IMAGE-BASED CLUSTERING OF MICROBIAL COLONIES

AUTHOR: JAN LÁNCOŠ
SUPERVISOR: KAREL BENEŠ
CONSULTANT: MICHAL ČIČATKA

MOTIVATION

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 the precise localization and segmentation of the colonies during image analysis. The state of the art solutions often employ machine learning models. However, these models tend to be heavily reliant on the existence of quality labels, which leads to a data scarcity problem, since professionally cultivated agar plates are hard to obtain.

SYNTHETIC DATA GENERATION

To mitigate the dependence on hard-to-obtain real data, I decided to focus on more sophisticated approaches towards data augmentation instead of the machine learning methods themselves. By using already existing labeled agar plate images and utilizing a custom made image processing pipeline – I created a sorted database of individual labeled semi-transparent colonies with their respective segmentation mask. Then, using a simple genetic algorithm I deploy these isolated colonies upon images of empty Petri dishes containing various kinds of uncultivated agar, which I acquired for this very purpose. After implementing this generator, I gained the ability to produce diverse and large datasets with corresponding labels for both segmentation and clustering fast and with no additional costs.

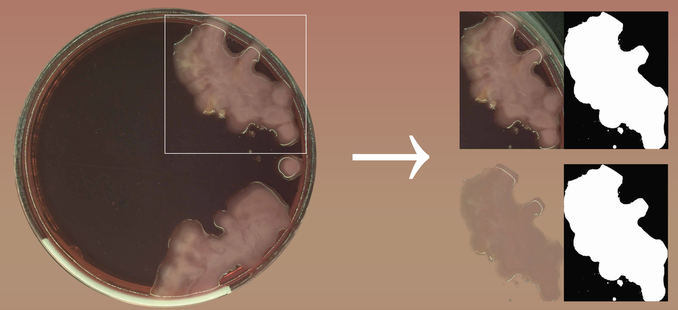

FIG. 1: COLONY EXTRACTION

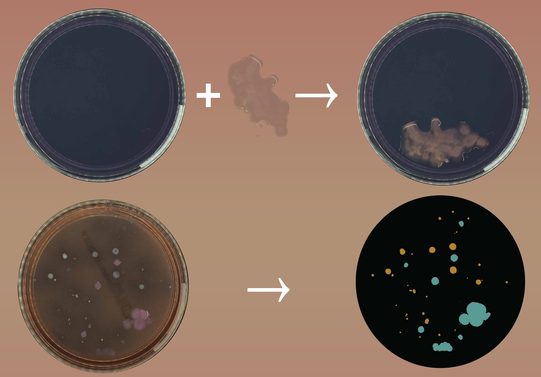

FIG. 2: COLONY SUPERIMPOSING

FIG. 3: SEGMENTATION AND CLUSTERING

THESIS GOAL

To further improve the performance of the state of the art segmentation models to allow for the introduction of colony clustering based on their perceived visual similarities, as that is the missing step towards implementing total laboratory automation in regards to agar plate image analysis.

TRAINING SEGMENTATION

To compare the effectiveness of the synthetic data when used for training colony segmentation models, I have compared U-net models trained on purely real data against ones trained on the same real data largely extended by the synthetic. All were evaluated on previously unseen real data. The overall performance of the segmentation expressed by the F1 score has increased from 0.51 to 0.73, corroborating the proposed solution as effective.

INTRODUCING CLUSTERING

To cluster the segmented colonies I used the K-Means algorithm combined with the knee/elbow detection for determining the optimal number of clusters. In terms of feature extraction, I attempted clustering the RGB pixel values; extracting features from a U-Net autoencoder; and specifyng the colonies' features manually. The manual approach reached a V-measure score of 0.91, which is fairly near the theoretical limit of 1.0 of this bounded metric.

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY