Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100103

Computerized Classification System for the Identification of Soil Microorganisms

Michał Kruk[1,*], *Ryszard Kozera*[1], *Stanisław Osowski*[2,3], *Paweł Trzciński*[4], *Lidia Sas-Paszt*[4], *Beata Sumorok*[4] *and Bolesław Borkowski*[1]

[1] Faculty of Applied Informatics & Mathematics, Warsaw University of Life Sciences, ul. Nowoursynowska 159, 02-776 Warsaw, Poland
[2] Faculty of Electrical Engineering, Warsaw University of Technology, Pl. Politechniki 1, 00-661 Warsaw, Poland
[3] Military University of Technology, Kaliskiego 2, 00-908 Warsaw, Poland
[4] Research Institute of Horticulture, ul. Pomologiczna 18, 96-100 Skierniewice, Poland

Received: 6 Jun. 2015, Revised: 4 Aug. 2015, Accepted: 5 Aug. 2015
Published online: 1 Jan. 2016

Abstract: This paper presents the method of soil microorganisms identification from the microscopic digital images. The proposed approach includes: segmentation of the image, feature generation, selection of the most important features and the final recognition stage applying five different solutions of classifiers. The paper presents and discusses the results concerning the recognition of several most popular soil microorganisms: *Bacillus subtilis, Paenibacillus glucanolyticus, Rachnella aquatilis, Scoleobasidium sp., Trichoderma sp., Pseudomonas fluorescens, Bacillus atrophaeus, Azotobacter sp., Streptomyces sp.* and other bacterias and fungi. The proposed system enables the recognition of the microorganisms with the accuracy close to 98%.

Keywords: microorganisms identification, image processing, SVM classifier

1 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 in turn to biosynthesize [2,3] the 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 exploited to expand the knowledge in the field in question (researchers and education sector) for development of microbiological preparations employed in horticultural and agricultural production by growers and commercial sector. The manual detection of microorganisms under the microscope by laboratory staff is slow and tedious. Therefore the major challenge in microbiology is to develop computerized tools that can extract the information from digital images of the population of microorganisms and to use it in automatic recognition. A lot of work has been already done in developing different tools that can be applied in the recognition process of these microorganisms. Among them one should mention the methods of automatic image processing like filtering, segmentation, transformation of the image into the describing features, etc. [4,5,6]. Special tools are used for classification. To the most important belong the Bayes classifiers, decision trees, neural networks and support vector machine (SVM) [7,8,9,10]. Combining the image preprocessing tools with the machine learning techniques enables to build the computerized system that is able to recognize different classes of microorganisms, not necessarily existing in soil environment. Examples of such solutions are presented in [11,12,13,14,15]. The declared accuracy of recognition of several microbial morphotypes is up to 97%. Most solutions presented in the papers are concentrating on extracting the individual organisms, describing them by the numerical descriptors and finally classifying by using different classifiers. This paper exploits various strategies. In practice the soil microorganisms appear in the agglomerations. To simplify the recognition task we consider here the recognition of the whole agglomeration instead of single

[*] Corresponding author e-mail: michal_kruk@sggw.pl

Table 1: The data base of images of the microorganisms

Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
12	49	19	48	48	17
Class 7	Class 8	Class 9	Class 10	Class 11	Class 12
59	21	15	43	78	32

individuals. In this way we transform the problem to the easier task. The numerical results presented in the paper will show high efficiency of our approach in application to the recognition of the beneficial and pathogenic soil microorganisms. The input data were prepared by the Agrotechnical Department, Research Institute of Horticulture, Skierniewice, Poland. All numerical algorithms were implemented by Warsaw University of Life Sciences - SGGW and Warsaw University of Technology.

2 Materials

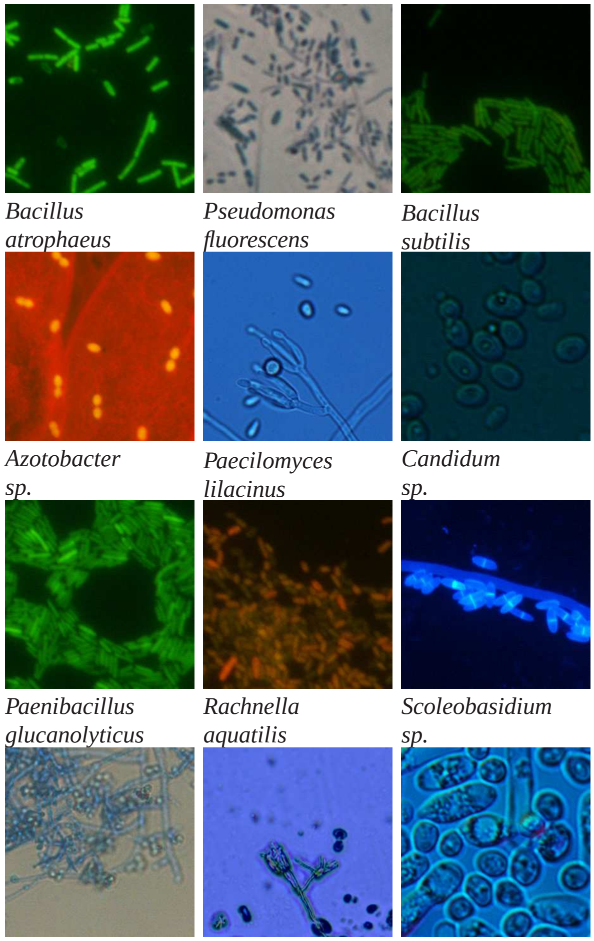
The microorganism samples were isolated from the soil samples in the Agrotechnical Department, Research Institute of Horticulture, Skierniewice, Poland. In this research we recognize 12 families of microorganisms. The recognized classes include: class 1 - *Bacillus atrophaeus*, class 2 - *Pseudomonas fluorescens*, class 3 - *Bacillus subtilis*, class 4 *Azotobacter sp.*, class 5 - *Paecilomyces sp.*, class 6 *Candidum sp.*, class 7 - *Paenibacillus glucanolyticus*, class 8 - *Rachnella aquatilis*, class 9 - *Scoleobasidium sp.*, class 10 - *Trichoderma sp.*, class 11 - *Gliocladium sp.* and finally class 12 *Trichosporon beigelii*. These classes are represented by 441 images. Table 2 illustrates the typical forms of all these families, and Table 1 presents the population size of images within each family, which took part in experiments.

To make the recognition problem more difficult we added the next mixed class, composed of 5 families of microorganisms. They include: *Bacillus megaterium*, *Bacillus megaterium endospors*, *Rhizobium meliloti*, *Serratia*, *Sesulfovibrio*. Table 3 presents the images of single representatives of these families. This mixed class is represented by 50 samples (10 representatives for each 5 species forming the single class).

3 Methods

In solving the microorganism class recognition problem we propose the computerized system composed of few stages. The first one is the segmentation of the original image aiming to separate the background from the region of interest (ROI) containing microorganisms. In the next step we generate the numerical descriptors (features) of

Table 2: The examples of images of the investigated classes of microorganisms



Bacillus atrophaeus

Pseudomonas fluorescens

Bacillus subtilis

Azotobacter sp.

Paecilomyces lilacinus

Candidum sp.

Paenibacillus glucanolyticus

Rachnella aquatilis

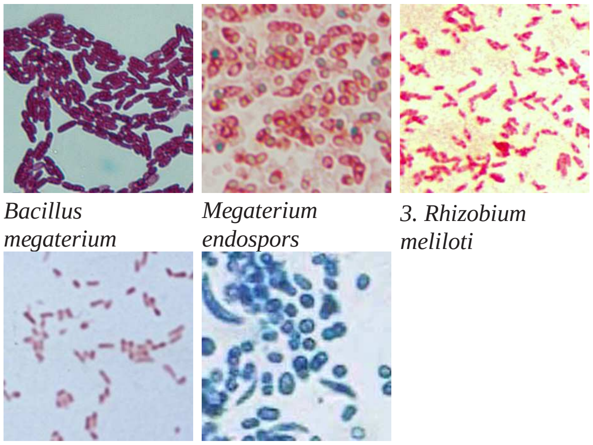
Scoleobasidium sp.

Trichoderma sp.

Gliocladium sp.

Trichosporon beigelii

Table 3: The examples of images of the investigated classes of microorganisms



Bacillus megaterium

Megaterium endospors

3. Rhizobium meliloti

Serratia

Desulfovibrio

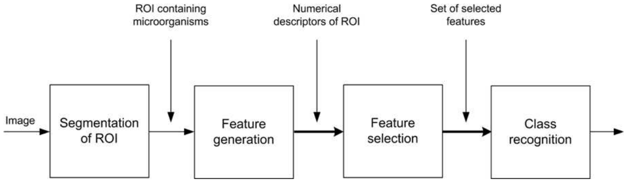


Figure 1: The proposed system for recognition of the microorganism classes.

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.

3.1 Segmentation of region of interest

The first step in image processing is identification and segmentation of the region of interest containing microorganisms, which is 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 is 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 doing so we can apply well known segmentation methods based on thresholding [6,16,17]. The thresholding is performed using method of Otsu [16] on the grey scale images. The Otsu method determines the threshold value in an automatic way on the basis of the variance of the objects (cells) with respect to the background of the image. It is a very reliable method and delivers stable results for large variation of the intensity of the image.

According to Otsu method many relevant threshold values t are tried and the respective mean and variance of both classes are calculated for each of them. The main task is to minimize the intra-class variance (the variance within the class). It is equivalent to maximizing the between-class variance defined as a weighted sum of variances of the two classes, which is expressed in terms of class probabilities $p_i(t)$ and class means m_i for i=1 and 2. The bimodal gray-level histogram is normalized and is regarded as a discrete probability distribution function $p_i(t)$ that is, $p(i) = n_i/M$, where where n_i is the frequency of the gray level i and M is the total number of pixels in the image. If the obtained histogram is divided into two classes by the gray-level intensity t (potential threshold), then the probabilities of the respective classes can be expressed as

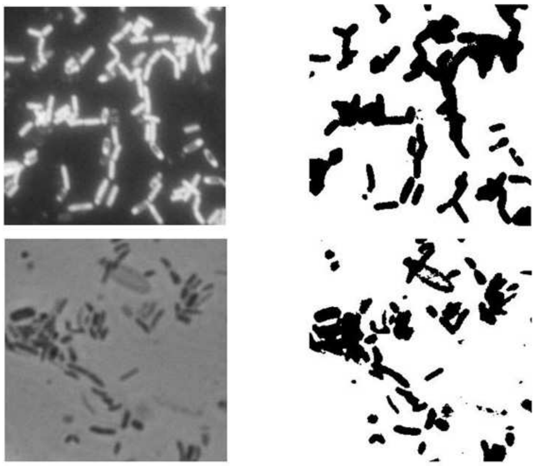


Figure 2: The results of application of Otsu method to the segmentation of 2 exemplary images of microorganisms. The upper images represent *Bacillus atrophaeus* and the lower *Pseudomonas fluorescens*.

$$p_1(t) = \sum_{i=0}^{t} p(i),$$

$$p_2(t) = \sum_{i=t+1}^{N-1} p(i).$$

Denoting by $m_1(t)$, $m_2(t)$ the means of these two classes at the actual threshold value t, where $(j = 1,2)$ the task is to maximize the value of between class variance defined in the form [16]

$$\max_t \{p_1(t)p_2(t)(m_2(t) - m_1(t))^2\}.$$

Application of Otsu criterion to the determination of the optimal threshold permits to separate the ROI from the background. In our experiments the Otsu procedure has resulted in proper segmentation of all classes of microorganisms from the background. Fig. 2 presents two examples of such segmentation. The left column shows the original images and the right one their binary segmented masks.

3.2 Generation of diagnostic features

To create the efficient classification system we need to generate the proper set of diagnostic features, forming the input attributes to the classifier [17,18]. To obtain high efficiency of class recognition one should 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 apply 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 [19], HSV and RGB).

3.2.1 Features based on cluster centroids

In the applied 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 [10,20] applied for different components of the chosen color representation. From the variety of color systems one selects here the CIELAB color space [19]. The CIELAB color space is derived from the CIE XYZ tristimulus values [6]. The L*a*b* space consists of a luminosity layer L*, chromaticity layer a* indicating where color falls along the red-green axis, and chromaticity layer b* indicating where the color falls along the blue-yellow axis. One of the most important advantages of the CIELAB model is its good approximation of the human vision. 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 [20]. 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 [20]. 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 tried different numbers of clusters. They reflect the fuzzy character of the images under recognition. The best results in classification are obtained for K=3. In this case the number of cluster features is equal 6 (three centres for a* and three for b* representations). Fig. 3 presents the results of clustering the CIELAB a* and b* components for 4 classes of microorganisms: *trichoderma*, *paenibacillus glucanolyticus*, *scoleobasidium* and *rahnella aquatilis*. They show the regions of the image belonging to three clusters represented by while, black and gray colors respectively. Each colored region is characterized by its mean (centroid), treated as the feature. For three clusters and two color components 6 features are generated. At the bottom of the images the cluster centroids are visible. Upon comparing the positions of these centroids for different microorganisms we can observe significant differences.

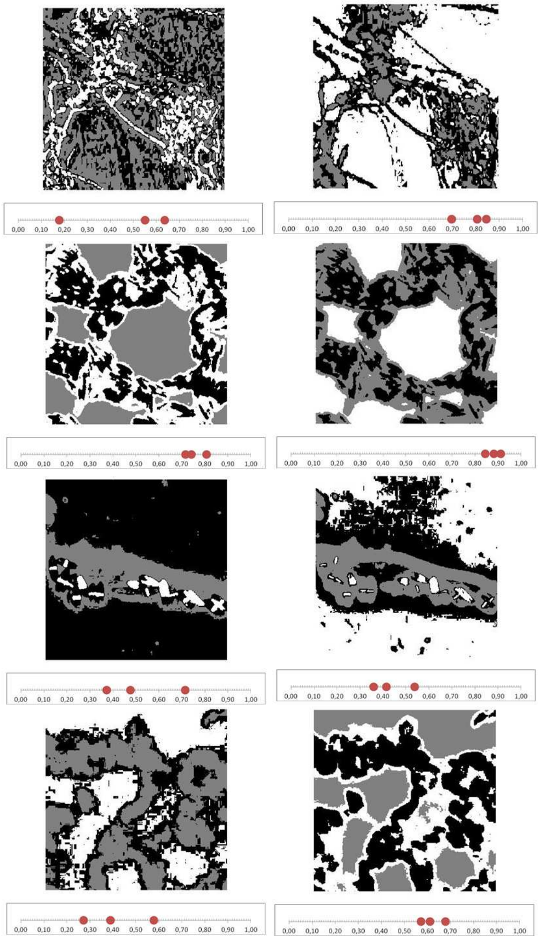


Figure 3: The results of clusterization of the CIELAB a* (left column) and b* (right column) components for 4 classes of microorganisms.

3.2.2 The colorimetric features

The next family of features is defined on the basis of the intensity of the ROI region of the image [6]. The colorimetric features are defined here on 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 use the mean and standard deviation of pixel intensities in the ROI containing microorganisms. They are calculated for each color component in these three color spaces. Up to 18 colorimetric features are created following the above procedure. Table 4 depicts the statistics (the mean values and standard deviation - std) of the colorimetric features of the 441 representations of all investigated classes of microorganisms.

Table 4: Statistics of the colorimetric features of all representations of 12 classes of microorganisms.

	RGB	RGB	LAB	LAB	HSV	HSV
Class	mean	std	mean	std	mean	std
1	0,181622	0,163551	0,523194	0,175936	0,640192	0,305631
2	0,904186	0,068065	0,708979	0,031441	0,463045	0,173982
3	0,045889	0,169714	0,606822	0,067512	0,519432	0,119947
4	0,077446	0,159681	0,78952	0,028751	0,416262	0,137755
5	0,170242	0,165967	0,803036	0,028877	0,397761	0,119995
6	0,163985	0,194397	0,524362	0,140453	0,634613	0,279942
7	0,172967	0,2875	0,780102	0,01125	0,479947	0,15594
8	0,259639	0,137761	0,460677	0,178878	0,839038	0,205619
9	0,902075	0,323865	0,68613	0,020105	0,499872	0,089103
10	0,117994	0,128049	0,393707	0,179513	0,771995	0,179716
11	0,231988	0,087825	0,6407	0,015314	0,724999	0,033775
12	0,279684	0,23601	0,662105	0,030159	0,570373	0,012613

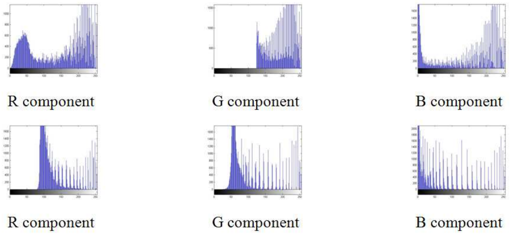

R component G component B component
R component G component B component

Figure 4: The histograms in RGB representation of *Bacillus atrophaeus* (upper row) and *Rachnella aquatilis* (lower row).

3.2.3 Features based on histogram in different color spaces

The next set of features is 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. The typical histograms in RGB representation of 2 classes of microorganisms (*bacillus atrophaeus* and *rachnella aquatilis*) are presented in Fig. 4. The investigations show that microorganisms of different classes are characterized by the histograms of various shapes. On the other hand the histograms of the microorganisms belonging to the same class are similar to each other.

Different numerical features can be defined on the basis of the histograms. In this work we use the mean, variance, skewness, kurtosis, energy and entropy [4,17]. Taking into account that the same image is represented in 3 different color spaces (each characterized by 3 components - RGB, HSV, LAB) we generate thus 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 simply 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.

3.3 Feature selection

The process of feature selection is an important step in developing the efficient procedure of microorganism recognition. A good feature should be characterized by the stable values for samples belonging to the same class and at the same time these values should differ significantly for different classes [10,18]. 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. The most important selectors include: the principal component analysis (PCA) [10,18], Fisher discriminant measure (FD) [4], sparse logistic regression (SLR) [18], correlation feature selection (CFS) [21], information gain (IG) [22], mutual information (MI) [18], statistical independence (SI) [23], fast correlation based filter (FCBF) [23], sparse Bayesian multinomial logistic regularization (SBMLR) [24], etc. On the basis of the numerical experiments we found the FCBF the most efficient and this method is chosen in our approach. It exploits the correlation measure based on the information-theoretical concept of entropy, defined for the variable x in the form

$$H(x) = -\sum_i P(x_i)\log_2(P(x_i))$$

and for variable x after observing the variable y

$$H(x|y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)).$$

The main idea is to select the features which are relevant to the class recognition but at the same time not redundant to any of the other relevant features. The correlation of the features x and y is measured through the information gain $IG(x|y)$, defined as

$$IG(x|y) = H(x) - H(x|y).$$

The feature y is more correlated to feature x than to feature z if $IG(x|y) > IG(z|y)$. To avoid the problem of biasing in favor of specific features with more values the normalized symmetrical uncertainty $SU(x,y)$ is used [23]

$$SU(x,y) = 2\left[\frac{IG(x|y)}{H(x)+H(y)}\right].$$

This measure normalizes the information gain into the range [0, 1]. The value 1 means that the knowledge of the value of either one predicts completely the value of the other one. The value 0 indicates that x and y are independent. The relevance of the feature x to the class c is decided by calculating the proper SU value indicating the symmetrical uncertainty $SU(x,c)$. By assuming a threshold SU value calculated for the feature x and the class c the user can eliminate all features which are below the selected threshold. In a similar way by the analysis of the pairwise correlations between all features the redundant features can be eliminated. In FCBF approach the concept of predominant correlation is applied to accelerate the process of removing the redundant features. In practical application of this algorithm we assume the threshold $SU(x,c)$ equal to 0.64 and $SU(x_i,x_j) = 0.50$. As a result we select 21 features treated as the most important for the class recognition. Table 5 presents the contents of this set, starting from the most and ending on the least important.

Table 5: The contents of the selected features.

No	Description of feature
1	First centroid of cluster of a* component
2	Second centroid of cluster of a* component
3	Second centroid of cluster of b* component
4	Third centroid of cluster of a* component
5	Mean of the blue component
6	Variance of the red component
7	Entropy of the histogram of the green component
8	Entropy of the histogram of the blue component
9	Mean of the hue component (HSV)
10	Variance of the histogram of the hue component (HSV)
11	Energy of the histogram of the hue component (HSV)
12	Entropy of the histogram of the hue component (HSV)
13	Entropy of the histogram of the saturation component (HSV)
14	Mean of the histogram of the luminance component (HSV)
15	Skewness of the histogram of the luminance component (HSV)
16	Entropy of the histogram of the luminance component (HSV)
17	Kurtosis of the histogram of the green component (HSV)
18	Mean of the luminance component (HSV)
19	Energy of the histogram of the saturation component (HSV)
20	Mean of the a* component (L*a*b*)
21	Mean of the b* component (L*a*b*)

3.4 Classification system

The selected features are served as the input attributes to the proposed classifier system. In order to reach the highest efficiency of classification process we try different solutions 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) [7,8,9,10, 25]. The first three belong to the family of neural based solutions. All of them were implemented in Matlab [26]. Multilayer perceptron [9] is a neural network applying the neurons of sigmoidal activation function organized in layers. In our implementation we use one hidden layer and output layer of the number of neurons equal to the number of recognized classes. Each output neuron represents one class of data. The MLP structure is trained using Levenberg-Marquard algorithm [27]. The radial basis function network [9,10] uses Gaussian neurons in the hidden layer and linear neurons on the output, each responsible for one class recognition. The positive value of the output neuron signal means recognition of the particular class. The RBF structure is adapted using the self-organization for finding the Gaussian function parameters and SVD for calculating the output weights. The SVM [7,25] developed by Vapnik is a linear machine, working in the high dimensional feature space formed by the non-linear mapping of the N-dimensional input vector $\mathbf{x}$ into a L-dimensional feature space ($L > N$) through the use of a kernel function $K(\mathbf{x},\mathbf{x}_i)$. The learning problem of SVM is formulated as the task of separating the learning vectors into two classes of the destination values either $d_i = 1$ (one class) or $d_i = -1$ (the opposite class), with the maximal separation margin. The SVM of the Gaussian kernel is used in our application. The hyperparameters (the regularization constant C and Gaussian kernel width) are adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one on the validation data sets. To deal with many classes one against one approach working on a principle of the majority voting [7] is applied. The Breiman random forest [8] is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. To improve the generalization property the randomness in selecting the learning data is implemented. Each tree is grown using randomly selected inputs or combinations inputs at each node. Random forest belongs to the most efficient classification systems. The k nearest neighbor classifier (kNN) belongs to the simplest solutions [9,10], in which learning is merely a question of encapsulating the training data. Classification of the new vector relies on locating its k nearest neighbors and letting the majority vote decide the outcome of the class labelling. The important problem is to choose proper values of k. In our approach it is done by trying different values and accepting one, which provides the best results of classification on the validation data (small portion of the learning data).

4 The results of numerical experiments

The data used in experiments are normalized by dividing each column of data matrix by the maximum value. In order to obtain the most objective results of experiments we apply the k-fold cross validation. In k-fold cross-validation, the original data set is randomly partitioned into k equal size subsets, each containing approximately equal part of data of the same class. Of the k subsets, a single subset is retained as the validation data for testing the model and the remaining $k-1$ are used as the training data. The cross-validation process is then repeated k times (the folds), with each of the k subsets used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. 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 apply 5 mentioned above classifiers. The MLP classifier structure is 21-32-12. In the RBF network we use 27 radial neurons and the final structure is 21-27-12. In the case of SVM networks we use 72 two-class SVM classifiers working in one-against-one mode [7]. The optimal value of regularization parameter C is 1000 and the width of the Gaussian kernel function is 0.7. Random forest (RF) is run in turn using 60% of data for training and the rest for validation. Five out of 21 features are used in each node of decision trees. The best results of kNN classifier is obtained at k=5. To compare the efficiency of different methods of feature selection we performed additional experiments of the image recognition by applying different methods of feature selection: fast correlation based filter (FCBF), sparse logistic regression (SLR), correlation feature selection (CFS), Fisher discriminant measure (FD), sparse Bayesian multinomial logistic regularization (SBMLR), information gain (IG), mutual information (MI) and statistical independence (SI). They are compared to the results of application of all features. The summary of these experiments in the form of 12 class recognition accuracy is presented in Table 6. It shows the mean values of the accuracy of classification by using all mentioned above classifiers in 5-fold cross validation experiments. The numbers in bold represent the best result for the particular selection method. As we can see the highest accuracy of recognition corresponds to the application of FCBF method of feature selection.

Table 6: The accuracy of recognition of 12 classes of microorganisms at application of different methods of feature selection (all values in percent).

Classifier	All features	FCBF	SLR	CFS	FD	SBMLR	IG	MI	SI
SVM	89.80	98.19	92,17	97,34	96,43	93,45	92,79	91,28	91,87
RBF	87.30	94,78	91,23	93,26	91,42	91,39	90,93	91,04	90,22
MLP	83.80	92,97	87,74	92,97	89,12	85,38	86,37	86,29	85,11
RF	89.57	98,41	92,17	97,78	95,31	94,56	93,14	92,41	92,11
kNN	86.58	95,24	90,53	94,67	92,17	88,74	87,35	86,98	87,09

We have repeated the same experiments by including the 13th mixed class composed of 5 other species, mentioned in section 2. The numerical results of recognition in this case are shown in Table 7.

Table 7: The accuracy of recognition of 12 classes of microorganisms in the presence of the additional mixed class of microorganisms.

Classifier	All features	FCBF	SLR	CFS	FD	SBMLR	IG	MI	SI
SVM	87,1	97,56	90,98	96,81	96,11	91,54	91,82	89,31	90,18
RBF	86,21	94,23	89,58	91,69	89,23	90,3	90,19	89,97	89,34
MLP	81,26	91,13	83,42	89,14	87,14	84,19	85,91	86,02	84,78
RF	89,91	97,96	93,2	96,97	96,03	94,17	92,16	91,18	91,95
kNN	85,32	94,79	88,63	93,9	90,89	87,73	87,85	85,68	85,19

In this particular case the class recognition accuracy is only slightly worse to the previous case. However, once again the best is the FCBF selection method. The other point of consideration is the misclassification problem existing among the classes. We present it in the form of confusion matrix [10] depicted as the actual number of data belonging to different classes in the process of recognition. The rows represent the real class membership of the data and the columns - the respective results of classification. The diagonal entries $(i = j)$ depict the number of properly recognized samples belonging to the ith class. Each entry outside the diagonal represents the misclassified case. The entry in the (i, j)-th position of the matrix means false assignment of ith class to the jth one. The confusion matrix is prepared for the data used only in testing. We show first the results corresponding to the best (SVM) learned system for all input attributes (Table 8) and then for the selected set of them (Table 9 and 10). The results in Table 9 are related to the recognition of 12 classes and in Table 10 to 13 classes, both at application of the best RF classifier.

Table 8: The confusion matrix of the best classifier (SVM) for all features without selection.

Classes	1	2	3	4	5	6	7	8	9	10	11	12
1	10	0	1	0	0	1	0	0	0	0	0	0
2	0	48	0	1	0	0	0	0	0	0	0	0
3	0	1	17	0	0	1	0	0	0	0	0	0
4	0	0	2	40	0	3	1	0	1	0	1	0
5	0	0	1	0	44	0	1	0	1	1	0	0
6	0	0	0	0	0	16	0	1	0	0	0	0
7	0	0	0	1	2	1	51	0	2	0	1	1
8	0	0	0	0	0	0	0	21	0	0	0	0
9	1	0	0	0	0	0	0	0	13	0	1	0
10	0	0	1	0	1	0	0	2	0	38	0	1
11	0	0	2	1	1	0	2	0	1	0	71	0
12	0	1	0	1	0	1	1	0	0	1	0	27

Table 9: The confusion matrix of the best classifier for the set of features selected by FCBF in recognition of 12 classes.

Classes	1	2	3	4	5	6	7	8	9	10	11	12
1	12	0	0	0	0	0	0	0	0	0	0	0
2	0	49	0	0	0	0	0	0	0	0	0	0
3	0	1	17	0	0	1	0	0	0	0	0	0
4	0	0	0	47	0	0	0	0	0	0	1	0
5	0	0	0	0	48	0	0	0	0	0	0	0
6	0	0	0	0	0	16	0	1	0	0	0	0
7	0	0	0	0	0	0	58	0	0	0	0	1
8	0	0	0	0	0	0	0	21	0	0	0	0
9	0	0	0	0	0	0	0	0	15	0	0	0
10	0	0	0	0	0	0	0	1	0	42	0	0
11	0	0	0	0	0	0	0	0	1	0	77	0
12	0	0	0	0	0	0	0	0	0	1	0	31

Table 10: The confusion matrix of the best classifier for the set of features selected by FCBF in recognition of 13 classes.

Classes	1	2	3	4	5	6	7	8	9	10	11	12	13
1	12	0	0	0	0	0	0	0	0	0	0	0	0
2	0	49	0	0	0	0	0	0	0	0	0	0	0
3	0	0	17	0	0	1	0	0	0	0	0	0	1
4	0	0	0	47	0	0	0	0	0	0	1	0	0
5	0	0	0	0	48	0	0	0	0	0	0	0	0
6	0	0	0	0	0	16	0	1	0	0	0	0	0
7	0	0	0	0	0	0	58	0	0	0	0	1	0
8	0	0	0	0	0	0	0	21	0	0	0	0	0
9	0	0	0	0	0	0	0	0	15	0	0	0	0
10	0	0	0	0	0	0	0	1	0	42	0	0	0
11	0	0	0	0	0	0	0	0	1	0	77	0	0
12	0	0	0	0	0	0	0	0	0	1	0	31	0
13	0	0	0	1	0	0	1	0	0	0	0	0	48

In the case of the selected features only single elements are located outside the main diagonal and the matrix is sparse. On the other hand, the application of all generated features as the input attributes resulted in many misclassifications (the matrix is far from diagonal). This fact confirms the superiority of feature selection in the classification procedure. Inclusion of the 13th class does not change the results in a significant way. There is a visible similarity of these two matrices. Only one sample of the third class is recognized as the mixed one. On the other hand two samples of the mixed class are wrongly recognized as the fourth and the seventh class.

5 Conclusion

This paper presents 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 is verified successfully in case of the recognition of 12 classes of microorganisms with the addition of the 13th mixed class. Five different solutions of the classifiers were tried. Irrespectively to the number of recognized classes the obtained accuracy is close to 98% for the data base containing 491 images. The results of the paper may be applicable 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 permits to alleviate 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 the continuation of the further study. In future work we intend to extend the solution to counting the number of individual microorganisms appearing 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 and Sustainability of Tropical Agriculture. Eds. K Mulongoy, M. Gueye and D. S. C. Spencer, 1992, John Wiley and Sons, Chichester, UK, pp. 27-40.

[3] S.K.A. Danso, G.D. Bowen, N. Sanginga, Plant and Soil **141**: 177-196, 1992

[4] R.O. Duda, P.E. Hart, P. Stork, Pattern Classification and Scene Analysis, Wiley, New York, 2003.

[5] B. S. Everitt, S. Landau, M. Leese, D. Stahl, Miscellaneous Clustering Methods in Cluster Analysis, Wiley & Sons, Chichester, 2011.

[6] R. Gonzales and R. Woods, Digital Image Processing, Prentice Hall, New Jersey, 2008.

[7] B. Schölkopf and A. Smola Learning with Kernels, MIT Press, Cambridge MA, 2002.

[8] L. Breiman, Machine Learning, 2001, **45**, No 11, pp. 532.

[9] S. Haykin, Neural Networks, a Comprehensive Foundation, Macmillan College Publishing Company, New York, 2000.

[10] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Education Inc., Boston, 2006.

[11] M. G. Forero, G. Cristbal, M. Desco, Journal of Microscopy, 2006, **223**, pp. 120-132.

[12] J. Liu, F. B. Dazzo, O. Glagoleva, B. Yu, A.K. Jain, Microbial Ecology, 2001, **41**, No 3, pp. 173-194.

[13] P. Ruusuvuori, J. Sepp, T. Erkkil, A. Lehmussola, J. A. Puhakka, O. Yli-Harja, Proceedings of the Nineteenth International Conference on Pattern Recognition, 2008, pp. 14.

[14] D.H. Theriault, M.L. Walker, J.Y. Wong, M. Betke, Machine Vision and Applications, 2012, **23**, pp. 659-673.

[15] P.S. Hiremath and P. Bannigidad, International Journal of Computational Biology and Drug Design, 2011, **4**, No 3, pp. 262-273.

[16] N. Otsu, IEEE Transactions on Systems, Man and Cybernetics, 1979, **9**, No. 1, pp. 62-66.

[17] M. Kruk, S. Osowski, R. Koktysz, Computers in Biology and Medicine, 2009, **39**, pp. 156-165.

[18] A. Guyon, Journal of Machine Learning Research, 2003, **3**, pp. 1158-1182.

[19] R. Hunter, Journal of the Optical Society of America, 1948, **38**, No 7, pp. 661 (Proceedings of the Thirty-Third Annual Meeting of the Optical Society of America).

[20] J. A. Hartigan and M. A. Wong, Applied Statistics, 1979, **28**, No. 1, pp. 100-108.

[21] M. A. Hall and L. A. Smith. Proceedings of FLAIRS Conference, pp. 235-239, 1999.

[22] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, 1991.

[23] H. Liu and L. Yu, Proceedings of the Twentieth International Conference on Machine Leaning (ICML-03), pages 856-863, Washington, D.C., 2003. ICM.

[24] C. G. Cawley, N. L. C. Talbot, M. Girolami, In NIPS, pp. 209-216, 2006.

[25] V. Vapnik, Statistical Learning Theory, New York: Wiley, 1998.

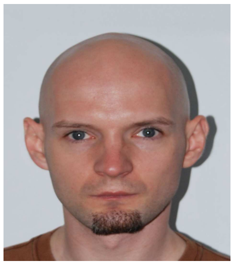
[26] Matlab User Manual Image Processing Toolbox, MathWorks, Natick, 2012.

[27] C. T. Kelley, Iterative Methods for Optimization, SIAM Frontiers in Applied Mathematics, Philadelphia, 1999

Michał Kruk is an adjunct at the Faculty of Applied Informatics and Mathematics at Warsaw University of Life Sciences - SGGW. He received the Ph.D. degree at Warsaw Univeristy of Technology (2008). His research interests are in the areas of image processing, artificial intelligence and data exploration. The main part of interest contains biomedical image processing especially building automatic systems applied in medical diagnostics in cancer recognition. More information can be found under M. Kruk's homepage (see URL: http://www.michalkruk.pl)

Ryszard Kozera is a Professor at the Faculty of Applied Informatics and Mathematics at Warsaw University of Life Sciences - SGGW. He is also the Adjunct Associate Professor at the School of Information Science and Software Engineering at The University of Western Australia in Perth. He received his M.Sc. degree in Pure Mathematics at Warsaw University, Poland (1985), a Ph.D. degree in Computer Science at Flinders University of South Australia, Adelaide, Australia (1991) and his D.Sc. degree in Computer Science (a Habilitation) at The Silesian University of Technology, Gliwice, Poland (2006). Professor Kozera was awarded three times Alexander von Humboldt Fellowships between 1995-2004. He delivered about 60 seminars at various prestigious international research centers, published 79 refereed journal and conference papers, organized international conferences and participated in the collaborative international grants. He has 21 postgraduate supervisions completed including Ph.D. and M.Sc. students. Professor Kozera's research interests include computer vision, numerical analysis, optimization, interpolation and approximation, partial differential equations and artificial intelligence. More information can be found under Professor Kozera's homepage (see URL: http://www.wzim.sggw.pl/ryszard_kozera or URL: http://www.csse.uwa.edu.au/˜ryszard).

Stanisław Osowski is a full professor of Electrical Engineering at Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Faculty of Electrical Engineering, Warsaw University of Technology (WUT). His academic record: 1972 - MSc Thesis at WUT - specialization Automatics, 1975 - Ph. D. Thesis at WUT: ”The algorithms of the investigations of the nonhomogenous ladder networks and long transmission lines”, 1981 D.Sc. (habilitation) at WUT: ”Minimization of the number of operational amplifiers in the synthesis of the multiport networks”, 1990 - Professor at WUT, 1995 - Professor title. Stanislaw Osowski is also member of IEEE (Section Computational Intelligence) - No 40346158, member of the Section of Electronic Signal and Systems, Committee on Electronics and Telecommunication, Polish Academy of Sciences and member of the Section of Theoretical Electrotechnics, Polish Academy of Sciences.



Paweł Trzciński is an Ph.D student at the Rhizosphere Laboratory at Institute of Horticulture in Skierniewice (Poland). His research is focused on plant-bacteria and interaction and identification of soil bacteria and fungi. His primary aim is isolation and identification of plant growth promoting microorganisms.

Lidia Sas-Paszt Current appointment: senior researcher, Head of Rhizosphere Laboratory, Research Institute of Horticulture (IO), Skierniewice, Poland. Research interests: mineral nutrition of horticultural plants, root & rhizosphere research, sustainable cultivation technologies, bio-friendly nutrient management strategies, physiology and biochemistry of fruit ripening, plant in vitro cultures. Author of more than 60 scientific publications. Experience in international cooperation: Coordinator or investigator of more than 12 European and national research projects, e.g. technical coordinator of the European CRAFT Project (Ensuring the quality of innovative crop growth inputs derived from biological raw materials, 2004-2006) and leader of the project financed from structural funds the Innovative Economy Program (Development of innovative products and technologies for organic fruit production, 2009-2014). Member of the Management Committee and Working Group, COST Action 631 Understanding and Modelling Plant-Soil Interactions in the Rhizosphere Environment. Member of Working Group of COST Action 836 Organization of the Integrated Research in Berries. Member of the Management Committee and Working Group of COST Action E38 Woody Root Processes. Expert in evaluation of EC project proposals in FP6, FP7 and Horizon 2020. Expert of the EC Programme Committee on Food Quality and Safety in FP6. Expert of the EC Programme Committee on Food, Agriculture and Fisheries, and Biotechnology in FP7. Expert to EFSA (European Food Safety Authority).

Bolesław Borkowski is a full professor of economy at the Faculty of Applied of Informatics and Mathematics Sciences at Warsaw University of Life Sciences-SGGW. He was first dean of this faculty. He is also a visiting Professor at the Department of Operations Research in Management at Warsaw University. Main fields of his scientific interests are mathematical methods in economy, especially in economics agricultural. He is an author of 110 scientific publications of which 11 books. Conjuration of econometrical models and methods of their parameters estimation are the main points of interest of the author. Prognosis of financial instruments prices and modeling of risk management aspects at farm level are of great interest to prof. Borkowski

Beata Sumorok is an adjunct at the Research Institute of Horticulture, Rizosphere Laboratory, Department of Microbiology. She received the Ph.D. degree at University of Lodz (2001). She conducts research on the biodiversity of arbuscular mycorrhizal fungi in fruit crops, especially the research on the role of arbuscular mycorrhizal fungi in the rhizosphere of plants, development of microbial inocula for horticultural production, identification of arbuscular mycorrhizal fungi with the use of classical methods. More information can be found under Rizosphere Laboratory homepage (see http://www.inhort.pl/pracowniarizosfery.html)