BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Image-based Clustering of Microbial Colonies

Jan Láncoš*, Michal Čičatka**

Abstract
In-lab analysis of microbial colonies grown on Petri dishes is on the frontier of efforts for total laboratory automation. The core of this issue lies in precisely localizing the colonies during image analysis. The state of the art solutions often use machine learning.
Machine learning is heavily reliant on quality labeled data, which is scarce. To address this issue, we have created a sample generator. The generator has multiple uses, we have successfully applied it in both our segmentation and colony clustering efforts, raising the F1 segmentation score from 0.518 to 0.737 and achieving a V-measure clustering score of 0.912.
This approach to generating synthetic data moves us one step closer towards total laboratory automation.

*xlanco00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*
**xcicat02@vutbr.cz, *Faculty of Electrical Engineering and Communications, Brno University of Technology*

1. Introduction

One of the uses for Petri dishes in laboratories is to multiply microbes on them and observe them to determine those worthy of further analysis by dedicated instruments. Despite this being a common and indispensable practice, it has not yet been fully automated. While the machines are capable of cultivating the dishes and analysing the retrieved samples, laboratory technicians still have to decide which of the possible hundreds of colonies are worthy of being analysed in a more sophisticated manner.

The task of automating this middle step boils down to two problems – precisely localizing the microbial colonies on the dish and grouping the detected colonies according to their presumed cultures. A robot could then select a minimal number of samples covering the whole diversity of a single dish without the need for human interaction.

While the localization part has to a certain degree been addressed by the scientific community, it still lacks a reliable one-size-fits-all solution. One of the more recent articles has for example utilized U-Nets using pre-trained encoder models [1]. The experiments were however conducted on small datasets with not a lot of variability in terms of cultures and agars.

Regarding clustering, one team of scientists has managed to predict two distinct kinds of colony forming units [2]. Another team has classified two species of bacteria based on their visual properties [3]. Clustering an arbitrary number of cultures has to our knowledge not yet been attempted, possibly again for the lack of data.

Since both these problems have a lack of data as a hurdle, we decided to create a synthetic data generator. By extracting labeled images of single colonies from the already possessed dish images, they can be redistributed upon newly obtained images of uncultivated Petri dishes to create original samples. The implemented generator can create large amounts of labeled data fast.

We have used this generated data to extend the training datasets of a U-Net segmentation model in an attempt to improve its performance and generalization abilities.

We have also used the synthetic data to evaluate and compare our attempts at clustering the colonies based on their common features.

2. Objective

A possible laboratory scenario goes as follows – the laboratory receives a biological sample. A Petri dish containing agar is automatically cultivated with microbes. Then a laboratory technician inspects the

number of cultures on the dish, selects their representatives, they are automatically picked up from the dish and transferred to be prepared for analysis by dedicated instruments Figure 1.

In order to automate the step featuring a human operator, the algorithm we propose has to be able to produce two masks – a segmentation mask Figure 2 and a clustering mask Figure 3.

The segmentation mask defines the areas where the microbial colonies are.

The clustering mask describes which colonies are grouped together based on their visual properties. To be produced, the segmentation mask is needed.

3. Generating synthetic data

Producing and labeling images of cultivated dishes is time-consuming and expensive. For this reason we thought up a way of repurposing already existing images into original composites.

By using the existing labeled images we created a database of individual labeled colonies Figure 4.a. Then, using a technique called color keying we tuned down the alpha channel in places where the pixels' color was similar to what the label states is the surrounding and also cut off the non-colony parts entirely, which resulted in a database of labeled semi-transparent colonies with transparent backgrounds Figure 4.b.

Using a genetic algorithm, we then deploy these colonies upon images of Petri dishes containing many kinds of uncultivated agar, which we acquired earlier for this very purpose Figure 5.

Lastly, for the sake of realism we also simulate a phenomenon developing near some colonies, which is a degree of agar alteration Figure 6.

After implementing this generator, we are able to produce diverse and large datasets with corresponding labels for both segmentation and clustering.

4. Improving segmentation

We have trained a baseline U-Net segmentation model on the original real dataset introduced as part of a master's thesis [4]. Additionally, four models of identical U-Net architecture were trained on four different datasets Figure 7. Each dataset contained the same original real data and each was also extended by some synthetic data.

To test if both the keying and the agar reactions simulation are working, we extended the prepared original datasets by four identically generated synthetic datasets. While the colonies, layouts and agar were always identical across corresponding samples, they differed in the use of keying and agar reactions simulation Figure 8.

Upon evaluating all the five trained models on the original testing dataset extended by previously unseen real images of different agar colors, it can be claimed that not only does this approach work and improve the performance of segmentation considerably, but also that both the realism tweaks do in fact improve the models by themselves. The overall performance of the segmentation expressed by the F1 score model has increased from 0.51 to 0.73 Table 1.

5. Introducing clustering

We then introduced colony clustering to allow for more complex information being extracted from the dish images. To cluster the segmented colonies we decided to use the K-Means algorithm. To find the optimal number of clusters we are using the knee/elbow detection method.

In terms of feature extraction, we propose three approaches: clustering the RGB values such as they are, which is to serve as a baseline; extracting features from a U-Net autoencoder; and formulating the features manually for each distinct isolated colony.

We have used an autoencoder trained on the original dataset of real images. The features are extracted from the output of the final concatenation layer after the last up-sampling is made within the net Figure 9.

Features manually extracted for the other approach are on the other hand just simple visual properties such as an average color, size or shape Figure 10.

Both of the more complex clustering approaches have surpassed the baseline and achieved a comparably better result. While the achieved results can not be compared to any other research, the V-measure is a bounded metric and the achieved score of 0.91 is fairly near the theoretical limit of ideal clustering, which is 1.0 Table 2.

Acknowledgements

I would like to thank my supervisor Ing. Karel Beneš for all his help, patience and valuable insights.

References

[1] Thomas Beznik, Paul Smyth, Gaël de Lannoy, and John A. Lee. Deep learning to detect bacterial colonies for the production of vaccines, 2020.

[2] Tanguy Naets, Maarten Huijsmans, Paul Smyth, Laurent Sorber, and Gaël de Lannoy. A mask r-cnn approach to counting bacterial colony forming units in pharmaceutical development, 2021.

[3] Shimaa A. Nagro, Mohammed Kutbi, Wafa M. Eid, Essam J. Alyamani, Mohammed H. Abutarboush, Musaad A. Altammami, and Bandar K. Sendy. Automatic identification of single bacterial colonies using deep and transfer learning. *IEEE Access*, 10:120181–120190, 2022.

[4] Michal Čičatka. Detection and localization of microbial colonies by means of deep learning algorithms. Master's thesis, Brno University of Technology, Faculty of Electrical Engineering and Communication, 2021.