



Universiteit  
Leiden  
The Netherlands

## Image analysis for gene expression based phenotype characterization in yeast cells

Tleis, M.

### Citation

Tleis, M. (2016, July 6). *Image analysis for gene expression based phenotype characterization in yeast cells*. Retrieved from <https://hdl.handle.net/1887/41480>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/41480>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/41480> holds various files of this Leiden University dissertation

**Author:** Tleis, Mohamed

**Title:** Image analysis for gene expression based phenotype characterization in yeast cells

**Issue Date:** 2016-07-06

# 6

## Discussion

“ *In this chapter, we will evaluate our research questions and make separate discussion for each of these research questions. In the first section, we specify the research problem and the research questions we addressed in this dissertation. In the following sections we discuss our image analysis pipeline, our segmentation algorithms, our machine learning approach to perform object recognition and pattern recognition using our combination of feature sets, and then the validation of our analysis system.* ”

## 6.1 Research Problem and Questions

**H**ERE we again clearly state our research problem (RP), which is at the center of our entire project and directly related to our goals and the associated research questions (RQs) that frame our research study. The answers to these questions creates an identifiable connections between them and the main research problem inspiring the study. Such connections is what makes our research meaningful. In the following sections we discuss these research questions.

- **RP:** There is no mature Pattern Recognition system to support objective analysis and phenotype characterization in single-cell image-based gene expression experiments.
- **RQ1:** Which components and processes are required to build a comprehensive image analysis pipeline for single-cell image based gene expression experiments?
- **RQ2:** Can a segmentation based on Hough-Transform and minimal path extraction algorithms improve the detection of ovoid-shaped objects in micro/cell biology images? Can we realize an optimization of the initial result using contour expansion algorithms?
- **RQ3:** What machine learning approach using sophisticated features can support in the object recognition process? and can it improve the identification of subtle patterns residing within the measurement data?
- **RQ4:** Can we study the role of 14-3-3 proteins and Nha1 antiporter based on imaging and image analysis? Do results from other analysis confirm our result and thereby validate our method?

## 6.2 Image Analysis Pipeline

**RQ1:** *Which components and processes are required to build a comprehensive image analysis pipeline for single-cell image based gene expression experiments?*

Fluorescent reporter proteins like GFP are widely used in biological research, also in research employing the model yeast *S. cerevisiae*. The analysis of these images is very time consuming and not completely objective. In this study we developed a novel and comprehensive image analysis platform for analysis of microscope images of cells expressing fluorescent proteins. Our application was focused on the model organism *S. cerevisiae* as a case study. The main outcome is a software platform that is currently used by our collaborators to assist them in their experiments.

The novelty of this system is the overall image analysis pipeline and the segmentation algorithm developed for yeast cell detection. This segmentation algorithm for bright-field channels has advantages over algorithms implemented in other software packages as additional staining is not required, the segmentation approach is optimized for *S. cerevisiae* cells and it is freely available.

To the best of our knowledge no similar platform exists, where the existing systems are not flexible as they either offer part of the pipeline such as the segmentation modules as in CellStat [Kva08] and CellSerpent [Bre11], or they are general and not able to offer segmentation for the image modalities that we have as in CellProfiler [Car06]. On the other hand, our proposed solution overcomes these issues and can be extendible and reusable. Our proposed system is a new and comprehensive tool that can generate measurement reports to confirm and validate experiments performed by biologists. *YeastAnalysis* system also extends previous work through providing a complete image analysis system instead of only parts of the pipeline. The strength of this platform relies on a novel segmentation algorithm, user friendly interface, automatic reports generation and selected features to be measured. The system also has flexibility in verifying the results, such as manual segmentation of individual cells, or manual correction of the measurement results and performing analysis of selected groups of cells, as well as choosing from various visualization charts. Such input from the user can be fed in a future work into the system to train it using the machine learning approach we developed. The already existing software packages that deal with *S. cerevisiae* yeast cells are usually developed for a specific task. For example, they are developed to achieve only the segmentation step [Kva08, Bre11, Pen13], to measure only a few features [Kva08, Bre11, Pen13], or to address a specific experiment [Maz13]. This complicates their ability to perform a complete analysis which is especially important for analysis of large datasets as required for systems biology. The *YeastAnalysis* platform with its novel segmentation algorithm and its complete pipeline is a promising tool for the analysis of large number of images generated in large scale experiments. This is further supported by the observation that the time required for segmenting an average 1024x1024 image with around 30 cells, using a quad core computer with *Ubuntu* operating system, is only around 5 seconds.

In conclusion, this platform contributes to improve the analysis of gene expression studies, it provides a comprehensive image analysis platform that can be used by cell biologists to analyze their experiments. This platform can be further improved and its usability can be extended in future work.

The model in Fig. 6.1 depicts an example of what can be done next to extend the usability of the system. This model can be used to help us know and understand the ideas we have for a future work. The primary objective of our model is to convey the fundamental principles and basic functionality of the system to be developed. It must be developed in such a way as to provide an easily understood system interpretation for the models users. In order to implement this model properly, it should satisfy four fundamental objectives [Str11]:

- Enhance an individual's understanding of the representative system.
- Facilitate efficient conveyance of system details between stakeholders.
- Provide a point of reference for system designers to extract system specifications.
- Document the system for future reference and provide a means for collaboration.

Our system is now a standalone application. However, for better maintainability, we can develop into a web application. In the model depicted in Fig. 6.1, a user can connect with its

account to a web application through a web interface. This application is connected to a network storage, where the users data are stored. Through this Application a user can perform either segmentation, measurement or data-analysis.

The segmentation module requires as input the images to be segmented and the segmentation method to be used in addition to the specific parameters for that method. The measurement module requires the binary masks or contour coordinates of the segmented images and as parameters all the features required to be measured and the image channels to be measured. The output of this module would be a CSV file holding all the individual cell objects measured for the requested features. The data analysis module requires the CSV file generated by a measurement process and keywords specifying what cell groups to be analyzed and what type of visualization charts to be output into the final PDF report. These steps can be performed one by one. However, when preferred, the parameters for all the modules can be fed in advance into an additional batch module that performs all these steps sequentially and generate automatically the required report.

At the lowest level in the model there is a Cell Analysis API that can communicate with each of the mentioned modules through a web services description language (WSDL). In future implementation, this model will play an important role in the overall system development life cycle.

## 6.3 Ovoid Objects Segmentation

**RQ2:** *Can a segmentation based on Hough-Transform and minimal path extraction algorithms improve the detection of ovoid-shaped objects in micro/cell biology images? Can we realize an optimization of the initial result using contour expansion algorithms?*

The segmentation algorithm is an improvement of an algorithm developed in an initial work [Kva08] by increasing the detection rate and the accuracy of the detected contours [Tle14]. Our approach consists of two steps, the first uses the Hough Transform to locate circular objects in an image, the second step is fine-tuning each detected object and extracting its exact contour where the center of the circle detected in the first step is taken as a seed point. From that point, a polar representation of the object (referred to as polar image) is generated and an algorithm is applied to extract the exact contour of the object by determining the minimal path from the first to the last column in the polar image.

The detection rate is dependent on the quality of the images. Images containing out-of-focus cells and images containing debris and dead cells are more difficult to segment. Furthermore, during confocal imaging, the intensity of the fluorescence drops dramatically when cells are out-of focus, making quantification difficult. In the case study on the role of 14-3-3 proteins and Nha1 antiporter mentioned in Chapter 5 the segmentation method was able to detect 92 percent of the cells. Additionally, the non-detected cells could be segmented using an interactive GUI by manually seeding (mouse click within) the cells. Detected artefacts such as debris or dead cells could easily be removed manually as well.

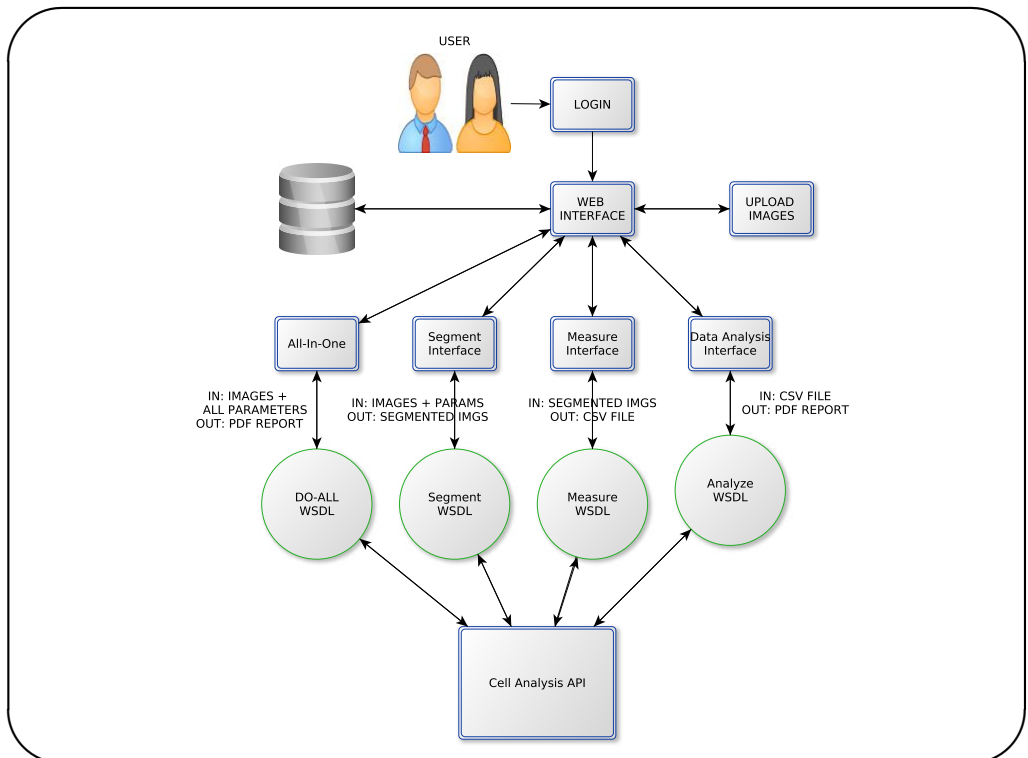
The novelty of this method is the threshold equations presented to control the Hough-

Transform to detect circular objects; in addition to the extraction of the circular shortest path method from the polar representation of the image object. Our novel segmentation method is better than the other methods and offers higher detection rate and more accurate contours. In a validation experiment the detection rate obtained an F1-score of 0.93 outperforming similar methods.

Since microscope settings are delicate and contours are biased to the inner part of the objects, the novel expansion algorithm comes into play. In one experiment the expansion algorithm could improve the accuracy of the extraction contour with an F1-score of 0.92 and a Pratt score of 0.58 outperforming other methods.

The strength of our segmentation algorithm is represented by the higher detection rate and the more precise contours compared with other methods. It is much better than many algorithms such as heavy active-contour based algorithms [Bre11]. The improved segmentation level means consequently an improved accuracy in the analysis of the measurement.

In conclusion, we presented a novel segmentation algorithm to extract the contours of ovoid-shaped and circular objects by using a variety of Hough Transform and a minimal path



**Figure 6.1: Model for further platform development.** At the lowest level in this model there is a cell analysis API that communicates with several modules through web services description language (WSDL). At the top level, users can interact with these modules through a web interface.

algorithm. In addition to a contour expansion algorithm that optimizes the extracted contour. As a future work we can apply parallel programming to make a real-time detection method and analysis a possibility. We can also improve the detection by using noise removal algorithms and segmentation that uses the information from all the image channels.

## 6.4 Machine Learning and Feature Sets

**RQ3:** *What machine learning approach using sophisticated features can support in the object recognition process? and can it improve the identification of subtle patterns residing within the measurement data?*

The popular feature sets we chose to describe the single-cell objects has played a significant role in raising the performance of the classification models. The Wavelet-based texture measurement has shown their superiority to discriminate cell objects in the top classifiers. In addition, the fusion of invariant moments and wavelet details in many dimensions obtained high weights in some classifiers such as Simple Logistics and Linear Model Trees revealing their discriminative power. The second order histogram features extracted from the co-occurrence matrix has also played a role in the discrimination of single-cell objects. However, a complete feature set gained the best classification results on two different datasets. The complete feature set combines features from basic shape descriptors, invariant moments, wavelet texture measurement, invariant moments on wavelet detail images, co-occurrence matrix derived features, basic texture measurement. The *AUC* metric was the main evaluation criteria used. The complete feature set got an *AUC* value of 0.91 and 0.92 on the two different datasets respectively. This suggests that the presentation of complex feature sets to describe the single-cell characteristic and morphology in a more sophisticated way is an advantageous step in building classification models that aims to recognize objects and identify subtle patterns.

The use of wavelet-based texture measurement, invariant moments and second order histogram features are not new in describing texture measurement in bioimaging. However, the novelty of our work is to combine these feature sets to describe the objects in addition to our novel approach to fuse moment invariants and wavelet details. Our results show optimal classification when combining various feature sets.

The strengths of these features is their ability to describe objects characteristics and morphology is a more sophisticated way enabling its recognition and identification of subtle patterns residing within these objects. As in many disciplines, there is a drawback. This drawback is the complexity of the constructed classification models based on these features. This makes it very difficult to describe the underlying processes of the classification model. Despite this difficulty, the classification results were excellent and the constructed model could easily classify the cell objects into its correct group when compared to groundtruth.

The feature sets contribute to the improvement of classifying single-cell objects and consequently improving cell object recognition and identification of subtle patterns residing within the measurement data; i.e. recognizing the patterns that are not obvious and hard to notice within

standard measurement methods. Nevertheless, further research is still possible. Interesting research could be to try more techniques and feature sets known to be used with texture measurement and trying to fuse different methods together. For example, if we were to redo this research we would consider methods based on Zernike moments, Discrete Cosine Transform and Linear Binary Patterns, as these methods are mentioned in the literature and are successful in texture retrieval [Sim04, Bae97, Mäe03]. Moreover, an interesting research topic would be to use fewer features that can gain a similar classification power. This would enable a more easy understanding of the underlying process that is generating the data.

Using the aforementioned features, we presented a machine learning system that was able to discriminate cell objects to identify essential characteristic differences between two groups treated under different conditions, and a similar approach that improves the object recognition based on the object's measured sophisticated features. This implies that we can apply the same approach in further biological studies and especially in high-throughput screening.

Machine learning systems are gaining popularity within bioimaging and biomedical studies because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. The idea of applying machine learning is not a new idea by itself. However, our novel machine learning workflow is the combination of selected sophisticated and complex features, the sampling of the dataset using *SMOTE* method (cf. Section 4.3.1) after measuring individual features, normalization of the data features using the *MV* scheme in addition to the evaluation of numerous linear and non-linear classification approaches in an automatic way to elect a classifier to construct our model.

The strength of this approach resides in the selected features as a powerful discriminative power. In addition, the evaluation of the best existing classification models assures the selection of the best model to be used as our classifier. One issue with this system is the parameters that most models utilize. With default parameters, an excellent classifier might not show its strength as we showed for the *SVM* classifier in Chapter 4. Despite this fact, the performance of the classification models were superior with an excellent classification level, i.e.  $AUC \geq 0.90$ .

With such machine learning approach we expect an easier identification of objects and patterns within these objects. This fact will contribute to improve the understandability of protein and gene behaviours in cell/micro biological studies. Further work in this field is still possible. Applying developmental techniques to create an optimal classifier is an interesting domain. Moreover, an automatic parameter optimization and selection for the existing popular classifiers is also a possibility for a future work. In addition, adding more feature sets as mentioned in the previous section and testing the classifiers on global existing datasets would be an important project. Next to that, an implementation of such approach in a real-time image analysis might have an additional value to biologists performing there experiments.

To summarize the discussion, we state that a feature set based on wavelet-based texture measurements, moment invariants, first and second order histogram features, basic texture measurement and fusion of moment invariants with wavelet details has an optimal classifying results in classifying single-cell based gene expression data. In addition, we offered a reusable machine learning workflow that offers an excellent performance to recognize biological objects

and identify subtle patterns within the measurement of these objects. There is still more work to explore, and the application of such approach in a highthroughput study is the obvious thing to do next.

## 6.5 Image Based Proteomics

**RQ4:** *Can we study the role of 14-3-3 proteins and Nha1 antiporter based on imaging and image analysis? Do results from other analysis confirm our result and thereby validate our method?*

We have used *YeastAnalysis* to address the role of 14-3-3 proteins and the *Nha1* transporter in the response of yeast cells under salt stress. 14-3-3 proteins are highly conserved eukaryotic proteins binding to hundreds of different mostly phosphorylated proteins. Also the *S. cerevisiae* 14-3-3 proteins, encoded by *BMH1* and *BMH2*, bind to hundreds of phosphorylated proteins [Heu95, Kak07, Heu09] and play a role in the regulation of many processes including tolerance to NaCl [Pos00, Zah12]. Deletion of *BMH1*, encoding the major 14-3-3 isoform, resulted in an increased sensitivity to  $\text{Na}^+$ ,  $\text{Li}^+$  and  $\text{K}^+$  and to cationic drugs [Zah12]. Testing the genetic interaction between BMH genes and genes encoding plasma membrane cation transporters revealed a genetic interaction between *BMH1* and *NHA1*. In addition, using bimolecular fluorescence complementation (BiFC) a physical interaction between 14-3-3 proteins and the Nha1 antiporter was shown. These results show that the yeast 14-3-3 proteins and an alkali-metal cation efflux system interact and that this interaction enhances cell survival upon salt stress [Zah12]. To further understand the role of 14-3-3 proteins and the Nha1 antiporter in salt stress resistance, the effect of high external NaCl concentration on the levels of these proteins is studied here. Cultivation in the presence of 0.5 M NaCl resulted in an increased level of both Bmh1-GFP and Bmh2-GFP. This observation is in line with a role of 14-3-3 proteins in tolerance to high environmental NaCl.

The results obtained by *YeastAnalysis* were further validated using Flow Cytometric analysis, which were in good agreement. This validation suggests the beneficial advantage of using this system in further gene expression analysis experiments.