Biological model representation and analysis

Cao, L.

Chapter 1

Introduction

Microscopes have enabled mankind to increase the resolution of their vision in the micro and nano-scales. Fabulous visualizations can be achieved in this manner, but there are still patterns in these visualizations that need to be addressed. The technical developments in microscope instrumentation are incredible and have made it possible to observe biological structures at near molecular scale up to the tissue and organismal level (cf. Figure 1). The developments of digital instruments and computers have further boosted the area of microscope analysis. Microscopes are equipped with digital cameras and researchers can produce large amounts of high quality digital images. In these images there are patterns to be analyzed and thus we need to look for efficient and correct ways to extract information from these images and find patterns in this information. This particularly holds for the description of biological specimens that are observed in one way or the other by microscopy. The research of this thesis contributes to the efforts to find solutions in working with large amounts of images and extracting information in a correct and comprehensive way.

In this thesis, we discuss solutions of phenotype description based on the microscopy image analysis to deal with biological problems both in 2D and 3D space. Our description of patterns goes beyond conventional features and helps to visualize the unseen in feature dataset. These solutions share several common processes which are based on similar principles. Furthermore, we notice that advanced features and classifier strategies can help us improve the performance of the solutions. The biological problems that we have studied include the endocy-

tosis routing using high-throughput screening in 2D and time and 3D geometrical representation from biological structures.
In order to have a general view of the solutions, we would first introduce the generic workflow as shown in Figure 1.2 which is applicable for both 2D and 3D objects. For 3D images additional techniques are required.

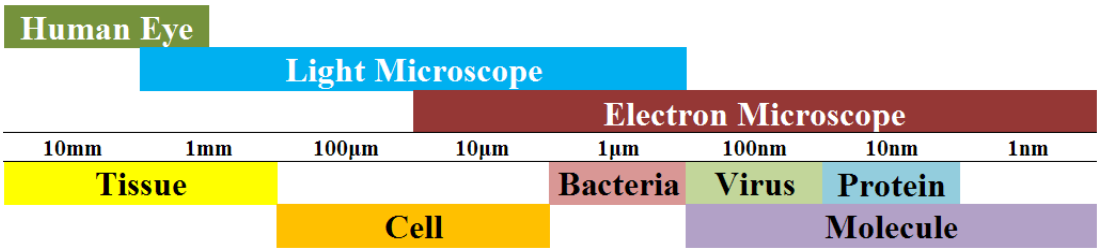

Figure 1.1. Resolution required for several objects (middle line) and the imaging equipment with which this resolution can be achieved (upper half) and the typical resolution of objects (lower half).

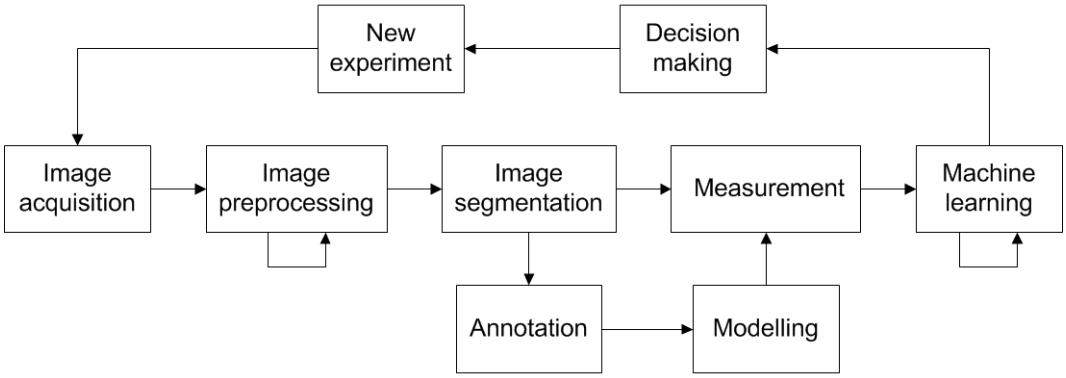

Figure 1.2. Image processing and image analysis pipeline. Image processing includes image preprocessing and image segmentation. Image analysis includes image annotation, modelling, measurement and machine leaning. The pipeline influences the decision making for a new experiment.

1.1 2D and 3D microscopy images acquisition

In the imaging pipeline, the first step is the image acquisition from microscopes. In this thesis our inputs come from different microscopes. We utilize the bright-field microscope for invasive sectioning of large structures in X,Y,Z space. For

non-invasive observation on cell level, we use confocal microscopy both for single slide screening in X,Y,T space and multiple slides sectioning in X, Y, Z space. Bright-field microscope is suitable for the imaging of stained tissue sections of specimens because the easiest way to deal with thicker specimens (3D images) is to slice the specimen into many consecutive thin sections. Given the resolution of the imaging system, they provide clear information in x and y axis (2D images), but limited information in z axis. This invasive sectioning enables the application of staining techniques so that molecular phenotype of the specimen under study can be revealed [Verbeek, 1999b]. This approach is useful for larger sections of tissue whose z-resolution is of the order of millimeters, but it will not work in the micro-nano range.

Confocal microscopy can be used in cellular high-throughput screening. It also enables observation of thick specimens by optical sectioning which eliminate the artifacts existed in specimen preparation by physical sectioning. However, optical penetration into the specimen has its limitations. The scope of confocal microscope is very good on the cellular level but less effective on the level of an organ or a tissue; concluded from Figure 1.1. Consequently, for 3D reconstruction of a larger embryo or a substantial part of it, confocal microscope is usually not always the most appropriate technique [Verbeek, 1999b].

Optical projection tomography (OPT) is another non-invasive sectioning technique for 3D biological specimens. It aims at producing high-resolution 3D images of both fluorescent and nonfluorescent biological material with a thickness of up to 15 millimeters. OPT microscopy allows the rapid mapping of the tissue distribution of RNA and protein expression in intact embryos or organ systems and can therefore be instrumental developmental biology studies for objective phenotype description. [Sharpe et al., 2002]

Subsequent to image acquisition, the process of image and data analysis starts. For both 2D and 3D microscopy images, the solutions for analysis are quite similar. In this thesis, the major focus is on the description of the phenotype measurement and data analysis. Therefore, we introduce a generalized solution for 2D and 3D images in the following sections.

1.2 Image processing and analysis

The aim of image processing and analysis is to accomplish image understanding and data reduction. The pipeline includes image enhancement or restoration and image segmentation. The first step of image processing and analysis for microscopy is to improve the quality of images by enhancing the foreground as well as suppressing the background. Image enhancement aims for improving the interpretability or perception of information in images for human viewers [Maini and Aggarwal, 2010]. We see it separated into two main categories: spatial domain filters and frequency domain filters. Spatial domain filters directly deal with the image pixels such as histogram enhancement. Frequency domain filters are performed using the Fourier transform of the image and include low-pass filters, bandpass filters and high-pass filters. Noise suppression algorithms often make a tradeoff between actual noise removal and preservation of real low-contrast detail. Most commonly used methods are linear filters such as *Gaussian filter*, *Wiener filer* [Wiener, 1964] and non-linear filters such as *median filter* and the filters based on the paradigm of its mathematical morphology [Serra, 1983].
Image segmentation is the technique dividing the image constituent parts most notably in foreground and background. So it results in a separation of foreground and background. In microscopy it is specifically used to detect objects, object regions or edges in an image. Basically the image segmentation is divided into two approaches: region-based segmentation and edge-based segmentation [Tripathi et al., 2012]. Region-based segmentation partitions an image into regions that are similar according to a set of predefined criteria [Gonzalez and Woods, 2001]. Some representative methods include *thresholding*, *clustering* and *region growing*. The thresholding operation converts a gray-scale image into a binary image by a set of thresholds. Popular methods include the *maximum entropy method* [Leung and Lam, 1994], *Bernsen's method* [Bernsen, 1986], *Niblack's method* [Niblack, 1985], *Isodata method* [Manakos et al., 2000], *Otsu's method* [Otsu, 1979]. Clustering partition the image into the sets or clusters of pixels which have similar feature space. Clustering methods can be further divided into *k-means clustering* [Kanungo et al., 2002] and *fuzzy clustering* [Naz et al., 2010]. Region growing extracts a region of the image that is connected based on predefined criteria [Chen

and Shen, 2010]. Region growing techniques are often used in noisy images where edges are extremely difficult to detect. Some well-known region based segmentation methods include *the level set method* [Qu et al., 2007], *watershed transformation* [Vincent and Soille, 1991] and *texture segmentation* [Ray et al., 2008]. In Edge-based segmentation, an image is partitioned based on abrupt changes in the intensity values [Gonzalez and Woods, 2001]. In Edge-based segmentation first the edges are identified. These are linked together to form consistent boundaries. Many edge operators are applied to locate edges in images such as *the Sobel operator*, *the Prewitt operator* and *the Canny operator* [Gonzalez and Woods, 2001]. The canny operator is used to find the edge pixels while eliminating the influence of noise. Other well-known edge-based segmentation method is *active contours* [Kass et al., 1988].
If the aim is to measure information on objects in the image then subsequent to segmentation a labeling operation is required. Each object from the segmentation process is attributed a label which can, if necessary, be given an annotation [Verbeek, 1999a] to provide biological context. An automatic annotation method can be regarded as a multi-class object classification which is based on image analysis to extract features and data analysis to train a proper classifier. In the next section, we introduce the necessary concepts and context for this thesis.

1.3 Phenotype measurement

In order to correctly annotate each object separated from the segmentation method, we need to quantify the object into all kinds of features describing the unique pattern of the object for further multi-class classification. This quantification step is, de facto, the measurement of the phenotype. Here we introduce two basic definition on phenotype measurement.

Definition 1.3.1. *"Phenotype is the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment."*

Definition 1.3.2. *"Phenotype measurements imply the measurement of observ-*

able attributes, reflecting the biological function of gene variants as affected by the environment." [Paulus et al., 2013]

For biological specimens, phenotype measurement is the next important step in image analysis. The phenotype measurements will be used to classify objects obtained from the segmentation into different categories that are meaningful with respect to the biology. Thus, it is crucial to measure representative features for each object in the image. These features, often, represent the characteristics of shape, intensity and texture of the objects.

Generally in 2D space, we can categorize the phenotype measurements into two groups: basic measurements and localized measurements. Basic measurements of the phenotype cover shape descriptors, texture patterns and invariant features. Localized measurements of phenotype describe the assessment of the correlation between multiple information channels. The information channels in the context of the research presented in this thesis are the imaging channels. The reason for splitting the channels in different parts of the color spectrum is that each channel contains individual characteristics of the object under investigation due to a specific staining method resulting in biological meaning.

For 3D images and 3D geometrical models, we need to look at different features. One group concerns the feature-based measurements including both global shape, such as volume and surface area, and local features such as surface curvature. Another part of phenotype measurements is graph based indicating the use of the geometrical and topological shape properties such as skeleton and centerline; in such a way that faithful and intuitive features can be derived.

For each study that we will introduce in this thesis, we intend to find the advanced and representative phenotype measurements from the biological image dataset so as to facilitate the solution of a specific biological question. There is no general standard for the selection of phenotype measurements and it is unrealistic to use all the extracted features for pattern classification. Normally, related phenotype measurements are based on the biological descriptions and further use feature selection methods to distill the feature dimensions for classification training. In order to extract related features, the knowledge combined from biologists and computer scientists is required. The image features as described by the biologists, are based on biological principles. These features are observable. The

computer scientists need to translate the feature description into features that can be derived through computation. We call it hidden features. Hidden features are those that can not be derived other than through computation from the digitized image. These features intend to use some advanced and invariant features to represent the biologists' description. In addition, the computer scientists can detect the variance of some other advanced measurements which can be further introduced in the phenotype measurements process.

1.4 Data analysis

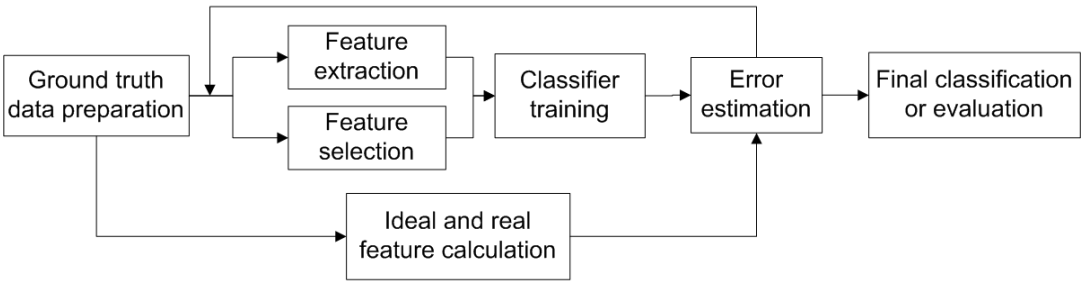

Figure 1.3. The model of phenotype data analysis.

Data analysis pipeline includes ground truth data preparation, feature reduction and classification as shown in Figure 1.3. For a description of phenotype we need ground truth data. For data analysis a good idea of ground truth is important; so for phenotype analysis we need good ground truth examples. But what is ground truth? Therefore we first give a definition.

Definition 1.4.1. *In machine learning, the term "ground truth" refers to the accuracy of the classification of the training for supervised learning techniques. This accuracy is used in statistical models to prove or disprove research hypotheses.*

In image processing, ground truth data could be derived from manual delineation by experts, synthetic images or analytical models based on mathematical expressions. These ground truth data set is used in both phenotype classification

and performance evaluation. Phenotype classification is discussed in 1.4.3.
In our workflow, after having ground truth data for reference, the next step is feature reduction including feature selection and feature extraction in the feature space. This step intends to find prominent features from the feature pool. Feature selection reduces the dimensionality of the feature set by selecting the subset of features from the original set. Feature extraction maps the original feature set into a new set with a reduced dimensionality. Next, classifier training tries different kinds of classifiers and uses an error estimation step to select the classifier with the lowest error.
We strive at using the best performing combination of feature reduction and classifier method for phenotype classification. For the process of performance evaluation, we start with preparing the ground truth data. Then, we measure the ideal features from the ground truth data and real features from the output of the methods. Next, we calculate the difference between ideal and real features by an error estimation.

1.4.1 Ground truth data

The verb "ground truthing" refers to the process of gathering the proper objective data for the test. Based on this ground truth, researchers train a suitable classification method to deliver probabilistic predictions for new observations [Kirchner et al., 2010]. For example, in next generation sequencing technology (NGS), ground truth data is used to train a standard supervised machine learning algorithm for the purpose of identifying somatic mutations from NGS data [Ding et al., 2011]. In classification of plant organs from laser scanned point clouds, the commercial software Geomagic Studio 12 is used to manually assign the ground truth data [Paulus et al., 2013].
Apart from classification training, ground truth is also used for performance evaluation of algorithms and methods. It checks whether the algorithm produces the right output or not. It is frequently used in the evaluation of image segmentation algorithms. The ground truth could be artificial or synthetic images which provide an unbiased ground-truth. Regarding a performance test with the microscopy images, the ground truth images are obtained by manual segmentation

performed by biologists through tracing on a digitizer surface. In order to reduce the observation bias during the manual segmentation, the experts need to repeat several times to obtain an idea of inter user variance [Yan and Verbeek, 2012b]. Similar ground truth production construction is used for algorithm evaluation including retinal vessel segmentation methods [Kaba et al., 2014], simultaneous recognition and segmentation (SRS) of cells [Qu et al., 2011] and microarray segmentation algorithms [Lehmussola et al., 2006]. The ground truth is also used in other methods evaluation discussed as follows: the study in [Lee et al., 2012] computes the distance between the reconstruction and the ground truth to analyze the accuracy of the 3D Neuronal Structure Reconstruction method; in next-generation sequencing, read mapping and genome-wide domain annotations are combined as the ground truth for evaluating the read classification sensitivity and specificity [Zhang et al., 2013]. For a better use of ground truth data, there are even further discussions on introducing a way to design ground-truthed data to compare and evaluate the performance of the real-world detectors [Vedaldi et al., 2010] and creating a ground truth database to evaluate algorithms in the field of mobile robots [Takeuchi et al., 2003].

1.4.2 Feature selection/extraction

In the pool of quantified features, some redundant or irrelevant features can occur within the feature set. Therefore, feature selection process is applied to select a subset of relevant features for further classification. Some popular feature selection methods include the *branch and bound procedure*, *sequential backward* and *forward selections*, *best individual feature selection.*

The branch and bound procedure is a top-down procedure without exhaustive search. It constructs a tree by deleting features successively based on the monotonicity property [Webb and Copsey, 2011].

Sequential forward selection is a bottom-up search procedure that starts with a null set and adds new features to the feature set one at a time until the final feature set is reached. An important disadvantage of the method is the lack of a mechanism of deleting features from the feature set once they have been added.

Sequential backward selection is the other way around. It is a top-down procedure starting with a complete feature set and deleting features one at a time until a predefined dimensionality of the set is reached. The disadvantage of the backward selection method is that it is computationally more demanding compared to forward selection during the criterion function evaluation.
The best individual feature selection [Webb and Copsey, 2011] is the simplest selection method, it might also be the one giving poorest performance; such occurs especially when the features are highly correlated [Webb and Copsey, 2011].
In addition, feature extraction is used to reduce the dimension of the feature set by combining the original features into reduced new features with functions.
Feature extraction is divided into supervised and unsupervised methods. *Principal component analysis* (PCA) is a typical unsupervised feature extraction method. This method aims at deriving new variables (in decreasing order of importance) that are linear combinations of the original variables and that are uncorrelated. *Principal component analysis* is a variable-directed technique and therefore is described as an unsupervised feature extraction technique [Webb and Copsey, 2011]. *Linear discriminant analysis* (LDA) is a supervised feature extraction method. It searches the directions for the maximum discrimination of classes in addition to the dimensional reduction. The criterion proposed by LDA is the ratio of between-class to within-class variances. It is generally believed that when it comes to solving pattern classification problems, LDA algorithms outperform PCA-based ones, since the former optimizes the low dimensional representation of the objects and focus on the most discriminant features, while the latter achieves simply object reconstruction [Youness and Hamid, 2013].
From analysis the microscopy images we can derive large amount of features. However, these features are not all prominently describing the phenotypical differences. Therefore we require a feature reduction method to control the redundancy and consistency that exist in our original feature set. If these two feature reduction process are carefully selected, the prominent information from original feature set could be extracted to perform a more efficient classification using this reduced feature set rather than using the complete original set of features.

1.4.3 Classification

Classification is a procedure in pattern recognition to identify objects in specific categories based on a training set of data containing labeled objects of known category. As for a supervised learning, the training set, in our case also regarded as ground truth data, is crucial for a correct classification. It needs to include a sufficiently large dataset with a variety of situations. Algorithms that implement classification schemes are called classifiers. Classifiers are divided into parametric and non-parametric categories. *The linear classifier* and *the quadratic classifier* belong to parametric category and *the k-nearest neighbor classifier* is in non-parametric category. Linear and quadratic discriminant functions are based on a normal distribution. The linear discriminant rule is quite robust and divides dataset from the normal distributions under the assumption of an equal covariance matrix. However, it is often better to use the quadratic rule if the sample distributions are not separated by the mean-difference but separated by the covariance-difference [Fukunaga, 1990; Webb and Copsey, 2011]. *The k-nearest neighbor classifier* assigns a point x to a particular class based on a majority vote among the classes of the k nearest training points to x. It is a simple and flexible classifier with a good classification performance. However, as the number of objects in the training set increases, it may lead to an excessive computational overhead [Fukunaga, 1990; Webb and Copsey, 2011].
After classification, the phenotypes in segmented objects from images are sorted into different categories. Subsequently we can analyze the changes of a specific category with different biological treatments or across a time line. These changes or trends are meaningful to proof a hypothesis in bio-medical research.

1.4.4 3D model representation

3D models can be represented in three ways: voxel, contour and surface [Verbeek et al., 1995]. voxel models use volume to represent the objects. These models are more realistic but more difficult to construct. The surface models use a surface element to represent the objects such as triangulated surface. These models are easier to deal with since the scale of the computing dataset is much smaller

than voxel models. Thus, surface models are often used to represent 3D models nowadays. Surface representation can contribute to the phenotype measurement considerably well. Many shape based features, such as surface area, volume, curvature, can be calculated well from a surface description. This requires a good surface description.

1.4.5 3D surface reconstruction

A large amount of research has been performed on surface reconstruction from a stack of 2D slices i.e. plan parallel sampled data. One direction is called contour based reconstruction methods. The existing approaches mostly fall into two categories: contour stitching and volumetric methods. Contour stitching directly connect the adjacent contours, while the volumetric methods need to interpolate intermediate gray-values firstly and extract the isosurface from the volumetric field.
The other, evenly popular direction is referred to as point cloud based reconstruction methods. In the literature, the proposed approaches are generally classified into two categories: explicit representation and implicit approximation. The major explicit representations include parametric surfaces and triangulated surfaces. Parametric surfaces attempt to represent all shapes with a set of elementary shapes such as super-quadratics, generalized cylinders, parametric patches, etc. In the explicit representation, all or most of the points are directly interpolated based on structures from computational geometry, such as *Delaunay triangulations* [Boissonnat, 1984], *alpha shapes* [Amenta et al., 2000], or *Voronoi diagrams* [Amenta et al., 1998]. The implicit approximation is based on a scheme which integrates characteristic of each point on the surface into a feature function, a.k.a. the implicit function such as *Fourier-based reconstruction scheme* [Kazhdan, 2005] and *Poisson reconstruction method*. The selection of a surface reconstruction method is important to precisely preserve the surface characteristics and show robustness in the presence of noise. This is addressed in this thesis to be able to come to good features derived from 3D images.

1.5 Structure of the thesis

The image data that are the basis of the phenotypical descriptions are the level of 2D dynamic images (x,y,t) and 3D images (x,y,z). Chapter 2 and Chapter 3 exemplify the image and data analysis of dynamic 2D image at cellular level as derived from high-throughput screening experiment. Chapter 4, Chapter 5 and Chapter 6 describe the 3D image representation and analysis at tissue/organ/organismal level.

The research in Chapter 2 illustrates the design and implementation of a system for automated high-throughput image and data analysis. The phenotypes are characterized according to a model that describes the process of endocytosis, i.e. the ability of cells to absorb molecules, in three characteristic stages. These stages are referred to as episodes and through image processing we try to establish these episodes and the vesicles involved in the ensocytosis are different for each episode. In the late process these vesicles are forming a cluster near the region of the nucleus. According to the model and the observations that it was conceived from, such cluster is larger, brighter and close to the nucleus. From the perspective of image processing, this requires to compuet the area, integrated intensisty of the vesicle and many more possible features derived from objects, i.e. vesicles, that are indentified in the image. From the computer scientists' point of view, it means to calculate the area, the intensity of the labeled object etc. Apart from standard phenotype measurements, we make use of the localized feature of phenotype such as closest object distance which is the distance between the object and the nuclei region so as to describe the correlation between two information channels. We obtained the ground truth data for classifier training by having the three characteristic episode groups manually delineated by biologists; this gives as binary mask. We make use of these binary masks for further phenotype measurements and derive the training set for the supervised classification procedure. Next, we use the model of phenotype data analysis for the classification of the three episodes (plasma-membrane, vesicle and cluster). We evaluate the performance of the combination of different feature selection and classification methods and select one with the lowest error estimation. The experimental results show that our analysis setup for high-throughput screening provides scalability and

robustness in the temporal analysis of an EGFR endocytosis model.
In Chapter 3, the results of Chapter 2 are further evaluated. Chapter 3 illustrates an integrated method employing a hierarchical classification strategy and wavelet-based texture measurements to further improve the recognition of phenotypic episodes of EGFR during endocytosis. During the previous single classifier training in Chapter 2, we find that the similarity between cluster and vesicle is higher than with plasma-membrane. As a result, we construct an advanced hierarchical classification strategy. This hierarchical classification strategy can construct the classes in a tree structure and train the classifier for each parent node to distinguish two child nodes that belong to the same parent node. We also introduced wavelet texture features to distinguish endosomes phenotype variation across timeline instead of the average intensity for each object, because a texture feature in a local patch is more discriminative than pixel intensities for candidate identification [Song et al., 2013]. The result of the hierarchical classifiers with wavelet-based texture measurements shows a noticeable improvement compared to the single classification strategy.
In Chapter 4, the work uses an analytical approach to evaluate four classical surface reconstruction methods. We make use of the ground truth concept for the evaluation of 3D surface reconstruction methods. In order to make an objective assessment of the surface quality, we utilize three synthetic objects for the error estimation. From mathematics, an analytical description of each synthetic objects is available. The three synthetic objects are the sphere, the ellipsoid and the ovoid. The parametrical mathematical representation of these synthetic surfaces helps us to compute the ideal surface features and provides a ground truth for the error estimation of surface. For the real surface feature calculation, we firstly deviate the ideal model by adding different levels of noise. Next, we use the noisy point cloud as our input for the reconstruction algorithms. Finally, we calculate the real surface feature from the reconstructed surface model. The aim of this evaluation study is to select the outstanding reconstruction method to improve reliability in surface reconstruction of biological models.
In order to apply the findings of Chapter 4, optimized 3D geometrical descriptions are requried. In Chapter 5, we therefore provide a pipeline to optimize the stack of biological images in 3D space and analyze the phenotypical difference by

extracting related shape features from the 3D biological model.
In Chapter 6 we applied the results from Chapter 4 and Chapter 5 for phenotype measurements; we extract centerline of the rodent newborn lactiferous duct to unfold the branch structure embedded in the duct rather than use standard surface descriptions. Next, we use the quantified features from the centerline to detect the morphological changes on the duct surface model. Furthermore, we extended the usage of ground truth for the simulation of mammary gland in Chapter 6. With the inspiration from the tree-like structure of mammary gland, we use a mathematical model: Lindenmayer systems (L-systems) which is a mathematical theory developed for the description of growth patterns in plants. We create a specified model for lactiferous duct of the new-born mouse from the L-system as our ground truth. With this mathematical model, we can simulate the phenotypical variation between various treatments by changing the parameters representing prominent features derived from phenotype measurements.
We conclude the discussion of this thesis in Chapter 7 with the insights that are obtained from the research describes in the Chapters 2-6.