Classification of Soil Bacteria Based on Machine Learning and Image Processing

Aleksandra Konopka[1[0000−0003−1730−5866]], Karol Struniawski[1[0000−0002−4574−2986]], Ryszard Kozera[1,2[0000−0002−2907−8632]], Paweł Trzciński[3[0000−0003−4568−6504]], Lidia Sas-Paszt[3[0000−0003−4076−4032]], Anna Lisek[3[0000−0002−3421−8759]], Krzysztof Górnik[3[0000−0002−6612−6779]], Edyta Derkowska[3[0000−0003−4108−336X]], Sławomir Głuszek[3[0000−0003−3158−6828]], Beata Sumorok[3[0000−0001−7688−2570]], and Magdalena Frąc[4[0000−0001−9437−3139]]

[1] Institute of Information Technology, Warsaw University of Life Sciences - SGGW, ul. Nowoursynowska 159, 02-776 Warsaw, Poland
{aleksandra_konopka, karol_struniawski, ryszard_kozera}@sggw.edu.pl
[2] School of Physics, Mathematics and Computing, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Perth, Australia
ryszard.kozera@uwa.edu.au
[3] Department of Microbiology and Rhizosphere, The National Institute of Horticultural Research, ul. Pomologiczna 18, 96-100 Skierniewice, Poland
{pawel.trzcinski, lidia.sas, anna.lisek, krzysztof.gornik, edyta.derkowska, slawomir.gluszek, beata.sumorok}@inhort.pl
[4] Institute of Agrophysics, Polish Academy of Sciences, ul. Doświadczalna 4, 20-290 Lublin, Poland
m.frac@ipan.lublin.pl

Abstract. Soil bacteria play a fundamental role in plant growth. This paper focuses on developing and testing some techniques designed to identify automatically such microorganisms. More specifically, the recognition performed here deals with the specific five genera of soil bacteria. Their microscopic images are classified with machine learning methods using shape and image texture descriptors. Feature determination based on shape relies on interpolation and curvature estimation whereas feature recognition based on image texture resorts to the spatial relationships between chrominance and luminance of pixels using co-occurrence matrices. From the variety of modelling methods applied here the best reported result amounts to 97% of accuracy. This outcome is obtained upon incorporating the set of features from both groups and subsequently merging classification and feature selection methods: Extreme Learning Machine - Radial Basis Function with Sparse Multinomial Logistic Regression with Bayesian Regularization and also k-Nearest Neighbors classifier with Fast Correlation Based Filter. The optimal parameters involved in merged classifiers are obtained upon computational testing and simulation.

Keywords: Soil bacteria · Machine learning · Image analysis · Shape and image texture extraction · Spline interpolation· Modelling and simulation· Computational optimization

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

1 Introduction

Soil bacteria despite of their small size may have a large impact on plant growth. Some of them are beneficial to agricultural sector, while the others are either harmless or pathogenic causing a vast diversity of plant diseases. Consequently, bacteria recognition becomes an important task for scientists equally as a research and agricultural problem. Bacteria identification is usually carried out using specific markers changing their color as a reaction to specific chemical compounds. The morphology of the bacteria colony is also usually analyzed by examining its shape, edges, color, colony distribution, consistency and surface structure [7]. This approach is usually laborious and depends on the subjective perceptiveness of the scientist. A natural step accelerating and facilitating the latter is to automate the process of microscopic image analysis. This paper[1,2] resorts to machine learning and image processing methods applied to soil bacteria recognition. In general, comparing images of bacteria belonging to certain species is difficult since they adopt similar morphologies [22]. Due to this reason, it is decided to distinguish here the input bacteria on the genera level. The microscopic images of soil bacteria examined in this paper (which are part of our data-set available in full resolution under the URL link: https://bit.ly/3qdDuHo) include pictures of *Enterobacter*, *Rhizobium*, *Pantoea*, *Bradyrhizobium* and *Pseudomonas* (see Fig. 1). The pictures of investigated bacteria are obtained from Symbio-Bank - the collection of microorganisms of The National Institute of Horticultural Research in Skierniewice. Some of *Enterobacter* are considered plant pathogens, whereas the others are conducive for plant growth [13]. The bacteria of the genus *Rhizobium* have a positive effect on increasing the yield of grains and the protein content in pea grains [23]. *Rhizobium* and *Bradyrhizobium* are nitrogen-fixing soil bacteria that live in symbiosis with legumes [3]. On the other hand, *Pantoea* causing plant infections [24] is also used in the production of antibiotics [2]. Some *Pseudomonas* are plant pathogens, while the others are used to stimulate plant growth and to remediate contaminated soil [30]. This paper discusses the identification of bacteria genera based on their morphological features. The calculated traits refer to bacteria shape and image texture. In order to automatize the entire recognition process a variety of feature selection and class recognition methods adapting the concept of machine learning are applied. On the basis of the supplied training data-set a classification model is built permitting to automatically categorize soil microorganisms.

2 Work-flow Scheme

The work-flow scheme adopted in this work consists of the following four consecutive steps: *segmentation of the Region of Interest, feature generation, feature selection* and *class recognition*.

[1] This research is financed by The National Centre for Research and Development of the BIOSTRATEG Project (Eco-Fruits) BIOSTRATEG3/344433/16/NCBR/2018.
[2] This work is a part of Polish National Centre of Research and Development research project POIR.01.02.00-00-0160/20.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

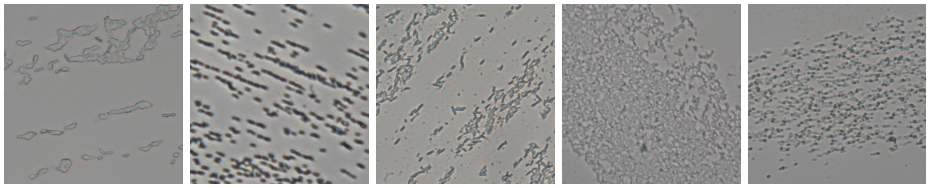

Fig. 1. Microscopic images of (from left): *Enterobacter*, *Rhizobium*, *Pantoea*, *Bradyrhizobium* and *Pseudomonas*. For more pictures see URL link: https://bit.ly/3qdDuHo.

2.1 Segmentation of the Region of Interest

The aim of this step is to extract bacteria and background image regions. Binary mask filter is applied yielding white pixels representing bacteria zones and black pixels corresponding to the background. To achieve the latter the image is first converted into gray-scale and then Otsu automatic image thresholding [18] with open and close morphological operations [28] is applied (see Fig. 2).

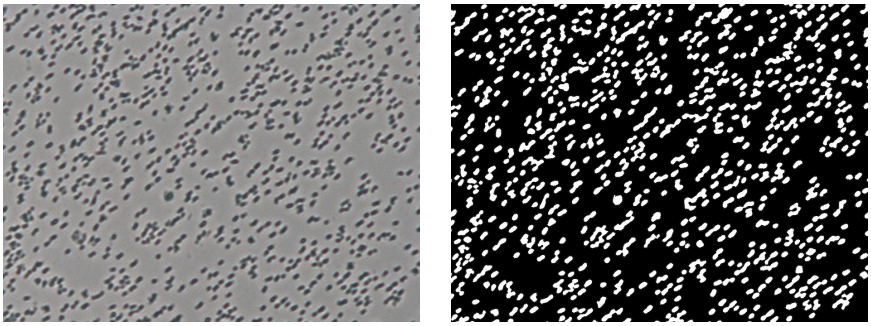

Fig. 2. Microscopic image of *Rhizobium* and its binary mask.

2.2 Feature Generation

Features that are considered in this paper refer to bacteria shape and texture of the input image. The determination of bacteria shape relies on estimating its boundary with the aid of cubic spline interpolation [5]. The latter permits to estimate the curvature of bacteria's boundary and to extract some correlation between selected distances and angles concerning the shape of bacteria in question. On the other hand, image texture features contain information about spatial relations between chrominance and luminance of the image pixels. To exploit such information, a statistical approach based on computation of the co-occurrence matrices is used [27]. The latter permits to estimate an image texture as a quantitative measure of luminance over the entire input image.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

2.3 Feature Selection

It is common that processing large set of generated features may yield some of them highly correlated with one another. Such potential redundancy usually impacts on the classification accuracy. In contrast, the other group of extracted features can be poorly correlated to the dependent variable affiliated to the respective class. Consequently, the reduced set of selected features (sifted from the full set of initially determined features) should consist of those which are strongly correlated to the image class and weakly associated with the remaining features. In order to accomplish the latter the following methods for improving feature selection are used here: FCBF (*Fast Correlation Based Filter*), SBMLR (*Sparse Multinomial Logistic Regression with Bayesian Regularization*) and CFS (*Correlation-based Feature Selection*) - see e.g. [21].

2.4 Class Recognition

This paper resorts to the machine learning classifiers such as RF (*Random Forest*), SVM (*Support Vector Machine*), kNN (*k-Nearest Neighbors*), MLP (*Multilayer Perceptron*) [17], ELM (*Extreme Learning Machine*) [26] and ELM-RBF (*Extreme Learning Machine - Radial Basis Function*) [10].

3 Features Based on Shape

3.1 Planar Cubic Spline Interpolation

Consider now the ordered set of $m+1$ planar points $\mathcal{Q}_m = \{q_k\}_{k=0}^m$ i.e. sequence of points $q_k = (x_k, y_k)$ contained in 2D-Euclidean space $\mathbb{E}^2$. In the context of this work $\mathcal{Q}_m$ represents selected points of bacteria's boundary $\partial\Gamma$. In a quest to extract some shape information of $\partial\Gamma$ (or its estimate) an interpolation based approach is applied here [5]. In the classical setting of fitting input data, $\mathcal{Q}_m$ is also supplemented with the associated parameters, called interpolation knots $\mathcal{T}_m = \{t_k\}_{k=0}^m$ subject to $t_k < t_{k+1}$, $t_0 = 0$, $t_m = T$ and $t_k \in [0, T]$. Here the unknown function $\gamma : [0, T] \to \mathbb{E}^2$ meeting the constraints $\gamma(t_k) = q_k$ is assumed to satisfy $graph(\gamma) = \partial\Gamma$. For a given pair $(\mathcal{Q}_m, \mathcal{T}_m)$ there is a variety of interpolation schemes $\gamma_I : [0, T] \to \mathbb{E}^2$ fulfilling $\gamma_I(t_k) = q_k$ - see e.g. [5] or [9]. Since the selected interpolant γ_I to fit $(\mathcal{Q}_m, \mathcal{T}_m)$ should be both twice-differentiable (for curvature calculation) and should not render too excessive variations of $graph(\gamma_I)$ (for arbitrary m) a cubic spline $\gamma_I = \gamma^{cs}$ is a natural choice [5]. The interpolant γ^{cs} is defined as a track-sum of cubics $\{\gamma_k^{cs}\}_{k=0}^{m-1}$ with each cubic $\gamma_k^{cs} : [t_k, t_{k+1}] \to \mathbb{E}^2$ depending on four 2D-parameters (as $a_k, b_k, c_k, d_k \in \mathbb{R}^2$)

$$\gamma_k^{cs}(t) = a_k + b_k(t - t_k) + c_k(t - t_k)^2 + d_k(t - t_k)^3. \tag{1}$$

Here $4 \times m$ coefficients $\{(a_k, b_k, c_k, d_k)\}_{k=0}^{m-1}$ are calculable from $4 \times m$ constraints:

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

1. $2 \times m$ interpolation conditions for $k \in \{0, 1, \ldots, m-1\}$:

$$\gamma_k^{cs}(t_k) = q_k \quad \text{and} \quad \gamma_k^{cs}(t_{k+1}) = q_{k+1}. \tag{2}$$

2. $m-1$ internal points' first-order smoothness for $k \in \{0, 1, \ldots, m-2\}$:

$$\dot{\gamma}_k^{cs}(t_{k+1}) = \dot{\gamma}_{k+1}^{cs}(t_{k+1}). \tag{3}$$

3. $m-1$ internal points' second-order smoothness for $k \in \{0, 1, \ldots, m-2\}$:

$$\ddot{\gamma}_k^{cs}(t_{k+1}) = \ddot{\gamma}_{k+1}^{cs}(t_{k+1}). \tag{4}$$

4. 2 boundary conditions complementing (2), (3), (4) to yield $4 \times m$ equations.

Usually, the last two equations are obtainable from extra conditions such as e.g. $\dot{\gamma}(0) = v_0$ and $\dot{\gamma}(T) = v_m$. Indeed, the latter yields two missing equations:

$$\dot{\gamma}_0^{cs}(t_0 = 0) = v_0 \quad \text{and} \quad \dot{\gamma}_{m-1}^{cs}(t_m = T) = v_m. \tag{5}$$

Although, in our setting both velocities v_0 and v_m are not *a priori* given, they can be still estimated from $(\mathcal{Q}_m, \mathcal{T}_m)$ following the concept of modified Hermite interpolation [14]. Indeed, a unique Lagrange cubic $\gamma_0^{L(3)} : [0, t_3] \to \mathbb{E}^2$ interpolating the first four points $\{q_k\}_{k=0}^3$ at $\{t_k\}_{k=0}^3$ (see [5]) yields some estimate of $v_0 \approx \hat{v}_0 = \dot{\gamma}_3^{L(3)}(0)$. Similarly, a Lagrange cubic $\gamma_{m-3}^{L(3)} : [t_{m-3}, t_m] \to \mathbb{E}^2$ interpolating the last four points $\{q_k\}_{k=m-3}^m$ at $\{t_k\}_{k=m-3}^m$ renders some approximation of terminal velocity $v_m \approx \hat{v}_m = \dot{\gamma}_{m-3}^{L(3)}(t_m)$. Consequently, taking into account the latter, condition (5) modifies into:

$$\dot{\gamma}_0^{cs}(t_0 = 0) = \hat{v}_0 \quad \text{and} \quad \dot{\gamma}_{m-1}^{cs}(t_m = T) = \hat{v}_m. \tag{6}$$

The scheme for selection $\mathcal{Q}_m$ from $\partial\Gamma$ is described in subsection 3.3. Note that in our setting to approximate $\partial\Gamma$ with the closed curve as $q_0 \neq q_m$ we extend $\mathcal{Q}_m$ to $\hat{\mathcal{Q}}_{m+1} = \{\hat{q}_k\}_{k=0}^{m+1}$ so that $\hat{q}_k = q_k$ (for $k = 0, 1, \ldots, m$) and $\hat{q}_{m+1} = q_0$.

Upon selecting the interpolation points $\mathcal{Q}_m$ (and thus $\hat{\mathcal{Q}}_{m+1}$) from the bacteria's boundary $\partial\Gamma$ the next step is to estimate the accompanying knots $\mathcal{T}_{m+1} \approx \hat{\mathcal{T}}_{m+1} = \{\hat{t}_k\}_{k=0}^{m+1}$ (as $\mathcal{T}_{m+1}$ is not available out of input images) from the distribution of $\hat{\mathcal{Q}}_{m+1}$. This permits to construct the interpolant $\hat{\gamma}^{cs} : [0, \hat{T}] \to \mathbb{E}^2$ as a track-sum $\hat{\gamma}^{cs} = \{\hat{\gamma}_k^{cs}\}_{k=0}^{m+1}$, with $\hat{\gamma}_k^{cs} : [\hat{t}_k, \hat{t}_{k+1}] \to \mathbb{E}^2$ satisfying (1), (2), (3), (4) and (6) along $\hat{\mathcal{Q}}_{m+1}$ with somehow estimated knots $\hat{\mathcal{T}}_{m+1}$. Addressing the latter, we resort here to the so-called *exponential parameterization* commonly used in computer graphics [19] and defined in accordance with:

$$\hat{t}_0 = 0, \qquad \hat{t}_{k+1} = \hat{t}_k + \|q_{k+1} - q_k\|^\lambda, \quad k = 0, 1, \ldots, m, \tag{7}$$

for some parameter $\lambda \in [0, 1]$, where $\|\cdot\|$ is a standard Euclidean norm. This paper selects $\lambda = 0.5$ in (7) yielding the so-called *centripetal parameterization* with $\hat{T} = \sum_{k=0}^m \|q_{k+1} - q_k\|^{1/2}$ (see [19]). Note that in order to preserve $t_k < t_{k+1}$ it is also assumed that $q_k \neq q_{k+1}$. More information on exponential parameterization (7) and other knots selection schemes can be found e.g. in [14–16, 19].

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

3.2 Curvature Calculation

Having found a cubic spline $\hat{\gamma}^{cs}$ approximating bacteria's boundary $\partial\Gamma$ one may extract some shape information of $\partial\Gamma$ by analyzing the geometry of $graph(\hat{\gamma}^{cs})$ forming the planar curve assumed also to estimate $\partial\Gamma$ (for m sufficiently big). In this work a curvature of $\hat{\gamma}^{cs}$ is computed to form a geometrical marker of $\partial\Gamma$ used later as one of the differentiating ingredients in classification process. Recall, that the curvature $\kappa(t)$ of a planar curve $\gamma : [a, b] \to \mathbb{E}^2$ at a given point $t \in [a, b]$ measures the amount by which such curve deviates from a tangent line to the curve at point $\gamma(t)$ - see [29]. The respective formula for the curvature $\kappa(t)$ of regular curve γ (i.e. for γ for which $\dot{\gamma}(t) \neq \mathbf{0}$ over $t \in [a, b]$) reads as:

$$\kappa(t) = \frac{\|\boldsymbol{T}'(t)\|}{\|\boldsymbol{r}\,'(t)\|}, \tag{8}$$

where $\boldsymbol{r}(t) = \dot{\gamma}(t)$ is a tangent vector to γ at t with its normalized vector $\boldsymbol{T}(t) = \boldsymbol{r}(t)/\|\boldsymbol{r}(t)\|$. In particular, for arc-length parameterization expressed as $s = \phi(t) = \int_a^t \|\boldsymbol{r}\,'(u)\|\, du$, for which reparameterized curve $\bar{\gamma}(s) = (\gamma \circ \phi^{-1})(s)$ satisfies $\|\dot{\bar{\gamma}}(s)\| = 1$ (yielding $\|\boldsymbol{T}(s)\| = 1$ - see [8]), the equation (8) reformulates into (with the respective derivative calculated for s-variable) $\kappa(s) = \|\boldsymbol{T}'(s)\|$.

3.3 Features Calculation

To estimate bacteria's shape (assumed here to be "more or less" convex), Region of Interest (ROI) mask is applied. In doing so, the following Matlab functions are exploited: *rgb2gray, imbinarize, imfill, bwareaopen* and *multithresh*. Upon localizing a single bacteria with ROI mask, the Laplacian filter is used to extract $\partial\Gamma$ of the analyzed object [4]. Next all computed boundary points $\mathcal{Q}_{\hat{m}} = \{q_j\}_{j=0}^{\hat{m}}$ are sorted out clock-wisely. To achieve the latter, we calculate and compare the angle between a given point $q_j \in \mathcal{Q}_{\hat{m}}$ and mean location of $\mathcal{Q}_{\hat{m}}$ i.e. the point $(\bar{x} = (1/(\hat{m}+1)) \sum_{k=0}^{\hat{m}} x_k, \bar{y} = (1/(\hat{m}+1)) \sum_{k=0}^{\hat{m}} y_k)$. It is assumed here that (x_k, y_k) represent Cartesian coordinates of the centers of bacterial boundary pixels (the center of the coordinate system is set in the upper left corner of an image). As a result the boundary of each single bacteria $\partial\Gamma$ is represented by a large set of points $\mathcal{Q}_{\hat{m}}$ which in turn is reduced to terser set $\mathcal{Q}_m$ with $m+1 = 10$ or $m+1 = 20$ points (and thus to $\hat{\mathcal{Q}}_{m+1}$ - see subsection 3.1) to be fitted with $\hat{\gamma}^{cs}$ and $\hat{\mathcal{T}}$ governed by (7). Such reduction is carried out upon selecting from $\mathcal{Q}_{\hat{m}}$ "more or less" equally spaced points with respect to their index distribution taken in clockwise order (e.g. for $\hat{m} = 54$ a possible reduction leads to $\{q_0, q_8, q_{17}, q_{26}, q_{35}, q_{44}, q_{54}\}$).

The feature extraction process aimed to determine some bacteria's shape information relies on curvature calculation from the estimated bacteria's boundary $\partial\Gamma$ - see [1]. In doing so, formula (8) is applied to cubic spline $\hat{\gamma}_k^{cs}$ (see subsection 3.1). More precisely, with the aid of (8) for each $\hat{\gamma}_k^{cs}$ we compute over $[\hat{t}_k, \hat{t}_{k+1}]$ the maximal and minimal values of the curvature function $\kappa(\hat{t})$ (i.e. κ^{max} and κ^{min}) yielding the corresponding knots $\hat{t}_k^{max}, \hat{t}_k^{min} \in [\hat{t}_k, \hat{t}_{k+1}]$ obtained from

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

$\kappa^{max} = \kappa(\hat{t}_k^{max})$ and $\kappa^{min} = \kappa(\hat{t}_k^{min})$, respectively. Note that if the pair of knots $(\hat{t}_k^{max}, \hat{t}_k^{min})$ is not uniquely determined one can choose e.g. the smallest two knots $t_k^{max,min} \in [\hat{t}_k, \hat{t}_{k+1}]$, respectively. This in turn, permits to determine two points $q_k^{max} = \hat{\gamma}_k^{cs}(\hat{t}_k^{max})$ and $q_k^{min} = \hat{\gamma}_k^{cs}(\hat{t}_k^{min})$ having maximal and minimal curvature $\hat{\gamma}_k^{cs}$ over the segment $[\hat{t}_k, \hat{t}_{k+1}]$. According to the order of all knots $\hat{t}_k^{max}$, $\hat{t}_k^{min}$ we re-index points q_k^{min} and q_k^{max} placing them into one sequence formula $\{q_k^{ext}\}_{k=0}^{2m+1}$ (where either $ext = max$ or $ext = min$). In the next step we determine the center of mass $q_c = (1/2(m+1)) \sum_{k=0}^{2m+1} q_k^{ext}$ needed as a reference point to compute the $2(m+1)$ distance values $a_k = \|q_k^{ext} - q_c\|$. In sequel, a family of triangles $\Delta_k(q_k^{ext}, q_c, q_{k+1}^{ext})$ with common apex at q_c (with the respective lengths of Δ_k sides: a_k, $b_k = \left\|q_{k+1}^{ext} - q_k^{ext}\right\|$ and a_{k+1}) form a polygonal approximation of $\partial\Gamma$. Given the lengths of all sides of triangle Δ_k, its respective angles $\alpha_k, \beta_k, \gamma_k = \pi - (\alpha_k + \beta_k)$ are easily computable from the cosine theorem - here $\alpha_k = \measuredangle(q_k^{ext}, q_c, q_{k+1}^{ext})$, $\beta_k = \measuredangle(q_k^{ext}, q_{k+1}^{ext}, q_c)$ and $\gamma_k = \measuredangle(q_{k+1}^{ext}, q_k^{ext}, q_c)$. Having determined the above distances and angles a given bacteria can be represented by the following four vectors (forming de facto its *polygonal shape descriptors*): $\boldsymbol{a} = (a_0, \ldots, a_{2m+1})$, $\boldsymbol{b} = (b_0, \ldots, b_{2m+1})$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{2m+1})$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{2m+1})$. At this point, we assume that we are given a reference bacteria (a kind of "geodetic benchmark" not necessarily belonging to any investigated herein soil microorganisms' classes) to which different five classes of examined bacteria are compared accordingly. Experiments carried out so far based on 5 generic representatives - one for each bacteria class - did not improve the results over selecting one reference bacteria. To juxtapose vectors representing an examined bacteria with the reference bacteria vectors $\boldsymbol{a}^{ref}$, $\boldsymbol{b}^{ref}$, $\boldsymbol{\alpha}^{ref}$, $\boldsymbol{\beta}^{ref}$ we calculate the cross correlation coefficient [6] between the respective pairs (i.e. $xcorr(\boldsymbol{a}, \boldsymbol{a}^{ref})$). For four cross correlation vectors one chooses their respective greatest values $a^{max}, b^{max}, \alpha^{max}$ and β^{max}. In each picture, we select from 5 to 50 bacteria whose surface area is the closest to the median surface area of all the bacteria in the input picture. The latter permits to select bacteria characterized by the average size and stage of growth. The less bacteria we select the less likely we qualify a group of overlapping bacteria as a single object. We considered 6 features (listed below) based on shape calculated for a fixed amount of points on one bacteria and the number of bacteria analyzed in a single image. We estimated the edge of bacteria using $m+1 = 10$ or $m+1 = 20$ points on one bacteria and compared $l = 5$, 10, 20, 25, 30, 40 and 50 bacteria on one image. The following 6 features based on shape information are considered:

1. *Mean bacteria arc-length* - which is a sum of all arc-lengths representing the perimeters of all selected bacteria divided by l.
2. *Mean curvature of l bacteria* - is a sum $(1/l) \sum_{k=1}^{l} \sum_{j=0}^{m} \int_{\hat{t}_j}^{\hat{t}_{j+1}} \kappa_j^k(\hat{t}) d\hat{t}$, where κ_j^k represents the curvature of k-th bacteria along j-th segment (see (8)).
3. *Mean maximal first distance correlation* - $(1/l) \sum_{k=1}^{l} a_k^{max}$.
4. *Mean maximal second distance correlation* - $(1/l) \sum_{k=1}^{l} b_k^{max}$.
5. *Mean maximal first angle correlation* - $(1/l) \sum_{k=1}^{l} \alpha_k^{max}$.
6. *Mean maximal second angle correlation* - $(1/l) \sum_{k=1}^{l} \beta_k^{max}$.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

4 Features Based on Texture

The second group of examined features relies on image texture analysis. In [12] Haralick introduced statistical measures resorting to the second order image histogram called GLCM (*Grey-Level Co-Occurrence Matrix*). In this paper we also used GLRLM (*Gray-Level Run-Length Matrix*) measures - see [11].

4.1 GLCM Features

GLCM is calculated for the following directional angles α: 0°, 45°, 90°, 135° and distance d on quantized image Ω to n levels that are represented in gray-scale. The co-occurrence matrix M of size $n \times n$ is initialized with all its coefficients set to zero. Assume the image Ω is represented by the pixel table $\bar{M}$ having m_1 rows and n_1 columns. Note that here pixel $(1,1)$ represents the top-left pixel in Ω, whereas pixel (m_1, n_1) corresponds to the bottom-right image pixel. In addition, let matrix $W^{k,l}$ have m_w rows and n_w columns. $W^{k,l}$ is used iteratively to extract the following pixels of $\bar{M}[i,j]$: $m_w(k-1) < i \leq km_w$ and $n_w(l-1) < j \leq ln_w$. The latter can be geometrically viewed as positioning top-left corner of $W^{k,l}$ at (k,l) pixel of Ω. The coverage of Ω with $W^{k,l}$ abides the following pattern. First Ω is horizontally covered by windows $W^{1,1}, W^{1,n_w+1}, W^{1,2n_w+1}, \ldots, W^{1,n_1-n_w+1}$, respectively. Next after vertical shift to $W^{m_w+1,1}$ we move horizontally up to W^{m_w+1,n_1-n_w+1}. This procedure of disjoint coverage of Ω is continued up until reaching $W^{m_1-m_w+1,n_1-n_w+1}$ window. Note that if either $m_1(mod\ m_w) \neq 0$ or $n_1(mod\ n_w) \neq 0$ (since in practice $n_w \ll n_1$ and $m_w \ll m_1$) one can supply extra missing pixels for the most right or bottom part of Ω by extrapolation techniques. Additionally for each of $W^{k,l}$ we iterate over the pixels in that window incrementing values in M based on the correlations between pixels for direction α and distance d that is explained below (for more details see [12]).

Assume we use a window of size $m_w = 5$ and $n_w = 5$. For every $W^{k,l}$ we go through each pixel in that window. Let $n = 5$, $d = 2$, $\alpha = 45°$ and $w_{11}^{k,l}, \ldots, w_{55}^{k,l}$ be certain pixel values in $W^{k,l}$ (see Fig. 3). Here we are in the iterative step that analyzes pixel $w_{42}^{k,l}$ (that is marked as dark gray), its value is equal to 0. Then we know that we increment co-occurrences in GLCM matrix in first row that is responsible for relationships between values of pixels 0 and $n_i = 0, \ldots, 4$. Next step is to check values of pixels located in direction $\alpha = 45°$ and maximum distance of $d = 2$. There are two pixels meeting these requirements: $w_{33}^{k,l}$ and $w_{24}^{k,l}$. Since $w_{33}^{k,l} = 0$ we have to increment value in first row and in first column, and since $w_{24}^{k,l} = 3$ we need to increment value in first row and in fourth column in GLCM. Finally after we moved our window through the entire image registering co-occurrences of the pixels we divide each of the values in $W^{k,l}$ by n^2 that gives us probabilities of co-occurrences between gray levels of pixels for direction α and maximum distance d. Based on GLCM computation, the following 8 statistical measures are calculated [32]: *Contrast, Correlation, Energy, Homogeneity, Autocorrelation, Cluster Prominence, Inverse Difference and Dissimilarity*. Note that as GLCM is calculated for 4 different directions, we

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

obtain measures such as Contrast 0°, Contrast 45°, Contrast 90° and Contrast 135°. The final value of Contrast is taken as mean of these four values. The 8 statistical measures from above determine 8 texture features based on GLCM.

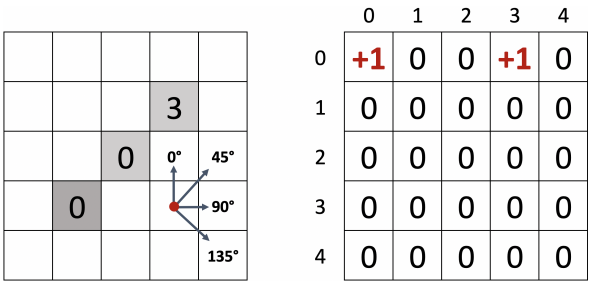

Fig. 3. Increment of the values in GLCM (right) based on the window $W^{k,l}$ (left) of size 5×5, pixel $w_{24}^{k,l}$ for $n = 5$, $d = 2$, $\alpha = 45°$ and a presentation of all possible directions α from the pixel $w_{44}^{k,l}$ marked with a red dot.

4.2 GLRLM Features

The calculations of GLRLM based features are very similar to determining GLCM. The computations involved are also carried out for same directional angles α: 0°, 45°, 90°, 135° and for the maximum distance d on quantized image to n levels that is represented in gray-scale. We initialize GLRLM of size $n \times d$ with zeros. Using window $m_w \times n_w$ we move through the image in the same manner as in GLCM and increment values in matrix according to the run-length of co-occurrence pixels for direction α and length that is equal to $d_i = 1, \ldots, d$. In the next step each of the values in GLRLM is divided by dn. The latter introduces matrix with probabilities of the respective co-occurrences of the gray level $n_i = 0, \ldots, n-1$ and run-length d_i.

As an example assume an input image is quantized to 5 gray levels, maximum run-length $d = 3$, window's size is 5×5, $\alpha = 90°$ and that we analyze pixels with values equal to zero. Let $W^{k,l}$ have sample entries as shown in left window of Fig. 4. The algorithm represents searching sequences of length d_i of pixels that have values equal to zero in $W^{k,l}$ e.g. two sequences of length 3 (marked as green - see Fig. 4) are found filling GLRLM for $n = 0$ with value $d_3 = 2$. For more details on GLRLM see also [11].

Again based on probabilities stored in GLRLM the following statistical measures are calculated [11]: *Short Run Emphasis, Long Run Emphasis, Grey Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Grey Level Run Emphasis, High Grey Level Run Emphasis Short Run Low Gray Level Emphasis, Short Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis* and *Long Run High Gray Level Emphasis*. As previously, these 11 measures are computed along 4 different directions and analogously to the case of GLCM

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

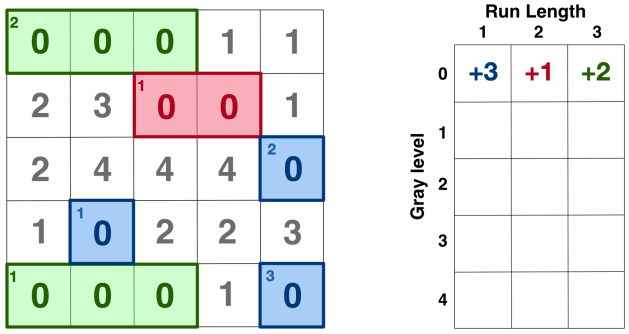

Fig. 4. Incrementing values in GLRLM (right) according to pixels in window $W^{k,l}$ (left) for $\alpha = 90^\circ$ and the gray level equal to zero.

the respective mean values of each measures are determined. Ultimately, this approach determines additional group of 11 features based on texture information. The latter together with the previously introduced GLCM based group yields 19 texture based features considered in this paper. They complement previously discussed 6 shape based features introduced to accomplish bacteria classification.

5 Selected Class Recognition Methods

5.1 KNN

K-Nearest Neighbors is a classification method [17] that permits to assign a new object to one of the constructed classes. We are supplied here with a data-set attributed with the existing features and its membership to the respective class. Based on the latter, a new object needs to be classified with respect to the same set of attached features. In doing so, the calculation of Euclidean distances between a new object and every single object from the data-set is performed. We sort these objects by distance in ascending order. Then we choose k objects (k value is set arbitrarily) whose distances are the smallest and conduct majority voting (see [17]) to decide to which class the new object should be attached.

5.2 ELM-RBF

Extreme Learning Machine - Radial Basis Function is feed-forward neural network with two hidden layers [10]. First one is responsible for the input vectors conversion to the distances based on Gaussian radial function to the closest centroid. Their amount is selected arbitrarily and they are computed using k-means method, where k defines number of centroids to be calculated. This procedure is very similar to RBN (*Radial Basis Network*) and brings much more robustness to the prediction [20]. Upon converting the input vectors into distances

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

in the first hidden layer, we treat them as input vector in ELM network that contains one hidden layer with experimentally chosen number of neurons and their activation function. The bias and weights between input and hidden layer are assigned randomly. Weights between hidden and output layer are calculated using Moore-Penrose matrix pseudo-inverse operation [25].

6 Experiments and Results

We report now on generated experimental results based either on combined set of shape and texture features or solely relying on shape or texture information.

First, the experiments incorporating a full set of shape and texture attributes juxtapose different feature selection and classification methods to reach satisfactory classification accuracy. Additionally, the tests determining the optimal number of interpolation points $\hat{\mathcal{Q}}_{m+1}$ along $\partial\Gamma$ together with gauging the amount of selected bacteria are carried out. The best classification results obtained yield $m+1=10$ and 50 bacteria. We present now the list of implemented methods (see also [18, 31]) with the experimentally picked up optimal parameters guaranteeing the highest possible classification accuracy: SVM, RF (using 200 trees), kNN (using $k=1$), MLP (using back-propagation learning method, topology of the net $22-20-22$ and *tanh* as an activation function on all hidden layers), ELM (with 2800 neurons in hidden layer units with *tanh* activation function) and ELM-RBF (900 neurons with linear activation function in hidden layer units and 40 centroids).

Table 1. Mean accuracy percentage of 50 tests using 10% cross validation and feature selection with classification method performed on shape and texture features.

Feature Selection Method	SVM	RF	kNN	ELM	MLP	ELM-RBF
None	93.84	92.76	95.30	92.50	92.03	94.95
FCBF	95.69	93.62	97.07	78.80	89.73	94.05
SBMLR	96.61	93.92	96.00	91.61	92.65	97.03
CFS	93.61	91.23	95.00	85.34	91.03	94.74

As shown in Tab. 1, the best result in bacteria classification amounts to 97% in accuracy recognition which is obtained upon either applying ELM-RBF with SBMLR or using kNN with FCBF. In this case ELM-RBF shows superiority over ELM method increasing accuracy by over 5% and due to a smaller number of neurons in hidden layer has a vastly shorter training and testing time. Here SBLMR selects 5 shape and 15 texture features whereas FCBF relies on using 3 shape and 4 texture traits. In order to justify merging features from both classes of examined attributes (i.e. shape and texture), we subsequently tested the bacteria classification accuracy when either only the set of shape or the set of texture features is admitted, respectively.

The best experimental result relying exclusively on shape features amounts to 78.92% in classification accuracy (see Tab. 2). It is derived with the aid of

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

Table 2. Mean accuracy percentage of 50 tests using 10% cross validation and feature selection with classification method performed on features based on shape.

Feature Selection Method	SVM	RF	kNN	ELM	MLP	ELM-RBF
None	75.69	72.61	77.00	49.84	73.84	76.92
FCBF	78.07	74.07	74.76	29.38	70.84	76.00
SBMLR	76.06	76.40	77.56	48.00	73.47	77.96
CFS	78.92	73.30	76.69	42.92	74.38	77.69

Table 3. Mean accuracy percentage of 50 tests using 10% cross validation and feature selection with classification method performed on features based on texture.

Feature Selection Method	SVM	RF	kNN	ELM	MLP	ELM-RBF
None	80.07	76.30	81.61	64.84	67.46	68.23
FCBF	75.76	79.76	76.61	41.61	68.30	74.76
SBMLR	78.93	77.06	82.27	63.46	67.81	69.64
CFS	79.00	78.84	78.53	55.46	70.53	74.38

SVM coupled with CFS. In contrast, the best accuracy using solely texture based features equals 82.27% (see Tab. 3) and is achieved upon combining kNN with SBMLR. Having juxtaposed results from Tab. 1, 2 and 3 it is transparent that merging shape and texture features improves classification accuracy by 15%. As shown in Tab.1 the mean accuracy from 50 tests using 10% cross validation amounts to 97% matching state of the art results.

7 Conclusions

Experiments based on 6 shape features render (for our data) the best accuracy reaching 78.92% correct classification for SVM combined with CFS. On the other hand class recognition based on 19 image texture traits yields up to 82.27% for kNN and SBMLR. In contrast, gathering together both shape and texture information (totalling 25 conjugated features) leads up to 97% correct classification upon coupling either kNN with FCBF or ELM-RBF with SBMLR. The iterative optimization of the classification model parameters including selection of the number of knots and the amount of bacteria analyzed in one picture, improves accuracy and reduces time execution of the implemented congregated classifier. These results seem to be unexpectedly satisfactory for our proposed *aggregated bacteria classifier* in the absence of incorporating color information. Still, within the setting of this work, there is a natural scope for further improvements. In particular, any method selecting characteristic benchmark bacteria for a given genus permitting to compare bacteria's curvature with the reference bacteria would be desirable. In this work originally, such five exemplary bacteria were selected arbitrarily but the results obtained did not improve significantly the case of fixing one reference bacteria for five considered genera. Another related issue refers to the task of selecting all significant points (and knots) on the bacteria's boundary (see e.g. [5, 14–16, 19]). This work assumes "more or less"

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

equally spaced points Q_m. The impact of convexity or non-convexity of the bacteria should also be analyzed with respect to ordering $\mathcal{Q}_{\hat{m}}$. Furthermore, the comparison of standard classification methods with deep learning methods and extending admissible set of features incorporating color and dispersion information forms potential research topics within the field of soil bacteria classification. Lastly, the robustness of all examined methods may also be tested against the varying number of bacteria genera (or their respective representatives) and possibly in regard to other dynamic factors such as time aspect impacting on shape, size or a color of the examined bacteria and/or its bacterial colony distribution.

References

1. Amirani, M.C., Gol, Z.S., Shirazi, A.A.B.: Efficient feature extraction for shape-based image retrieval. J. Appl. Sci. **8**, 2378–2386 (2008). https://doi.org/10.3923/jas.2008.2378.2386
2. Anderson, L.M., Stockwell, V.O., Loper, J.E.: An extracellular protease of Pseudomonas fluorescens inactivates antibiotics of Pantoea agglomerans. Phytopathology **94**(11), 1228–1234 (2004). https://doi.org/10.1094/PHYTO.2004.94.11.1228
3. Beeckmans, S., Xie, J.: Glyoxylate cycle. In: Reference Module in Biomedical Sciences. Elsevier (2015). https://doi.org/10.1016/B978-0-12-801238-3.02440-5
4. Bhairannawar, S.S.: Chapter 4 - efficient medical image enhancement technique using transform HSV space and adaptive histogram equalization. In: Soft Computing Based Medical Image Analysis, pp. 51–60. Academic Press (2018). https://doi.org/10.1016/B978-0-12-813087-2.00003-8
5. de Boor, C.: A Practical Guide to Splines. Springer (2001)
6. Buck, J.R., Daniel, M.M., Singer, A.: Computer Explorations in Signals and Systems Using MATLAB. Prentice Hall (2002)
7. Caprette, D.R.: Describing colony morphology, https://bit.ly/324cqkA
8. do Carmo, M.P.: Differential Geometry of Curves and Surfaces. Prentice Hall (1976)
9. Das, B., Chakrabarty, D.: Lagrange's interpolation formula: representation of numerical data by a polynomial curve. Internat. J. Math. Trends Technol. **34**, 64–72 (2016). https://doi.org/10.14445/22315373/IJMTT-V34P514
10. Dhini, A., Surjandari, I., Kusumoputro, B., Kusiak, A.: Extreme learning machine-radial basis function (ELM-RBF) networks for diagnosing faults in a steam turbine. J. Ind. Prod. (2021). https://doi.org/10.1080/21681015.2021.1887948
11. Ferro-Flores, G., Zhou, Y., Ma, X.L., Pu, L.T., Zhou, R.F., Ou, X.J., Tian, R.: Prediction of overall survival and progression-free survival by the 18F-FDG PET/CT radiomic features in patients with primary gastric diffuse large B-Cell Lymphoma. Contrast Media & Molecular Imaging **2019** (2019). https://doi.org/10.1155/2019/5963607
12. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man. **SMC-3**(6), 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314
13. Iversen, C.: Encyclopedia of Food Microbiology: Enterobacter. Academic Press, 2nd edn. (2014)
14. Kozera, R.: Curve modeling via interpolation based on multidimensional reduced data. Studia Informatica **25**(4B), 1–140 (2004)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23

15. Kozera, R., Noakes, L., Wiliński, A.: Generic case of Leap-Frog Algorithm for optimal knots selection in fitting reduced data. In: Computational Science - ICCS 2021. pp. 337–350 (2021). https://doi.org/10.1007/978-3-030-77970-2_26
16. Kozera, R., Noakes, L., Wilkołazka, M.: Parameterizations and Lagrange cubics for fitting multidimensional data. In: Computational Science - ICCS 2020. pp. 124–140 (2020). https://doi.org/10.1007/978-3-030-50417-5_10
17. Kramer, O.: K-Nearest Neighbors, pp. 13–23. Springer Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-38652-7_2
18. Kruk, M., Kozera, R., Osowski, S., Trzciński, P., Paszt, L.S., Sumorok, B., Borkowski, B.: Computerized classification system for the identification of soil microorganisms. Applied Mathematics and Information Sciences **10**(1), 21–31 (2016). https://doi.org/10.18576/amis/100103
19. Kvasov, B.: Methods of Shape-Preserving Spline Approximation. World Scientific (2000)
20. Lee, C.C., Chung, P.C., Tsai, J.R., Chang, C.I.: Robust radial basis function neural networks. IEEE Trans. Syst. Man Cybern. **29**(6), 674–685 (1999). https://doi.org/10.1109/3477.809023
21. Lefakis, L., Fleuret, F.: Jointly informative feature selection made tractable by Gaussian modeling. J. Mach. Learn. Res. **17**(182), 1–39 (2016), http://jmlr.org/papers/v17/15-026.html
22. Lim, Y., Shiver, A.L., Khariton, M., Lane, K.M., Ng, K.M., Bray, S.R., Qin, J., Huang, K.C., Wang, B.: Mechanically resolved imaging of bacteria using expansion microscopy. PLOS Biology **17**(10), 1–19 (2019). https://doi.org/10.1371/journal.pbio.3000268
23. Malhi, S.S., Sahota, T.S., Gill, K.S.: Chapter 5 - potential of management practices and amendments for preventing nutrient deficiencies in field crops under organic cropping systems. In: Agricultural Sustainability, pp. 77–101. Academic Press (2013). https://doi.org/10.1016/B978-0-12-404560-6.00005-8
24. Morin, A.: Encyclopedia of Food Microbiology: Pantoea. Academic Press, 2nd edn. (2014)
25. Rao, C.R.: Generalized inverse of a matrix and its applications, pp. 601–620. University of California Press (1972). https://doi.org/10.1525/9780520325883-032
26. Satoto, B.D., Utoyo, M.I., Rulaningtyas, R., Koendhori, E.B.: Classification of features shape of Gram-negative bacterial using an Extreme Learning Machine. IOP Conference Series: Earth and Environmental Science **524**(1), 012005 (2020). https://doi.org/10.1088/1755-1315/524/1/012005
27. Shapiro, L.G., Stockman, G.: Computer Vision. Pearson, 1st edn. (2001)
28. Soille, P.: Morphological Image Analysis: Principles and Applications. Springer-Verlag (1999)
29. Sokolov, D.D.: Encyclopedia of Mathematics. EMS Press (2001)
30. Sorensen, J., Nybroe, O.: Pseudomonas: Volume 1 Genomics, Life Style and Molecular Architecture, chap. Pseudomonas in the Soil Environment, pp. 369–401. Springer US (2004). https://doi.org/10.1007/978-1-4419-9086-0_12
31. Toprak, A.: Extreme Learning Machine (ELM)-based classification of benign and malignant cells in breast cancer. Medical Science Monitor **24**, 6537 – 6543 (2018). https://doi.org/10.12659/MSM.910520
32. Yang, X., Tridandapani, S., Beitler, J.J., Yu, D.S., Yoshida, E.J., Curran, W.J., Liu, T.: Ultrasound GLCM texture analysis of radiation-induced parotid-gland injury in head-and-neck cancer radiotherapy: an in vivo study of late toxicity. Medical physics **39**(9), 5732–5739 (2012). https://doi.org/10.1118/1.4747526

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_23