Journal of Innovative Optical Health Sciences
Vol. 11, No. 1 (2018) 1850007 (10 pages)

DOI: 10.1142/S1793545818500074

Rapid bacteria identification using structured illumination microscopy and machine learning

Yingchuan He*, Weize Xu†, Yao Zhi‡, Rohit Tyagi†,‡, Zhe Hu‡,‖,†† and Gang Cao†,‡,§,¶,**,††

*College of Engineering
Huazhong Agricultural University
Wuhan 430070, P. R. China

†College of Veterinary Medicine
Huazhong Agricultural University
Wuhan 430070, P. R. China

‡State Key Laboratory of Agricultural Microbiology
Huazhong Agricultural University
Wuhan 430070, P. R. China

§Bio-Medical Center
Huazhong Agricultural University
Wuhan 430070, P. R. China

¶Key Laboratory of Development of Veterinary Diagnostic Products
Ministry of Agriculture, College of Veterinary Medicine
Huazhong Agricultural University
Wuhan 430070, P. R. China
‖huzhe@mail.hzau.edu.cn
**gcao@mail.hzau.edu.cn

Received 28 July 2017
Accepted 26 August 2017
Published 22 September 2017

Traditionally, optical microscopy is used to visualize the morphological features of pathogenic bacteria, of which the features are further used for the detection and identification of the bacteria. However, due to the resolution limitation of conventional optical microscopy as well as the lack of standard pattern library for bacteria identification, the effectiveness of this optical microscopy-based method is limited. Here, we reported a pilot study on a combined use of Structured Illumination Microscopy (SIM) with machine learning for rapid bacteria identification. After applying machine learning to the SIM image datasets from three model bacteria (including *Escherichia coli*, *Mycobacterium smegmatis*, and *Pseudomonas aeruginosa*), we obtained a

††Corresponding authors.

classification accuracy of up to 98%. This study points out a promising possibility for rapid bacterial identification by morphological features.

Keywords: Structured illumination microscopy; bacterial classification; principal component analysis; support vector machine; random forest.

1. Introduction

Bacteria are microorganisms with typical length of several micrometers and different shapes (sphere, rod, spiral, etc.).[1] Some bacteria are harmful to man by causing serious infections and diseases (thus called pathogenic bacteria). Bacteria detection and identification are critical for the diagnosis and treatment of infectious diseases. Currently, pathogenic bacteria are usually identified by: morphological features, physiological and biochemical characteristics (such as nutritional type and antibiotic sensitivity), immunological markers (bacterial antigen, capsular antigen, etc.), chemical composition characteristics (for example, fatty acid composition, ribosomal protein) and genetic markers (such as 16S rDNA).[2] With the advancement in biochemical analysis technology and the progress of nucleic acid sequencing, a number of bacteria identification methods have been commercialized, leading to generation of commercial products such as assay kits, equipment and technical services. Although the accuracy of bacteria identification has been improved drastically, there are still some limitations in the practical applications of these methods. For example, it requires the use of microbiological techniques for isolation of pure culture before applying the physiological and biochemical identification methods. On the other hand, the methods based on high-throughput sequencing technology are usually expensive, complicated and time-consuming. Hence we want to eliminate these issues by directly applying optical microscopy for simplicity and cost effective procedure with high accuracy.

The traditional method based on microscopic morphology seems to be a simple, fast and economical way for bacteria identification, especially for some bacteria with unique structural features.[3] However, the development of this method is slow, mainly due to the limited morphological features visualized by conventional optical microscopy and the absence of standard pattern image database. Furthermore, this microscopy-based method usually relies on manual bacteria identification which suffers from time-consuming and training-dependent identification.

In recent years, the advent of super-resolution microscopy techniques, such as Stimulated Emission Depletion (STED) Microscopy, Stochastic Optical Reconstruction Microscopy (STORM), Photo-activated Localization Microscopy (PALM) and Structured Illumination Microscopy (SIM), has extended the application range of conventional optical microscopy beyond the diffraction limit and achieved more structural details for different applications.[4] It is noteworthy that SIM technology is advantageous for imaging bacterial morphology without any further requirements of biological sample preparation. In the meantime, the rapid development of machine learning is helpful for a lot of applications, including but not just limited to the applications in the biomedical field.[5] Therefore, it is highly possible that combining SIM technology with machine learning could provide a rapid and automatic way for bacterial identification with higher accuracy than the conventional microscopy-based method.

Here we reported a pilot study of combining SIM technology with machine learning for rapid bacteria identification. We firstly used SIM technology to image the fine structures of three model bacteria, including *Escherichia coli* (*E. coli*), *Mycobacterium smegmatis*, and *Pseudomonas aeruginosa*. Then, we applied classical algorithms in the field of machine learning to extract morphological features of these bacteria. Finally, we established a machine learning system for rapid bacteria detection and identification. This study might open a new avenue for rapid clinical diagnosis of pathogenic bacteria by addressing the limitation in available morphological features and identification accuracy.

2. Materials and Methods

2.1. *Bacterial culture and sample preparation*

Three different bacteria, *E. coli* MG1655, *Mycobacterium smegmatis* MC155, and *Pseudomonas*

aeruginosa PAO1, were used in this study. MG1655 and PAO1 were cultured in Luria–Bertani (LB) broth. MC155[2] was grown in 7H9 (Middlebrook) supplemented with 10% OADC(BD). Cells were grown overnight to attain OD600 of 0.5, then 200 μl of broth culture harvested, and suspended in 50 μl PBS. Staining of the bacterial membrane was performed by incubating with NanoOrange (Invitrogen, 1/10 v/v) for 30 min at room temperature.[6] Because NanoOrange exhibits very-weak fluorescence when it is not binded to membrane, we directly spot 3 μl of this suspension onto a poly-L-lysine-treated glass coverslip without washing.

2.2. *SIM imaging*

A Nikon Structured Illumination Microscope (N-SIM) was used for super-resolution microscopy imaging of the bacteria. Images were captured with an EMCCD camera (Andor iXon DU-897) and a 100 × 1.49 NA TIRF objective (Nikon CFI Apo TIRF). The fluorescence was excited by a 488 nm laser and cleaned by a bandpass emission filter (500–545 nm). Image acquisition and reconstruction were performed with Nikon NIS-Elements software in SIM and wide-field mode, respectively.

2.3. *Methods for machine learning*

2.3.1. *Image segmentation and negative samples generation*

Firstly, we used the watershed algorithm[7] in the open-source computer vision library — OpenCV[8] to segment the SIM images into several target bacterium regions. Then, we reproduced standard images with a size of 250 × 250 pixels, consisting of a segmented target bacterium in a noise-free background, for model training. In addition, since some nonbacterial images were needed in the negative regions during the model training process, we manually selected some sub-regions in the SIM images as a reference area. These sub-regions included as much noise types as possible and do not contain any bacteria. Finally, a sufficient number of negative samples with the same size (250 × 250 pixels) were generated by random selection from these sub-regions.

2.3.2. *Algorithm for feature extraction*

Feature extraction determines the efficiency of model selection. If an image is input directly as a vector rather than extracted features in the classifier training process, extra computing time and resources will be required due to the high data dimension. And, for most classification models, high data dimension usually reduces the efficiency of classification. After considering the characteristics of our SIM images, we selected Principal Component Analysis (PCA)[9] method to extract the algebraic features of the images. This method reduces the dimension of the SIM images to acceptable sizes for classifier training.

2.3.3. *Algorithm for classification*

We selected three classifier models in this study: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest. SVM is a widely used classifier model in computer vision with excellent classification performance.[10] KNN is a relatively simple classifier, where the main idea is to classify a new data point from the nearest K data points.[11] Random Forest is based on voting from a combination of multiple decision trees,[12] and is capable of reducing the impact of noise and the possibility of over-fitting.[13]

Among the three classifiers, Random Forest and KNN support multi-classification, while the standard SVM is a two-class model and thus needs to combine with a suitable strategy to become applicable for multi-classification tasks. Here we apply one-vs-rest[14] strategy to SVM for this purpose. All of the three classifiers are derived from the open-source Python machine learning library — sklearn[15] that provides the classifier codes.

2.4. *Evaluating the classification models*

Accuracy and F1-Score were used to evaluate the classification performance of the classifier models. For binary-classification, the Accuracy and F1-Score were calculated from Eqs. (1)–(4), where "TP", "FP", "TN", and "FN" represent True Positive, False Positive, True Negative, and False Negative, respectively:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \qquad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}. \tag{4}$$

For multi-classification, the evaluation can be derived from the binary-classification. Assuming that k types of data need to be classified, and that "TP_i", "FP_i", "TN_i", and "FN_i" represent the True Positive, False Positive, True Negative, and False Negative of the ith data, respectively, we can calculate Accuracy, Precision and Recall with Eqs. (5)–(7), and then F1-score using Eq. (4):

$$\text{Accuracy} = \frac{\sum_{i=1}^{k} \text{TP}_i + \sum_{i=1}^{k} \text{TN}_i}{\sum_{i=1}^{k} \text{TP}_i + \sum_{i=1}^{k} \text{TN}_i + \sum_{i=1}^{k} \text{FP}_i + \sum_{i=1}^{k} \text{FN}_i}, \tag{5}$$

$$\text{Precision} = \frac{\sum_{i=1}^{k} \text{TP}_i}{\sum_{i=1}^{k} \text{TP}_i + \sum_{i=1}^{k} \text{FN}_i}, \tag{6}$$

$$\text{Recall} = \frac{\sum_{i=1}^{k} \text{TP}_i}{\sum_{i=1}^{k} \text{TP}_i + \sum_{i=1}^{k} \text{FN}_i}. \tag{7}$$

3. Result and Discussion

3.1. *Resolution estimation*

We used 140 ± 5 nm "GATTA-SIM" nanorulers (Gattaquant) to characterize the performance of SIM (Fig. 1). This kind of nanorulers carries two fluorescent markers at each end and is an ideal sample to quantify the lateral resolution of our SIM system. As shown in Figs. 1(a) and 1(b), SIM can clearly resolve the fine structure of the nanorulers. In contrast, conventional fluorescence microscopy provides only blurry, undistinguishable images (Fig. 1(c)). The distance between the two fluorescent spots in Fig. 1(b) was estimated to be 138 nm (Fig. 1(f)), which is consistent with the size of the nanoruler (140 ± 5 nm). We also performed direct experimental comparison between the SIM and conventional fluorescence microscopy imaging of a bacterium, and observed significant improvement of resolution in the SIM image (Figs. 1(d)–1(e)). With SIM imaging, we can obtain more morphological features from SIM images which are beneficial for subsequent machine learning.

3.2. *Preparation for machine learning*

3.2.1. *Standard images for machine learning*

The SIM images of three types of bacteria, including *E. coli* MG1655 (178 images), *Mycobacterium smegmatis* MC155 (168 images), *Pseudomonas aeruginosa* PAO1 (202 images), were acquired with a Nikon N-SIM with 50–100 ms exposure and 100 EM gain. Representative SIM images are shown in Fig. 2(a). The bacteria in the SIM images were segmented into individual positive images containing only one bacteria (Figs. 2(b) and 3). Negative images (Fig. 2(b)) were also generated using the procedures described in Sec. 2.3.1. Both the positive and the negative images had the same size of 250 × 250 pixels. Table 1 shows the number of raw images and positive images in this study.

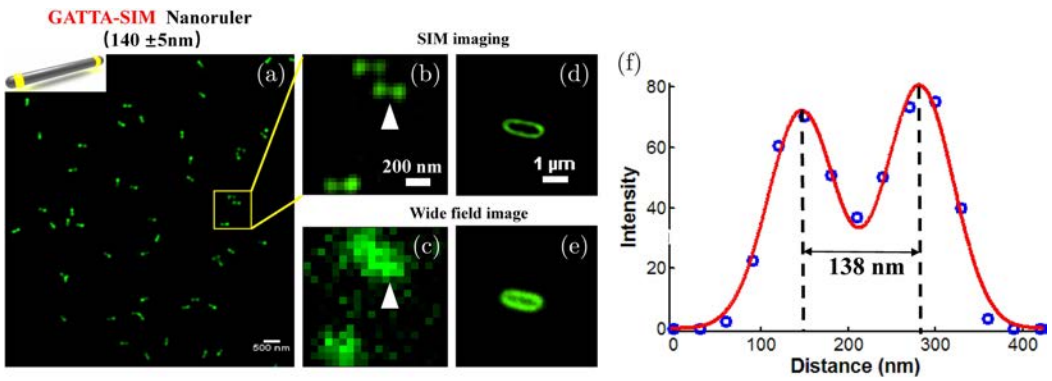


Fig. 1. (Color online) SIM and conventional fluorescence microscopy images of GATTA-SIM nanoruler and *E. coli*. (a) SIM image of GATTA-SIM nanorulers. (b) Enlarged SIM and (c) conventional fluorescence images of the boxed regions in (a). (d) SIM and (e) conventional fluorescence microscopy images of a representative bacterium. (f) Line profile of the nanoruler image indicated by the arrowhead in (b). The blue dots show the experimental data, and the red line is for Gaussian fits.

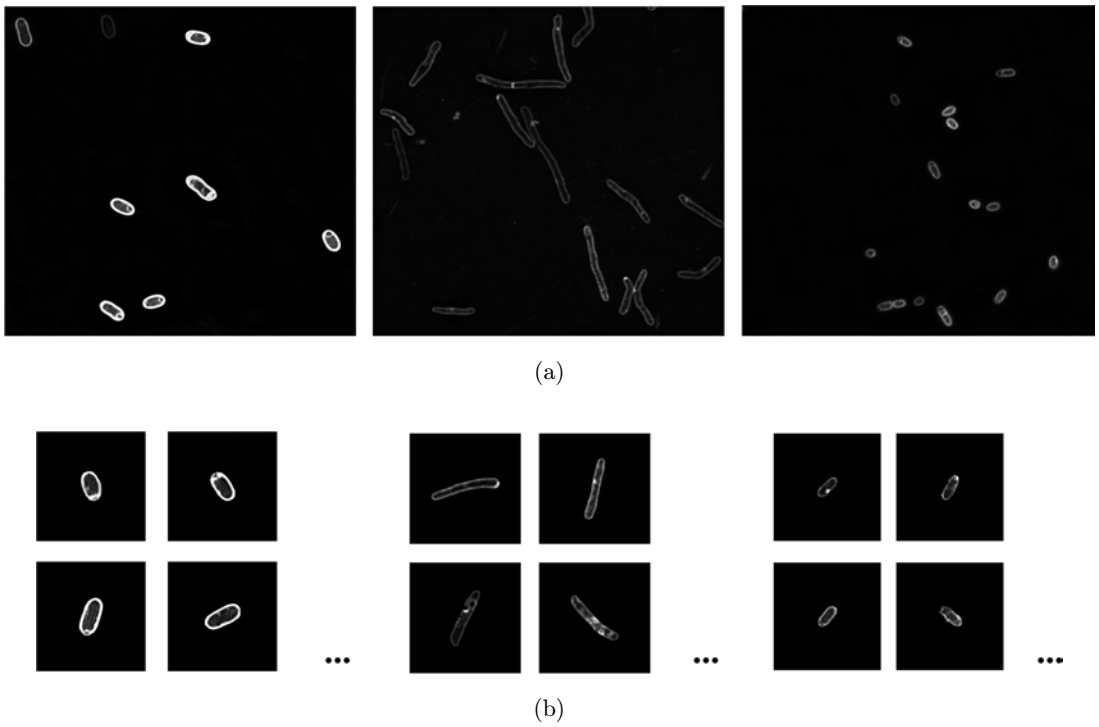


Fig. 2. Representative SIM images of the bacteria. (a) Raw SIM images of *E. coli* (left), *Mycobacterium smegmatis* (middle) and *Pseudomonas aeruginosa* (right). (b) Resized positive images for a standard training library.

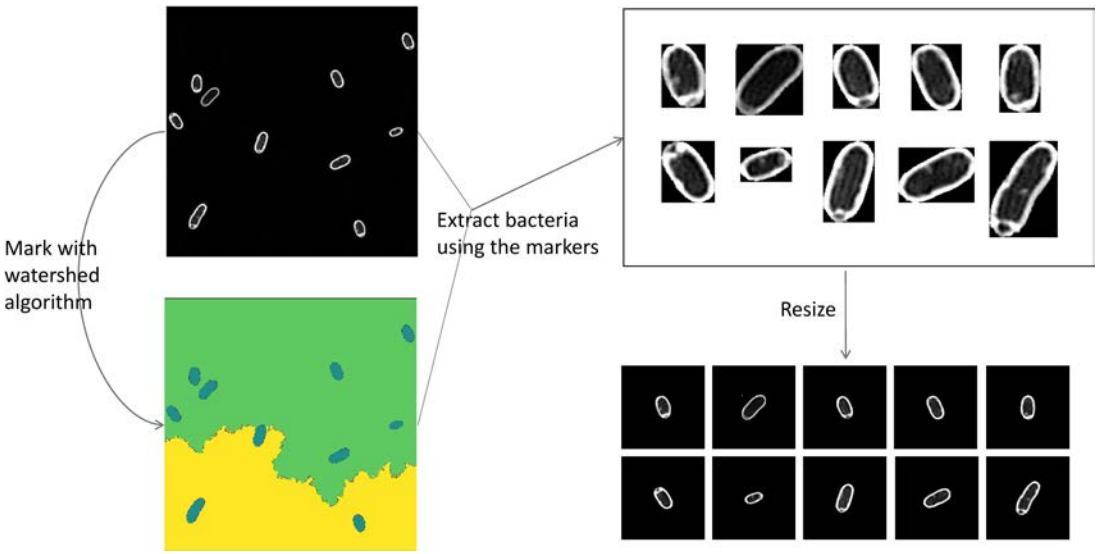


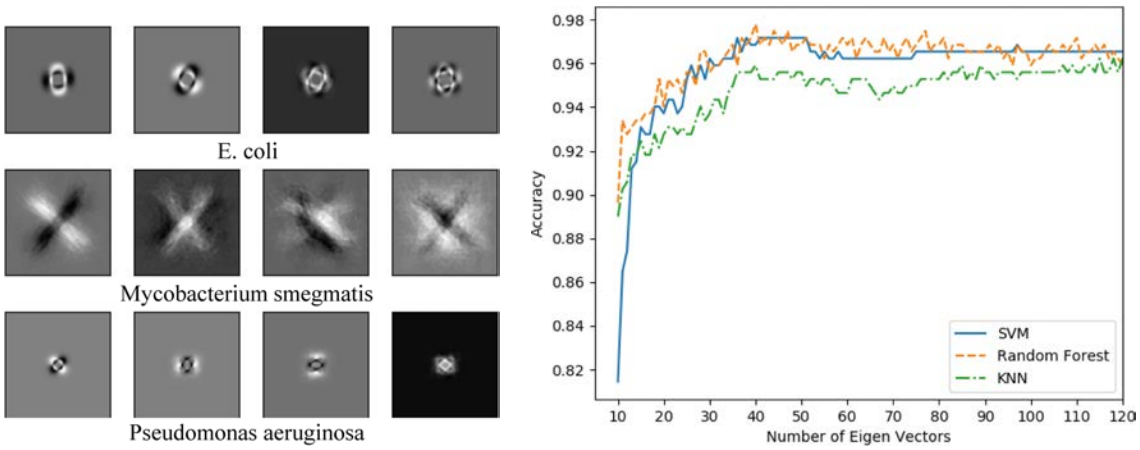
Fig. 3. Flowchart of strategy for generating positive samples.

Table 1. A summary of the number of raw images and positive images.

	Raw images	Positive images
E. coli	178	953
Mycobacterium smegmatis	168	538
Pseudomonas aeruginosa	202	1441

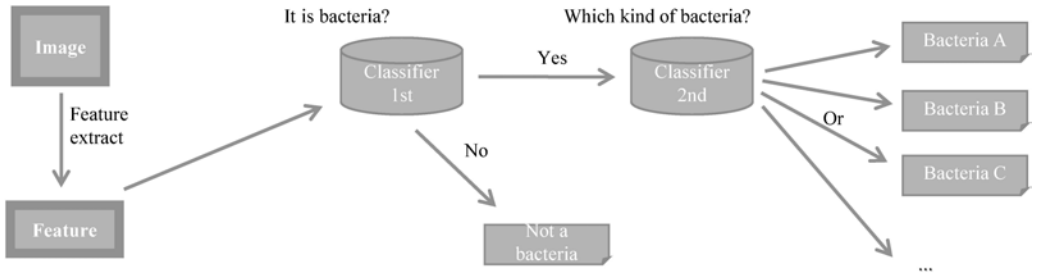
3.2.2. *Structural features for bacteria identification*

In this study, we used PCA to extract the structural features of the bacteria and obtained the eigenvectors for each type of bacteria. Figure 4(a) shows four of the most important eigenvectors with the

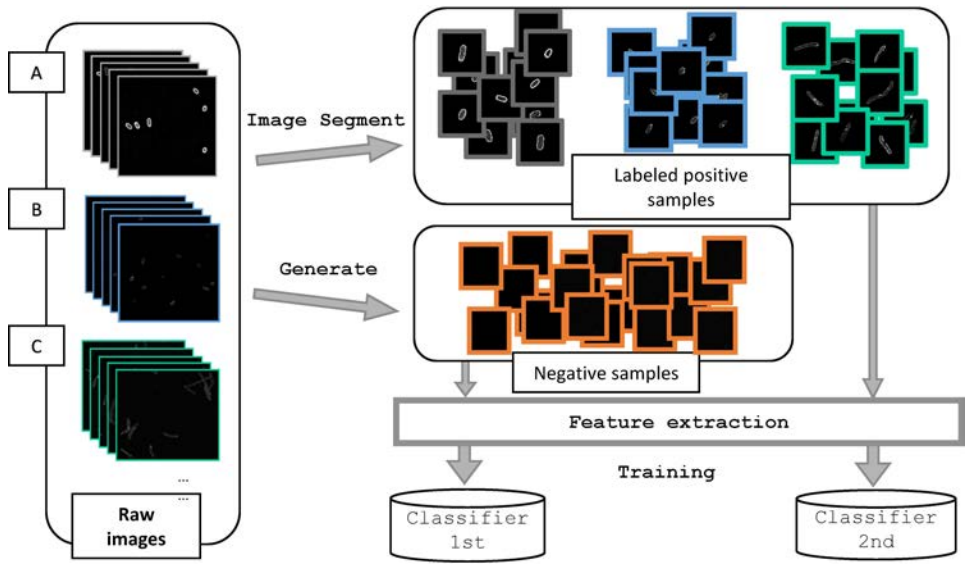


(a) (b)

Fig. 4. (a) The representative eigenvectors of the bacteria and (b) the relationship between the number of eigenvectors and the classification accuracy.



(a)



(b)

Fig. 5. Flowchart of strategy used for model classification and training. (a) The strategy of classification. (b) The pipeline of classifier training.

largest contribution to the variances during PCA. The eigenvectors for each type of bacteria are different, and thus can be used to identify the type of the bacteria. Furthermore, to find out the best number of eigenvectors for bacteria classification, we quantified the dependence between the number of eigenvectors and the classification accuracy (Fig. 4(b)). The classification accuracy improved rapidly by increasing the number of eigenvectors, and then became stable after the eigenvectors increased to 40. In this study, we determined to set the number of eigenvectors to 100 to obtain highly stable results.

3.3. *Bacteria identification strategy*

For the bacteria identification, it is important to determine an optimal algorithm for feature extraction and a suitable classification classifier. Figure 5(a) shows the strategy to classify bacteria image: Firstly the structural features of the image is extracted, then the features are sent to a first classifier which is used to determine whether the image belongs to any kind of bacteria. If the conclusion is "Yes", the structural features are further sent to a second classifier for further determination of the types of the bacteria.

The pipeline shown in Fig. 5(b) is used for classifier training. First of all, SIM images of bacteria were segmented and labeled as positive samples. Then, negative samples are generated from the same raw images. Finally, the structural features for both positive and negative samples were extracted and used for classifier training.

3.4. *Bacteria identification performance*

3.4.1. *Identification performance*

The strategy of cross-validation[16] was used to test the effect of classifier. We found that the SVM algorithm used in this study (Classifier one) was sufficient to distinguish the positive images from the negative images. In a five-fold cross validation testing, the accuracy and F1-score of classification were both above 99%.

Classifier two was responsible for identifying the type of different bacteria. We tested the identification performance of three classifiers: SVM, KNN and Random Forest. The results for multi-classification were shown in Table 2. The parameters used in the classifier models were presented in Table 3. After carefully optimizing the parameters, all of the classifiers provided excellent Accuracy and F1-Score ($> 95\%$), while SVM presents the best performance.

Table 2. Five-fold cross validation testing results.

	Accuracy	F1-Score
SVM	0.9836 ± 0.0065	0.9826 ± 0.0050
Random Forest	0.9703 ± 0.0096	0.9718 ± 0.0068
KNN	0.9683 ± 0.0139	0.9723 ± 0.0102

The confusion matrices for a representative multi-classification test are shown in Fig. 6(a), which allows a clear visualization on the identification performance of the classifiers. To further understand the classifiers' capability on differentiating the bacteria types, we performed a binary-classification test. From the results in Fig. 6(b), we concluded that the classifiers have no specificity for the bacteria.

3.4.2. *Time performance*

The time performance is also an important factor for choosing a good classifier. Here, with the same training and test datasets, the time performance of the classification models is similar (shown in Table 4), but Random Forest seems to be less efficient than the other two classifiers.

Table 3. Parameters of each classifier model.

	Parameters
SVM	Gamma: 0.0001 C: 6000 Kernel: RBF
Random Forest	Min samples leaf: 1 Min samples split: 4
KNN	Neighbor: 4

Notes: (1) Gamma represents Gamma function value, C is for Penalty parameter value, Kernel indicates Kernel function type, and RBF is abbreviation for Radial basis function; (2) Min samples leaf is for leaf nodes with the least number of samples, and Min samples split is for the least number of samples when the internal node is divided; (3) Neighbor is for the number of nearest points.

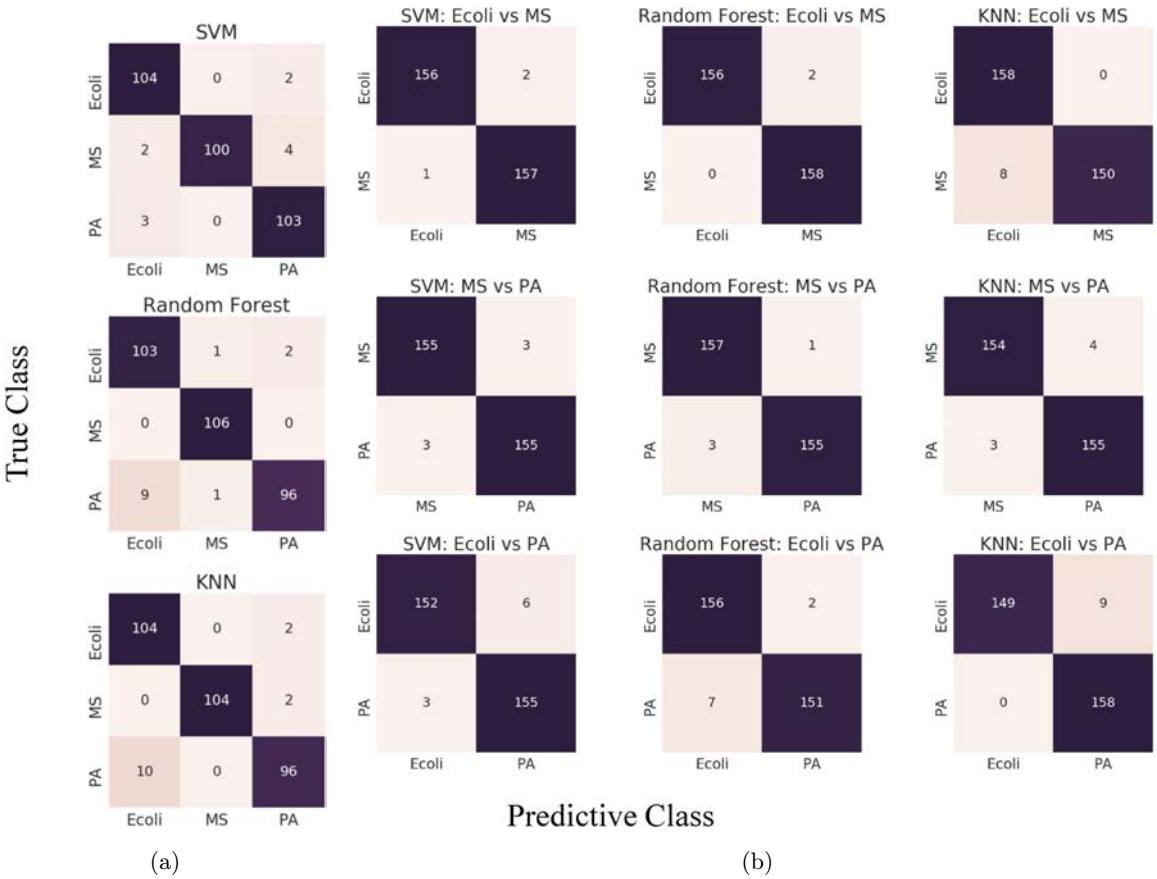


Fig. 6. The representative confusion matrices (e.g., in SVM confusion matrix, of the 106 actual *E. coli*, the classifier predicted that two were PA, and of the 106 MS, it predicted that two were *E. coli* and three were PA). (a) Confusion matrices for the classifiers in the multi-classification test. (b) Confusion matrices in the binary-classification test. *E. coli* is an abbreviation for *E. coli*, MS is for *Mycobacterium smegmatis* and PA is for *Pseudomonas aeruginosa*.

3.4.3. *Robustness performance*

In real applications, the SIM images of the bacteria may contain different level of noises. In this regard, we tested the robustness of the classifiers under three types of noises: bar mask, square mask and Gaussian noise (shown in Fig. 7(a)). We firstly added these noises to original images and then performed the same bacteria identification processes to the new images containing noises. Figure 7(b) shows the testing results. We observed that SVM was sensitive to the bar and Gaussian noises, KNN was sensitive to the Gaussian noise, and Random Forest was robust to all of the noises.

Table 4. Time performance of different classification models.

	Training time (s)	Prediction time (s)
SVM	17.5802 ± 1.8315	0.1886 ± 0.0059
Random Forest	18.2690 ± 1.2241	0.2289 ± 0.0302
KNN	16.3916 ± 1.2523	0.2183 ± 0.0167

3.5. *Comparison of bacteria identification performance*

To investigate the superiority of SIM's high resolution, we trained classification model using the images captured by normal fluorescence microscopy. Figure 8 shows that for original image SIM improved accuracy slightly. However, for the deficient images, SIM is better than normal fluorescence microscopy in most instances. This shows that SIM images improved the robustness of all three kinds of classification models and the classification accuracy.

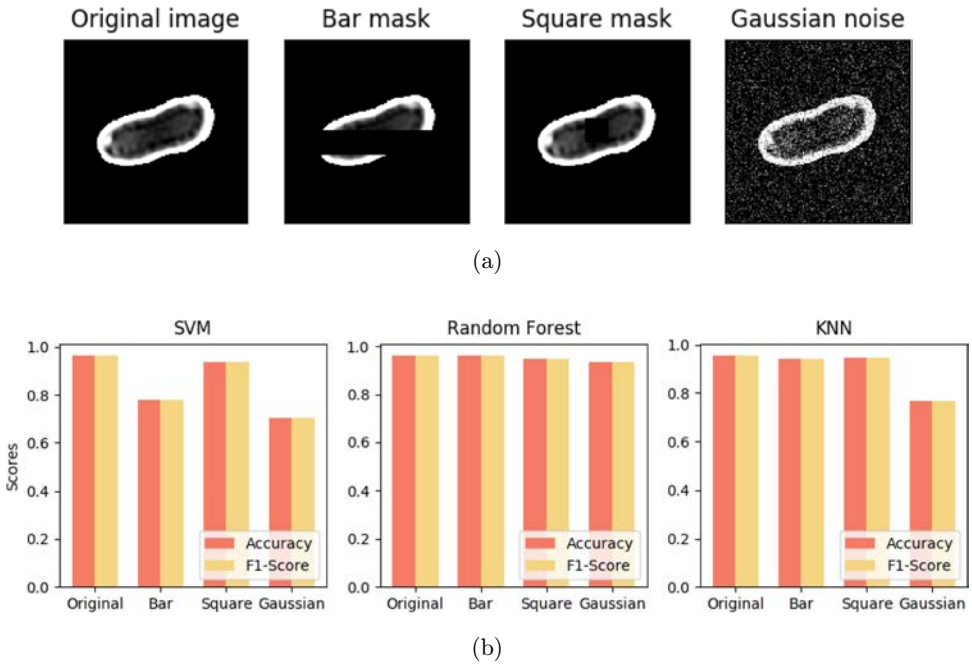


Fig. 7. (Color online) The robustness of different classifiers under three kinds of noises. (a) Original images and the images with different noises. (b) The testing results.

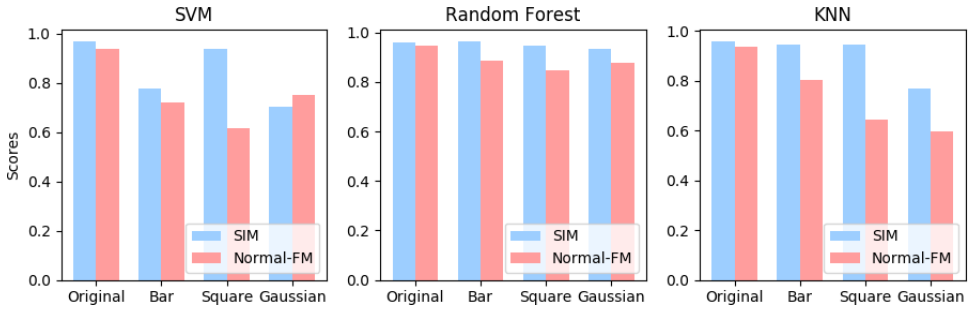


Fig. 8. (Color online) Comparison of classification accuracy between SIM images and normal fluorescence microscopy (Normal-FM) images.

3.6. *Cost performance*

The cost performance is an important reference factor for rapid clinical diagnosis of pathogenic bacteria. Based on local market, we provided a comparison between biochemical, genotypic analysis method and the SIM-based method (shown in Table 5). It shows that this study is cost effective, less time-consuming and less technological demanding.

Table 5. Comparison between biochemical and genotypic analysis method and the SIM imaging-based method in this study.

	Timing (day)	Cost* (¥)	Identification level	Technology required
16S rDNA	5–7	500	Strain	+++
Ribosomal protein	1–2	500	Strain	+++
Fatty acid composition	1–2	800–1000	Species	++
Nutritional type	5–6	800–1000	Species	++
This study	1–2	50	Species	+

Notes: *The cost is expected according to local market prices.

4. Conclusion

In this study, we report a new method for bacterial identification. This method is based on SIM technology which is capable of providing more morphological features than conventional fluorescence microscopy. After applying a machine learning strategy to the SIM images, we obtain an identification accuracy up to 98%. This study opens new possibility for rapid bacteria identification, especially after further training of more bacteria types and optimizing labeling strategies and machine learning algorithms.

Acknowledgments

We thank Qi Zong, Fangkui Wang, and Shaoran Zhang in State Key Laboratory of Agricultural Microbiology for their useful suggestions. This work was supported by the National Key Research and Development Program of China (Grant No. 2017-YFD0500303), the National Natural Science Foundation of China (Grant Nos. 31371106, 91640105), the China Agriculture Research System (No. CARS-36) and the Huazhong Agricultural University Scientific and Technological Self-innovation Foundation (Program No. 52204-13002).

Y. He and W. Xu are contributed equally to this work.

References

1. S. C. Gopinath, T. H. Tang, Y. Chen, M. Citartan, T. Lakshmipriya, "Bacterial detection: From microscope to smartphone," *Biosens. Bioelectron.* **60**, 332–342 (2014).
2. J. E. Clarridge III, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clin. Microbiol. Rev.* **17**(4), 840–862 (2004).
3. S. A. Bandh, A. N. Kamili, B. A. Ganai, "Identification of some Penicillium species by traditional approach of morphological observation and culture," *Afr. J. Microbiol. Res.* **5**(21), 3493–3496 (2011).
4. B. Huang, H. Babcock, X. Zhuang, "Breaking the diffraction barrier: Super-resolution imaging of cells," *Cell* **143**(7), 1047–1058 (2010).
5. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* **542**(7639), 115–118 (2017).
6. J. L. Alonso, S. Mascellaro, Y. Moreno, M. A. Ferrus, J. Hernandez, "Double-staining method for differentiation of morphological changes and membrane integrity of *Campylobacter coli* cells," *Appl. Environ. Microbiol.* **68**(10), 5151–5154 (2002).
7. D. L. Pham, C. Xu, J. L. Prince, "Current methods in medical image segmentation, *Annu. Rev. Biomed. Eng.* **2**(1), 315–337 (2000).
8. G. Bradski, "The OpenCV Library," *Dr. Dobb's J.* **25**(11), 120–123 (2000).
9. W. Zhao, R. Chellappa, A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," *Third IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 336–341, IEEE, New Year (1998).
10. C. Cortes, V. Vapnik, "Support vector machine," *Mach. Learn.* **20**(3), 273–297 (1995).
11. L. E. Peterson, "K-nearest neighbor," *Scholarpedia* **4**(2), 1883 (2009).
12. A. Liaw, M. Wiener, "Classification and regression by random forest," *R News* **2**(3), 18–22 (2002).
13. T. M. Khoshgoftaar, M. Golawala, J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," *19th IEEE Int. Conf. Tools with Artificial Intelligence*, Vol. 2, pp. 310–317, IEEE, New York (2007).
14. J. Weston, C. Watkins, "Support vector machines for multi-class pattern recognition," *ESANN* **99**, 219–224 (1999).
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.* **12**(October), 2825–2830 (2011).
16. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *IJCAI* **14**(2), 1137–1145 (1995).