clustering algorithms on 2D datasets (**https://scikit-learn.org/stable/auto_examples/ cluster/plot_cluster_comparison.html**). The labels in the bottom right corner of the pictures show the running times of the algorithms for the related data set.


Source: https://scikit-learn.org/stable/modules/clustering.html
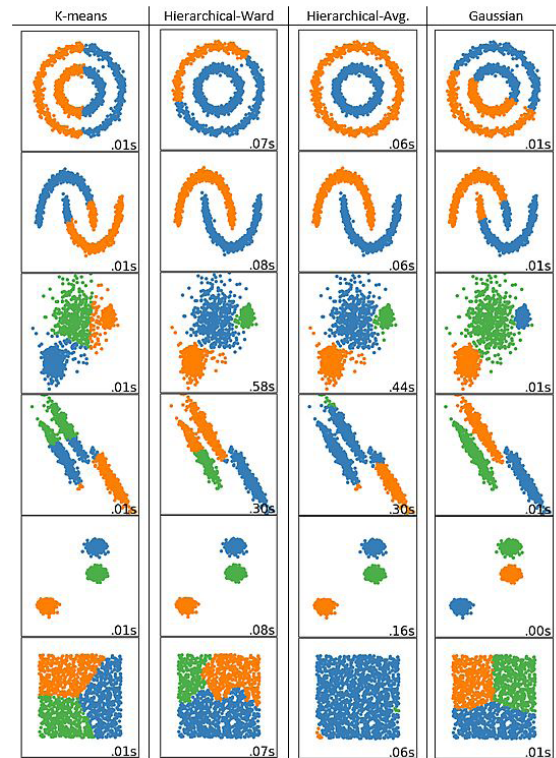
Figure 4. Results of k-means, hierarchical-ward, hierarchical-avg and gaussian mixture clustering algorithms on 2D datasets.

The K-means algorithm clusters data by separating data into k clusters, in which each datum belongs to the cluster with the nearest mean (cluster center). The aim of the algorithm is to minimize the sum-of-squares within cluster (Sabo, 2014). Since it is suitable for large amounts of data, it has been used in a wide variety of application areas in many different fields. Hierarchical clustering algorithms organize data into a hierarchical structure in proportion with the similarity matrix. These algorithms build nested clusters by merging or splitting them successively (Rokach and Maimon, 2005). There are 4 common metrics used for the merge strategy; ward, maximum(complete), average and single linkage (Ward, 1963; Defays, 1977; Sibson, 1973). In this this study, ward and average linkages are utilized. While ward minimizes the sum of squared differences within all clusters, average minimizes the maximum distance between observations of pairs of clusters. Gaussian mixture model is a probabilistic model for representing the presence of subpopulations within an overall population (Yu et al., 2011). It assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions.

Ground truth data

To generate the ground truth dataset, one sample scan file was randomly selected from the dataset. Each scan file consists of 12 trays. According to the pre-processing steps, trays were separated from each other, the soil data were cleaned, and the beans data belong to each tray were saved as a separate data file. Then, all the beans in each data file were separated manually and saved. Colorized example of a manually created ground truth bean data file is given in Figure 5.
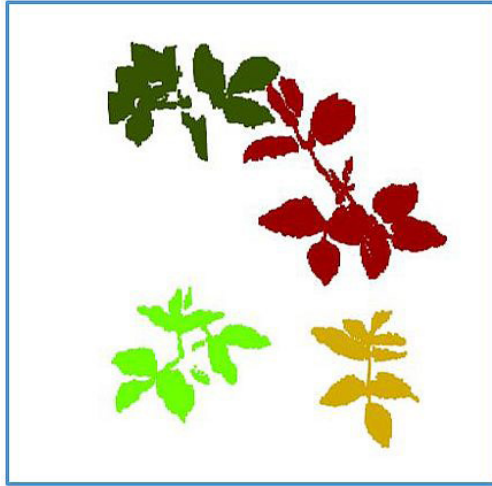
Source: own data processing

Figure 5: Colorized example of a manually created ground truth bean data.

Results and discussion

In order to investigate the performance of the clustering algorithms on bean data segmentation process, 12 ground truth data files

were employed. All the algorithms were coded in Python environment using scikit-learn library. Before starting the test, cluster counts were set depending on the beans counts in each tray.

After the clustering operations were performed, results files were colorized and saved similar to the ground truth data given in Figure 5. This visualization made it easy to evaluate the performance of the algorithms and to see their shortcomings. Additionally, the success rates of the algorithms were calculated by comparing the result images produced by the algorithms with the ground-truth images. During this process, the ratio of pixels correctly clustered by the algorithms to the total number of pixels was calculated as in Eq. 1. White pixels forming the background in the image were not included in the calculation process, the obtained results were used to quantitatively assess the accuracy of the clustering algorithms in segmenting bean plants. The formula for the calculation of success rate (SR) is as follows:

$$SR = \frac{M_{pixel}}{T_{pixel}} \qquad (1)$$

Where; M_{pixel} and T_{pixel} are the number of matching (correctly clustered) pixels and number of total pixels (except white pixels), respectively. According to this function, SR values of the k-means, Gaussian mixtures and two different versions of Hierarchical clustering algorithms are given in Table 1., respectively.

In Table 1, while the results of all the clustering algorithms over all the data are given separately, the average results are also given at the last row.

| Tray	K-means	Gaussian	Hier.-Avg.	Hier.-Ward
1	77.54	77.87	88.37	78.38
2	93.79	94.98	88.39	86.01
3	96.54	80.26	80.92	80.92
4	94.91	94.88	99.87	99.87
5	89.51	74.44	88.31	84.8
6	84.45	74.55	54.91	59.54
7	89.18	88.97	85.81	99.88
8	93.78	97.01	99.98	99.98
9	98.26	94.22	99.76	99.76
10	83.36	80.91	93.19	85.12
11	95.67	81.38	93.96	94.05
12	97.09	82.05	87.54	98.85
Average	90.13	85.81	87.95	87.43

Source: own data processing

Table 1: SR results of clustering algorithms.

The results reveal that, all the clustering algorithm produces the acceptable results for the segmentation process. When the results are examined in detail, it is seen that the maximum SR differences among the algorithms occurs in the 6th row/data. In order to analyse the performance of the algorithms in more detail, the colorized results of the algorithms on the first 6 data are given below.

When the first test data was examined, it is clearly seen that although hierarchical-avg clustering algorithms produces the best SR, this algorithm presents the worst performance in the visualized result. This is because this algorithm is very sensitive to noisy data. The same can be seen in the 6th result. Since this datafile has a small group of noisy data, the hierarchical-avg algorithm considers this noisy data as a separate cluster and tries to produce 3 clusters from the remaining data. The results of Gaussian algorithm illustrate that this algorithm is not robust when the plants are very close to each other and the leaves are overlapped. This weakness can be obviously seen
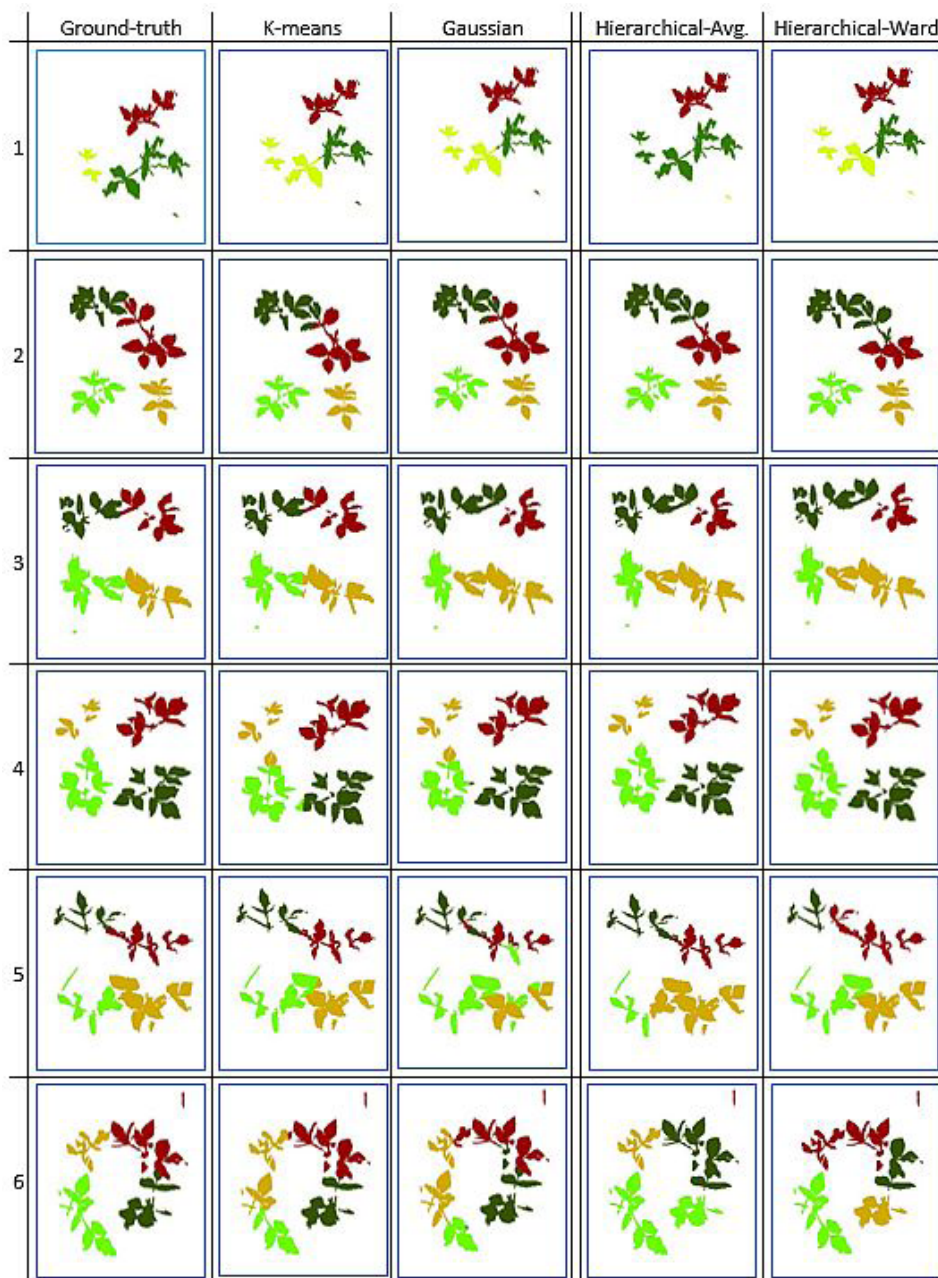
Source: own data processing

Figure 6: 2D visualization of ground truth data and results of k-means, gaussian mixture, hierarchical-avg and hierarchical-ward clustering algorithms, respectively.

in the 5th and 6th images given in the Figure 6. Therefore, just like the Hierarchical-avg. clustering algorithm, it can be misleading to evaluate the performance of the Gaussian algorithm just by looking at the quantitative results. By taking these considerations into account, we can conclude that performance of the Hierarchical-avg and Gaussian clustering algorithms are not good enough to separate the bean crops.

In accordance with the results comparison it can be said that the k-means clustering algorithm produces more consistent and reliable results than other algorithms. Although not as good as k-means, we can say that the Hierarchical-ward algorithm can be an alternative to k-means by looking at its general performance. Moreover, if algorithms are examined in terms of running time, as it seen in the Figure 4., k-means algorithm executes much faster than hierarchical algorithms. Therefore, the k-means algorithm provides advantages in separating bean plants in terms of both SR and running time.

Conclusion

The purpose of this study was to investigate the performance of clustering algorithms in bean-plant separation and to find the most suitable one. Before the clustering process, required pre-processing steps for tray segmentation and the soil cleaning were developed and employed. After the clustering process, the results were evaluated according to success rates and visualized clustering results. All algorithms performed the separation process with a success rate of over 85%. Despite this, it can be seen that the k-means algorithm is more advantageous in terms of both running-time and robustness against noise data. As a result, a reliable 3D bean-plant segmentation algorithm was obtained. This process will provide opportunities to extract the phenotypic traits belongs single plant such as height and digital biomass in the future works.

Acknowledgements

Corresponding authors
Serkan Kartal, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague, Kamýcká 129, Prague - Suchdol, 165 00, Czech Republic
E-mail: kartal@pef.czu.cz, skartal@cu.edu.tr

References

[1] Bannayan, M. and Sanjani, S. (2011) "Weather conditions associated with irrigated crops in an arid and semi arid environment", *Agricultural and Forest Meteorology*, Vol. 151, No. 12, pp. 1589-1598. ISSN 0168-1923, DOI 10.1016/j.agrformet.2011.06.015.

[2] Cook, R. J. and Veseth, R. J. (1991) "Wheat health management", St. Paul. MN. American Phytopathological Society.

[3] Defays, D. (1977) "An efficient algorithm for a complete link method", *The Computer Journal*, Vol. 20, No. 4, pp. 364-366. E-ISSN 1460-2067, ISSN 0010-4620. DOI 10.1093/comjnl/20.4.364.

[4] Domínguez, J. A., Kumhálová, J. and Novák, P. (2015) "Winter oilseed rape and winter wheat growth prediction using remote sensing methods", *Plant, Soil and Environment*, Vol. 61, No. 9, pp. 410-416. E-ISSN 1805-9368, ISSN 1214-1178. DOI 10.17221/412/2015-PSE.

[5] Gebbers, R. and Adamchuk, V. I. (2010) "Precision agriculture and food security", *Science*, Vol. 327, No. 5967, pp. 828-831. ISSN 1095-9203. DOI 10.1126/science.1183899.

[6]   Guerrero, J. M., Pajares, G., Montalvo, M., Romeo, J., and Guijarro, M. (2012) "Support vector machines for crop/weeds identification in maize fields", *Expert Systems with Applications*, Vol. 39, No. 12, pp. 11149-11155. ISSN 0957-4174. DOI 10.1016/j.eswa.2012.03.040.

[7]   Itakura, K. and Hosoi, F. (2018) "Automatic individual tree detection and canopy segmentation from three-dimensional point cloud images obtained from ground-based lidar", *Journal of Agricultural Meteorology*, Vol. 74, No. 3, pp. 109-113. ISSN 1881-0136, ISSN 0021-8588. DOI 10.2480/agrmet.D-18-00012.

[8]   Jannoura, R., Brinkmann, K., Uteau, D., Bruns, C. and Joergensen, R. G. (2015) "Monitoring of crop biomass using true colour aerial photographs taken from a remote controlled hexacopter", *Biosystems Engineering*, Vol. 129, pp. 341-351. ISSN 1537-5110. DOI 10.1016/j.biosystemseng.2014.11.007.

[9]   Kamilaris, A., Gao, F., Prenafeta-Boldu, F. X. and Ali, M. I. (2016) "Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications", In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pp. 442-447). DOI 10.1109/WF-IoT.2016.7845467.

[10]   Kamilaris, A. and Prenafeta-Boldú, F. X. (2018) "Deep learning in agriculture: A survey", *Computers and Electronics in Agriculture*, Vol. 147, pp. 70-90. ISSN 0168-1699. DOI 10.1016/j.compag.2018.02.016.

[11]   Khan, M. A., Akram, T., Sharif, M., Awais, M., Javed, K., Ali, H. and Saba, T. (2018) "CCDF: Automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep CNN features", *Computers and Electronics in Agriculture*, Vol. 155, pp. 220-236. ISSN 0168-1699. DOI 10.1016/j.compag.2018.10.013.

[12]   Kitzes, J., Wackernagel, M., Loh, J., Peller, A., Goldfinger, S., Cheng, D. and Tea, K. (2008) "Shrink and share: humanity's present and future Ecological Footprint", *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 363, No. 1491, pp. 467-475. ISSN 09628436. DOI 10.1098/rstb.2007.2164.

[13]   Kumhálová, J., Kumhála, F., Novák, P. and Matějková, Š. (2013) "Airborne laser scanning data as a source of field topographical characteristics", *Plant, Soil and Environment*, Vol. 59 No. 9, pp. 423-431. E-ISSN 1805-9368, ISSN 1214-1178. DOI 10.17221/188/2013-PSE.

[14]   Kumhálová, J., Zemek, F., Novák, P., Brovkina, O. and Mayerová, M. (2014). "Use of Landsat images for yield evaluation within a small plot", *Plant, Soil and Environment*, Vol. 60, No. 11, pp. 501-506. E-ISSN 1805-9368, ISSN 1214-1178. DOI 10.17221/515/2014-PSE.

[15]   Muhammad, K., Mastuki, N., Darus, F. and Ghani, E. K. (2019) „Accounting information system change in an agriculture company: Examination using burns and scapens framework", *Journal of International Studies*, Vol. 12, No. 1, pp. 105-118. E-ISSN 2306-3483, ISSN 2071-8330. DOI 10.14254/2071-8330.2019/12-1/7.

[16]   Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L. and Mouazen, A. M. (2016) "Wheat yield prediction using machine learning and advanced sensing techniques", *Computers and Electronics in Agriculture*, Vol. 121, pp. 57-65. ISSN 0168-1699. DOI 10.1016/j.compag.2015.11.018.

[17]   Payne, A. B., Walsh, K. B., Subedi, P. P. and Jarvis, D. (2013) "Estimation of mango crop yield using image analysis–segmentation method", *Computers and Electronics in Agriculture*, Vol. 91, pp. 57-64. ISSN 0168-1699. DOI 10.1016/j.compag.2012.11.009.

[18]   Raišienė, A. G., Podviezko, A., Skulskis, V. and Baranauskaitė, L. (2019) „Interest-balanced agricultural policy-making: Key participative and collaborative capacities in the opinion of NGOs' experts", *Economics and Sociology*, Vol. 12, No. 3, pp. 301-318. E-ISSN 2306-3459, ISSN 2071-789X. DOI 10.14254/2071-789X.2019/12-3/20.

[19]   Pourreza, A., Lee, W. S., Etxeberria, E. and Banerjee, A. (2015) "An evaluation of a vision-based sensor performance in Huanglongbing disease identification", *Biosystems Engineering,* Vol. 130, pp. 13-22. ISSN 1537-5110. DOI 10.1016/j.biosystemseng.2014.11.013.

[20] Ramos, P. J., Prieto, F. A., Montoya, E. C. and Oliveros, C. E. (2017) "Automatic fruit count on coffee branches using computer vision", *Computers and Electronics in Agriculture*, Vol. 137, pp. 9-22. ISSN 0168-1699. DOI 10.1016/j.compag.2017.03.010.

[21] Rokach, L. and Maimon, O. (2005) "Clustering methods", In *Data mining and knowledge discovery handbook*, pp. 321-352. Springer, Boston, MA. ISBN 978-0-387-24435-8. DOI 10.1007/0-387-25465-X_15.

[22] Sabo, K., and Scitovski, R. (2014) "Interpretation and optimization of the k-means algorithm", *Applications of mathematics*, Vol. 59, No. 4, pp. 391-406.   E-ISSN 1572-9109, ISSN 0862-7940. DOI 10.1007/s10492-014-0063-5.

[23] Sakamoto, T., Gitelson, A. A., Nguy-Robertson, A. L., Arkebauer, T. J., Wardlow, B. D., Suyker, A. E., ... and Shibayama, M. (2012) "An alternative method using digital cameras for continuous monitoring of crop status", *Agricultural and Forest Meteorology*, Vol. 154, pp. 113-126. ISSN 0168-1923. DOI 10.1016/j.agrformet.2011.10.014.

[24] Sibson, R. (1973) "SLINK: an optimally efficient algorithm for the single-link cluster method", *The Computer Journal*, Vol. 16, No. 1, pp. 30-34. ISSN 1460-2067, ISSN 0010-4620. DOI 10.1093/comjnl/16.1.30.

[25] Suh, H. K., Ijsselmuiden, J., Hofstee, J. W. and Van Henten, E. J. (2018) "Transfer learning for the classification of sugar beet and volunteer potato under field conditions", *Biosystems engineering*, Vol. 174, pp. 50-65. ISSN 1537-5110. DOI 10.1016/j.biosystemseng.2018.06.017.

[26] Vega, F. A., Ramirez, F. C., Saiz, M. P., and Rosúa, F. O. (2015) "Multi-temporal imaging using an unmanned aerial vehicle for monitoring a sunflower crop", *Biosystems Engineering*, Vol. 132, pp. 19-27. ISSN 1537-5110. DOI 10.1016/j.biosystemseng.2015.01.008.

[27] Ward Jr, J. H. (1963) "Hierarchical grouping to optimize an objective function", J*ournal of the American Statistical Association*, Vol. 58, No. 301, pp. 236-244. E-ISSN 1537-274X, ISSN 0162-1459. DOI 10.2307/2282967.

[28] Yuehua, C., Xiaoguang, H. and Changli, Z. (2007) "Algorithm for segmentation of insect pest images from wheat leaves based on machine vision", *Transactions of the Chinese Society of Agricultural Engineering*, Vol. 23, No. 12.pp. 187-191. ISSN 10026819.

[29] Yu, G., Sapiro, G. and Mallat, S. (2011) "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity", *IEEE Transactions on Image Processing*, Vol. 21, No. 5, pp. 2481-2499. ISSN 10577149. DOI 10.1109/TIP.2011.2176743.

[30] Zhang, J., He, L., Karkee, M., Zhang, Q., Zhang, X. and Gao, Z. (2018) "Branch detection for apple trees trained in fruiting wall architecture using depth features and Regions-Convolutional Neural Network (R-CNN)", *Computers and Electronics in Agriculture*, Vol. 155, pp. 386-393. ISSN 0168-1699. DOI 10.1016/j.compag.2018.10.029.

[31] Zhang, W., Wan, P., Wang, T., Cai, S., Chen, Y., Jin, X. and Yan, G. (2019) "A novel approach for the detection of standing tree stems from plot-level terrestrial laser scanning data", *Remote sensing*, Vol. 11, No. 2. ISSN 2072-4292. DOI 10.3390/rs11020211.