Computerized classification system for the identification of soil microorganisms

Michał Kruk, Ryszard Kozera, Stanisław Osowski, Paweł Trzciński, Lidia Sas Paszt, Beata Sumorok, and Bolesław Borkowski

Citation: AIP Conference Proceedings **1648**, 660018 (2015); doi: 10.1063/1.4912894
View online: http://dx.doi.org/10.1063/1.4912894
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1648?ver=pdfcov
Published by the AIP Publishing

Computerized Classification System for the Identification of Soil Microorganisms

Michał Kruk[1, a], Ryszard Kozera[1, b], Stanisław Osowski[2,3, c], Paweł Trzciński[4], Lidia Sas Paszt[4, d], Beata Sumorok[4] and Bolesław Borkowski[1]

[1]*Warsaw University of Life Sciences*
[2]*Warsaw University of Technology*
[3]*Military University of Technology, Warsaw*
[4]*Research Institute of Horticulture, Skierniewice*
[a]*michal_kruk@sggw.pl,*
[b]*ryszard_kozera@sggw.pl*
[c]*sto@iem.pw.edu.pl*
[d]*lidia.sas@inhort.pl*

Abstract. The paper presents the method of soil microorganisms identification in the microscopic digital images. The solved task includes: segmentation, feature generation, selection of the most important features and the final recognition stage applying 5 different solutions of classifiers. The paper presents and discusses the results concerning the recognition of several most popular soil microorganisms. The proposed system is able to recognize the microorganisms with the accuracy around 99%.

Keywords: image recognition, artificial inteligence
PACS: 07.05.Pj, 07.05Mh

INTRODUCTION

Soil microorganisms perform important functions in nature, affecting the soil properties [1]. They are responsible for the process of nitrogen fixation, which is the conversion of atmospheric nitrogen into nitrogen-containing compounds, used next to biosynthesize [2, 3] basic building blocks of plants. Microorganisms in soil are important because they affect the structure and fertility of different soils. Identification of beneficial strains and species of microorganisms can be used to expand the knowledge in the field of question (researchers and education sector) for development of microbiological preparations employed in agricultural production (growers and commercial sector).

Most solutions presented in the papers [4-8] are concentrating on extracting the individual organisms, describing them by the numerical descriptors and classifying by using different classifiers. The declared accuracy of recognition of several microbial morphotypes is up to 97%. This paper exploits different strategy. In practice the soil microorganisms appear in the agglomerations. To simplify the recognition task we will consider the recognition of the whole agglomeration instead of single individuals. In this way we transform the problem to the easier task.

MATERIALS

The microorganism samples have been isolated from the soil samples in the Agrotechnical Department, Research Institute of Horticulture, Skierniewice, Poland. All numerical algorithms were implemented and developed by Warsaw University of Life Sciences and Warsaw University of Technology. In this research we will consider 12 families of microorganisms. The recognized classes include (the number of images in parentheses): class 1 (12) - *Bacillus atrophaeus*, class 2 (49) - *Pseudomonas fluorescens*, class 3 (19) - *Bacillus subtilis*, class 4 (48) – *Azotobacter sp*., class 5 (48) - *Paecilomyces sp*., class 6 (17)– *Candidum sp.*, class 7 (59) - *Paenibacillus glucanolyticus*, class 8 (21) - *Rachnella aquatilis*, class 9 (15) - *Scoleobasidium sp.*, class 10 (43) - *Trichoderma sp.*, class 11 (78) - *Gliocladium sp.*, class 12 (32) – *Trichosporon beigelii*. These classes have been represented by 441 images. The examples of the images are available on http://michalkruk.pl/ea2014/examples.jpg.

Proceedings of the International Conference on Numerical Analysis and Applied Mathematics 2014 (ICNAAM-2014)
AIP Conf. Proc. 1648, 660018-1–660018-4; doi: 10.1063/1.4912894

METHODS

In solving the microorganism class recognition problem we have proposed the computerized system composed of few stages. The first is the segmentation of the original image aiming to separation of the background from the region of interest (ROI) containing microorganisms. In the next step we generate the numerical descriptors (features) of the ROI, which represent the potential input attributes to the classifier system. These features undergo the assessment of their class discrimination ability (selection process). The selected features are treated as the input attributes to the classification stage, responsible for final class recognition. The general scheme of the proposed system is presented in Fig. 1.

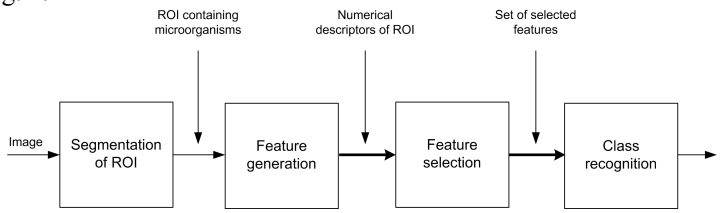

FIGURE 1. The proposed system of recognition of the microorganism classes.

Segmentation of Region of Interest

The first point in image processing is identification and segmentation of the region of interest containing microorganisms, which will be called ROI. The individual microorganisms are agglomerated and clumped in certain regions of the image. The segmentation and then recognition of the particular individuals from any group is a complex task. To simplify it we exploit the fact that they appear as a compact group in the image. Our approach to microorganism recognition will be based on proper description of the whole group instead of single individuals.

In this way the segmentation task is limited to recognition of the background from the region of interest (ROI) containing microorganisms which are treated as one object. In this way we can apply well known segmentation methods based on thresholding. The thresholding is performed using method of Otsu on the grey scale images. It is very reliable method and delivers stable results for large variation of the intensity of the image. In our experiments the Otsu procedure has resulted in proper segmentation of all classes of microorganisms from the background.

Generation of Diagnostic Features

To create the efficient classification system we have to generate the proper set of diagnostic features, forming the input attributes to the classifier. To obtain high efficiency of class recognition we have to define features which assume similar values for the objects belonging to the same class and are different for different classes. In the proposed solution we have applied various approaches to the feature generation. They are created on the basis of clusterization of the image, description of the color distribution and characterization of the histograms in different color descriptions (CIELAB, HSV and RGB). We used a few main groups of features:

- **Features Based on Cluster Centroids** In this approach we group the pixels of the image into few clusters which gather the similar objects. The clusterization is done using the K-means method applied for different components of the chosen color representation. We have used the CIELAB color space. As all color information is contained in the a* and b* layers in further processing only a* and b* components are used. All pixels of the image described by the pair of a* and b* are assigned to K clusters by the K-means algorithm. The aim of the K-means algorithm is to divide the image area into clusters so that the within-cluster sum of squares of values of chromaticity parameters is minimized. The algorithm seeks for locally optimal solution such that no movement of a point from one cluster to another will reduce the within-cluster sum of squares. When all pixels are assigned to clusters the means of the clusters are computed as the cluster centroids. These means are used as the numerical features describing the image. In our experiments we have tried different number of clusters. They reflect the fuzzy character of the images under recognition. The best results in classification have been obtained for K=3. In this case the number of cluster features is equal 6 (three centres for a* and three for b* representations).

- **The Colorimetric Features** The next family of features has been defined on the basis the intensity of the ROI region of the image. The colorimetric features have been defined on the basis of the intensity of pixels for each R, G and B components in RGB representation, L*, a* and b* in CIELAB and H, S and V in HSV. As colorimetric features we have used the mean and standard deviation of pixel intensities in the ROI containing microorganisms. They have been calculated for each color component in these three color spaces. Up to 18 colorimetric features have been created in this way.
- **Features based on histogram in different color spaces** The next set of features has been defined on the basis of histogram of the image. The histograms are created only for the ROI containing microorganisms (the image without background) in all 3 applied color spaces: RGB, CIELAB and HSV. In this work we have used the mean, variance, skewness, kurtosis, energy and entropy. Taking into account that the same image is represented in 3 different color spaces (each characterized by 3 components - RGB, HSV, LAB) we get 54 histogram features.

The features defined on the basis of cluster centroids, color characterization and the histogram form the set of 78 components. Some of them have no class discrimination ability or represent the noise. Therefore the next step is directed to the assessment of their quality and selection of the final set, which is used in the classification stage.

Feature Selection

The process of feature selection is an important step in developing the efficient procedure of microorganism recognition. Good feature should be characterized by the stable values for samples belonging to the same class and at the same time they should differ significantly for different classes. Thus the main problem in the classification and machine learning is to find out the features of the highest importance for the problem solution. Elimination of features of the weakest class discrimination ability leads to the reduction of the dimensionality of the feature space and improvement of generalization ability of the classifier in the testing mode for the data not taking part in learning. There are many known techniques of feature selection. In our analysis we will use the fast correlation-based filter (FCBF) solution [9]. It is an approach exploiting the correlation measure based on the information-theoretical concept of entropy. As a result we have selected 21 features treated as the most important for the class recognition.

Classification System

The selected features are served as the input attributes to the classifier system. To get the highest efficiency of classification process we have tried different solution of the classifiers: the multilayer perceptron (MLP), radial basis function network (RBF), support vector machine (SVM), random forest of decision trees (RF) and k nearest neighbor classifiers (kNN). All of them have been implemented in Matlab.

THE RESULTS OF NUMERICAL EXPERIMENTS

The data used in experiments have been normalized by dividing each column of data matrix by the maximum value. To get the most objective results of experiments we have applied the k-fold cross validation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. In the experiments we have applied 5 classifiers: Multi Layer Perceptron (MLP), Radial Basis Function (RBF), k-Nearest Neighbours (kNN), Support Vector Machine (SVM), Random Forest Tree (RF). The MLP classifier structure was 21-32-12. In the RBF network we used 27 radial neurons and the final stricture was 21-27-12. In the case of SVM networks we have to use 72 two-class SVM classifiers working in one-against-one mode. The optimal value of regularization parameter C was 1000 and the width of the Gaussian kernel function was 0.7. Random forest was run using 60% of data for training and the rest for validation. Five out of 21 features have been used in each node of decision trees. The best results of kNN classifier have been obtained at k=5. The statistical results of testing the classifiers in 5-fold cross validation mode are presented in Table 1. They depict the mean values of the accuracy of classification by using 5 mentioned above classifiers. The classifiers have been supplied by two sets of input attributes (the set selected by us and the set containing all features). The second column refers to the results at application of all features and the last column for the set of 21 selected features.

TABLE 1. The average accuracy of application of different classifiers in recognition of 12 classes of microorganisms.

Classifier	All features	FCBF selection (21 features)
SVM	89.8%	98.19%
RBF	87.3%	94,78%
MLP	83.8%	92,97%
RF	89.57%	98,41%
kNN	86.58%	95,24%

CONCLUSIONS AND FURTHER WORK

The paper has presented the automatic method of the recognition of different classes of microorganisms existing in the soil. The solved problems include: segmentation of the image directed to the localization of regions of interest containing the microorganisms, generation of the numerical descriptors related to the segmented ROI, selection of the most important descriptors as the diagnostic features and finally the recognition of the microorganisms using few classifier solutions. The proposed approach has been checked successfully in the recognition of 12 classes of microorganisms. Five different solutions of the classifiers have been tried. The best random forest classifier has obtained the accuracy of 98.41% on the data base containing 441 images. The results of the paper may find application in accelerating the research aimed on recognition of the soil microorganisms investigated in horticulture to improve the quality of the soil. Application of the developed system in this research will allow to free experts from the tedious manual work at the microscope and contribute significantly to the progress in this area. The problem solved in the paper is the first stage of microorganisms image processing. The presented results are encouraging and motivate continuation of the study. In future work we extend the solution to counting the number of individual microorganisms existing in the analyzed image.

REFERENCES

1. S. M. David, J. J. Fuhrmann, P. G. Hartel, D.A. Zuberer, Principles and applications of soil microbiology, Prentice Hall, Upper Saddle River, 1998.
2. G. H. Elkan, Biological nitrogen fixation systems in tropical ecosystems: an overview. In Biological Nitrogen Fixation and Sustainability of Tropical Agriculture. Eds. K Mulongoy, M Gueye and D S C Spencer. pp 27–40. 1992. John Wiley and Sons, Chichester, UK.
3. S. K. A. Danso, G. D. Bowen, N. Sanginga, Biological nitrogen fixation in trees in agro-ecosystems. Plant and Soil 141: 177-196, 1992
4. M. G. Forero, G. Cristóbal, M. Desco, Automatic identification of mycobacterium tuberculosis by Gaussian mixture models, Journal of Microscopy, 2006, vol. 223, pp. 120 – 132.
5. J. Liu, F. B. Dazzo, O. Glagoleva, B. Yu, A. K. Jain, CMEIAS: A computer-aided system for the image analysis of bacterial morphotypes in microbial communities, Microb Ecol., 2001, vol. 41, No 3, pp. 173-194.
6. P. Ruusuvuori, J. Sepp, T. Erkkil, A. Lehmussola, J. A. Puhakka, O. Yli-Harja, Efficient automated method for image-based classification of microbial cells, Proceedings of the 19th Inter-national Conference on PatternRecognition, 2008, pp. 1–4.
7. D. H. Theriault, M. L. Walker, J. Y. Wong, M. Betke, Cell morphology classification and clutter mitigation in phase-contrast microscopy images using machine learning, Machine Vision and Applications, 2012, vol. 23, pp. 659–673.
8. P. S. Hiremath, P. Bannigidad, Identification and classification of cocci bacterial cells in digital microscopic images, Int. J. of Computational Biology and Drug Design, 2011, vol. 4, No 3, pp. 262-273.
9. Y. Lei, L. Huan, Feature selection for high-dimensional data: a fast correlation-based filter solution, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.