Image analysis for gene expression based phenotype characterization in yeast cells

Tleis, M.

1

Introduction

“This chapter presents the reader with our research objective, the problem that we address and the goals of this research. In addition, it offers a basic background and definitions necessary to follow up in this thesis. Specifically, it introduces the necessary background about cytomic studies and pattern recognition methods followed within our research. The final section illustrates the scope or structure of the thesis.”

1.1 Research Objective

THIS thesis addresses the main research problem on how pattern recognition systems can support objective analysis and phenotype characterization of single-cell in image-based gene expression experiments. Hence, we address the crucial research questions directly connected to this problem. Our starting point is spherical/ovoid shape cells. First we address the components and processes required to build a comprehensive image analysis pipeline for single-cell image based gene expression. Second we address the best approach to segment ovoid-shaped cells that are common in micro/cell biology such as *Saccharomyces cerevisiae*. In addition, we address how a machine learning approach can aid in the object recognition process, and how it can improve the 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. Another directly related research question is about the features used by the machine learning approach, where we address the extraction of relevant and meaningful feature sets. Finally, we address the question on whether the recognition system can be validated on a yeast study experiment to discriminate various cell groups.

1.2 Cytomics and *Saccharomyces cerevisiae*

Cytomics is the study of cell systems (cytomes) at a single cell level. It combines all the bioinformatics knowledge to attempt to understand the molecular architecture and functionality of the cell system. Much of this is achieved by using molecular and microscopic techniques that allow the various components of a cell to be visualized as they interact in vivo [Bra11]. In this section, we define and discuss the gene expression and measurement followed within cytomics. Subsequently, we define and discuss fluorescent proteins used to study cell behaviours. Finally we present and discuss the eukaryote model organism used commonly in cell/micro biological studies, and which we take as a case study in our research, i.e. the *Saccharomyces cerevisiae* yeast cells.

1.2.1 Gene Expression and Measurement

In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype, i.e. observable trait. The genetic code stored in DNA is "interpreted" by gene expression, and the properties of the expression give rise to the organism's phenotype. Such phenotypes are often expressed by the synthesis of proteins that control the organism's shape, or that act as enzymes catalysing specific metabolic pathways characterising the organism. Protein-coding genes are transcribed into messenger RNA (mRNA), which is an information carrier coding for the synthesis of one or more proteins.

Measuring gene expression is an important part of many life sciences. The ability to quantify the level at which a particular gene is expressed within a cell or organism can provide a huge amount of information. Similarly, the analysis of the location of expression protein is a powerful

tool and this can be done on an organism or cellular scale. Investigation of localisation is particularly important for study of development in multicellular organisms and as an indicator of protein function in single cells. Ideally, measurement of expression is done by detecting the final gene product (for many genes this is the protein); however, it is often easier to detect one of the precursors, typically mRNA, and infer gene expression level. Levels of *mRNA* can be quantitatively measured by northern blotting, which provides size and sequence information about the *mRNA* molecules.

For expression profiling or high-throughput analysis of many genes within a sample, quantitative *PCR* may be performed for hundreds of genes simultaneously in the case of low-density arrays. A second approach is the hybridization microarray, which is a popular approach in gene expression studies. Microarrays reveal expression profiles for a large number of genes at different time points. A single array or "chip" may contain probes to determine transcript levels for every known gene in the genome of one or more organisms [Wel07]. Alternatively, "tag based" technologies can be used, such as Serial analysis of gene expression (*SAGE*), which can provide a relative measure of the cellular concentration of different mRNAs. Next-generation sequencing (*NGS*) such as *RNA-Seq* is another approach, producing vast quantities of sequence data that can be matched to a reference genome. Although *NGS* is comparatively expensive, and resource-intensive, it can identify single-nucleotide polymorphisms, splice-variants, and novel genes, and can also be used to profile expression in organisms for which little or no sequence information is available.

For genes encoding proteins, the expression level can be directly assessed by a number of means with some clear analogies to the techniques for *mRNA* quantification. The most commonly used method is to perform a Western blot against the protein of interest. The gel-based nature of this method makes quantification less accurate although it has the advantage of being able to identify later modifications to the protein [Nei00, Ama08]. Moreover, Mass spectroscopy is developing fast and allows the quantification of a large part of the proteome, which directly addresses the level of gene products present in a given cell state and can further characterize protein activities, interactions and subcellular distributions [Ong05]. Mass spectrometry (MS) is an analytical technique that ionizes chemical species and sorts the ions based on their mass to charge ratio. In simpler terms, a mass spectrum measures the masses within a sample.

Analysis of expression is not limited to only quantification; localisation can also be determined. *mRNA* can be detected with a suitably labelled complementary *mRNA* strand and protein can be detected via labelled antibodies. The probed sample is then observed by microscopy to identify where the *mRNA* or protein is.

By tagging the gene with a reporter gene, i.e. by replacing the gene with a new version fused to the reporter gene expressing fluorescent proteins as markers, expression may be directly quantified in live cells. It is very difficult to clone a reporter gene into its native location in the genome without affecting expression levels so this method often cannot be used to measure endogenous gene expression. It is, however, widely used to measure the expression of a gene artificially introduced into the cell; for example, via an expression vector. It is important to note that, in some cases, by fusing a gene to a fluorescent reporter the expressed protein's behaviour, including its cellular localization and expression level might change. The analysis

of reporters is initiated by imaging using a confocal laser scanning microscope and by flow cytometry, discussed hereafter.

1.2.2 Fluorescent Microscopy and Flow Cytometry

There are two popular cell analysis techniques including fluorescent microscopy and flow cytometry, these are discussed and subsequently their complementarity is considered.

Microscopy

Proteins that are tagged with fluorescent molecules can be studied through imaging. The most common techniques is by using fluorescent microscopy. Laser scanning microscopy in general is preferred for fluorescence imaging because of their resolution, higher contrast, ability to reconstruct 3-D images, the absence of artefacts induced by conventional microscopy, and most importantly its ability to penetrate into the specimen and obtaining an image of a specific focal plane [Mas01]. The multi-photon photo-luminescence microscopes have high spatial resolution and reduced background [Zho10] and also permit additional structures to be observed [Mas01]. However, multi-photon microscopes do not contain pinhole apertures, which give confocal microscopes their optical sectioning quality [Kam13]. In addition, modern confocal laser scanning microscope *CLSM* can do fast scanning that prevents photo-toxicity due to thermal damage [Paw10], as well as minimizing photo-bleaching. In confocal microscopy, the focus plane of illumination is the same as the focal plane of detection. In other words, the focus plane of illumination and the focal plane of detection are confocal [Row00].

Flow Cytometry

In cell biology, flow cytometry is a laser-based, biophysical technology employed in cell counting, cell sorting, biomarker detection and protein engineering. It suspends cells in a stream of fluid and passes them by an electronic detector. Flow cytometry allows simultaneous multi-parametric analysis of the physical and chemical characteristics of up-to thousands of particles per second [Yan15]. A flow cytometer is similar to a microscope; however, it doesn't produce an image of the cell but offers high-throughput automated quantification of the set parameters for a high number of single cells during each analysis session [Tho06].

Complementarity of Flow Cytometry and Fluorescence Microscopy

Flow cytometry and fluorescence microscopy both provide single-cell analysis using different but complementary sets of data, essentially population-based target intensities versus target morphology in relatively small sample sizes. Both approaches employ optical filters to analyze fluorescence emissions and have to overcome some of the same physical limitations including spectral overlap of dyes and the dynamic range limits of measuring systems. Hence, flow cytometry and confocal fluorescence microscopy technologies both have specific characteristics and limitations. In microscopy, photostability is a more critical issue. Flow cytometry is

limited by its requirement that analyzed cells are in suspension, making information on tissue architecture and cell-cell interactions inapplicable. On the other side, fluorescence microscopy is well suited to the resolution of cell and tissue architecture, and to following kinetic and trophic responses in single cells. In flow cytometry, cell subpopulations with similar marker expression are difficult to differentiate, and analyses that employ more fluorophores are subject to signal spillover [Jah12]. In addition, there is the inability of flow cytometry to recognize morphologically analyzed cells [Mur08]. Flow cytometry rapidly quantifies small differences between cell populations using statistically significant numbers of events. Flow cytometry can represent a “black box” when looking at the magnitude of a population response; fluorescence microscopy can help verify that measured results represent meaningful biological effects.

In general, the microscopist may arrive at quantitative data. However, the cooperative use of both flow cytometry and microscopy can provide more robust numerical description of biological phenomena [God05].

1.2.3 Cytomics and Fluorescent Proteins

Proteins are vital parts in living organisms. Many important proteins in human biology were understood by studying their homologs in yeast; such proteins include cell cycle proteins, signalling proteins, and protein-processing enzymes [Wal04]. The large-scale study of the structures and functions of such proteins is called Proteomics [Bla99]. In Proteomics, Fluorescent proteins such as green fluorescent protein (*GFP*) and its derived variants are widely used. One of the most exciting applications is the generation of a library of *S. cerevisiae* strains in which each coding sequence is tagged with green fluorescent protein (*GFP*) [Huh03]. This library enables the determination of the localization of more than 70 percent of the *S. cerevisiae* proteins. In addition, levels of these proteins can be quantified after cultivation under different conditions.

Tagging genes with the reporter expressing green fluorescent protein (*GFP*) is a highly specific and sensitive technique for studying the inter-cellular dynamics of proteins and organelles [Sha97]. *GFP* expression is an excellent marker to monitor the gene expression [Phi01] and protein localization in the living yeast cells. The biggest advantage of the intracellular *GFP* is that it is heritable, since it can be transformed with the use of DNA-encoding *GFP*. Additionally, visualizing *GFP* is non-invasive as it is detectable by just shining light on it. Furthermore, it is a relatively small and inert molecule that does not appear to interfere with cell growth and function. Moreover, if *GFP* is used with a monomer it can diffuse readily throughout cells [Cha09].

The green fluorescent protein (GFP) was first isolated from the jellyfish *Aequorea victoria* [Sha97, Phi01]. It is a protein composed of 238 amino acid residues (26.9 kDa) that exhibits bright green fluorescent when exposed to light in the blue to ultraviolet range [Wal04, Moy08, Cha94]. It has a beta-barrel structure consisting of eleven β-strands. The beta barrel structure is a nearly perfect cylinder, 42 Å long and 24 Å in diameter, creating what is referred to as a "β-can" formation, which is unique to the GFP-like family [Yan97]. In GFP the fluorophore is formed inside the protein globule by modification of amino acids [Shi79].

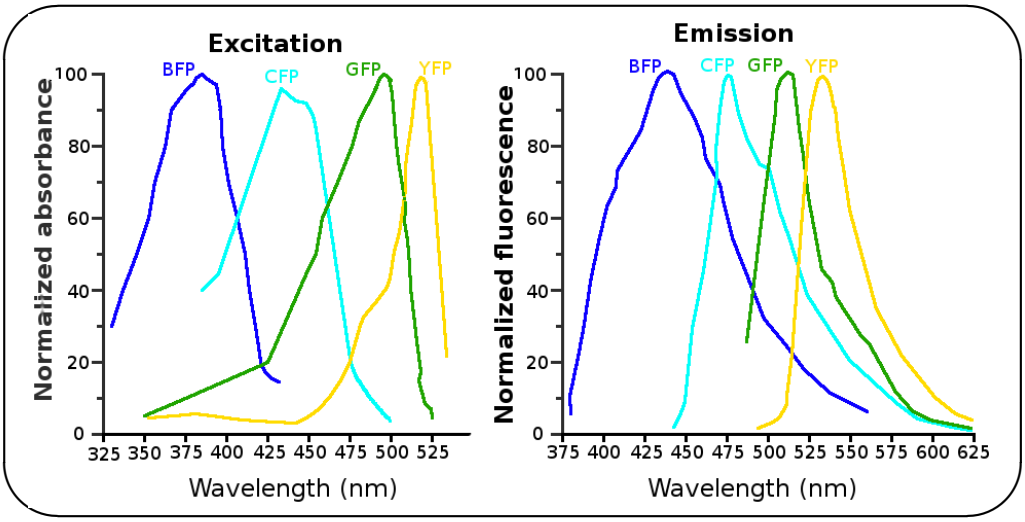

Figure 1.1: *Excitation and Emission of GFP and some variants. Data from [Hei96].*

The *GFP* gene is widely used as a reporter gene and can be introduced into organisms and maintained in their genome through breeding, injection with a viral vector, or cell transfection. The *GFP* gene has been introduced and expressed in the *S. cerevisiae* yeast cells as well as other types of yeast cells, bacteria, fungi, fish (such as zebrafish), plant, fly, and mammalian cells, including human [Cha09].

GFP requires low excitation light intensity to prevent photo-bleaching and photo-toxicity at a light wavelength of 490 nm or less [Sha97]. The *GFP* from *A. victoria* has a major excitation peak at a wavelength of 395 nm and a minor one at 475 nm. Its emission peak is at 509 nm, which is in the lower green portion of the visible spectrum [Phi01]. This fluorescence is very stable, and virtually no photo-bleaching is observed [Cha94].

Many different mutants (variants) of the *GFP* have been engineered, with improved spectral characteristics of *GFP*, resulting in increased fluorescence, photo-stability, and a shift of the major excitation peak to 488 nm. *EGFP* and *Superfolder GFP* are examples of such variants. Many other mutations have been made as well including color mutants; in particular, blue fluorescent protein (*EBFP, EBFP2, Azurite, mKalama1*), cyan fluorescent protein (*ECFP, Cerulean, CyPet, mTurquoise2*), and yellow fluorescent protein derivatives (*YFP, Citrine, Venus, YPet*). They exhibit a broad absorption band in the ultraviolet spectrum (cf. Fig. 1.1).

Knowing how much of a protein is expressed is not sufficient to understanding its behaviour. It is particularly important to also know its subcellular location because changes in protein subcellular location can cause dramatic effects on cell behaviour. Changes in location within a cell type may also cause or result from disease [Mur05]; however, subcellular location has received less attention than many other aspects of gene and protein behaviour. The major exception is in yeast, in which almost all proteins have been assigned to a set of major subcellular structures using fusion of DNA, with the coding sequence of fluorescent proteins such as the

green fluorescent protein. For example, Huh et al [Huh03] used green fluorescent protein tagging of DNAs and visual examination to assign proteins to 12 categories: cell periphery, bud, bud neck, cytoskeleton, microtubule, cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, vacuole, vacuolar membrane, and punctate. They then used colocalization with red fluorescent protein markers to divide the cytoskeleton class into two classes, actin cytoskeleton and spindle pole, and to add nine new categories: nucleolus, nuclear periphery, golgi apparatus, three types of transport vesicles, endosome, peroxisome, and lipid particle. In all, 4,156 proteins were assigned to these 22 categories in their study [Huh03, Mur05].

In our experiments on *S. cerevisiae*, the model organism is tagged with fluorescent proteins and is visualized by confocal laser scanning microscope. The subsequent sub-section introduces this organism that is used to understand gene expression and genetic networks in cytomics.

1.2.4 Saccharomyces cerevisiae in cytomics

Saccharomyces derives from Latinized Greek and means "sugar-mold" or "sugar-fungus", *saccharo* being the combining form "sugar" and *myces* being "fungus". *Cerevisiae* comes from Latin and means "of beer". This organism is also known as Baker's yeast, Brewer's yeast, Ale yeast, Top-fermenting yeast and Budding yeast [Stă13].

S. cerevisiae is a well-known yeast species and used since ancient times in wine-making, brewing and baking. It was originally isolated from the skin of grapes and has been one of the most intensively studied eukaryote model organisms in molecular and cell biology. [Fel10]. *S. cerevisiae* cells are round to ovoid with a diameter between 2 to 10 micrometers. [Par97].

Many cell processes in the yeast model eukaryote cell are similar to that in plants and mammalians including humans. This fact makes yeast an excellent model organism to understand the behaviour of proteins involved in such processes. Several traits in the *S. cerevisiae* drive researchers to look for this organism. Among these traits is its size, generation time, accessibility, manipulation, genetics, conservation of mechanisms, potential economic benefits [Gov11], cell's transparency, the fact that its genome sequence was completed in 1996, and the availability of a library of strains in which each individual coding sequence is tagged with green fluorescent protein (GFP) [Huh03] in addition to a library of knock-out strains [Win99] and a library of TAP-tagged strains [Gha03]. These traits make *S. cerevisiae* a significant tool in biological research. Studying DNA damage and repair mechanisms is one example [Nic01].

S. cerevisiae yeast cells can survive and grow in two forms; as haploid or diploid cells where they undergo a simple life-cycle of mitosis and growth. Haploid cells usually die under stress conditions, while diploid can undergo sporulation, entering meiosis and producing four haploid spores, which can proceed to mate. This reproduction process is known as budding and there where Budding yeast get their name from. With adequate nutrients, yeast cells can double in numbers within 100 minutes [Her88]. The mean replicative life span of the *S. cerevisiae* is about 26 cell divisions [Kae05, Kae10].

In molecular genetics, *S. cerevisiae* is used as a model system in the understanding of gene expression and genetic networks, because it combines considerable variation in key cell characteristics such as protein levels and expression, cell size, shape and age. It also has short

generation time and immobility [Don13b]. Moreover, it can be manipulated and genetically engineered.

Genetic and hereditary diseases are still incurable because we lack the understanding behaviours of many proteins. This fact drives research groups to use the *S. cerevisiae* yeast cell to understand the behaviour of proteins responsible for many processes in the cells. Such as the 14-3-3 family of proteins which is considered vital to cell life. In order to prepare such cells for the experiments, the gene expressing a protein under study is attached (tagged) to a report gene that expresses fluorescent protein such as the popular green fluorescent protein (*GFP*) or any of its variants. How we use this model organism to recognize patterns is the topic of next section.

1.3 Pattern Recognition

Pattern recognition aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space. A complete pattern recognition system consists of: (i) a sensor that gathers the observations to be classified or described, (ii) a feature extraction mechanism that computes numeric or symbolic information from the observations and (iii) a classification or description scheme that does the actual job of classifying or describing observations relying on the extracted features [dic16].

In our research, the sensor used is the fluorescence microscope that gathers the observations, i.e. the yeast cell images. Such image observations are acquired by *CLSM* microscopy and are discussed in the first sub-section. From such observations image analysis techniques are applied to extract the features. Such techniques involve image processing, image segmentation, and object measurement techniques; which will be the topic of the second sub-section. The last part of the pattern recognition discussion is dedicated to the data analysis and classification of the measured features obtained by image analysis.

1.3.1 Image Acquisition

All the images created in this work were analyzed by confocal laser scanning microscopy (CLSM) to view fluorescent tagged proteins expressed within the cells. Images are acquired as two or three channel images. One or two channels of the reporter construct (*GFP, YFP, CFP*). The *GFP* is excited at 488nm and *YFP* at 514nm, and emission is at 500-550 for *GFP* or 530-600nm for *YFP* [Zah12]. In addition, a bright-field channel is acquired; the bright-field image depicts the structure of the cells.

The Bright-field image channel is acquired through a bright-field technique embedded within the confocal microscope. Since it depicts the yeast structure, this channel of the microscope images of yeast *S. cerevisiae* cells is used primarily, in our research, to detect the cell contours. For optimal detection, optimal microscope settings have to be set and hence the parameters to be used have to be determined. To determine these parameters, an experiment was done where several images were generated under different microscope settings.

Table 1.1 lists all the different settings used, where each setting is labelled with a different letter (A...E). This label represents a different category of images. To choose the optimal image category that works best with segmentation algorithms. We applied a segmentation algorithm to try the detection of cell objects in these images. A score of true positive cell detections was computed for each category.

Image Label	Resolution	Master Gain	Digital Offset	Laser
A	512x512	320	-1.20	18%
B	1024x1024	320	-1.20	18%
C	1024x1024	340	-1.90	20%
D	1024x1024	261	-0.05	18%
E	1024x1024	301	-0.50	18%

Table 1.1: *microscope settings*

The result in Table 1.2 shows different labelled images with their acquired scores. Each image label corresponds to images acquired with the microscope settings listed in Table 1.1. The score of true positive rate indicates the number of cells correctly detected. The highest score corresponds to category labelled as B, i.e. the microscope settings of 1024x1024 resolution, 320 master gain, -1.20 digital offset and 18% laser power.

Image Label	True Positive
A	67%
B	70%
C	52.6%
D	48.4%
E	47.6%

Table 1.2: *Percentage of True Positives*

The availability of *GFP* and its derivatives has thoroughly redefined fluorescence microscopy and the way it is used in cell biology and other biological disciplines [Yus05]. Such Fluorescent (photon-emitting) molecules are introduced into yeast cells because they have a helpful property of fluorescing when in the presence of non-fluorescent molecules or structures under study. When the *S. cerevisiae* specimen is illuminated using laser techniques in confocal microscopy, the fluorophores (fluorescent molecules) absorbs the light photons raising them to an excited state with a wavelength (energy) specific to the fluorophore itself. The fluorophore then returns to its ground state and may emit a photon with lower energy (longer wavelength). This photon might then strike the detector with the proportion of light entering the objective lens of the microscope. The charges of electrons produced by photons striking the detector are quantified and from these quantifications, pixel values are determined. These pixel values correspond to the number of detected photons (cf. Fig. 1.2). From the number of photons detected at each pixel, interpretations can be made about the presence or absence of some feature, the size and shape of a structure, or about the relative concentration of a molecule [Ban13]. The excitation and emission wavelength used to detect the fluorescent molecules varies with different types. For example the *GFP* is imaged with excitation at 488 nm and emission at 505-530 nm.

Now that the image acquisition is performed, we shift our discussion to the image analysis phase.

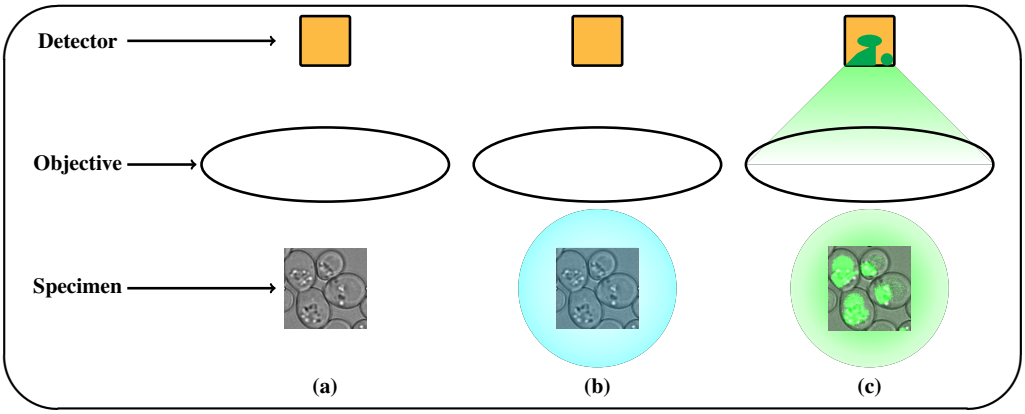

Figure 1.2: *Schematic Image Formation of Fluorescence Microscopy. (a) - basic setup. (b) - the specimen is illuminated in higher energy light, fluorophores become excited. (c) - Lower energy Light is emitted; some enters the objective lens and is detected.*

1.3.2 Image Analysis

When digital cameras where introduced to microscopy, digital image analysis has developed as an established complement, allowing routine quantification of microscope observations [Tho96]. Currently, flow cytometry is one of the methods used to measure levels of fluorescent proteins; this can, however, not provide the quantification that image analysis can. Hence, the use of image analysis was probed to accomplish further progress. Digital image analysis, widely known as image analysis, is when a machine automatically studies an image to obtain useful information from it. The applications of digital image analysis are continuously expanding through all areas of science and industry. It has been successfully applied in a wide variety of fields ranging from astronomical observations to cell analysis. To name few other fields: nuclear medicine, medical diagnostics, industry, lithography, microscopy, lasers, biological imaging, remote sensing, law enforcement, radar images, geological exploration [Gon08], sports management [Fer99] and chemical imaging [Sch95]. There are many different techniques used in automatically analyzing images. Each technique may be useful for a small range of tasks. However, there still are not any known methods of image analysis that are generic enough for wide ranges of tasks, compared to the abilities of a human's image analyzing capabilities [Sol11]. In our domain, we considered few of these techniques including image processing, image segmentation and feature extraction discussed hereafter.

Image Processing

Image processing is usually used to refer to digital image processing. It is a process that accepts an image as an input, and the output may be either image or a set of characteristics or parameters related to the image. It is the use of computer algorithms to perform image processing on digital images. Since images are defined over two dimensions (perhaps more),

image processing may be modelled in the form of multidimensional systems. Image processing is the only practical technology for classification, feature extraction, multi-scale image analysis and pattern recognition [Gon08].

Image Segmentation

Crucial to image analysis is image segmentation, which is defined as subdividing an image into its constituent regions or objects [Gon08]. The result of image segmentation is objects representing connected regions of similar intensity. From these regions, measurements can be conducted. Some practical applications of image segmentation are image processing, computer vision, face recognition, medical imaging, digital libraries, image and video retrieval, and cell image analysis to measure gene expression [Tle13, van07].

Segmentation of individual cells relies on the ability to detect cell boundaries and classifying all pixels in a given image as foreground or background. The differentiation between foreground and background pixels can be accomplished by a threshold function determined by a simple intensity based method, or by more complex functions such as graphical models, pattern recognition, deformable templates, cell contours or the watershed algorithm [Don13b].

There are a number of refining segmentation algorithms, tracking algorithms, morphology characterization, and protein localization. However, we lack a robust approach for the segmentation and tracking of budding yeast [Don13b]. The result of such a robust segmentation algorithm enables us to extract binary masks and contours of cells. These obtained masks and contours allow us to measure various features of those objects, i.e. cells. More details on object measurement follow hereafter.

Measurement and Features

Herein, we will define the concepts of measurement, features, textures and feature extraction.

Measurement

In its classical definition throughout physical sciences, measurement is the determination or estimation of ratios of quantities [Mic99]. However, information theory recognises that all data are inexact and statistical in nature. Thus the definition of measurement in information theory is "a set of observations that reduce uncertainty where the result is expressed as a quantity" [Hub07]. In general, we can state that measurement is the assignment of a number to a characteristic of an object or event, which can be compared with other objects or events [Ped91]. The science of measurement is pursued in the field of metrology, which includes all theoretical and practical aspects of measurement [BIP08].

Measurement is a cornerstone in Bio-Imaging and pattern recognition as well as in most science fields. It is an important step in the discovery process. In Biological image analysis, measurement is strongly related to features and textures extracted from images. Hereafter, we will define what a feature is? what a texture is? and we will highlight on the well-known feature extraction techniques followed within bio-imaging.

Features

In computer vision and image processing, a feature is a piece of information which is relevant for solving the computational task related to a certain application. This is the same sense as feature in machine learning and pattern recognition generally, though image processing has a very sophisticated collection of features. Features may be specific structures in the image such as points, edges or objects. Features may also be the result of a general neighborhood operation or feature detection applied to the image. In general definition, an image feature is a representation or an attribute of an image describing certain special characteristics of the pattern of interest.

In our application it is not sufficient to extract only one type of feature to obtain the relevant information from the image data. Instead multiple different features are extracted, resulting in multiple feature descriptors at each image point. The information provided by all these descriptors is organized as the elements of one single vector, referred to as a feature vector. The set of all possible feature vectors constitutes a feature space. A common example of feature vectors appears when each image point is to be classified as belonging to a specific class. Assuming that each image point has a corresponding feature vector based on a suitable set of features, meaning that each class is well separated in the corresponding feature space, the classification of each image point can be done using standard classification methods [wik15]. Chapter 4 applies such classification on our feature space.

Textures

An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. Image texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image [Sha01]. In other words, it is defined as the visual effect which is produced by spatial distribution of total variations over relatively small areas [Bar95]. Image textures are believed to be a rich source of visual information. They are complex visual patterns composed of entities, or sub-patterns, that have characteristic brightness, colour, slope, size, etc... Thus a texture can be regarded as a similarity grouping in an image. Image textures can be artificially created or found in natural scenes captured in an image. Image textures are one way that can be used to help in segmentation or classification of images. To analyze an image texture in computer graphics, there are two ways to approach the issue: Structured Approach and Statistical Approach. A structured approach sees an image texture as a set of primitive texels in some regular or repeated pattern. This works well when analyzing artificial textures. However, since natural textures are made of patterns of irregular sub-elements as is the case in our application, statistical approach is used. In general, statistical approach is easier to compute and is more widely used. It sees as image texture as a quantitative measure of the arrangement of intensities in a region.

Feature extraction

Feature extraction is defined as locating those pixels in an image that have some distinctive characteristics [Gub09]. The most known feature extraction techniques in image analysis are classified into first-order histogram based features, invariant moment features, co-occurrence matrix based features, and multi-scale features [Mat98]. More details about these techniques will be discussed in Chapter 4.

1.3.3 Data Analysis

After performing the measurement, data analysis is necessary to make the measurement meaningful for users. Data analysis can be defined as the process for obtaining raw data and converting it into information useful for decision making and suggesting conclusions [Jud11]. The workflow depicted in Fig. 1.3 illustrates the general process followed within bioinformatics. Initially images are acquired from the imaging device during the image acquisition phase; then these images undergo segmentation to locate the individual objects. Following the segmentation is the measurement phase where objects are measured for various features. The obtained data must be processed or organized for analysis [Sch13], but the data could be incomplete, contain duplicates, or contain errors. Such issues are presented and corrected through data cleaning [Ara15]. Once the data is processed and cleaned, it can be analyzed. Analysts apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data [Few04]. Descriptive statistics are generated to help understand the data. These data are examined in graphical format using data visualization techniques to communicate key messages contained within the data through these graphical means and charts [Fri08]. Machine learning models and algorithms can be applied to identify relationships among the variables or to classify the instances of the data based on these variables. Data mining and machine learning plays a major role here. They focus on modelling and knowledge discovery for predictive rather than purely descriptive purposes. The output of such models is data products fed back to the user such as conclusions or identified subtle patterns residing within the data.

1.4 Thesis Structure

In the remaining chapters we explain the research we have set out to perform.

Chapter Two: Image Analysis Platform. This chapter presents a complete framework for biological experiments. It demonstrates how an automated platform based on a complete image analysis pipeline assist biologists in their experiments? Moreover, this chapter discusses the individual modules that form the complete framework, including the segmentation, measurement, data analysis and the GUI that combines these modules together.

Chapter Three: Hough Transform Based Contour Extraction and Optimization. In this chapter, we show how a new approach based on Hough transform and minimal path algorithms can improve the segmentation of ovoid objects, i.e. yeast cells. We start by defining Hough transform and minimal path algorithms. Subsequently we present our general approach to detect ovoid objects in microscope images by detecting circular arcs using a variety of the Hough transform. In addition, we discuss the application of minimal path algorithms to extract the exact contour of detected objects from a polar representation of the image surrounding the object. Furthermore, this chapter presents an additional novel algorithm to expand the extracted contours of ovoid objects. Such expansion is necessary for some settings due to the inherently fuzzy nature of edges and delicate microscope settings. This chapter explains how the polar

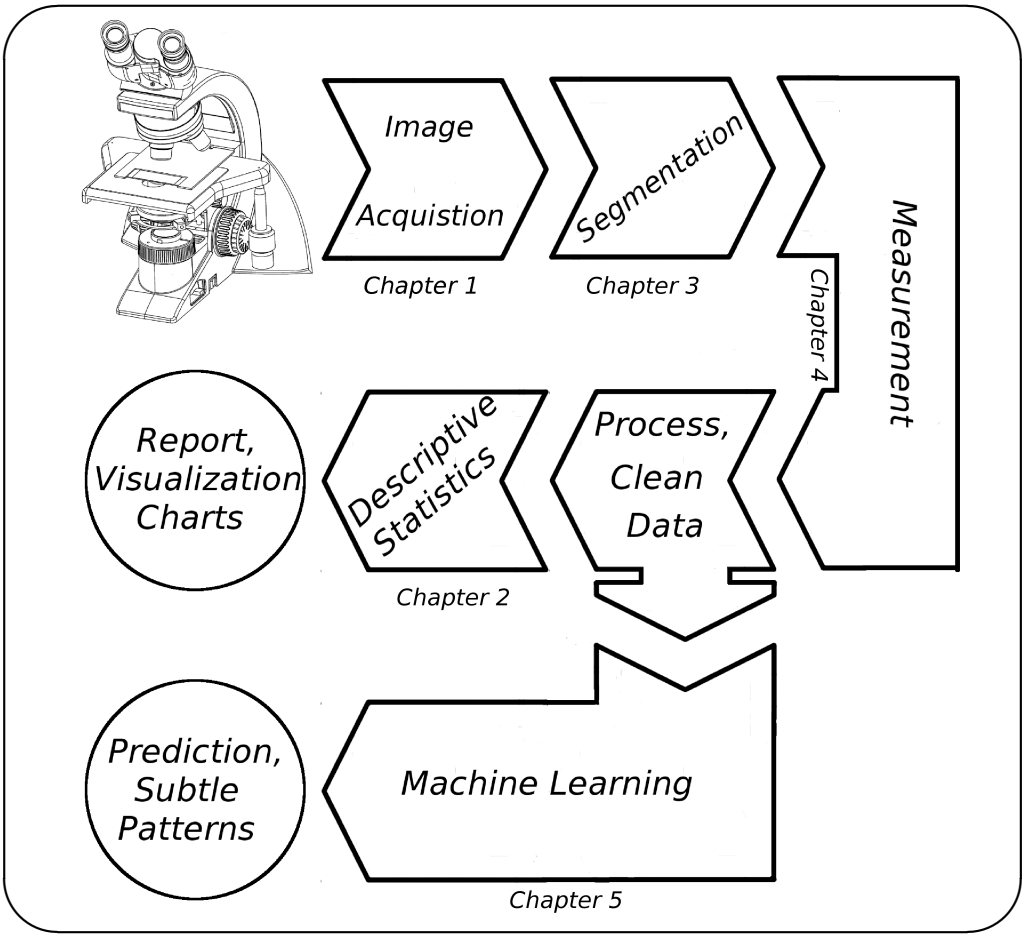

Figure 1.3: *Flow Chart of the data analysis process*

representation of images is used to expand the initially detected contours by applying circular shortest paths. In addition, it explains the three introduced parameters to control the expansion process. These parameters are *resistance, limit* and *convergence*. The expansion of a sample contour is demonstrated as well in this chapter. Finally, results and comparison with other methods are evaluated using a dataset of *S. cerevisiae* yeast cells.

Chapter Four: Machine Learning to Improve Object Recognition and Discrimination of Cell Groups Using Sophisticated Features. This chapter specifically address machine learning where we introduce features to be used in a machine learning approach to automatically identify cell groups cultivated in two different media. We use the same approach to classify cell objects from artefacts. First we discuss the feature extraction techniques including first-order histogram features, texture measurement, moment invariants, co-occurrence matrix based features and multi-scale wavelet-based texture measurement. Subsequently, various

classification methods are evaluated to build a model imported into the yeast analysis platform to be trained for the automatic discrimination of cell groups. This discrimination helps showing that there are different gene expression patterns between cells cultivated under different stress levels. Moreover, the same classification methods are evaluated to build another model for the identification of cell objects. This model is used to discriminate the segmented objects in images into intact cell objects or artefacts such as debris and dead cells.

Chapter Five: The Effect of NaCl on 14-3-3 Proteins and Nha1 antiporter. In this chapter, the designed image analysis platform is used in a case study to determine the effect of sodium chloride on 14-3-3 genes including Bmh1 and Bmh2 in addition to the *NHA1* encoding an antiporter. The study also includes a mutant of *BMH1 (Δbmh1)* to study the expression of Nha1 under different osmotic stress levels. The result obtained from using the yeast analysis software is also validated with that obtained from flow cytometry.

Chapter Six: Discussion. In this chapter, we draw out conclusions learned from this dissertation and give scope for further research and applications.