­Lehrstuhl für Medizinische Biotechnologie

Technische Fakultät

Friedrich-Alexander-Universität Erlangen-Nürnberg

Semi-Supervised Z-stack segmentation using Random Forest Classifier

Mini Thesis

*In the program Advanced Optical Technologies*

*Submitted by*

Sergei Dobrovolskii

Matrikelnummer: 22745251

Erlangen, date

*Supervisors:*

Dr. Sebastian Schürmann

Dr. Daniel F. Gilbert

Prof. Dr. Dr. Oliver Friedrich

Zusammenfassung

Biologie ist eine quantitative Wissenschaft, die eine erhebliche Menge an Daten benötigt, um eine bestimmte Hypothese zu beweisen. Bilder von Zellen und Geweben sind die große Datenquelle, aber Daten allein reichen nicht aus, sie müssen verbessert, verarbeitet, analysiert und quantifiziert werden, um aussagekräftige Erkenntnisse zu gewinnen.

Eines der Werkzeuge, die bei der Verarbeitung der Daten helfen, sind Deep Neural Networks. Diese Netzwerke können hervorragende Ergebnisse bei Aufgaben wie der Bildklassifizierung und -segmentierung zeigen, erfordern jedoch eine große Menge an Trainingsbeispielen, um daraus zu lernen. Die Bildklassifizierung lässt sich recht einfach von Hand beschriften, dies ist jedoch bei der semantischen Segmentierung nicht der Fall. Die manuelle Beschriftung zur Bildsegmentierung erfordert viel Zeit, um die Umrisse für jeden Pixelbereich zu zeichnen und ihnen Beschriftungen zuzuordnen. Die Hauptidee dieser Arbeit ist es, einen halbautomatischen unüberwachten Zellsegmentierungsalgorithmus für die 3D-Zellsegmentierung zu finden.

Abstract

Biology is a quantitative science, that requires a significant amount of data to prove a certain hypothesis. Images of cells and tissues are a great source of data, but data itself is not enough, it must be enhanced, processed, analyzed, and quantified to get important insights.

One of the tools that help to process the data are Deep Neural Networks. These networks can show outstanding results in tasks like image classification and segmentation, although require a huge amount of training samples to learn from. Image classification is quite easy to label by hand, but that's not the case for semantic segmentation. Manual labeling for image segmentation requires a lot of time to draw the outline for each pixel area and assign labels to them. The main idea of this thesis is to find a semi-automated unsupervised cell segmentation algorithm for 3D cell segmentation.

Table of Content

[1 Introduction 8](#_Toc93250899)

[2 State of the Art 9](#_Toc93250900)

[2.1 Imaging in Medical Diagnostics 9](#_Toc93250901)

[2.1.1 Histopathology H&E 9](#_Toc93250902)

[2.1.2 Volumetric Data 9](#_Toc93250903)

[2.2 Motivation 10](#_Toc93250904)

[2.2.1 Human immune cells in colon tissue 10](#_Toc93250905)

[2.2.2 Multiphoton Microscope 10](#_Toc93250906)

[2.3 Aims of Image Analysis 11](#_Toc93250907)

[2.4 Machine learning in image processing 12](#_Toc93250908)

[2.5 Decision Tree 13](#_Toc93250909)

[2.6 Random forest 14](#_Toc93250910)

[2.7 Feature extraction 14](#_Toc93250911)

[2.7.1 Convolutions 15](#_Toc93250912)

[2.7.2 Feature sets 15](#_Toc93250913)

[3 Methods 17](#_Toc93250914)

[3.1 Weka segmentation plugin 17](#_Toc93250915)

[4 Results and Discussion 18](#_Toc93250916)

[4.1 Data cleaning 18](#_Toc93250917)

[4.2 Data labeling 19](#_Toc93250918)

[4.3 Classifier features selection 19](#_Toc93250919)

[5 Conclusion 21](#_Toc93250920)

[6 References 22](#_Toc93250921)

[7 Appendix 23](#_Toc93250922)

Abbreviation

|  |  |
| --- | --- |
| MPM | Multiphoton Microscope |
| CNN | Convolutional Neural Network |
| H&E | Hematoxylin and Eosin |
| RF | Random Forest |
| DTC | Decision Tree Classifier |
|  |  |
|  |  |
|  |  |

# Introduction

Getting insights from volumetric biological data is a sophisticated task, that depends upon the modality that was used to acquire images, sample preparation, and software that will quantify the raw data and give significant results, that can be interpreted. Each month new software algorithms are published, serving new approaches for segmentation and classification tasks, but most of them are Deep Learning algorithms, that require gigantic human-labeled data. Human time is expensive, hence assistance for labeling would be a helping hand to reduce the time for labeling. In this Thesis classical machine learning approach using Random Forest (RF) will be explained and utilized for human immune cells segmentation.

# State of the Art

## Imaging in Medical Diagnostics

### Imaging modalities in medicine

The human body is a highly complex system with billions of cells that group together to make parts that we call organs. These parts, as well as their cells, interact with each other and support the life of the body. Like any complex system, this one tends to fail too – any disease can make life complicated or impossible. The first step of curing any disease is diagnostics. There are multiple imaging modalities in medicine these days:

1. X-ray
2. Ultrasound
3. Computed tomography
4. Optical coherence tomography
5. Magnetic resonance imaging
6. Microscopy

All these imaging modalities, except for microscopy, are used for large structures and cannot be used for cellular imaging. Microscopy is highly used for visualizing single cells and tissue. Different microscopic modalities can deliver volumetric data such as Multiphoton Microscopy or Confocal laser Microscopy.

### Volumetric Data

Cells are usually nonplanar and tend to live in three dimensions. The examination under the coverslip is not accurate enough and can hide some of the cellular behavior. To receive more data from the tissue it is required to record volumetric images also known as Z-stacks or stacks. This is a three-dimensional image with multiple color channels. It can be expressed as:

Stack allows seeing the cells in volume which is oftentimes necessary. For example, direct observation of skin cells together with immune cells lying under the skin layer, like in paper [2]. Volumetric data is more difficult to work with, due to the high memory consumption and visualization tradeoffs that we must do to observe the data.

## Motivation

In biological research, it is highly important to understand, prepare and detect a specimen to conduct an experiment. Data acquired for this thesis was taken from Human colon tissue using Multiphoton Microscope.

### Human immune cells in colon tissue

Any human tissue can be infected. But the human body is not defenseless, its immune system is always aware of the incoming infection and permanently fights it back. The immune cells are present everywhere, but if the inflammation occurs, then the immune cells will accumulate in that area, hence their concentration will rise. The types of cells in the infected area and their relative concentration can tell a lot to the doctor about the inflammation process.

A high concentration of immune cells is concrete evidence of inflammation in human colon tissue. To detect the presence, quality, and quantity of immune cells in tissue techniques like MPM might be very useful.

### Multiphoton Microscope

Multi-photon microscopy (MPM) is a powerful tool that allows three-dimensional mapping of samples that have a measurable nonlinear optical response such as second harmonic generation (SHG), third-harmonic generation (THG), or fluorescence induced by multiphoton absorption. MPM provides a way to see the nonlinear microscopic world with high resolution, in 3D. It reduces the scattering from non-focal planes, by excitation only at the focal plane.

Fluorescence is the process when the electron of the fluorophore absorbs the excitation photon and settles on a higher energetic level of the atom. Then it drifts lower and drops down to the ground state, with the emission of a light photon of a different wavelength (Figure 1A). The 2 PE works the same way, but instead of a single excitation photon, it requires 2 incident photons each with half the energy of the required one. For this effect to happen, these photons must hit the same atom simultaneously. To achieve this seldom event the density of photons must be high.

The principle of this modality differs from the fluorescent microscope. In a fluorescence microscope, the excitation response of the fluorophore is linear, more excitation light - more fluorescence response. But with the 2 PE, it is different (Figure 1B). Excitation is visible only at the focal point – the place where the photon density will be the highest.

Diagram

Description automatically generated

Figure 1 A: Multiphoton fluorescence energy diagram comparison of a single-photon (1 PE) fluorescence and two-photon (2 PE) fluorescence. In two-photon fluorescence 2 photons of the energy twice lower than required are exciting electron together. This is achieved by higher energy density which results in higher probability of excitation event to happen. B: The comparison of the excited volumes – 1 PE has a lot of excited molecules out of the focal plane which results in worse image quality, compared to 2 PE.

MPM has found important applications in nonlinear materials characterization, biological research, and diagnosing medical conditions. The data in this thesis was acquired using this modality.

## Aims of Image Analysis

For the experiment conduction, the data is acquired by any given sensor (camera matrix in our case). This raw information doesn’t give a lot to scientists. To prove the hypothesis this information must be cleaned, distilled, and processed. All these tasks are faced by image analysis – the field of science and the mathematical toolset. Image analysis involves processing images into fundamental components to extract important information. It may involve tasks such as finding shapes, detecting edges, removing noise, counting objects, texture analysis, etc. In this Thesis, only the segmentation procedure will be explained.

## Machine learning in image processing

Machine learning is a set of sophisticated mathematical operations performed on data to receive an expressive result from it. These algorithms are conventionally generalized in form of a pattern recognition pipeline (Figure 2). At first, the data is acquired using any type of electronic devices such as a camera, microphone, or microscope. Then this data is stored and preprocessed. In preprocessing step data is filtered, enhanced, and prepared for further steps. Next goes a feature extraction, which extracts representative features from data using a set of mathematical operations. These features can represent a simplified version of original data or can create new data, that allows for algorithms to find intra-data correspondences. At a later stage these features are used for so called learning or training step. During training Algorithm will try to find the best possible split to classify the data with minimal error, based on extracted features.

Graphical user interface, diagram

Description automatically generated

Figure 2 pattern recognition pipeline [3]. It is divided in two parts: Test phase and Training phase. Sensor, preprocessing and feature extraction steps are common for both phases.

These algorithms in image processing are conventionally divided in two groups: Classical and Deep Learning approaches. Difference between those groups can be compared using the pattern recognition pipeline (Figure 2):

* Classical machine learning approaches follow this pipeline. They require small amount of data and small computational costs. Lack on generalization and precision.
* Deep Learning approaches bypass part of this pipeline by combining feature extraction with classification. There are no predefined feature extraction procedures in neural networks, and these procedures are estimated, during the training process. They require huge amount of human labeled data, high computational costs, and time. They are good at generalization and most of the state-of-the art methods now are using neural networks.

Deep Learning approaches require gigantic amount of human labeled data. Due to that it is always preferable to find a way to minimize the labeling time from human because it is tedious and expensive. For that reason, it is important to assist human operator with labeling task. One of the promising methods is to use classical machine learning approaches like Random Forests for easier data labeling.

## Decision Tree

Before speaking about Random Forest, it is important to explain decision tree classifier (DTC). Decision Trees are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-regression). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The selection process can be described as a sequence of binary selections corresponding to the traversal of a tree structure. One limitation of decision trees is that the division of input space is based on hard splits in which only one model is responsible for making predictions for any given value of the input variables.

Diagram

Description automatically generated

Figure 3 Classification Tree example. On the left is the example of the classification tree: orange is the root node, blue is the branch node, green is the leaf node (decision node). On the right is the sample space division based on the classification tree from the left. Red samples correspond to class 1 and blue to the class 2.

Tree model consists of nodes and each node will ask a true false question about one of the features () (Figure 3). And in response to this question, the data is split into two subsets. These subsets then become the input to two child nodes that are added to the tree. And the goal of the question is to unmix the labels as proceeding down. Or in other words, to produce the purest possible distribution of the labels at each node. The quantification of a split uncertainty at a single node can be achieved using a metric called Gini impurity:

– number of classes, – probability of correctly classifying class . And we can quantify how much a question reduces that uncertainty using a concept called information gain. Given that question, we'll recursively build the tree on each of the new nodes. We'll continue dividing the data until a limit called a tree height will occur, at which data is no longer divided and a class is assigned based on the majority of samples present in the last subset of points. This height is set up manually to prevent a tree from overfitting.

A single classification tree is a powerful algorithm that lacks on generalization and prone to overfitting. Ensembling of multiple classification trees is a common strategy to achieve higher generalization and accuracy.

## Random forest

Random forest is a classical machine learning method for data classification. It is based on ensembling of multiple Decision Trees and “decides” based on the majority voting of all decision trees. Each tree is trained on a random subset of data, which leads to a random set of uncorrelated trees. This algorithm is used for classification of pixels for annotation of digital stacks.

## Feature extraction

Intensity values on their own are a weak data representation, they do not to represent the neighborhood regions, also they cannot be used with RF, because it will only divide the image dataset based on intensity threshold, which is ineffective. Spatial information is much more valuable and has a tendency for better generalization. How to blend the information from a neighborhood region and extract important information about it? It can be done using a mathematical operation, called convolution.

### Convolutions

Convolution is a mathematical operation, performed on two functions, that produces a third one, that expresses how a shape of one function will modify the shape of another one:

Once used for discrete domain, it must be reformulated for image stack application like:

Here are the pixel coordinates in image stack and are the pixel-wise iterators. From these two definitions it is defined, that a single pixel intensity is a function of the pixel’s neighborhood and a new function, called kernel.

### Feature sets

In image processing different types of functions or kernels are convolved with image to obtain different spatial effects:

1. Box kernel – spatial linear filter, each pixel in the resulting image has a value averaged of its neighborhood pixels.   
   ,   
   where is a size of a box kernel. This kernel blurs the image.
2. Gaussian kernel – , where is a standard deviation. Blur the image with an effect of a circular aperture – the kernel has a radial symmetry.
3. Sobel operator – used in image processing for edge detection. Composed of two operations: Finding the spatial x and y image derivatives and combining their results with
4. Difference of Gaussians – a kernel composed as a difference of two different gaussian kernels is a close approximation of and is used as a replacement for – blurring and edge detection procedure, due to decreased computational cost.
5. Median filter – nonlinear filter, that assigns the median intensity value of the neighborhood pixel area to a given pixel. Allows to minimize a presence of high frequency noise.

All these kernels with different parameters are used to create image features, necessary for preserving the neighborhood correspondence and extraction of important features from it, like edges, textures, and shapes.

# Methods

## Weka segmentation plugin

Open-source FIJI [3] (based on ImageJ) has a built-in plugin for human aided image segmentation, using random forests. This plugin called Weka Segmentation [4]. It utilizes the random forest approach together with the feature extraction described in the previous chapters. It can be accessed from the toolbar Plugins menu -> Segmentation -> Trainable Weka segmentation 3D.

A screen shot of a map

Description automatically generated with medium confidence

Figure 4 Example image of the Weka Segmentation plugin. Red segment is the background, green is the cell neighborhood, purple is the cell.

Usage of this plugin is simple, but some preparations are required for the given task. Setup requires:

1. Select number of classes: it is necessary to segment 3 classes – Background, Cells, Cell borders.
2. Select a set of the feature types: these features will be used for classification by a RF. It should consist of Gaussian blur and Difference of Gaussian filters with the sigma range of 1.0 - 8.0.

# Results and Discussion

In this section the exact image processing pipeline will be explained, and the question of cell segmentation will be answered.

The 3d cell segmentation will be solved as example on one of the stacks, provided by the Institute of Medical Biotechnology FAU Erlangen. This is the stack taken from the patient’s colon tissue. This sample has a severe inflammation, which can be seen as a presence of the yellow immune cells (Figure 5).

A screenshot of a map

Description automatically generated with low confidence

Figure 5 A: Image of the human colon tissue taken from a stack, acquired with the multiphoton microscope. B: Magnified area. C: Collagen fluorescence channel. D: Green channel. E: Red channel

One of the typical human immune cell types is Lymphocyte. Its size lies in the range from 6 to 14 µm, which can be seen in B. The segmentation and measuring of these cells is the aim of this work. But first, it is necessary to simplify the data and get rid of the unnecessary parts.

## Data cleaning

The stack is a 3-channel image. Each channel represents a signal from a certain fluorophore. The yellow color of the immune cells is a combination of green and red channels. The green and blue areas are the colon crypts and the collagen matrix, that regions are not interesting in the scope of the research Figure 5 C, D. Therefore, to highlight the immune cells, it is sufficient to use only the red channel Figure 5 E. Hence crypts and collagen matrix signals will be suppressed. Next, the segmentation plugin will be used for the labeling of data.

## Data labeling

The number of classes for segmentation is equal to 3: Background, Border, Cell. The background class shows the signal, which is assumed to be a background signal. Border class defines the extracellular boundaries between cells. This class is aimed to improve the accuracy of background classification and to separate cells from each other.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 6 A: Weka segmentation 3D plugin window. Colored segments are the human input for the training procedure. All of them are listed on the right part of the window. B: Enlarged labeling example of cells (purple) and borders (green) – the separation of the cells is required to be labeled.

## Classifier features selection

Random forest is a suitable semantic segmentation tool for 3-dimensional images. More filters are used for classification the more stable and reliable results can be acquired. Generalization gets higher (the model does not overfit) and precision rises. On the other hand, the memory consumption rises too together with the computation time required for feature computation. To increase processing time main goal is to find the golden ratio between computation time and model precision. For smaller stacks, more different features can be computed in comparison to larger stacks with better precision and speed. For larger stacks, it is important to select the most adequate feature set as a precision trade-off.

The performance comparison will be performed using the IoU metric (Intersection over Union). It goes to 0 when there is a small overlap between the ground truth area and the predicted area.

The segmentation result can be seen in Figure 5. It was slightly filtered after classification, using a median filter of size 1 and cancellation of all blobs with size < 10 pixels.

A picture containing text

Description automatically generated

Figure 7. Examples of immune cells segmentation using Random Forest classifier. A: overlay of mask and real data. Red volume refers to cells and Dark to the original image. B: The segmentation results only

As mentioned in section 2.7 features are important, but heavy to compute and store in memory. Therefore, finding the right set of features is crucial. The influence of the feature selection on the classification accuracy will be shown in Figure 7. The training data was labeled and then a classifier was trained only using a single feature set and then the IoU (Intersection over Union) was calculated and compared to the ground truth (IoU = 1.0 Figure 7A). It can be observed that the Mean-Variance feature set alone will perform better than the rest (IoU = 0.68). Then comes the Max-Min-Median set, that is comparable to previous one with IoU = 0.66. The worst ones are the Derivatives and Gaussian Blur sets with IoU = 0.39 and 0.51 units. Any single feature set, used for training, didn’t allow to fit the model well – model underfitted. Hence the amount of training data must be extended by combining of feature sets together during the training Figure 7B. Even a combination of two worst feature sets together outperforms a single best set (). Combination of better performing sets will give even a better result (up to IoU = 0.82 and more). But with increased accuracy comes the downside – increased computational time and memory consumption.

Chart, box and whisker chart

Description automatically generated

Figure 8. Feature importance comparison to ground truth in RF classification using intersection over union metric. From this plot, the importance of the features for this dataset can be rated. Feature combination comparison – a combination of multiple features performs better than a single feature.

# Conclusion

Performance of a Random Forest classifier assumed to be suitable for this application. But how different is it in comparison with the deep learning approach? The DL approach will require hundreds of completely hand-labeled images, a powerful workstation, and hours of training. The preliminary result can be observed only after a day. But segmentation of unseen data can be done relatively fast, with good generalization and accuracy. With the RF approach, you can get the stack-wise result in less than a minute, depending on the stack size and number of features. The amount of data required can be equal to 3-4 labeled cells on 1 slice! The rest it generalizes itself. But with simplicity comes underperformance – RF lacks generalization for different images. The capacity of this classifier is limited, and features are not optimized for the application and are selected mostly intuitively. DL approach on the other hand optimizes a huge number of feature extractors to be most efficient for selected applications and hence performs better.

RF can be used as a labeling aid for DL dataset creation

Human manual labeling of 3d microscopic images is a tedious task. It requires a great amount of time and money to be invested. This routine can be simplified using machine learning algorithms. Human aided software might label data with higher speed than conventional methods. And it requires less training time in comparison to deep-learning methods. Common open-source software like FIJI provides users with a convenient tool to segment 3d images and to classify the respective pixels.

# References

|  |  |
| --- | --- |
| [1] | C. K. A. S. M. V. T. P. B. C. F. K. J. C. A. C. K. J. Z. P. B. S. M. S. S. O. F. M. J. W. Kristina Scheibe, "Inhibiting Interleukin 36 Receptor Signaling Reduces Fibrosis in Mice With Chronic Intestinal Inflammation," *Gastroenterology,* vol. 156, no. 4, pp. A1-A22, e1-e18, 827-1224, 2019. |
| [2] | C. S. T. L. C. R. Andreas Maier, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik,* vol. 29, no. 2, pp. 86-101, 2019. |
| [3] | J. A.-C. I. F. E. e. a. Schindelin, "Fiji: an open-source platform for biological-image analysis," *Nature Methods,* vol. 9, p. 676–682, 2012. |
| [4] | M. Hall, "The WEKA data mining software: an update Share on," *ACM SIGKDD Explorations Newsletter,* vol. 11, no. 1, pp. 10-18, 2009. |
| [5] | K. A. J. J. R. a. R. Z. Andrew H. Fischer, Hematoxylin and Eosin Staining of Tissue and Cell Sections, Cold Spring Harbor Laboratory Press, 2008. |

# Appendix

Curriculum vitae

|  |  |
| --- | --- |
| Personal Details |  |
| Name | **Sergei Dobrovolskii** |
| Date and place of birth | 03.02.1996 in Moscow |
| Adress | Koldestraße 10, 91052 Erlangen |
| Email | dobrik4ever@gmail.com |
| Telefon | 1516 3610896 |
| Education |  |
| 10/2019 – present | Master program Advanced Optical Technologies,  Friedrich Alexander Universität Erlangen-Nürnberg |
| 09/2017 – 07/2018 | *Bachelor of Science (B.Sc.)*  Bachelor Thesis: “Phase microscope based on multi-step phase calculation method“ |
| 09/2014 – 05/2018 | Bachelor program Optical engineering,  MIREA – Russian Technological University |
| 09/2002 – 07/2014 | School №875, Moscow |
|  |  |
| Professional Career |  |
| 08/2020-present | Interherence GmbH – Optical engineer |
| 10/2019 – 07/2020 | JSC "AMTEO M" - Software Developer. Developing the automated histopathology scanning brightfield microscope. |
| 10/2018 – 10/2019 | JSC "AMTEO M":   * Chief Engineer. Developing the automated histopathology scanning brightfield microscope. * Field service engineer. Repairment of medical devices. * Crisis manager. B2B communication, factory processes repairment. |
| 03/2016 – 08/2018 | State budget vocational educational institution of Moscow "Polytechnic College No 47 named after VG Fedorov" – teacher (subjects: microcontrollers, electronics, robotics, programming) |
|  |  |
|  |  |

Erklärung

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegeben Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Declaration

I confirm that I have written this thesis without any external help and

not using sources other than those I have listed in the thesis. I confirm also

that this thesis or a similar version of it has not been submitted to any

other examination board and has not been previously accepted as part of a

exam for a qualification.

Erlangen, den xx.xx.xxxx

(sign here)

Sergei Dobrovolskii

Acknoledgement