Shape analysis for phenotype characterisation from high-throughput imaging

Guo, Y.; Guo Y.

Chapter 1

Introduction

Computational approaches are important to characterise phenotypes in modern life-science research, in developmental biology, (patho)physiology, toxicology, pharmacology, etc.. According to the definition, the phenotype is recognised as an observable trait in the whole appearance of an organism from molecular to organism scale [1]. The study of gene expression can be used to visualise the phenotype. The process of gene expression operates on the level of synthesis and structuring of proteins which subsequently contribute to the organism's appearance in any form. In order to obtain an understanding of a phenotype, imaging is used to explore the phenotype characteristics. Thus, a readout of the phenotype by means of images is accomplished. Consequently, the phenotype characteristics need to be extracted from images and this can be realised through image analysis. Therefore, computational approaches for image analysis have to be developed. Recently, some approaches have been developed for this purpose (see [2, 3, 4]). In this thesis, we aim at developing dedicated computational approaches which may achieve an *efficient* and *robust* performance in generalising the whole description for phenotype characterisation.

1.1 Importance of shape

In this thesis, a *shape* in an image is formally defined as the quality of an object that depends on the relative position of all points on its surface (adapted from [5]). With respect to the whole phenotype of an organism, shape appears to be directly relevant to the physical development of that organism (or the organ) under investigation. For example, some shapes can directly reflect apparent variations of an object. The shape variations are often caused by the exposure of the samples in a compound for screening or by the timing control for wild type individuals.

However, in some circumstances, the shape variation is subtle, so that the variation is difficult to observe and analyse. In light microscopy, for example, it is difficult to compare the whole-mount of a sample with another from the same model organism in the same developmental stage. Yet, subtle shape variations can play an important role in toxicology, since the size of the organism reflects its response to certain drugs. Indeed, the shape variations in some experimental settings can be observed, but empirical interventions may introduce and propagate subjective errors. For instance, interactive annotation of key points in images is

often used to analyse delicate structures such as the skeleton of an organism [6], but different annotators may have their own assessments for the objectivity.

As a result, accurate shape analysis of an organism will lead to reliable characterisation of phenotypes. To this end, we first need a good shape representation to describe the shape of an object precisely. In order to validate the performance of such shape representation, the scalar primitives for a shape, e.g., (a) perimeter and area in 2D shape representation, and (b) volume and surface area in 3D shape representation, can be used. With a validated shape representation, the geometrical primitives for the shape can be further involved for more sophisticated shape analysis. In practice, the simple scalar primitives for the shape are very important as they provide us with intuitive, stable, and accurate 2D/3D measurements, for a delicate exploration of subtle shape variations in the phenotype characterisation. Therefore, we are motivated, in this thesis, to develop new computational approaches based on images (1) to promote precise shape description and (2) to make reliable and accurate 2D/3D measurements possible.

1.2 High-throughput imaging

We should be aware that an unbiased shape analysis is available in a population of the model. In practice, this is conditioned by the choice of the sampling size for the population. As a result, a sufficiently large sampling size may reflect the general properties of the samples, which will result in an accurate statistical assessment. In order to obtain adequate sampling size, high-throughput (HT) screening was initiated for the applications of cytomics and toxicology, and has been applied on the application on organism scale such as zebrafish [7]. The HT screening facilitated the fast development of high-throughput imaging (HTI). A feasible HTI architecture can easily acquire a sufficiently large volume of data represented by images of the subject under study. This HTI architecture has the following advantages: (1) we can use bright-field microscopy producing the images, representing an overall shape for a specimen; (2) fluorescence microscopy can be used to produce the images presenting the fluorescently marked components such as detailed inner structures of a specimen; (3) synchronization of bright and fluorescence microscopy results in the so-called multi-modal microscopy producing fused multi-modal images which can be used to represent and evaluate the comparison between different modes; (4) we can potentially obtain shape analysis

at a very high-resolution with the help of a better quality objective lens, in which structures such as tissues and cell type can be reflected from micro-scale texture in images.

1.3 Model organism

In modern life-science, a feasible and convenient model organism also plays an important role since many human diseases can be cultivated in a model organism. A good understanding of the model organism can obtain insights into the disease or treatment, and the obtained knowledge can be transferred to research on humans. In practice, we have many options for the model system. Invertebrate models such as fruit fly [8] and c. elegans [9] are intensively used in molecular genetics. As a comparison, vertebrate models like mouse are suitable and commonly used in the research of human diseases [10]. However, the growth of mouse is slow and it is difficult to get access to a large sampling size. Alternatively, in the last decade, zebrafish has been increasingly used for human disease studies as they present many remarkable characteristics [11], among which the most significant one is its 70% genome equivalence to human [12]. The development of zebrafish is pretty fast as its organs develop within 36 hours post fertilisation (hdf) [13]. The zebrafish are fertile, and one adult couple can easily produce 300 eggs per week [14]. In early larval stages, zebrafish are quite small ($< 1mm$) and optically transparent thus the whole body of a zebrafish as well as partial inner organs are observable using microscopy [15]. In particular, with the availability of many transgenic lines, the zebrafish is genetically modified with fluorescent markers like green fluorescent protein (GFP) [14, 16]. Fluorescence microscopy can be employed to visualise the fluorescently marked structures within a specimen such as organs and infectious diseases. Taking all the properties into consideration, zebrafish is very suitable to be used in high-throughput applications. Therefore, in this thesis, we will consistently use the zebrafish larvae as our model system to illustrate and validate our approaches.

1.4 Problem statement and research questions

In section 1.1, we have discussed the importance of shape analysis for phenotype characterisation. The HTI architecture will serve as the basis for the production

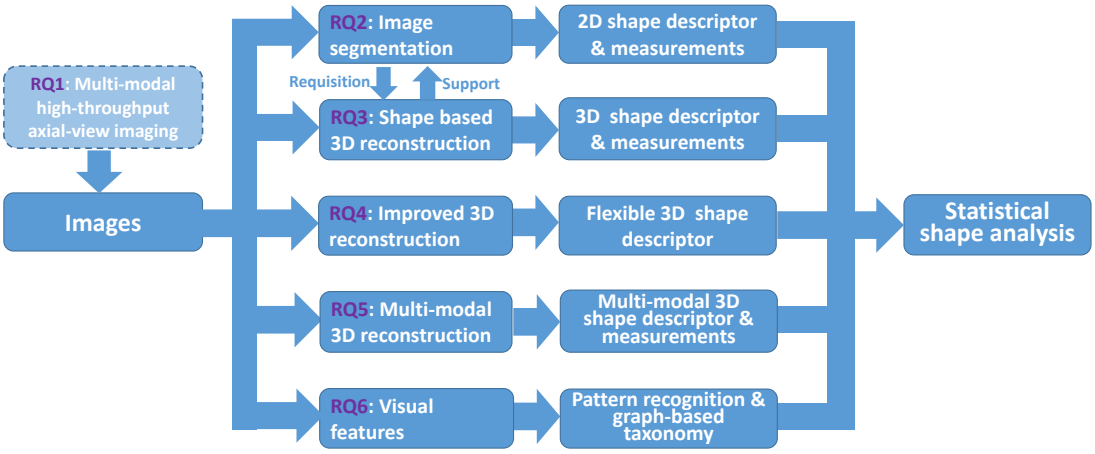

Figure 1.1: A schematic representation of a unified system for shape analysis in support of phenotype characterisation using an MM-HTAI architecture. The six RQs formulated in the text are indicated in the boxes.

of adequate data. We expect to obtain an *efficient* and *robust* shape analysis to characterise the phenotype from HTI. This idea has inspired the formulation of our problem statement (PS).

> **PS**: *To what extent can we develop a stable HTI architecture and produce a robust and accurate shape analysis for the phenotype characterisation from the HTI architecture?*

To address the PS, we are motivated to design a unified system integrating multiple functional modules. They should correspond to new computational approaches for shape analysis in support of the phenotype characterisation from HTI. We start remarking that the system needs to deal with the problem of a precise shape representation. Then a validation of the shape representation is required. This will imply a delicate shape analysis using geometrical primitives. The validation can be realised through scalar primitives for the shape in the form of 2D/3D measurements. To reach our goal, we specify and investigate six research questions (RQs). In Fig. 1.1, we provide a schematic representation corresponding to the six RQs.

As stated above, we need to develop a feasible HTI architecture for image acquisition of the zebrafish larvae. In practice, the zebrafish larvae are always positioned along their longitudinal axis when they are in an imaging modality. In this manner they are easy to manipulate. Moreover, most of its features are

then observable and can be visualised. This gives rise to the high-throughput axial-view imaging (HTAI) architecture. Since the HTI can adopt multi-modal microscopy, we can obtain a multi-modal high-throughput axial-view imaging (MM-HTAI) architecture. Our first RQ is thus formulated as follows.

> **RQ 1**: *To what extent is it possible to develop an MM-HTAI architecture for the zebrafish larvae?*

From the axial-view images of the zebrafish acquired by the MM-HTAI architecture, one should first obtain a well-defined 2D shape descriptor and then derive accurate 2D measurements such as the perimeter and the area of an object. With respect to zebrafish, we can see the following properties. On the one hand, an accurate description (both for a whole-shape and its organs on microscopic level) will ensure a reliable quantitative assessment in the applications, such as drug targeting. On the other hand, we have to realise that the part of zebrafish tail is particularly relevant to some diseases since the hemopoietic stem cells in the zebrafish are found predominantly in the tail. An detailed description for every section of the zebrafish such as its tail will ensure accurate localisation, tracking and evaluation of some infectious diseases, such as the spread and development of cancer cells [17, 18]. However, in applications using the zebrafish as model system, the partial transparency across the whole specimen and the weakly defined boundaries which are mainly distributed around the tail in early larval stages, are commonly existing and observable. Hence, we formulate our second RQ as follows.

> **RQ 2**: *To what extent is it possible to obtain an accurate 2D shape description for the zebrafish from the MM-HTAI architecture?*

With respect to shape analysis, a 3D description is more reliable due to the 3D nature of an object's shape than a 2D descriptor. The confocal laser scanning microscope (CLSM) can obtain 3D imaging for a fluorescently labelled structure in an organism. 3D reconstruction has been produced from the images acquired by CLSM using the TDR-3Dbase software [19, 20]. However, there are two obstacles. (1) It is difficult to depict an overall shape of the whole organism for the CLSM, and (2) the CLSM has a low efficiency of image acquisition. Recently, an attempt for 3D imaging has been reported as optical projection tomography (OPT) [21, 22], whilst rather dense scanning is required and extra processing like 3D image segmentation should be employed for further image analysis. In OPT, the sample preparation is also rather time-consuming. To the best of our

knowledge, there are few systematic assessments of 3D measurements, e.g., the volume and surface area, in real metrics for the zebrafish in phenotype research. Actually, these 3D measurements are essential in many applications. For example, statistical representations of the 3D measurements for the whole-mount of zebrafish in various developmental stages will give insights into the accuracy of shape analysis which enables HT compound screening. In fact, the axial-view images acquired by the MM-HTAI provide sufficient information for a good 3D description which is a prerequisite for reliable estimation of 3D measurements. This observation results in our third RQ.

> **RQ 3**: *To what extent is it possible to obtain a precise 3D shape description and derive accurate 3D measurements that are statistically relevant for the zebrafish from the MM-HTAI architecture*?

In our study, translucency and transparency often occur in light microscopy. Admittedly, in some cases, the boundaries of an organism are weakly defined. Yet, the qualities will still present a good 3D description for the whole-mount of the zebrafish, even without accurate 2D shape descriptions. Hence we formulate our fourth RQ as follows.

> **RQ 4**: *How can we efficiently deal with the translucency and transparency of specimen in light microscopy and still obtain a good 3D shape description from the MM-HTAI architecture*?

In life-science research such as toxicology, quantitative endpoints like organ size or growth retardation play significant roles. This requires an accurate 3D shape description and rather precise measurements on organ scale, such as the evaluation of organ susceptibility of toxicology in the zebrafish larvae [23]. The MM-HTAI is capable of producing multi-modal images including (1) bright-field images presenting the overall shape of the zebrafish and (2) fluorescence images presenting the detailed inner structure like zebrafish liver. So, we formulate our fifth RQ as follows.

> **RQ 5**: *How can we obtain a multi-modal 3D description and the corresponding measurements for the zebrafish from the MM-HTAI architecture*?

If we go to a higher resolution scale, i.e., on the cellular or tissue level, textures such as detailed fibrous structures in the specimens can contribute to even better shape analysis. In this case we will represent the shapes as well-defined features

according to geometrical and textural information extracted from an image [24] or even representative features such as the convolutional neural networks (CNN) [25]. Hence we formulate RQ 6 as follows.

> **RQ 6**: *To what extent is it possible that the classification models (or regression models) are able to validate the performance of the image features to characterise the phenotypes in support of shape analysis?*

1.5 Research methodology

The research methodology in this thesis consists of (1) literature study and analysis, (2) development and implementation of new computational approaches, (3) performance validation and evaluation for the approaches. The literature study is realised by reading and investigation; the analysis by a comparison with the development of state-of-the-art. The new computational approaches are inspired by ideas from other well-developed research fields, such as computer vision and machine learning. The performance validation is achieved by applying the methodologies on datasets and comparing the results with state-of-the-art. We elaborate the methodologies in the analysis of the six RQs. We do so as follows.

In RQ 1, we employ the Vertebrate Automated Screening Technology (VAST BioImager) [26] and light microscopy to develop the MM-HTAI architecture. The VAST BioImager is used to manipulate the input. The positioning module of the VAST BioImager consists of a delicate capillary which is held by a pair of stepper motors. The stepper motors can manipulate the positioning module to revolve for 360 degrees. A pumping action system is loading a zebrafish larva into the positioning module; a mounted camera which we refer to as VAST camera is used to detect and localise the object and then manage the system to position the specimen in the view of an observer from an arbitrary axial-view. The observer can be either the VAST camera or a microscope camera. In this manner, the MM-HTAI can be accomplished and a sequence of axial-view images including bright-field and fluorescence of the specimen can be acquired.

In RQ 2, we take inspiration from the field of computer vision [27]. We consider an advanced method as image segmentation or edge detection for the acquisition of the 2D shape description of an organism represented in the images obtained from

the MM-HTAI architecture. The methodology focusses on an *efficient* and *robust* segmentation method that incorporates conventional segmentation methods, such as variational based methods [28] and unsupervised learning based methods [29].

In RQ 3, a 3D shape representation for a scene or an object can be obtained from a range of multi-view images using multi-view stereo [30]. Our axial-view image sequence is a particular case for the multi-view, which is commonly referred to as turn-table data [31]. We can resort to a shape-based 3D reconstruction method to solve the problem of 3D shape description from the MM-HTAI architecture. Subsequently, 3D measurements can be directly derived from the results of the shape-based 3D reconstruction method. Importantly, we show that we can implement this 3D shape acquisition in an efficient manner.

In RQ 4, we need to solve the problem of 3D shape description for an organism that is partial translucent/transparent. It might result in weakly-defined boundaries. Therefore, we require an improved 3D reconstruction method which does not require the most accurate initial 2D shape descriptions. The method can be considered as an extension of the shape-based 3D reconstruction. We incorporate texture information from the axial-view images to infer a more flexible volumetric representation. We use probabilistic models and further validate accurate 3D measurements.

In RQ 5, the multi-modal images acquired by the MM-HTAI are used to obtain a multi-modal 3D shape description by the fusion of the 3D shapes both on organism and organ scale for the zebrafish larvae. This is supported by the shape-based and improved 3D reconstruction methods. The 3D shape description on organism scale presents a shape reference for the normalisation of the 3D description on organ scale. It requires an alignment of the multiple 3D shape descriptions resulting in a natural visualisation and a high quality 3D image fusion for the organism and its organs.

In RQ 6, we first extract features in the images from annotated datasets, and then apply classification (or regression) models to validate the performance of the features [32, 33]. Currently our research is hampered by the availability of sufficiently large annotated datasets for the zebrafish. Therefore, we choose to study the behaviour of the features and classification models in a collection of datasets concerning phenotypes/gene expression, e.g., humans, animals, and plants. We do so in order to obtain a balanced understanding of the methodologies. We

believe that it is possible to transfer the knowledge in these datasets towards the application area of the zebrafish for the phenotype characterisation. The reason is that we are able to provide sufficient evidence for the generality and accuracy of the system in similar domains. Therefore, we will investigate various texture features and develop a graph-based local-global strategy for taxonomy prediction in several datasets. One of the particular cases would be kinship recognition in humans by facial analysis; other cases will refer to a diverse collection of datasets including butterflies, orchids, and wood species.

1.6 Thesis structure

The structure of this thesis is as follows. We address at least one research question in a chapter by presenting a new approach. Careful discussions and thorough inferences will be given. Chapter 1 provides the PS, the six RQs and the research methodologies.

Chapter 2 aims to answer RQ 2 (For RQ 1, see Chapter 3). To this end, we present an efficient and robust hybrid method for zebrafish image segmentation for bright-field microscopy of the MM-HTAI architecture. We integrate the merits of conventional segmentation methods, i.e., the variational based segmentation method and the unsupervised learning based segmentation method. Then we propose a sequential refinement on the hybrid segmentation, resulting in a better 2D shape description. The results present an overview for the zebrafish larvae.

Chapter 3 addresses RQ 1 and RQ 3. We first specify the MM-HTAI architecture based on the VAST BioImager and the light microscopy. From the acquired images, we address the problem of 3D shape acquisition through a shape-based 3D reconstruction method. The method uses the 2D shapes obtained in Chapter 2. An accurate camera motion estimation is the basis for this method. We solve the problem by presenting a novel method as the voxel residual volume (VRV) maximisation algorithm. We validate our method through particles of known size. In this chapter we also report a 3D shape reference using statistical distributions from 3D measurements of the zebrafish for three commonly used larval stages, i.e., 3, 4, 5 days post fertilisation (dpf). According to the best of our knowledge, this is the first validated and justified report on the topic in this field; the results have already been successfully used in pharmacokinetics and toxicology [34, 35].

Chapter 4 answers the RQ 4. We improve the 3D reconstruction by the incorporation of texture information from the original axial-view images, since in some cases, a 2D shape is difficult to obtain due to partial transparency. So, we take the texture distribution sampled from the images into consideration to estimate a more flexible 3D volumetric representation with a confidence score as entry. We demonstrate the successful application of the method in the MM-HTAI architecture.

Chapter 5 presents a solution for RQ 5. We propose the methodology of multimodal 3D reconstruction for the zebrafish larvae on both organism and organ scale. We use the feature of our MM-HTAI architecture to produce images for both the whole organism in bright-field and detailed organ structures in fluorescence. We take the zebrafish larvae and its liver as examples to explain our method. The shape-based 3D reconstruction method is applied to obtain the multiple 3D shape description; an alignment and a fusion of the multiple 3D shapes are integrated to obtain a good visualisation of the results.

Chapter 6 concerns RQ 6. We apply a hand-crafted feature, the Local Binary Patterns (LBP), on human facial appearance. Then we propose a graphical model to predict the taxonomy (kinship) for genetic related family members. We also apply a CNN architecture to acquire representative features. Subsequently, we design a multi-output layer to enable taxonomy prediction for a set of datasets of biological specimens, i.e., butterflies, orchids and wood species. As a result, we have successfully applied our methods in the applications mentioned earlier. The experiment provides suitable knowledge and understanding for the behaviour of the method when transferring the knowledgeable items from current applications to the phenotype characterisation using the zebrafish.

Chapter 7 summaries the answers to the six RQs and answers the PS. We list a few limitations of the whole work and propose possible solutions. Finally, we offer six recommendations for further research.

Table 1.1: The structure of the thesis

Chapter	RQ 1	RQ 2	RQ 3	RQ 4	RQ 5	RQ 6	PS
1	✓	✓	✓	✓	✓	✓	✓
2		✓					
3	✓		✓				
4				✓			
5					✓		
6						✓	
7	✓	✓	✓	✓	✓	✓	✓