IDENTIFICATION OF PACEMAKING REGION IN ZEBRAFISH HEART FROM OPTICAL MAPPING DATA

by

Weiguang Ding

B.Eng., Beihang University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in the
School of Engineering Science
Faculty of Applied Sciences

© Weiguang Ding 2013
SIMON FRASER UNIVERSITY
Spring 2013

APPROVAL

Name: Weiguang Ding

Degree: Master of Applied Science

Title of Thesis: Identification of Pacemaking Region in Zebrafish Heart from Optical Mapping Data

Examining Committee: Dr. John Jones
Associate Professor, P.Eng.
School of Engineering Science
Chair

_____

Dr. Mirza Faisal Beg, Senior Supervisor
Associate Professor, P.Eng.
School of Engineering Science

_____

Dr. Marinko V. Sarunic, Supervisor
Associate Professor, P.Eng.
School of Engineering Science

_____

Dr. Glen F. Tibbits, Supervisor
Professor, Canada Research Chair
Department of Biomedical Physiology and Kinesiology

_____

Dr. Jie Liang, SFU Examiner
Associate Professor, P.Eng.
School of Engineering Science

Date Approved: _____

Partial Copyright Licence

SFU

Abstract

Junctional Ectopic Tachycardia (JET) is a cardiac arrhythmia which occurs immediately after open heart surgery in young children. In specific populations, JET has extremely high incidence, of up to 50%. There has not been a specific mechanism elucidated by clinical data or basic science. As a widely used vertebrate cardiovascular biological model, zebrafish heart is being studied to reveal the leading reasons of JET. Optical mapping (OM) techniques provide an effective approach to observe cardiac functionality by recording zebrafish heart action potential propagation. However, processing of vast amount of OM data also poses challenges on fast and accurate processing, measurement and interpretation.

This thesis presents novel automated pipelines for processing zebrafish heart OM data and identifying pacemaking regions from it through signal analysis. We first introduce a preprocessing pipeline for enhancing very low signal-to-noise ratio original OM data, which involves spatial-temporal smoothing, cycle averaging, drifting correction and scaling. After that, we present a computer assisted OM signal manually labeling pipeline, which reduces the manual workload significantly by clustering spatially adjacent similar signals. Furthermore, we make physiologically relevant measurements on OM data and do statistical analysis comparing different labeled regions. Finally, we propose a two-step signal clustering based method to divide atrium into different functional regions followed by pacemaking region identification. We present the formulation of these methods and discuss their validity and performance in various aspects. The work presented in this thesis could lead to significantly faster and larger scaled experimentation in optical mapping related physiology research.

Acknowledgments

First, I would like to thank my senior supervisor Dr. Mirza Faisal Beg, who is an excellent researcher, supervisor and teacher. He supports my research interests, offers me various research opportunities and gives me good suggestions. Working with him is enjoyable.

I would also like to thank my supervisors Dr. Marinko Sarunic and Dr. Glen Tibbits for their crucial directions and helpful suggestions during my work and study in the past two years. I thank Dr. Jie Liang for spending time to read my thesis and be my defence examiner. I thank Dr. John Jones for chairing my defence in spite of his busy schedule.

Next, I would like to thank my collaborators Dr. Eric Lin and Amanda Ribeiro for their work on data acquisition and help on understanding physiology knowledge. Especially, Eric is a good mentor, who is supportive and flexible to me, at the same time keeps the right direction. It is pleasant to work with them.

I wish to thank all my colleagues and friends in Medical Image Analysis Lab (MIAL) and Biomedical Optics Research Group (BORG): Evgeniy Lebed, Pradeep Reddy Ramana, Sieun Lee, Amanmeet Garg, Andy Bo Wu, Wenqi Sun, Jingyun Chen, Michelle Cua, Yifan Jian, Jing Xu, Mei Young, Lukas Merhi, Ali Issaei and Lukasz Szczygiel for their generous help, sharing and the wonderful time we had. Particularly, I thank Evgeniy for thesis proofreading and suggestions. I would also like to thank all the other colleagues of Medical Image Computing and Analysis (MICA) Lab for the interesting discussions and the memorable experience with you guys.

Finally, I would like to thank my family. I thank my parents Shuli Ding and Yanhua Xie, as well as my parents-in-law Yanmian Liu and Jianzhong Wang for their unconditional love, support, encouragement and understanding. I thank Jingya Wang, my wife and best friend, for her love, which is beyond words.

Contents

List of Tables

List of Figures

List of Abbreviations

JET	Junctional Ectopic Tachycardia
OM	optical mapping
SAN	sinoatrial node
SNR	signal-to-noise ratio
Di	Dataset i, where i is the dataset number
ROI	region of interest
FCM	fuzzy c-means
RFCM	robust fuzzy c-means
AN	normal atrial region
AC	corrupted atrial region
NN	normal nodal/pacemaking region
NC	corrupted nodal/pacemaking region
NMC	normal coefficient representing the percentage of identified uncorrupted area
NDC	nodal coefficient representing the percentage of identified nodal region area
OT	other unidentified region
DoD	discretization of derivative
DSC	Dice similarity coefficient
BPDSC	best possible Dice similarity coefficient

Chapter 1

Introduction

Junctional Ectopic Tachycardia (JET) [50] is a cardiac arrhythmia occurring immediately following open heart surgery in young children. JET has extremely high incidence in specific populations, yet its mechanism has not been elucidated by clinical data or by basic science. Abnormality in cardiac pacemaker cells is one of the possible reasons that may cause JET and is currently being studied by physiologists. To study the mechanism of JET, we use zebrafish as animal model and observe its cardiac activity using optical mapping techniques. Although pacemaker cells are able to be identified by histology, they are not visible from optical mapping data or identifiable by eye when they are still alive and functioning. To solve this problem, we need to develop methods that are able to identify pacemaker regions from optical mapping data when the heart is still alive.

1.1 Background

1.1.1 Zebrafish heart

Zebrafish is one of the most successful biological models [40] for studying human physiology and diseases. As a vertebrate cardiovascular model, the zebrafish heart has been used extensively to study cardiac physiology, development and human cardiac disease [3] due its genetic screenability, high permeability to small molecules and excellent developmental observability [40]. For example, scientists have created genetically encoded, optically controlled pacemakers and used them to stimulate heart diseases including tachycardia, bradycardia, atrioventricular blocks, and cardiac arrest [1]. Besides its flexible maneuverability and good
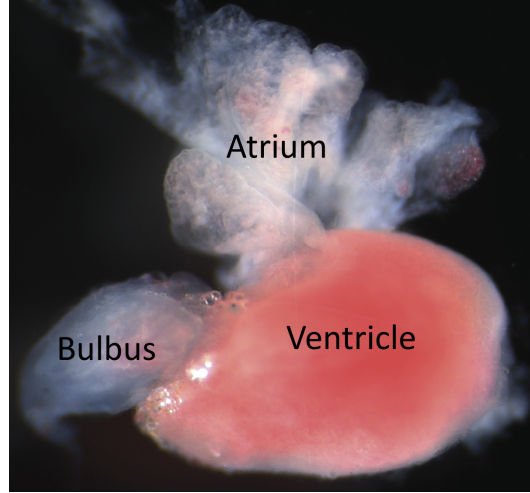
Figure 1.1: Zebrafish heart image from [16]. The real scale of this photo is about 1∼2mm by 1∼2mm.

observability, more importantly, compared to mouse and other small mammals, zebrafish cardiac electrophysiology has been found to be considerably more representative of human adult electrophysiology in various aspects including individual ionic currents, whole-heart electrophysiology, drug responses and cardiac contractility [2, 24, 32, 31, 30, 42, 35].

The zebrafish heart is comprised of cells of different types, including nodal, atrial and ventricular cells. Each of these cell types has distinct mechanical and electrical characteristics that determine the heart's ability to provide adequate blood flow to the rest of the body. Accordingly, each cell type has a distinct action potential [39] waveform, a specific time-dependent pattern of membrane depolarization that closely relates to its physiological function. To understand the cardiac activity of zebrafish heart, we aim to identify the location and measure relevant properties of these different regions, especially pacemaking regions, in other words nodal regions. Figure 1.1 shows an image of excised zebrafish heart [16] with labeled atrium, ventricle and bulbus.

1.1.2   Optical mapping

Traditionally, electrophysiology has been used to measure action potential value in various biological tissues, by directly measuring voltage with microelectrodes in contact with biological tissues. More recently, optical mapping (OM) [12] techniques have been used to study membrane potentials when microelectrodes encounter difficulties. In optical mapping,

tissues are soaked in fluorescence voltage sensitive dye, which emits light with different spectrum under different voltages. Therefore, the emitted fluorescence change is directly related to the transmembrane action potential. Utilizing this fact, the fluorescence intensities in a given bandwidth can be recorded over time to reflect the temporal action potential changes. Generally, optical mapping has three advantages over the traditional electrophysiology [40]:

- optical mapping is less affected by physical constraints

- optical mapping enables simultaneous multiple sites recording, which greatly increases the spatial resolution

- optical mapping does not stimulate tissue, which may distort the action potential

Optical mapping in cardiac research

Optical mapping techniques have also been used extensively in cardiac research and proven to be a successful tool in cardiovascular function and disease research [17, 12, 23, 9, 13]. Generally, having ex-vivo animal hearts immersed in solution with fluorescence voltage sensitive dye, cardiac action potential waveforms are recorded as the intensity change in different pixels. Each pixel contains a time series signal related to the actual action potential of the corresponding location.

As cardiac activity involves both action potential propagation and physical contractile motion, a single pixel in the recorded video does not always correspond to the same location on the heart. This will introduce motion artifacts in the acquired signals. There are three common methods [40] for dealing with motion artifacts in cardiac research: mechanical restriction, pharmacological approaches and ratio metric imaging. For zebrafish heart, pharmacological approaches and ratio metric imaging are commonly used, because mechanical restriction is difficult due to the fragility of zebrafish heart. Recent comprehensive reviews of optical mapping techniques can be found in [17, 12, 23].

Analysis of cardiac optical mapping data

The raw cardiac optical mapping data is usually noisy and not suitable for calculation of different measurements. Existing common preprocessing approaches have been summarized and introduced in [26], including spatial average filter, temporal average filter, temporal

frequency domain filters, ensemble average and drifting removal. After preprocessing, different approaches have been used on cardiac optical mapping data to get various types of measurements and analyses [26], including activation map, conduction velocity map, action potential duration map, repolarization map, upstroke analysis, arrhythmia phase map analysis, dominant frequency map analysis and etc. These are measurements on individual pixels, which are then combined to give local and global information about cardiac activities. In addition, the functional properties acquired from processing optical mapping data are combined with structural data to reveal cardiac structural-functional correlations [18].

For zebrafish heart, optical mapping data has not been extensively measured, probably due to its small size, which causes several problems like small field of view, low resolution and low signal-to-noise ratio. In previous work, voltage dynamics have been visualized and activation maps have been manually drawn based on the visualization [8, 29, 47]. Manual methods are not suitable for large dataset analyses and introduce rater bias and errors. This motivates the need to develop automated methods.

1.1.3 Identification of pacemaking region

Among different types of heart regions, identification of pacemaking (or nodal) cells is of particular interest. Unlike atrial and ventricular cells, nodal cells do not compose an individual chamber and are part of atrium. They act as pacemakers, which lead the depolarization of whole heart, are responsible for the heart's intrinsic rate and respond to extrinsic factors that modulate heart rate [48].

Existing method for identification of pacemaking region in zebrafish heart

Previously, a few efforts have been made on detection of pacemaker in zebrafish heart. In terms of functional analysis, pacemaker was identified as the earliest depolarization regions on activation map in different developmental stages in [8]. In terms of molecular and structural anlayis, pacemaker of adult zebrafish heart was identified as a ring around the venous pole [45] based on information from complex combinations of microscopic examination, gene expression pattern reconstruction, reporter transgenics and electrophysiology.

These methods are complicated, expensive, time consuming and still needs to be validated by other, equally-involved, independent methods. More importantly, as detection methods they only provide approximate locations with qualitative descriptions.

Necessity of developing new detection method based on optical mapping data

For relatively fast quantitative identification of pacemaking region, new identification methods solely based on optical mapping data need to be developed. There are five main reasons for us to work on this:

- Previously proposed pacemaker locations need to be validated.

- Since anatomical structure is hard to identify in optical mapping data due to the tissue transparency, even pacemaker location is known on the zebrafish heart. Thus, it is hard to mark and track it in optical mapping data. Therefore, identification of pacemaker location purely based on optical mapping data is desired.

- Accurate quantitative identification and localization of pacemaker are needed for precise local and global cardiac activity analysis.

- Since pacemaker location might change under different experimental conditions, we need reliable methods to detect pacemaker under different conditions, and even with disappearance, moving and multiple sites of nodal regions.

- Most importantly, detection methods need to be automated for high throughput experiments, which enables the utilization of the biological advantages of zebrafish.

1.2   Research problems

To achieve methods with desired properties described in Section 1.1.3, we have several difficulties to overcome and problems to solve:

- How to enhance the low signal-to-noise (SNR) ratio and low dynamic range data, but still keeping the necessary characteristics of the signal?

- What properties does pacemaking region have? How are they different from other regions?

- How do we perform computer assisted manual identification on optical mapping data?

- How do we automate the detection of pacemaking regions?

1.3   Primary contributions

In order to solve the above mentioned research problems, this thesis presents the following contributions on identification of optical mapped zebrafish heart.

- A novel pipeline is proposed and implemented for preprocessing low SNR raw zebrafish optical mapping data.

- A novel, fast, manual identification pipeline is proposed and implemented.

- Physiologically measurements are done and statistically analyzed.

- Novel automatic detection methods are proposed and validated by manual identification.

1.4   Organization of the thesis

This thesis is organized into the following chapters, focusing on the primary contributions:

- Chapter 2 describes the data acquisition method, OM data characteristics, and preprocessing pipeline.

- Chapter 3 discusses the manual identification process and results.

- Chapter 4 presents the signal measurements and their statistical analysis.

- Chapter 5 presents slope value based pacemaker identification method and automatic signal grouping method. These methods are also evaluated using the manually labeled data.

- Chapter 6 concludes the thesis, summarize the primary contributions and discuss future works.

1.5   Publications arising from the thesis

Several journal and conference publications are expected to be arising from the thesis. Among them, one conference paper has been accepted:

Weiguang Ding, Eric Lin, Amanda Ribeiro, Marinko Sarunic, Glen F. Tibbits, and Mirza Faisal Beg, "On Identification of Sinoatrial Node in Zebrafish Heart Based on Functional Time Series from Optical Mapping", 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), July 2013

This paper describes contents from Chapter 5.

Chapter 2

Data Acquisition and Preprocessing

2.1 Data acquisition

Optical mapping (OM) data acquisition was performed by our colleagues in Department of Biomedical Physiology and Kinesiology, Simon Fraser University. The process is briefly introduced below.

We acquired datasets from 5 individual experiments using different hearts with similar procedure. We numbered them as Dataset 1 to 5 (D 1 to 5) corresponding to their acquisition dates Jan_15 (D1), Jan_24 (D2), Oct_23 (D3), Oct_24 (D4) and Oct_25 (D5). In each experiment, zebrafish heart was isolated and then soaked in the fluorescent voltage sensitive dye RH-237 [14] by immersion in 8 μM solution. 10 μM blebbistatin excitation-contraction uncoupler [20] was used to inhibit motion artifact caused by contraction. Excitation illumination was provided by a 200 mW 532 nm DPSS laser (Laserglow Technologies). Spectral filtering was provided by a 560 nm long pass dichroic and a 710 nm long pass emission filter (Omega Optical). Fluorescent images of the heart were acquired using two GE680 (Allied Vision Techologies) cameras at 205 frame per second (fps) and resolution of 640×480 pixels. Typical optical mapping image is presented in Figure 2.1 with typical atrium and ventricle signals.

For each of the experiments, the temperature started from 28 $°C$, then dropped 1 $°C$ at each step until it reached 18 $°C$. After reaching 18 $°C$, the temperature increased

(a) Optical mapping image
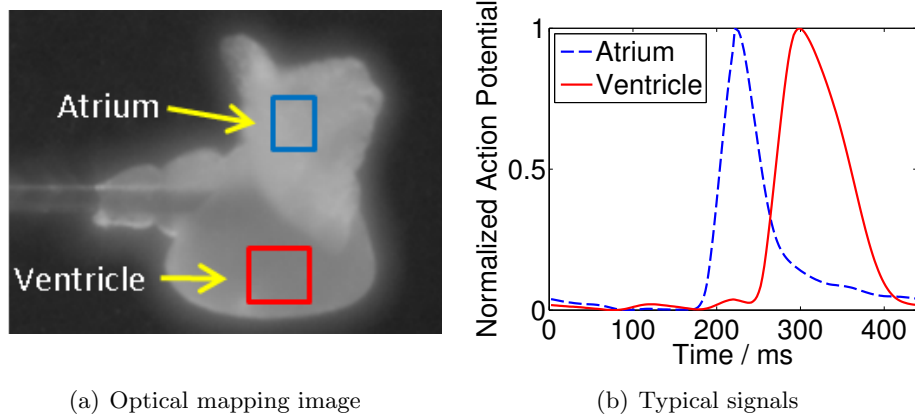
(b) Typical signals

Figure 2.1: Optical mapping zebrafish heart image and typical signals: (a) labeled optical mapping image zebrafish image, (b) plots of typical atrial and ventricular signals from bounding boxes in (a).

1 $^\circ C$ at each step until it got back to 28 $^\circ C$. There were several small exceptions in the dataset regarding to this temperature protocol: in D2 dataset, temperature was only decreased; in D1 dataset, temperature increase ended at 27 $^\circ C$, so there was no upward 28 $^\circ C$ data. In summary, D1 5 have 94 recordings in total from 5 different hearts at 11 different temperatures; in each recording, there are 4000 frames ($\sim$19.5 sec).

After data acquisition, boundaries of atrium and ventricle were delineated manually for each recording by a trained expert. The subsequent analysis and processing were done on this region of interest (ROI).

2.2 Characteristics of optical mapping zebrafish heart data

Many problems exist in the original optical mapping data which cause difficulties of data processing and interpretation.

1. The camera used in this study is slow compared to other systems [25, 23, 27], which typically have frame rate ranging from 333 fps to 3000 fps. This low temporal resolution makes timing information comparatively inaccurate and unreliable.

2. The optical mapping data has low signal-to-noise ratio (SNR) because of several reasons. The main reason for low SNR is the low dynamic range of voltage induced optical intensity change. The peak to peak value of voltage induced optical signal intensity

is only 2 or 3 unit in the output signal. After quantization, it loses the original signal shape and appears more like a staircase signal. In addition, the optical mapping signal has been corrupted by readout noise with $\sigma \approx 0.2$ unit [19] before going through the quantization.

3. Signal drifting happens due to various reasons. RH-237 has a well-known photobleaching effect, a property intrinsic to all fluorescence voltage-sensitive dyes. Fluorophore has photochemical destruction effect [15] when it is excited by laser. This causes a gradually decrease of recorded pixel intensity over time. Besides photobleaching effect, there are also 'upward' driftings potentially caused by laser intensity drifting and other slow change in experimental conditions. In our optical mapping data, these driftings are noticeable even with in a short time of recordings.

4. Non-uniform intensity makes the signal to signal comparison hard. The voltage sensitive fluorescence dye has different concentration in different locations of the heart. This results in large differences in signal magnitude across pixels, which makes comparison between signals hard to perform.

5. Various artifacts exist in zebrafish heart OM data. Although the ex-vivo zebrafish heart is contraction-inhibited by blebbistatin, the concentration of blebbistatin is not high enough to fully stabilize the heart, since higher concentration will tend to change the biological characteristics of cardiac activity. These small, even hardly visible motions are combined with non-uniform concentration of voltage sensitive dye to cause changes in pixel intensity. These motion inducted intensity changes are mixed up with voltage inducted intensity change. This way, the acquired signal contains both voltage and motion artifact information. Except for the internal motion from the heart, external motions caused by the vibration of laser, power supply and working platform also bring motion artifacts into the observed signals.

6. Signal inside a pixel is a convolution of signals from many different cells inside that pixel. Different from electrophysiology measurement using microelectrodes, which directly measures the action potential of single cell, the OM time series recorded for one pixel is actually composed of fluorescent light emitted from multiple cells inside that pixel. Furthermore, due to the translucent and scattering nature of zebrafish heart tissue, signal from single pixel also contains light from different depth and lateral

locations. This effect changes the nature of the signal and makes the electrophysiology signal characteristics for different cells less obvious and effective in OM data [26].

2.3 Preprocessing

Dealing with the above mentioned problems, a set of preprocessing steps are need to be performed before any measurements, automatic identification and even manual identification can be performed. Different methods for preprocessing cardiac optical mapping data have been reviewed in [26]. Due to size of zebrafish heart and our equipment limitations, our OM data is corrupted more heavily than OM data from other animals including mouse, rabbit and swine in previous published results. The very low SNR makes large amount of averaging operation necessary.

2.3.1 Full length signal enhancement

First, we performed a 4 by 4 binning on the original 640×480 original recording to make it a 160×120 image sequence. For any single image $I^($t$)original$ at time t, we transformed it to $I^($t$)binned$ by

$$I(t)^{binned}_{i,j} = \frac{1}{mn} \sum_{s=mi+1}^{(m+1)i} \sum_{t=nj+1}^{(n+1)j} I(t)^{original}_{s,t} \,, \qquad (2.1)$$

where i, j, s and t are image coordinates; m and n are binning size and $m = n = 4$ in this case. This binning (or down sampling) process was considered as the initial step in average smoothing. It also reduced the computation workload in following processing steps.

Then, both spatial and temporal Gaussian smoothing were performed on this binned data. To compensate for drifting caused by photobleaching and other effects, a quadratic curve was fitted to each individual pixel's signal [26] by least square fitting. For any single signal $y(t)$, we fitted a curve $q(t) = \beta_0 + \beta_1 t + \beta_2 t^2$ to represent the drifting trend of $y(t)$. As $y(t)$ was represented by $Y = [y_1, y_2, ..., y_3]^T$, we calculated $[\beta_0, \beta_1, \beta_2]^T$ by

$$[\beta_0, \beta_1, \beta_2]^T = (X^T X)^{-1} X^T Y \qquad (2.2)$$

where X is the matrix contains different powers of different t's used for the fitting.

$$X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ t_1^2 & t_2^2 & \cdots & t_n^2 \end{pmatrix}^T$$

This quadratic curve $q(t)$ was then subtracted from the original signal to remove the drifting effects.

Finally, the signals were reversed and scaled to $[0, 1]$. Here, the signal reversion is required because of the light intensity emitted by voltage sensitive dye RH-237 is negatively correlated with the value of action potential. The scaling eliminates the signal magnitude difference caused by uneven blebbistatin concentration. This entire processing pipeline for a single signal is illustrated in Figure 2.2.

After the enhancement, the new image sequence is a visualization of the action potential propagation. It is shown and compared with the original recordings in Figure 2.3. The action potential is invisible in original image sequence but clearly visualized in the enhanced data.

In addition, Figure 2.4, 2.5 and 2.6 shows the action potential propagation of same zebrafish heart under different temperature $18°C$, $28°C$ and $34°C$. We can see that high temperature leads to higher heart rates, since in the time period that $18°C$ heart finishes one cardiac cycles, the $28°C$ one finished about one and half cardiac cycles and $34°C$ heart finishes two cardiac cycles.

2.3.2 Cycle averaging

Even after full length signal enhancement described in Section 2.3.1, signals are still heavily corrupted by noise, which hampers further measurements and analysis. To increase the SNR even higher, we averaged of the signals inside each cardiac cycle and got a single one cycle signal and take it as the final output of preprocessing.

Previously, cycle averaging has been performed to increase SNR of OM data in [33]. The cycles were synchronized with recorded stimuli markers. Here, we propose a method that averages cardiac cycles solely based on OM data and does not involve any synchronization input from external sources.

We assume that the heart does not have second-degree atrioventricular block [43], which means both atrial impulses might fail to conduct to ventricle. But the heart may have first-degree atrioventricular block [4] or other problems that will not cause skipping cycles in ventricle. For example, different cardiac cycles have different periods.

For cycle identification, we used image similarity $\rho(t)$ to sync different cycles. First, a reference image was chosen. We took the average enhanced signal of atrium, and choose the frame correspond to the largest average signal value as reference image $I(t_0)$. Then, we calculated cosine similarity [10] between every frame $I(t)$ and the reference frame $I(t_0)$ as

(a) Single pixel signal from original data

(b) Single pixel signal from 4x4 binned data

(c) Spatial and temporal smoothed signal from a single pixel and overlaid with the drifting curve fitted

(d) De-trended, reversed and scaled signal inside a pixel
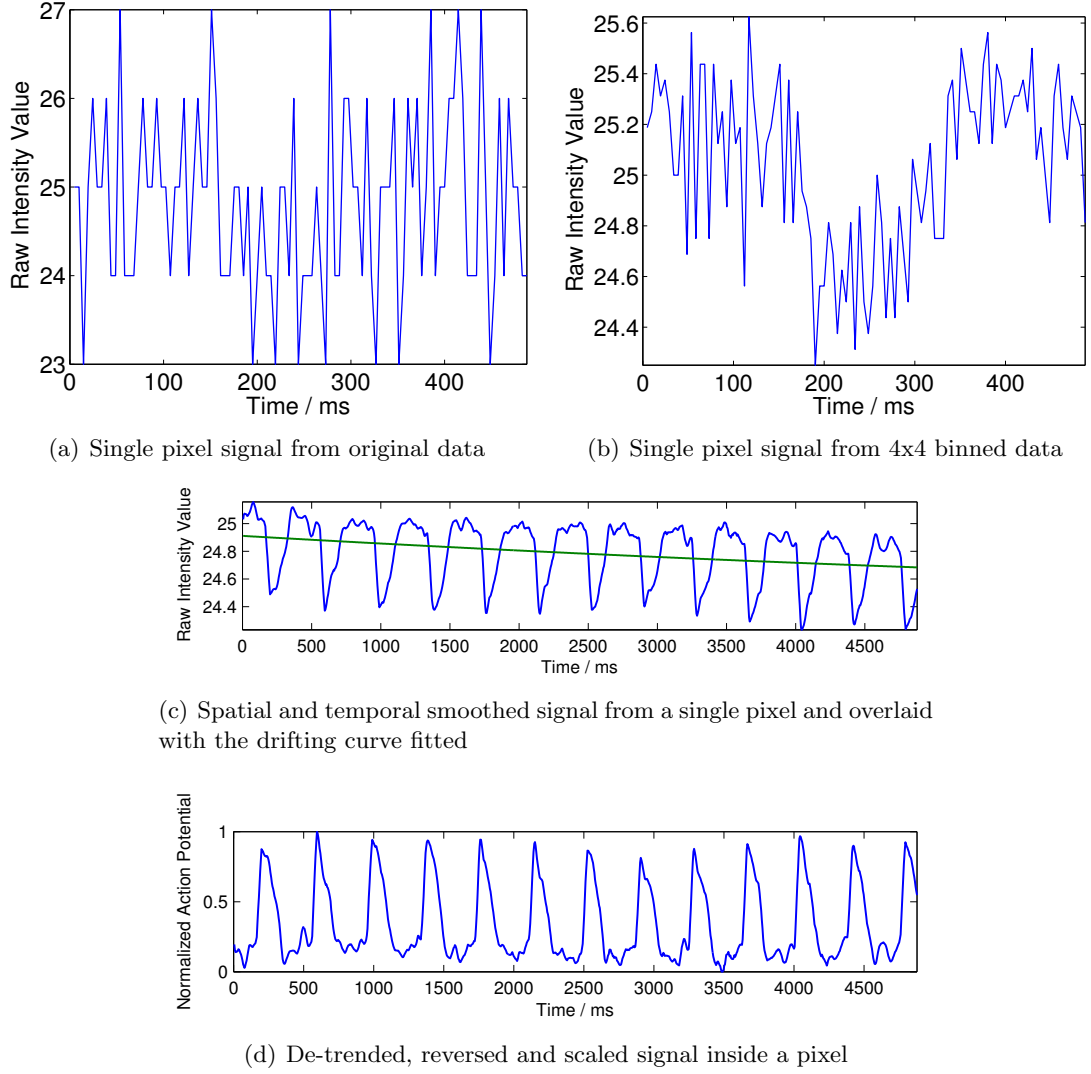
Figure 2.2: Preprocessing pipeline illustration: (a) Raw signal from single pixel of original data, (b) Raw signal from single pixel of 4 by 4 binned data, (c) spatial and temporal smoothed signal; the green line is the fitted drifting trend, (d) drifting is removed, and signal is reversed and scaled

image similarity:

$$\rho(t) = \frac{I(t) \cdot I(t_0)}{\|I(t)\|\|I(t_0)\|} = \frac{\sum_{i,j} I_{ij}(t) \times I_{ij}(t_0)}{\sqrt{\sum_{i,j} I_{ij}^2(t)} \times \sqrt{\sum_{i,j} I_{ij}^2(t_0)}} \ , \tag{2.3}$$

where $I_{ij}(t_0)$ and $I_{ij}(t)$ represent the pixels values in $I(t_0)$ and $I(t)$ respectively.

(a) Original optical mapping image sequence

(b) Enhanced optical mapping image sequence
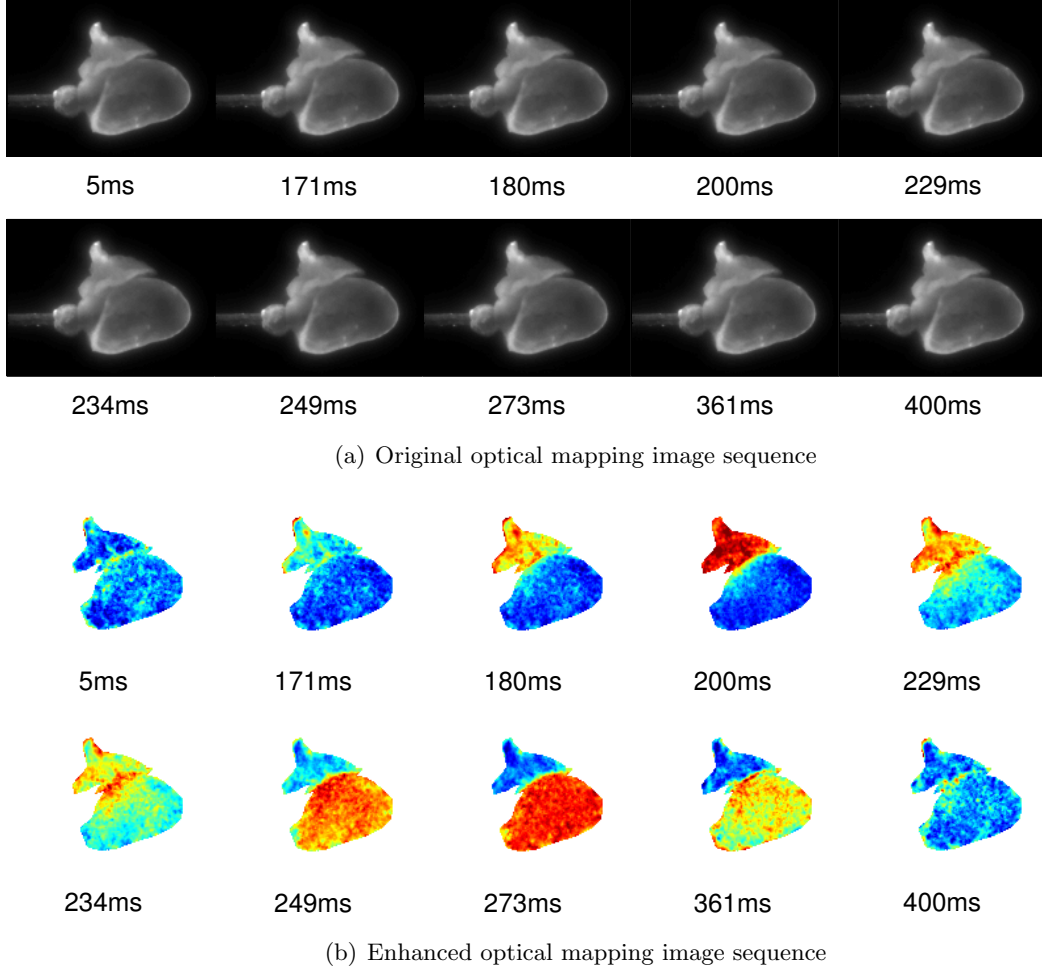
Figure 2.3: Illustration of action potential propagation after preprocessing: (a) Original image sequence, (b) Enhanced image sequence with pseudo jet color map showing the action potential propagation

To divide the full length signal into individual cardiac cycles, peaks in $\rho(t)$ were taken as the separators. After that, each individual cycle between 2 adjacent separators were interpolated to vectors with same length presenting time scaling.

To perform cycle averaging, we need to align each individual cycle first and then average those aligned signals. The 1-D diffeomorphisms were acquired in the 2 steps: First, a 'mean' one-cycle image cosine similarity was generated by algorithm proposed in [22, 44]. It treats different signals as random time-warpings of the same original signal. This original signal is estimated as the Karcher mean on the quotient space of equivalence classes formed by action
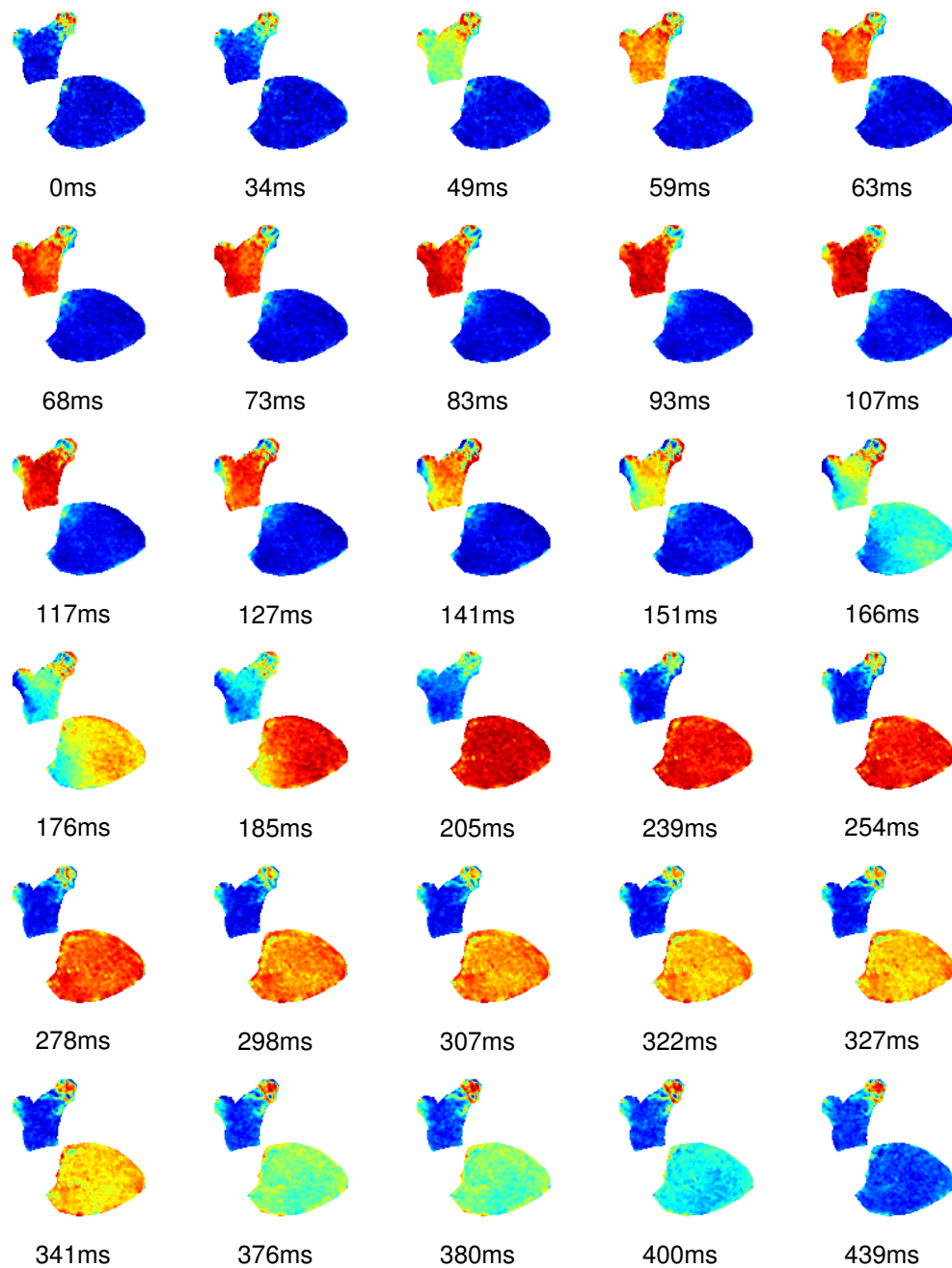
Figure 2.4: Illustration of action potential propagation under 18 °C. This is from a separate dataset (other than D1∼5) which contains temperature 34 °C.
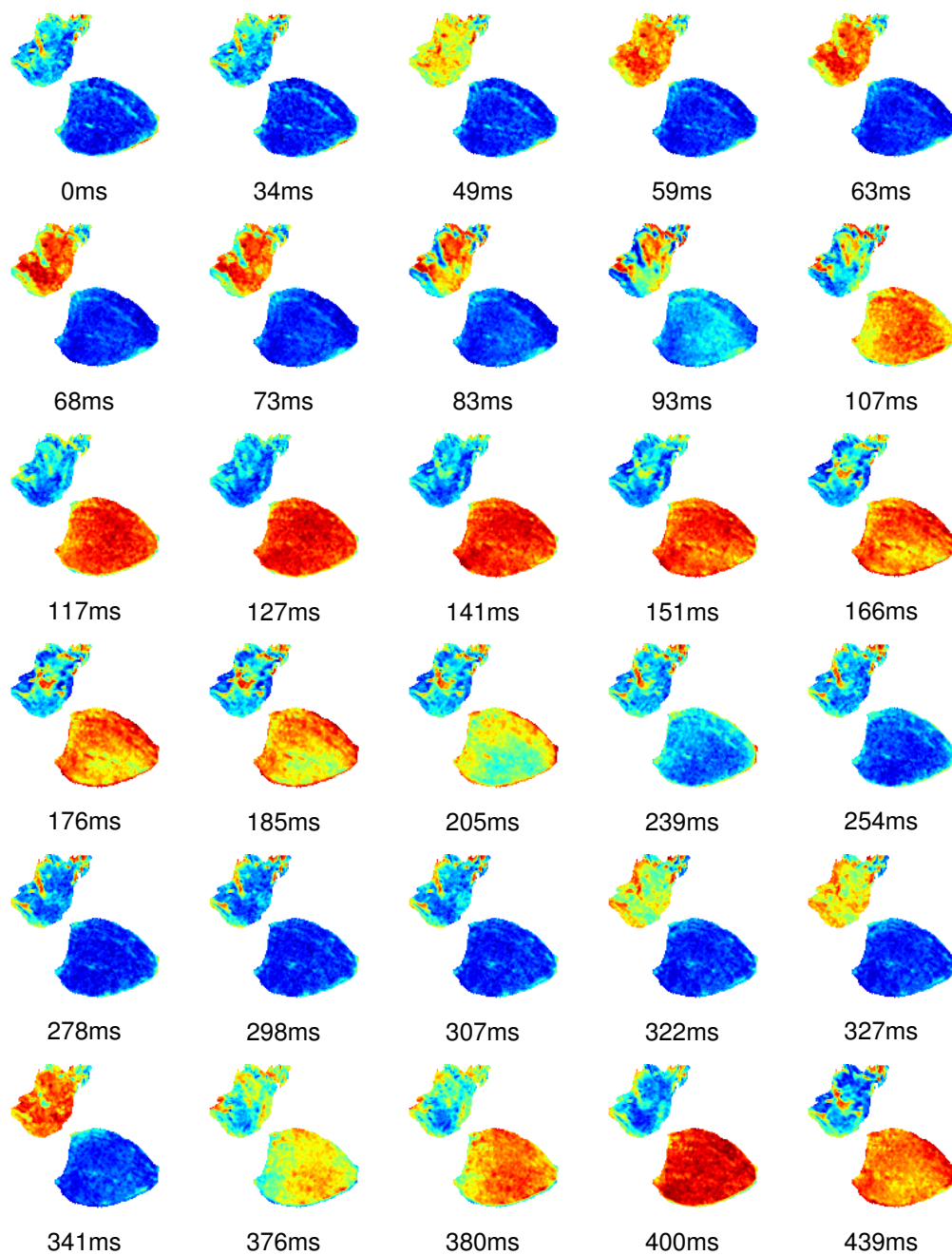
Figure 2.5: Illustration of action potential propagation under 28 °C. This is from a separate dataset (other than D1∼5) which contains temperature 34 °C.
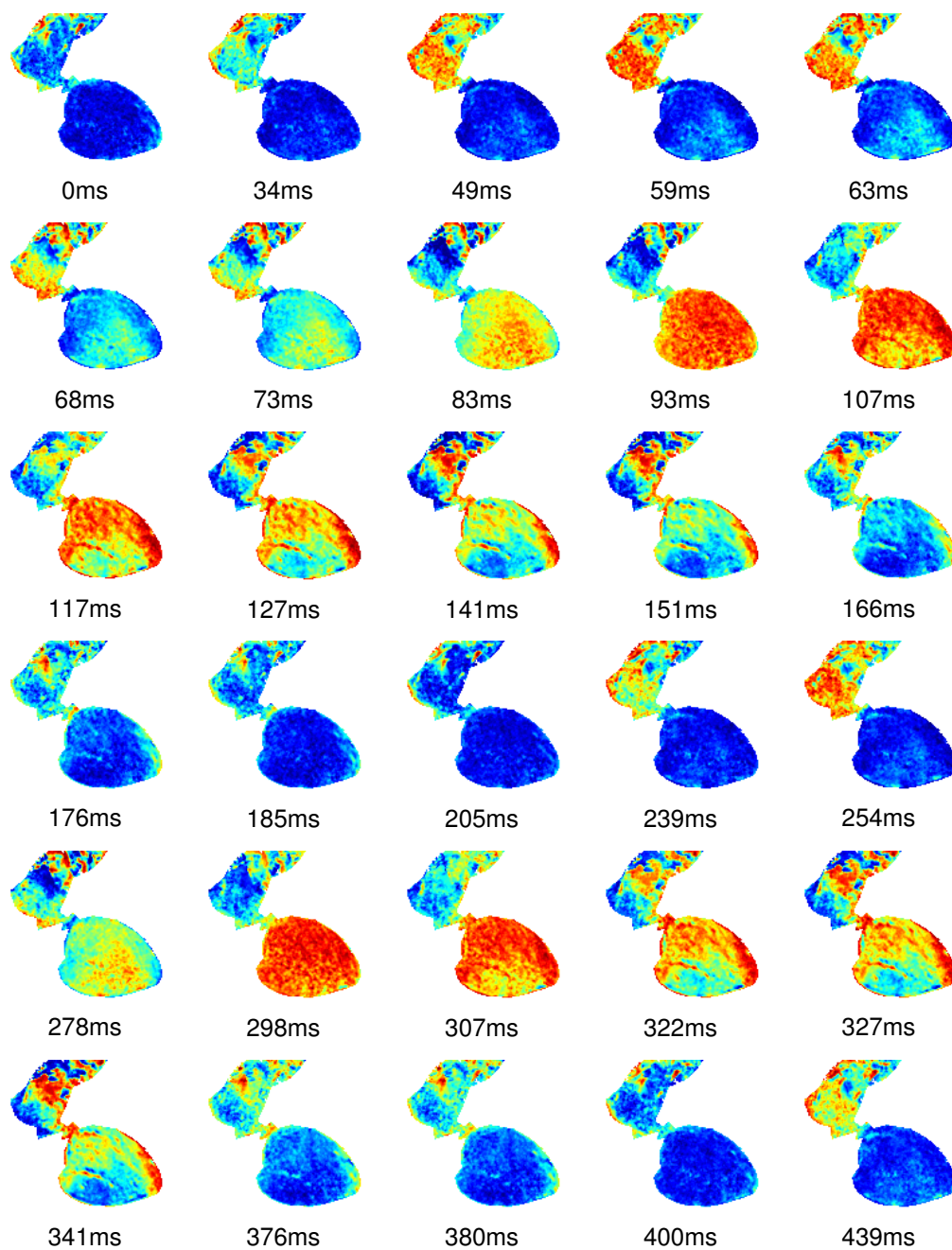
Figure 2.6: Illustration of action potential propagation under 34 °C. This is from a separate dataset (other than D1∼5) which contains temperature 34 °C.

of warping group. Second, every individual one cycle image cosine similarity is registered to the acquired 'mean' signal. The registration is done with time-warping algorithm based on dynamic programming [49]. It first smooths noisy signals with kernel estimation. Then, it defines a proper cost function between the target signal and the template signal. Finally, this cost function is minimized by shifting and warping the target signal using dynamic programming.

Having acquired cycle separators and time-warping functions based on image cosine similarity, we segmented individual cardiac signals in to cycles, interpolated them into same length, warped them with the warping functions and averaged them to get a single average one cycle signal. The process of cycle averaging is illustrated in Figure 2.7 and Figure 2.8. These 2 figures compares the linearly cycle averaging with the proposed cycle averaging after time-warping. We can see that applying time-warping function gives less standard deviation and less deformed average signal cycle, and hence is a superior method than linear averaging.

Here, the time-warping algorithm was added into the pipeline at a relatively later stage. The results discussed in the following chapters are therefore based on linearly averaged one-cycle signals. However, as the time-warping based cycle averaging has now been added to the tool chain, it can be used for analysis of new data.

2.4 Conclusion

In this chapter, we introduced the optical mapping data acquisition process, data characteristics and its preprocessing steps. The originally acquired data has (1) low frame rate, (2) very low SNR, (3) slow drifting, (4) non-uniform spatial intensity, (5) motion artifacts and (6) mixing of signal locations. The proposed preprocessing steps turned the very low SNR original data into clearly interpretable signals and videos. The proposed methods solved the problem of (2), (3), (4). However, as a price we pay, it aggravated the location mixing problem. The novelty of this preprocessing pipeline is in the cycle averaging method. Unlike previous cycle averaging method [33], the proposed method utilize the information from the optical mapping data itself and does not need external information. need to external information In summary, we illustrate the preprocessing pipeline in Figure 2.9.

(a) Cosine similarity over time

(b) warping functions

(c) Similarity cycles before warping

(d) Similarity cycles after warping

(e) Average similarity before warping

(f) Average similarity after warping
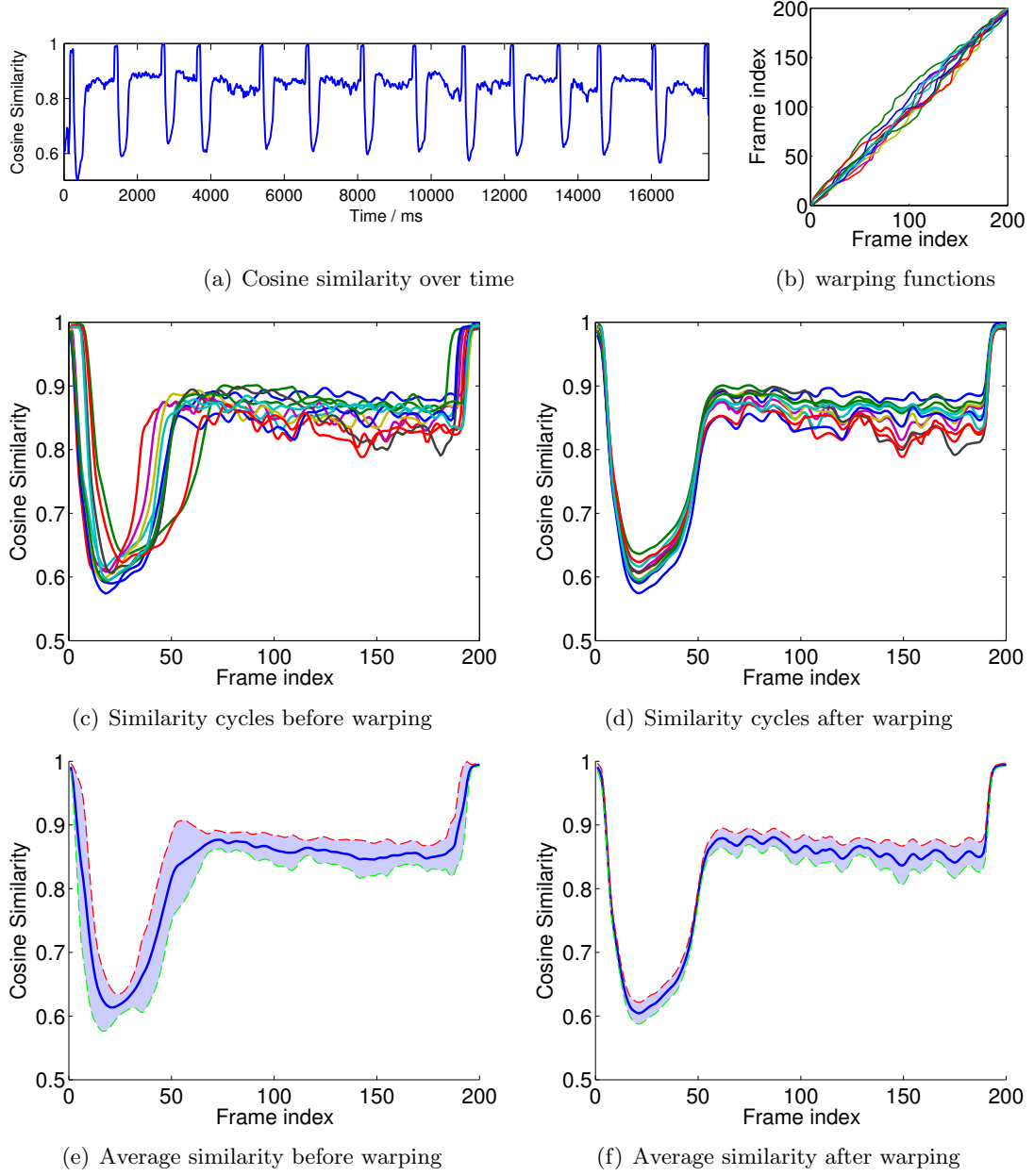
Figure 2.7: Acquiring warping functions from image cosine similarity: (a) image cosine similarity over time, (b) calculated warping functions, (c) segmented and interpolated cycles of similarity, (d) similarity cycles after applying warping functions, (e) mean and standard deviation before applying warping, (f) mean and standard deviation after applying warping.

(a) Signal from single heart pixel

(b) Signal cycles before warping

(c) Signal cycles after warping

(d) Average signal before warping

(e) Average signal after warping
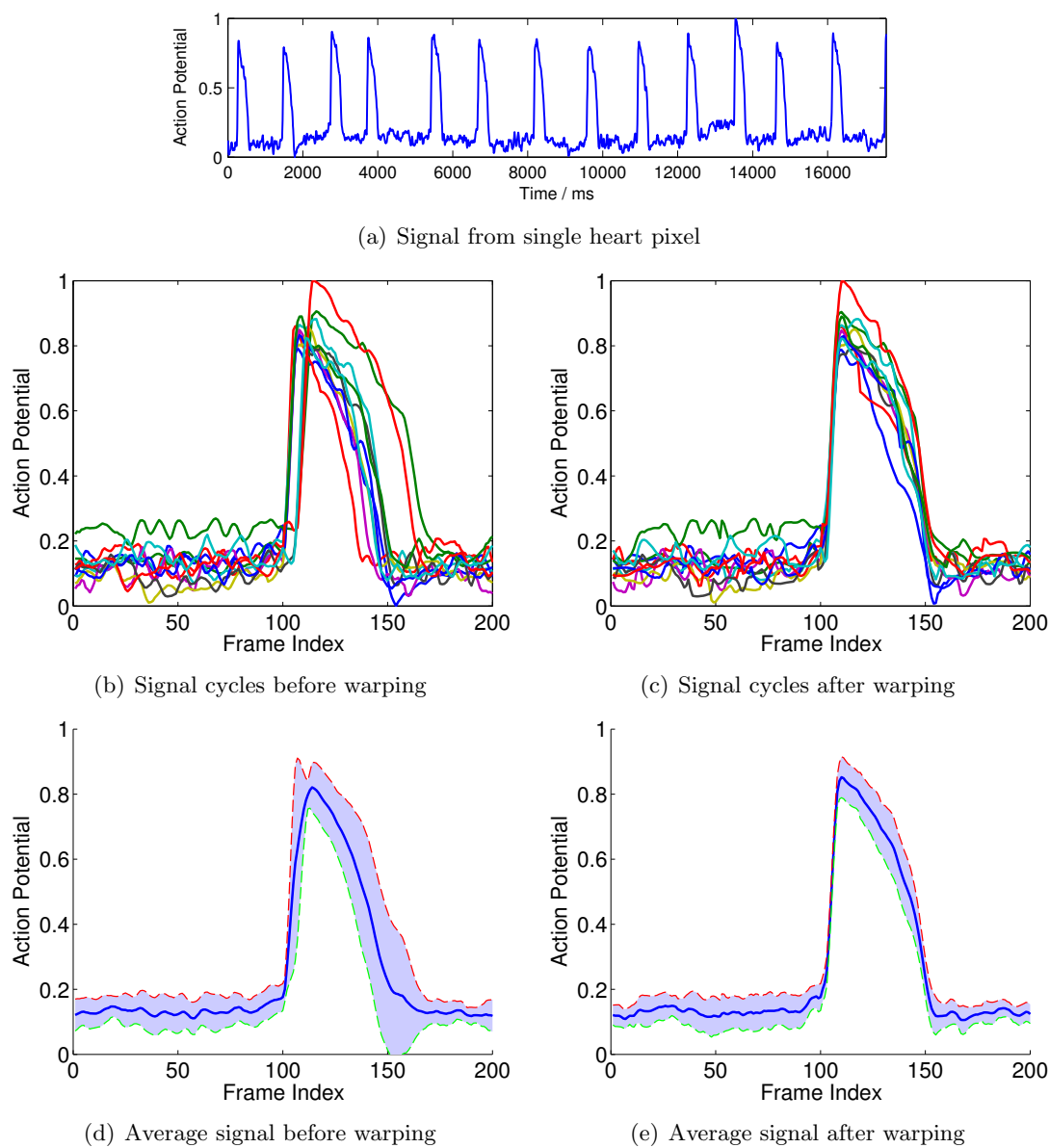
Figure 2.8: Use acquired warping functions on individual signal: (a) action potential signal from single pixel of heart, (b) segmented and interpolated cycles of action potential signal, (c) action potential signal cycles after applying warping function, (d) mean and standard deviation before applying warping, (e) mean and standard deviation after applying warping.
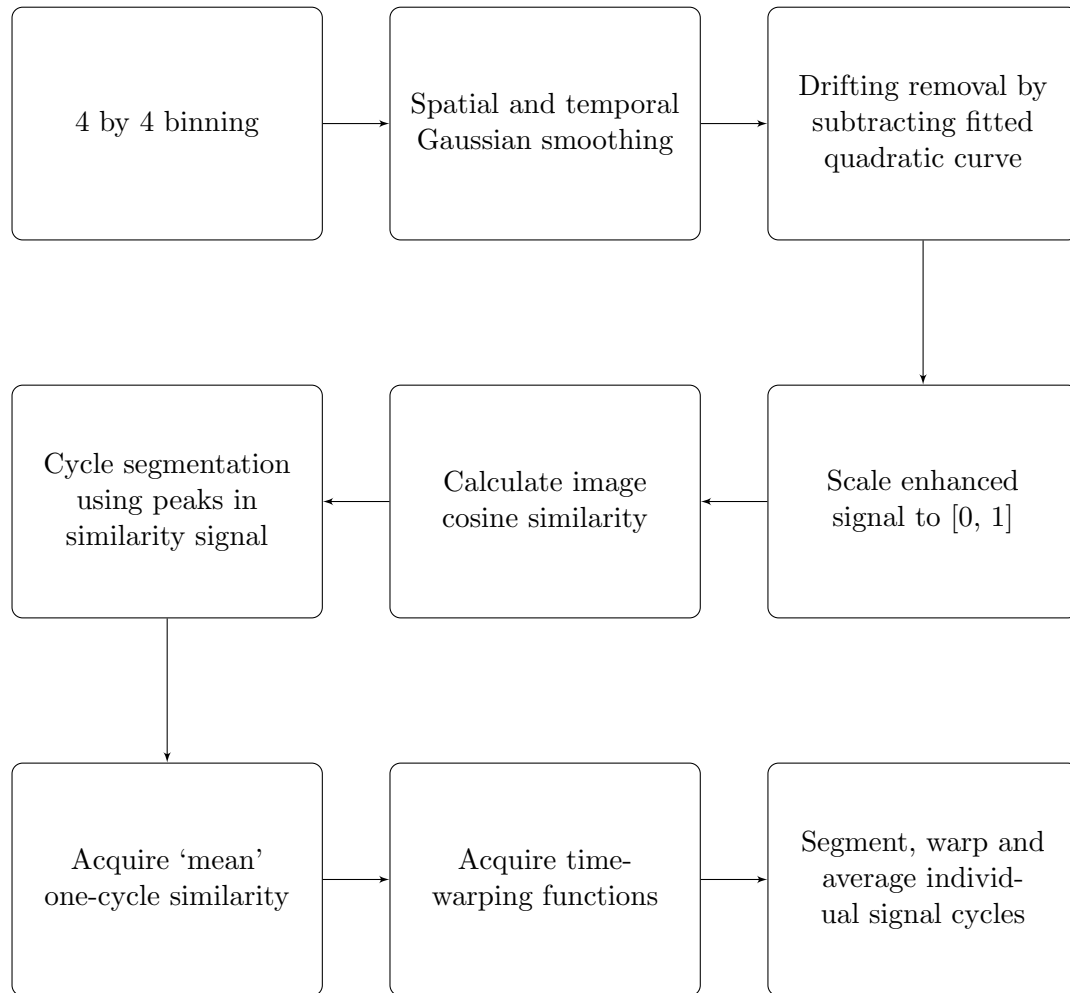
Figure 2.9: The preprocessing pipeline

Chapter 3

Manual Identification

Without clearly-stated, logical rules of how to detect pacemaker in zebrafish heart, the best approach is to have experienced physiologist manually label the optical mapping signals. This would benefit us in two aspects: (1) it provides information of how human expert identifies pacemaking signals; (2) it provides groundtruth data for automatic methods to be compared with. However, the number of pixels to label is very large, which makes this process very slow and tedious.

In this chapter, we propose and implement a manual pacemaker identification pipeline by first grouping signals into small clusters automatically, then perform expert labeling on the average signals of these clusters. This 2-step process significantly reduces the manual workload.

3.1 Methods

3.1.1 Pre-labeling Clustering

Ideally, to acquire expert opinion on whether one pixel is constituted by pacemaking cells, we need to have the expert to label the signal on every pixel inside the atrium. This work is tedious, since for each double-side recorded data we have about 1500~2500 pixels to label, given the frame size is 160×120. Even if we label one pixel every 2 seconds, we need an hour to label one recording. Labeling about 100 recordings would need 100 hours, which is not affordable. Furthermore, if the data resolution increases, this process will take even longer to finish.

Therefore, methods reducing the manual labeling workload is desired. Using clustering to reduce the number of signals to label is a natural way to achieve this goal. Common clustering methods include K-Means, fuzzy c-means (FCM) and expectation-minimization algorithms [51]. These clustering algorithms are iterative processes, in which the cluster centers and membership function of each sample are calculated based on each other alternatively until convergence.

In pacemaker identification problem, we try to identify different connected regions. It is reasonable to assume that cells of similar function will likely be spatially adjacent. Hence, a method that considers both signal space and spatial adjacency is desired. Here, we used the robust fuzzy c-means (RFCM) [37, 38], which incorporates spatial information by penalizing spatially scattered clusters. FCM and RFCM are briefly introduced below, under our application scenario. For detailed explanation of RFCM, we refer the reader to [37, 38].

Fuzzy c-means

RFCM considers additional spatial information based on the traditional FCM. In traditional FCM, we try to minimize an objective function with respect to the membership function of each data sample u_{jk} and the centroid of each cluster $\boldsymbol{v}_k$.

$$J_{FCM} = \sum_{j \in \Omega} \sum_{k=1}^{C} u_{jk}^q \|\boldsymbol{y}_j - \boldsymbol{v}_k\|^2 \ . \tag{3.1}$$

Here, j is the pixel index, $\boldsymbol{y}_j$ is the signal/vector on each pixel, C is the number of classes, k is class index, Ω is our region of interest (ROI) and q is a parameter that controls the 'fuzziness' of FCM clustering. The membership functions are constrained to be positive and need to sum up to one for each individual sample:

$$\sum_{k=1}^{C} u_{jk} = 1 \ . \tag{3.2}$$

J_{FCM} is the sum of squared difference inside each class. By minimizing it, the difference between signals inside each class are minimized. The algorithm starts from C random initialized centroids in the signal vector space, and use an iterative process to update u_{jk}'s and $\boldsymbol{v}_k$'s, until the convergence of J_{FCM}.

Robust fuzzy c-means

Based on Equation 3.1, RFCM adds spatial information related penalty into the objective function:

$$J_{RFCM} = \sum_{j \in \Omega} \sum_{k=1}^{C} u_{jk}^q \|\boldsymbol{y}_j - \boldsymbol{v}_k\|^2 + \frac{\beta}{2} \sum_{j \in \Omega} \sum_{k=1}^{C} u_{jk}^q \sum_{l \in N_j} \sum_{m \in M_k} u_{lm}^q \ , \tag{3.3}$$

where N_j is the set of j's neighboring pixels, and $M_k = \{1, ..., C\} \backslash \{k\}$ is the classes except for k. β is an introduced parameter that controls the trade off between original FCM objective function and smoothness of membership function.

The first part of Equation 3.3 is the original FCM objective function. The second term penalizes the spatial membership function inconsistency. In other words, for a specific pixel, this term is minimized when one of its class membership values is large and membership values for other classes at its neighboring pixels is small (and vice versa).

Similar to FCM, RFCM starts from C random initialized centroids and does an iterative process to update u_{jk}'s and $\boldsymbol{v}_k$'s. The updating rules derived using Lagrange multipliers are derived in [37] as:

$$u_{jk} = \frac{(\|\boldsymbol{y}_j - \boldsymbol{v}_k\|^2 + \beta \sum_{l \in N_j} \sum_{m \in M_k} u_{lm}^q)^{-1/(q-1)}}{\sum_{i=1}^{C} (\|\boldsymbol{y}_j - \boldsymbol{v}_i\|^2 + \beta \sum_{l \in N_j} \sum_{m \in M_i} u_{lm}^q)^{-1/(q-1)}} \ , \tag{3.4}$$

$$\boldsymbol{v}_k = \frac{\sum_{j \in \Omega} u_{jk}^q \boldsymbol{y}_j}{\sum_{j \in \Omega} u_{jk}^q}, \ \ k = 1, \cdots, C \ . \tag{3.5}$$

Here, we set the number of classes to be 100, which means that all the signals are grouped into 100 clusters. The final number of clusters are usually less than 100, since some of the clusters contain no pixel at the end. After the RFCM clustering, we take the average signal of each region and perform manual labeling on these average signals. This process reduced the manual work by at least 20 times compared to performing pixel by pixel manually labeling. For the RFCM clustering, we do 100 iterations, which takes about 2 minutes to run on our computer (Intel Core2 duo 2.66 GHz, 6GB of RAM).

3.1.2 Manual Labeling Process

After the clustering process, averaged signal in each cluster was presented to and labeled by a trained expert, EL. EL, a physiologist by training with a doctoral degree, is currently
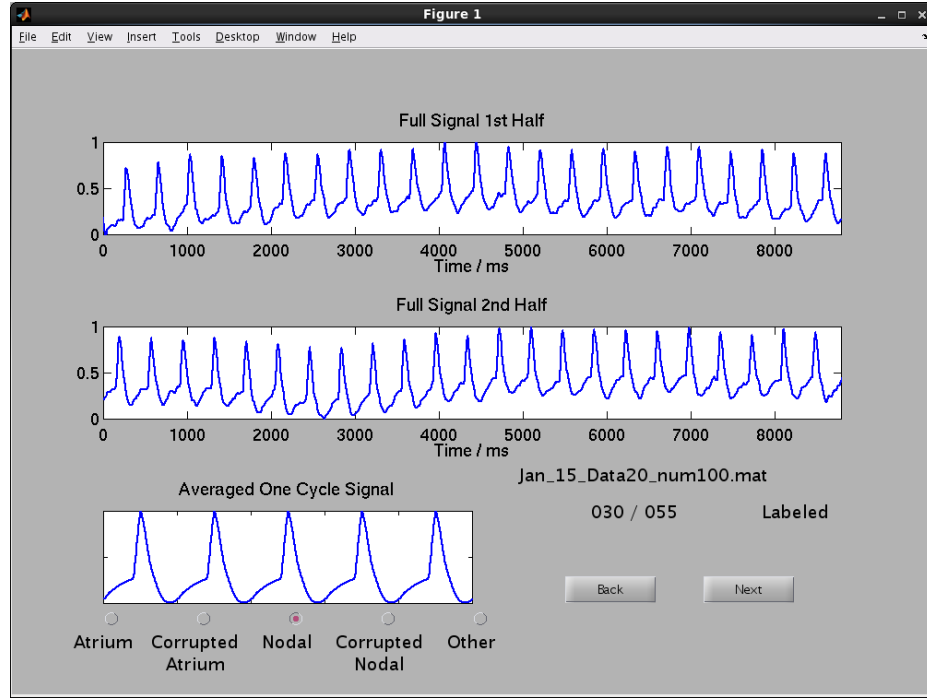
Figure 3.1: Graphical user interface for manual labeling

investigating cardiac pacemaking nodes in mouse, rabbit and zebrafish hearts with a variety of techniques including optical mapping.

When performing manual labeling, averaged full length signals and averaged one cycle signals were presented to the expert. As we thought the expert's opinion should only focus on identifying the waveform with physiologically meaningful interpretation, we did not provide him other information like location and neighborhood signals. Figure 3.1 is the screen shot of the manual labeling graphical user interface (GUI). It shows the averaged full length signal and one cycle signal for each clustered region.

User can choose to label the signal and its corresponding region to five different labels: normal atrial signal (AN), corrupted atrial signal (AC), normal nodal signal (NN), corrupted nodal signal (NC) and other (OT). Here corrupted means that the given signal has visible distortion, but the expert thinks he can still label it according to some features. On the GUI, it also shows the status of this signal (Labeled/Unlabeled), the data name and number of total and left signals in this data.

3.2 Results

3.2.1 Results of manual identification

Figure 3.2 shows an example of RFCM clustered the averaged signals of each cluster with their standard deviations. The manual labeling result on the same example is shown in Figure 3.3, where differently labeled signals and regions are color coded. We can observe some features that the expert is looking for. For example, nodal signals usually have positive diastolic depolarization slope. Atrium signal usually have flat resting period before the main upstroke. Decreasing slope in resting period and second peak are features for corrupted signals.

Figure 3.4 shows more labeling results with labeled images and grouped signals. From labeled images, we can see that regions with same label are mostly connected to each other as large image patches. Also, corrupted atrial and nodal region are usually connected to normal atrial and nodal region respectively. In addition, regions labeled from anterior and posterior sides of heart coincide with each other well.

Furthermore, part of our manual identification results on OM data agree with the pace-making region positions reported in literature. According to the labeled images, the potential nodal regions appear consistently in atrium on top side (at the opposite end of AV junction) and usually near the bulbs side. The identified pacemaking region on top side of atrium coincides with the 'ring around the venous pole' reported in [45] and the result acquired with activation time measurement in [29].

For the second pacemaking region on one side of atrium, we speculate their appearance is possibly caused by 2 reasons: (1) the deformability of tissue makes atrium stay in relatively different shapes and positions in experiments with different individual hearts; (2) current surgical procedure damages the SAN which allows other pacemaker cells starting to dominate.

3.2.2 Ranking data according to manual label

Suitability and quality of data plays an important role in development of pacemaker identification method. We mainly care about two things: (1) whether the data is largely corrupted by motion and (2) whether it contains identifiable nodal regions. Therefore we propose two measures which represent the percentage of uncorrupted region and percentage of nodal
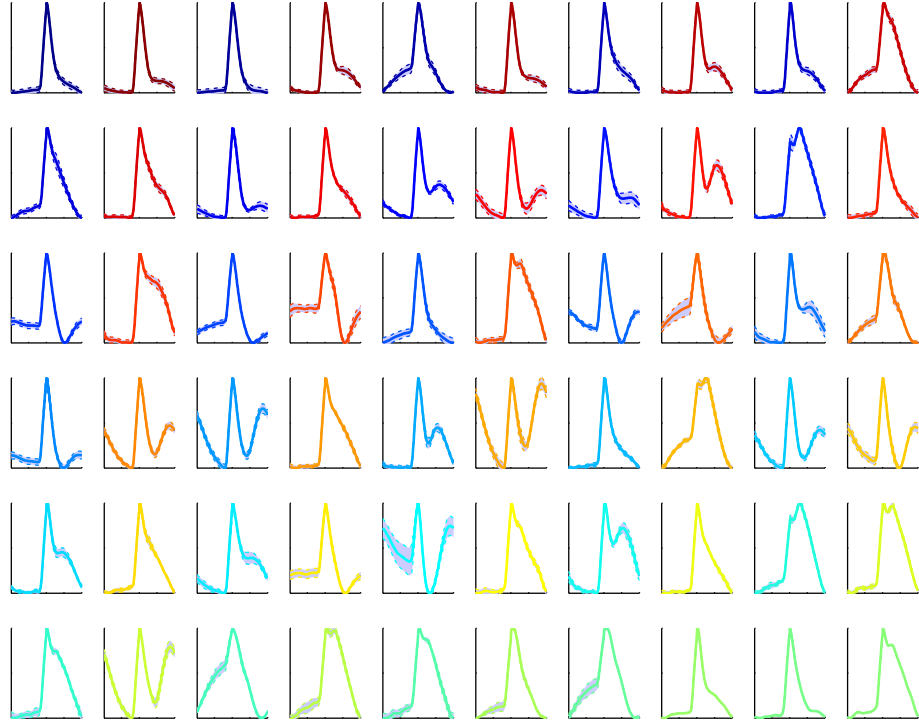
(a) Averaged signals from different clusters in atrium
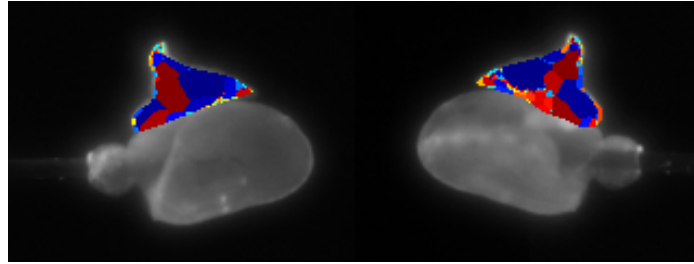
(b) Area classfication on atrium

Figure 3.2: Result of the first stage of clustering: (a) averaged signals from each clusters, (b) zebrafish heart with color coded patches labelled. Each signal in (a) represent the the average in the cluster of the same color in (b). (b) contains images from both anterior and posterior sides.

region respectively. The 'normal coefficient' (NMC) takes the form of:

$$NMC = 1 - \frac{area(OT) + 0.5(area(NC) + area(AC))}{Total\ area}. \tag{3.6}$$

By considering OT as totally corrupted region, NC and AC as half corrupted region, NMC calculates the percentage of uncorrupted region in the total area. Generally, it represents
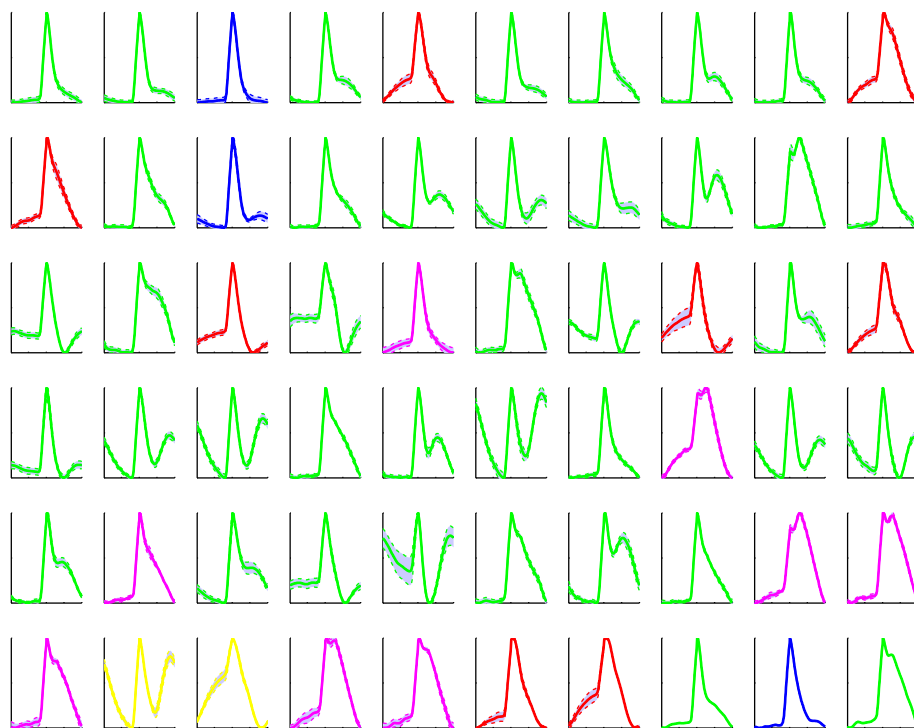
(a) Color coded labeled average signals
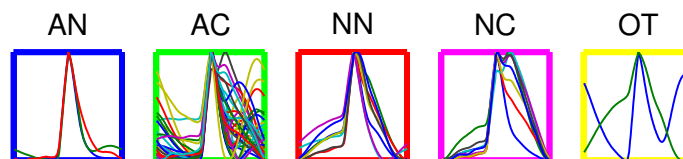
(b) Signals grouped according to label
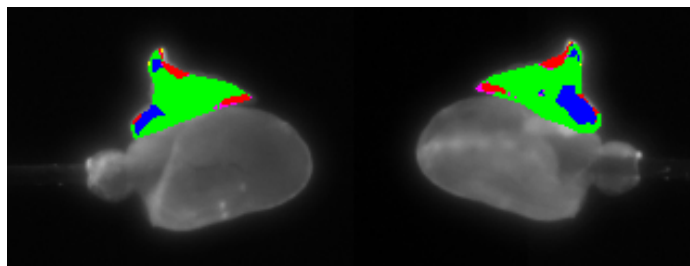
(c) Manual area classification on atrium

Figure 3.3: The color coded area partition after manual labeling: (a) color coded labeled average signals; (b) signals grouped according to label; (c) manual area classification on atrium.
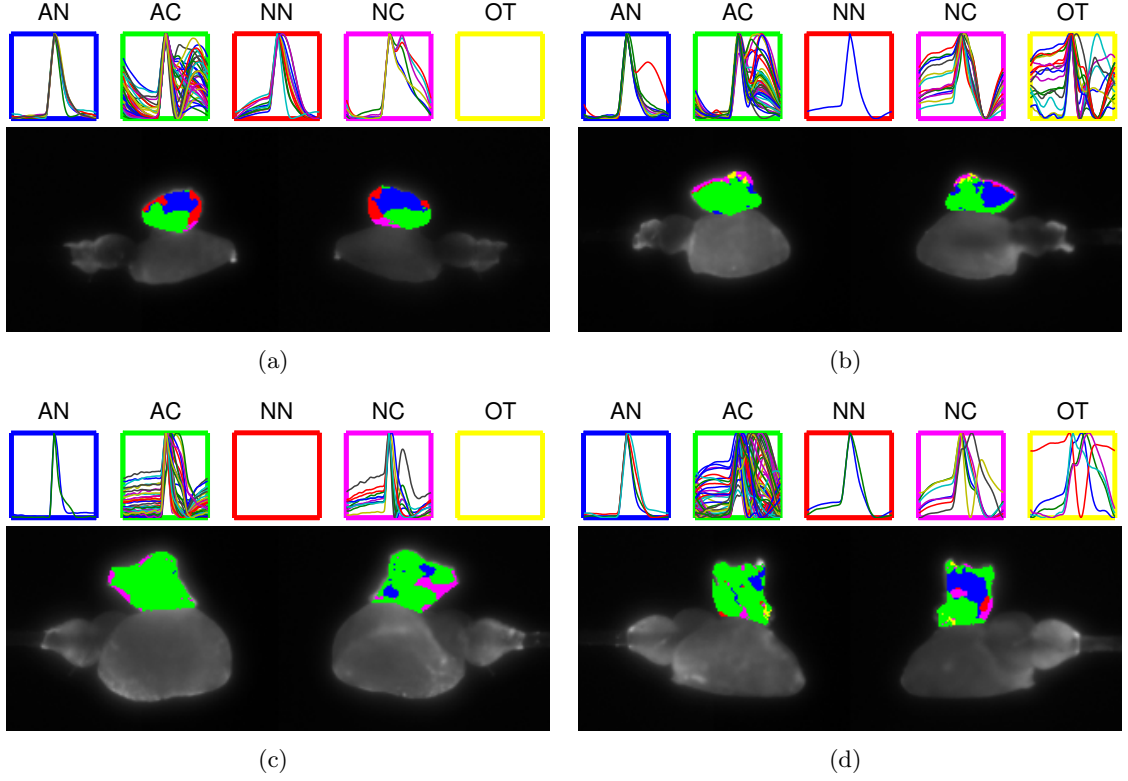
Figure 3.4: More image and signal labeling results from different zebrafish hearts. The colors of bounding boxes and image labels correspond to each other. Empty bounding box without signals means there no such type of signal labeled in this data.

the 'intactness' of the acquired data.

The 'nodal coefficient' (NDC) takes the form of:

$$NDC = \frac{0.5area(NC) + area(NN)}{0.5(area(NC) + area(AC)) + (NN + AN)}.$$

(3.7)

Similarly as NMC, NDC considers NC and AC as half corrupted region. It represents the percentage of nodal region on among the uncorrupted region.

Ideally, for 'good' data, we expect both NMC and NDC to be high. In Figure 3.5, we can see that only D1 dataset has both relatively high NMC and NDC. While for other datasets, either or both of NMC and NDC is low. Based on this observation, we separate these datasets into two groups: high quality data which contains D1 and low quality data which contains the rest of the datasets. To reflect the true performance under different data condition, some methods and measurements will be analyzed separately on high and low
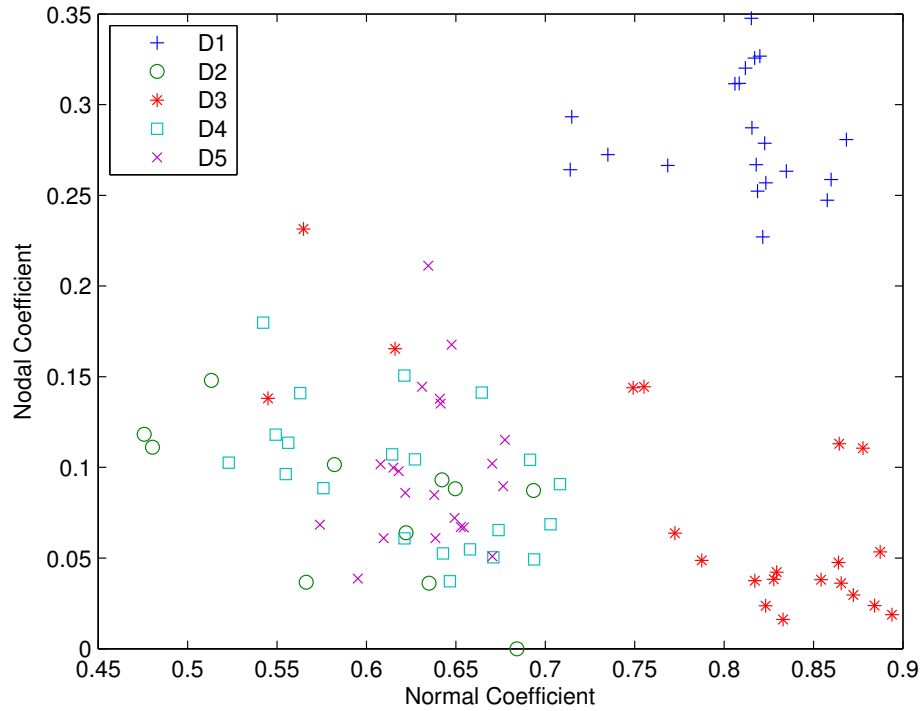
Figure 3.5: The normal coefficient and nodal coefficient for different datasets

quality data in the following chapters of this thesis.

3.3 Conclusion

This chapter introduced a computer assisted pipeline for labeling atrium pixels on optical mapping data. The advantages of this pipeline are: (1) the RFCM procedure groups signals together, the average signal of each cluster has higher SNR than the signals from single pixels; (2) it reduces the manual labeling workload significantly. The disadvantages include: (1) RFCM has randomness, which causes slight inconsistency of labeling if performed multiple times; (2) this procedure is not as accurate as labeling every single pixel signal and will cause inaccuracy when using it as a groundtruth for evaluation.

The manually labeled signals and image regions are verified in different aspects: (1) the pixels with same label are connected regions; (2) normal and corrupted regions are usually adjacent; (3) the labeled pacemaking sites correspond to those described in literature.

Chapter 4

Measurements and Analysis

Physiologically meaningful measurements provide useful information for identification of different functional areas on zebrafish heart. They also provide a way of validating the expert manually labeled data according to physiological properties. Early activation time and presence of diastolic depolarization are important criteria for determining pacemaker cells in cardiac electrophysiology [21]. In this chapter, activation time and diastolic slope value are measured on enhanced one-cycle optical mapping signals and analyzed statistically with the expert labeled data.

4.1   Methods

Early activation time and positive diastolic slope value are the two key features for distinguishing nodal signal from atrial signal. Here, we use different terms, including 'resting slope', 'slow slope' and 'diastolic slope' to refer the diastolic slope of nodal signal and resting period slope of atrial signal. Contrasting to 'slow slope', we refer to the main depolarization upstroke as 'fast slope'.

   The signal used for measurements is different from the scaled signals used in Chapter 2. In Chapter 2 we normalize the signal value to [0, 1], which eliminates the proportional correlation between different pixels. Here, the signal $f(t)$ is used for calculating diastolic slope value, thus need to be an 'absolute value' in the sense that it can be compared across pixels. Under the assumption that voltage induced intensity variation is proportional to the

mean intensity of that pixel, we define:

$$f(t) = \frac{F(t)}{\mathbb{E}(F(t))} \ ,$$ (4.1)

where $F(t)$ is the enhanced one-cycle signal without scaling to $[0, 1]$ and $\mathbb{E}(F(t))$ is the expectation (average value) of $F(t)$.

4.1.1 Measurements

Measurements on a single signal

Activation time is defined by time point when the signal has the fastest rising speed [26]:

$$t_{activation} = \arg\max_t \frac{df(t)}{dt} \ ,$$ (4.2)

where $f(t)$ is the recorded fluorescence intensity. OM signals are sometimes noisy, which makes the t that satisfies Equation 4.2 not the expected as correct activation time point. In order to deal with this problem, we utilized the single peak property of normal cardiac signal. The peak location was first extracted, and activation time was searched in the local range before peak time:

$$t_{activation} = \arg\max_{t_{peak}-t_{local}<t<t_{peak}} \frac{df(t)}{dt} \ ,$$ (4.3)

where t_{peak} is the peak time and t_{local} is the local range in which $t_{activation}$ is searched.

To calculate the diastolic depolarization slope value, we need to identify the starting and ending point of the 'slow slope'. Given t_{end} represents the diastolic slope ending point and t_{1st} represent the first point of the averaged one cycle signal, t_{end} was defined as the point that is farthest from the straight line connecting the 1st point $(t_{1st}, f(t_{1st}))$ and peak point $(t_{peak}, f(t_{peak}))$ and satisfy $t_{1st} < t_{end} < t_{activation}$:

$$t_{end} = \arg\max_{t_{1st}<t<t_{activation}} \frac{|(f(t_{peak}) - f(t_{1st}))t - (t_{peak} - t_{1st})f(t) - t_{1st}f(t_{peak}) + t_{peak}f(t_{1st})|}{\sqrt{(f(t_{peak}) - f(t_{1st}))^2 + (t_{peak} - t_{1st})^2}} \ .$$ (4.4)

Under this definition, the value of t_{end} is not affected by the scaling of both $f(t)$ and t, which is a desired property. Detailed derivation of Equation 4.4 and its scaling invariance property can be found in Appendix A.

Having t_{end} as reference, the starting point used to calculate diastolic slope was defined as $t_{start} = t_{end} - t_{offset}$. The parameter t_{offset} is an empirical temporal length to capture

enough points in diastolic period while avoiding including the corrupted interval. After having t_{start} and t_{end}, we calculated the slope value $s_{diastolic}$ by fitting a straight line to the points between t_{start} and t_{end} with least square regression [36]:

$$s_{diastolic} = \frac{\sum_{t_{start} \leqslant t_i \leqslant t_{end}} (t_i - \mathbb{E}(t_i))(f(t_i) - \mathbb{E}(f(t_i)))}{\sum_{t_{start} \leqslant t_i \leqslant t_{end}} (t_i - \mathbb{E}(t_i))^2} \ , \tag{4.5}$$

where $\mathbb{E}(t_i)$ is the average of t_i's and $\mathbb{E}(f(t_i))$ is the average of $f(t_i)$'s.

Activation map and slope map

Using methods described in Section 4.1.1, we measured the activation time $t_{activation}$ and diastolic depolarization slope $s_{diastolic}$ for every single pixel or clustered region in atrium. By plotting these values using pseudo colormap, we acquired the activation map and slope map.

4.1.2 Statistical analysis

T-test

After acquiring measurements from original enhanced one-cycle signals on both single pixels and average signals on manual labeled patches, we performed statistical analysis to reveal the correlation between these measurements and the manually labeled regions.

Pacemaker cells have early activation time and higher positive diastolic slope value. Hence, for both activation time and slope value, we performed one-sided hypothesis testing [36]. As nodal regions (or pixels) and atrial regions (or pixels) are not paired, we did one-sided two-sample t-test to examine our hypothesis on activation time and slope value.

For activation time, we tried to accept the alternative hypothesis H_A that activation time is earlier in nodal regions than in atrial regions:

$$H_A : \mu^t_{nodal} < \mu^t_{atrial} \ , \tag{4.6}$$

by rejecting the null hypothesis H_0 that states the opposite:

$$H_0 : \mu^t_{nodal} \geqslant \mu^t_{atrial} \ . \tag{4.7}$$

For diastolic slope value, we tried to accept the alternative alternative hypothesis H_A that diastolic slope value is larger in nodal regions than in atrial regions:

$$H_A : \mu^s_{nodal} > \mu^s_{atrial} \ , \tag{4.8}$$
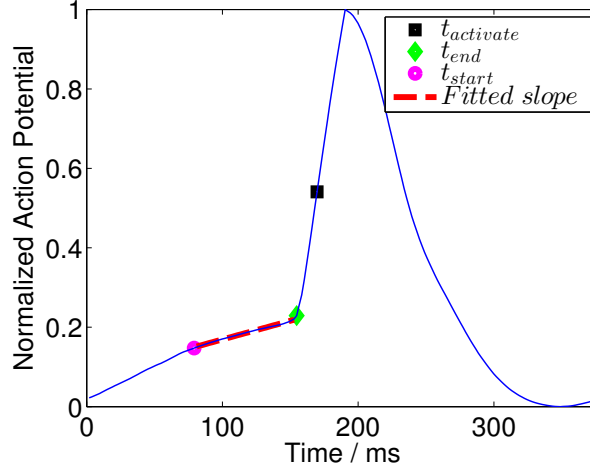
Figure 4.1: Measurements on single signal

by rejecting the null hypothesis H_0 that states the opposite:

$$H_0 : \mu^s_{nodal} \leqslant \mu^s_{atrial} \ . \tag{4.9}$$

4.2 Results

4.2.1 Examples of measurements

Figure 4.1 shows an example of found $t_{activation}$, t_{start} and t_{end} in a single typical nodal signal, which contains a positive diastolic depolarization slope (slow upward slope), a depolarization upstroke (fast upward slope) and a repolarization slope (downward slope). After measuring activation time and slope value on signal of every pixel, they can be plotted as activation map and slope map respectively as shown in Figure 4.2.

4.2.2 P-value of measurements

We calculated the activation time and slow slope value for both single pixel signals and the cluster average signal used in manually labeling process. For each single data, we did a t-test to compare both "AN vs NN" and "(AN + AC) vs (NN + NC)", to examine how are activation time and slope value related labeled atrial and nodal region. Results are shown in Figure 4.3 and Table 4.2.2.

(a) Activation map   (b) Slope map
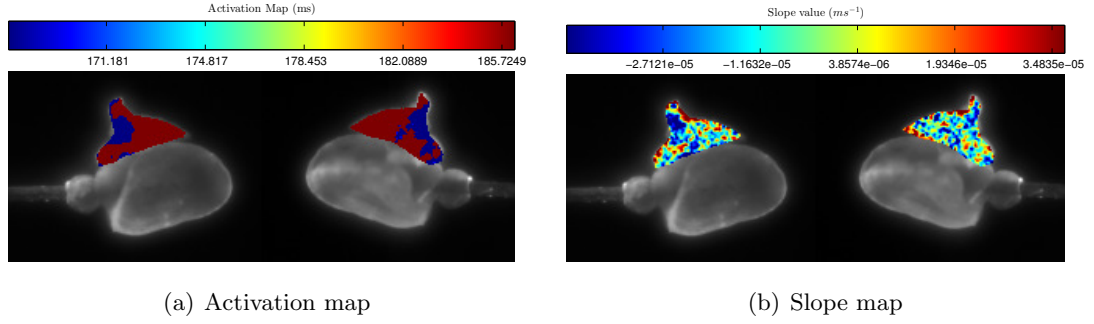
Figure 4.2: Example of Measurement maps: (a) activation map; (b) diastolic depolarization slope map
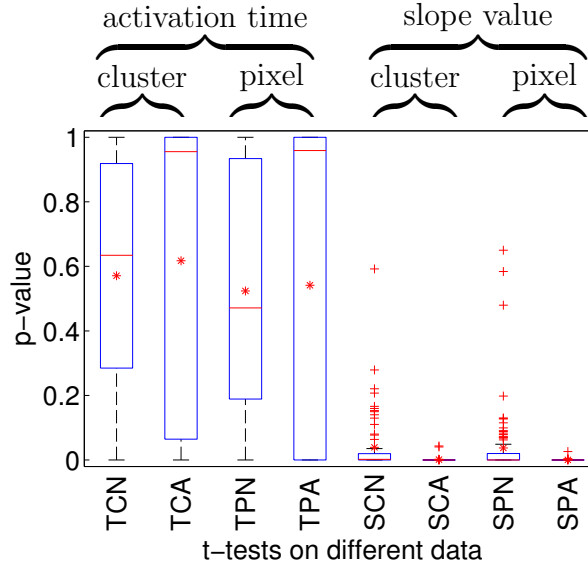
Figure 4.3: Boxplots of p-values: For the x-axis labels, T means we are comparing activation time; S means slope value; C means we do comparison on cluster average signals; P means that comparison are on pixel signals; N means we only use normal regions that is "AN vs NN"; A means we use all regions which include "(AN + AC) vs (NN + NC)". For the boxplot, the red horizontal bar is the median; red star is the mean; the edges of the box are the 25th and 75th percentiles; the upper and lower whiskers correspond to $\pm 2.7\sigma$ of the data; red plus signs are single outliers outside of the whiskers.

From Figure 4.3 and and Table 4.2.2, we can see that the behavior of slow slope value

		Percentage of $H_A = 1$		Average of p	
		Normal	All	Normal	All
Activation time	Pixel	0.2254	0.3978	0.6174	0.5413
	Cluster	0.0735	0.0968	0.5712	0.5236
Slope value	Pixel	1	1	0.0012	0.0004
	Cluster	0.8028	0.8065	0.0393	0.0373

Table 4.1: Statistics of activation time and slow slope value in terms of percentage of $H_A = 1$ and average p value

in most of the data satisfies our hypothesis in Inequality 4.8:

$$H_A : \mu^s_{nodal} > \mu^s_{atrial} \ ,$$

where μ^s_{nodal} and μ^s_{atrial} are the mean slope values in nodal and atrial regions, respectively. However, activation time does not behave as we expected. The potential reasons are discussed in Section 4.3

4.2.3   Data variability

In t-tests, slow slope value shows significant difference between nodal and atrial regions. We now examine the variability of slow slope values.

Figure 4.4 shows the boxplots of slope values in AN, NN, AN + AC and NN + NC respectively given the temperature is 27 $°C$. Although generally slope values in nodal regions are larger than those in atria regions, these values still shows large variance across different hearts and acquisitions even for the same temperature and same labeled regions.

4.2.4   Correlation between slope and temperature

Besides positive slope value, we expect to see the increase of slope value as temperature increases in nodal regions. Therefore, we expect to see this trend in nodal regions, especially on normal nodal regions (NN) and on the 'high quality' dataset D1 because of its large uncorrupted area and nodal area.

In Figure 4.5 and Figure 4.6, NN and NN + NC slope value boxplots are plotted with different acquisitions over temperatures. We can see increasing trends in more than half of the NN data and almost all of NN + NC data. Among plots of NN, D1 down, D1 up, D3 up and D5 up show clearly linear increasing trends. Here, 'up' and 'down' represents the temperature trend when acquiring the data. Among plots of NN + NC, all data except for
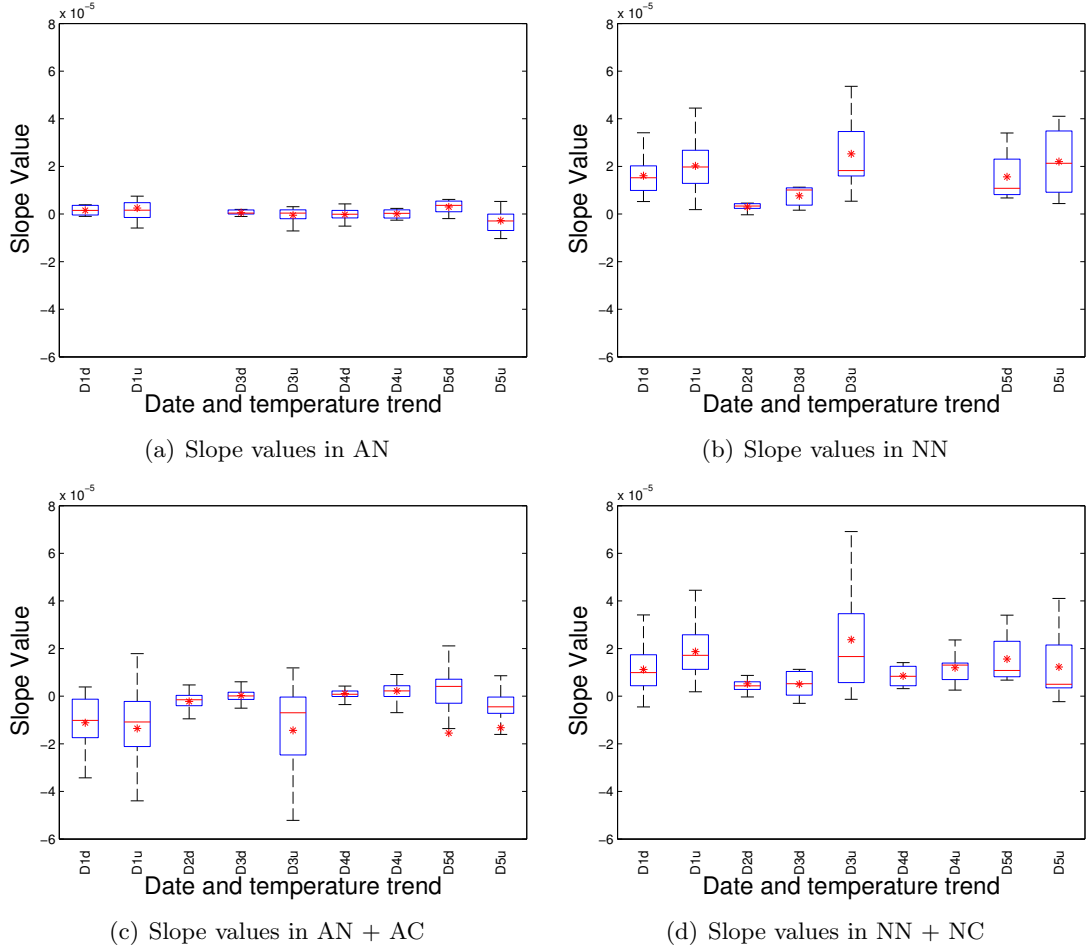
Figure 4.4: Example showing slope value variability given same temperature (27 $^\circ C$) and same manually labeled regions (for X axis labels, u means up and d means down): (a) Slope values of average cluster signals labeled as AN, (b) Slope values of average cluster signals labeled as AC, (c) Slope values of average cluster signals labeled as AN or AC, (d) Slope values of average cluster signals labeled as NN or NC

D3 down, D5 down and up show clearly linear increasing trends. The temperature relation is more obvious on NN + NC than NN on 'low quality' data, which is probably due to the number of labeled NN pixels in 'low quality' data is too small to give a stable expected trend. In addition, we plot the signals and slope values of representative NN data together in Figure 4.7.

(a) D1, down

(b) D1, up

(c) D2, down

(d) D3, down

(e) D3, up

(f) D4, down

(g) D5, down

(h) D5, up
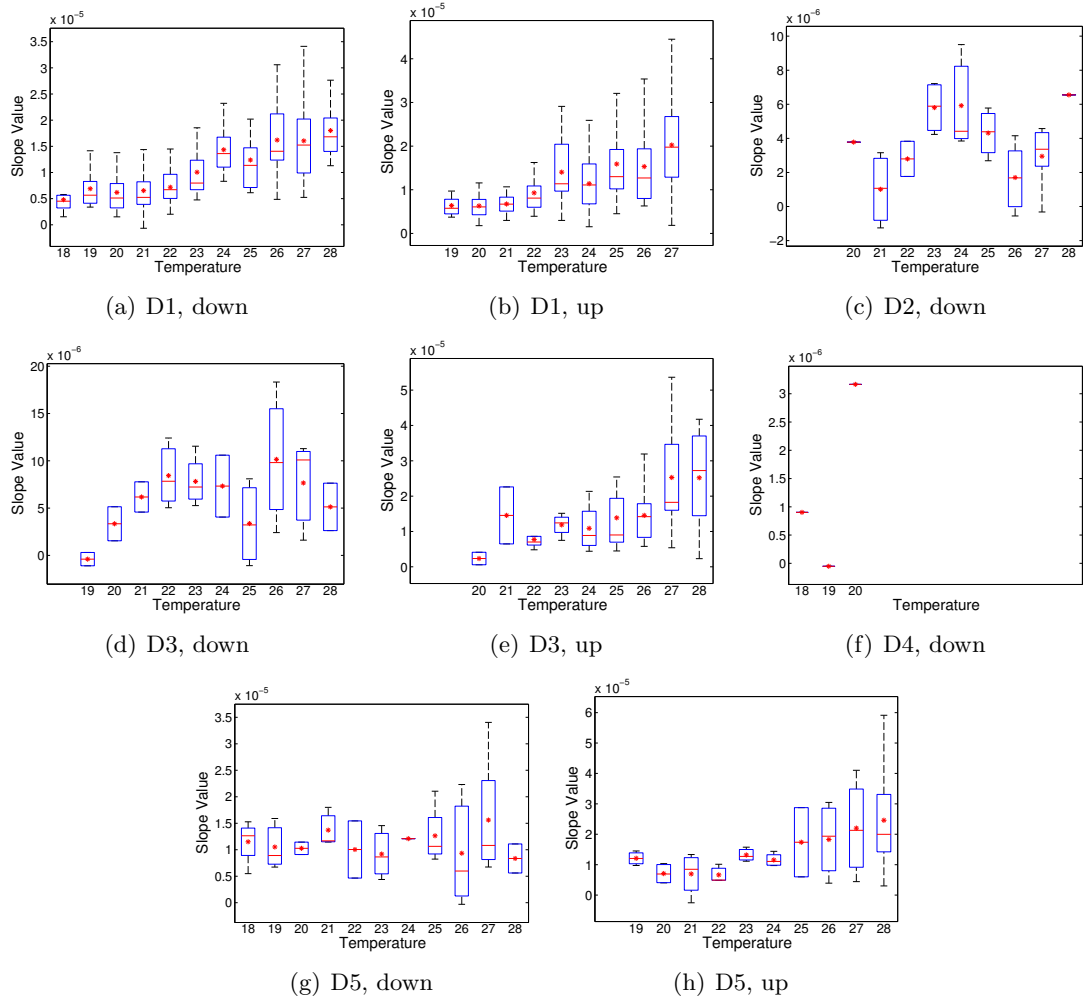
Figure 4.5: Boxplots of slope value of temperature on NN: 'up' and 'down' represents the temperature trend when acquiring the data. Empty column means there is no NN region labeled for that temperature. 'D4, up' has no NN region labeled for any single data, so it does not appear in this figure.
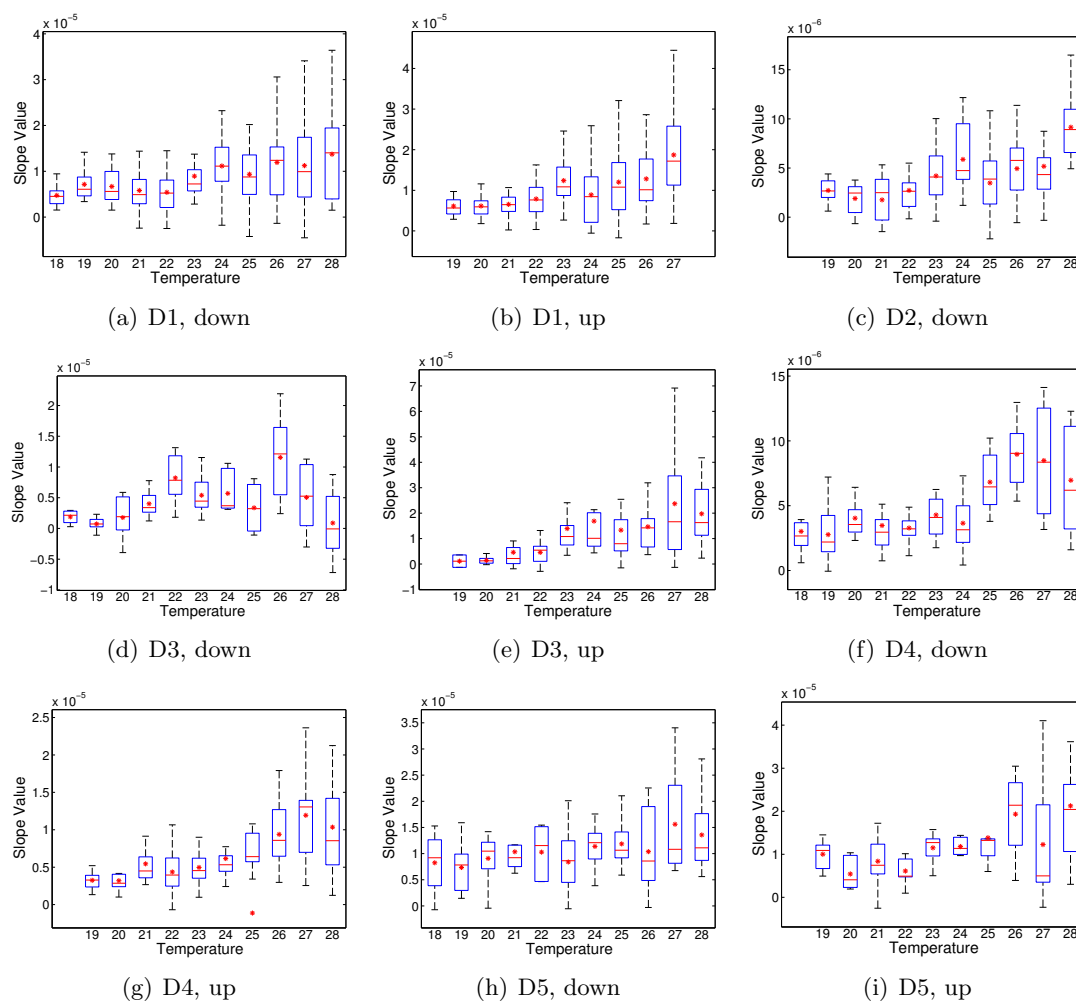
Figure 4.6: Boxplots of slope value of temperature on NN + NC: 'up' and 'down' represents the temperature trend when acquiring the data. Empty column means there is no NN + NC region labeled for that temperature.

(a) D1, down

(b) D1, up

(c) D3, up

(d) D5, up
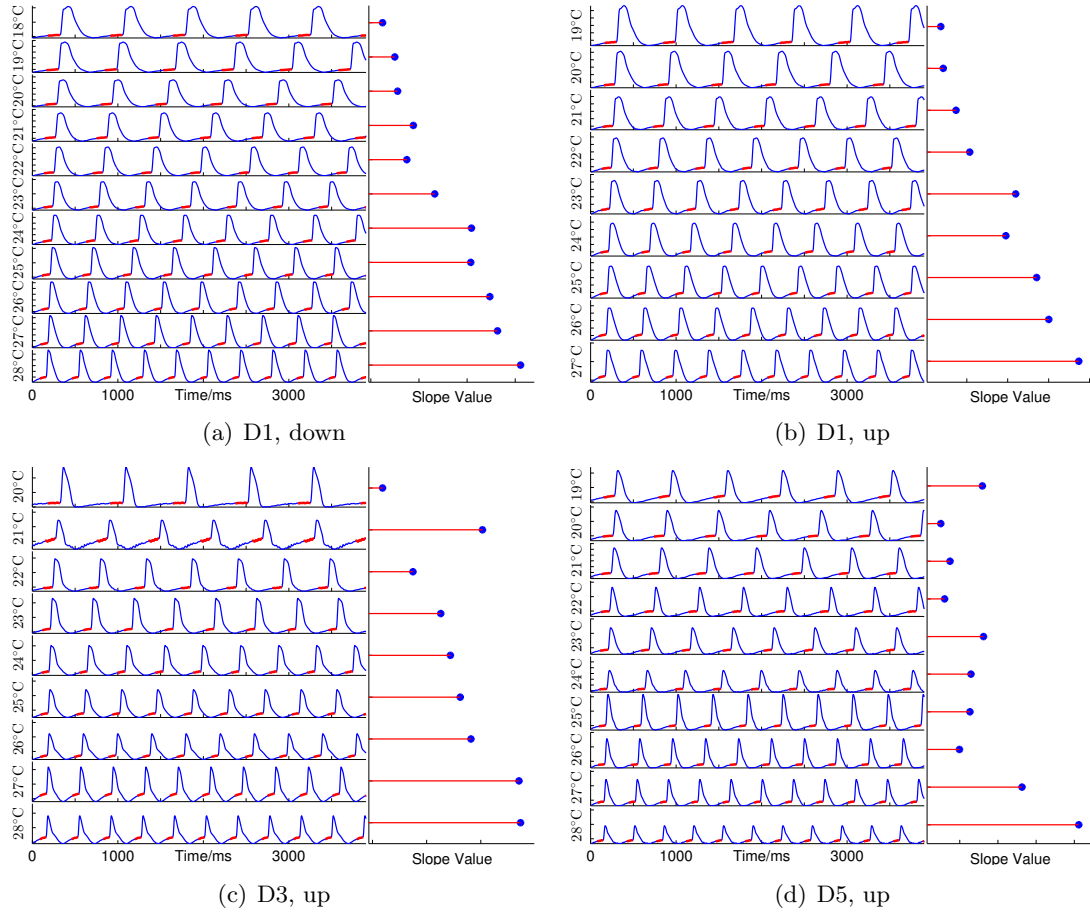
Figure 4.7: Signals and slope values over change of temperatures in the upward (up) or downward (down) direction: representative average normal nodal region signal is plotted over temperatures, and their measured slope is on the right side. The thick red curve segments in signals represent the intervals used for slope value calculation.

4.3 Conclusion and discussion

In this chapter, we measured and analyzed activation time and diastolic slope values. For activation time, labeled nodal region was typically found to possess later activation time than labeled atrial region. This is counter to what is expected from physiological properties. However, discrepancy between observation and expectation is not surprising. Activation time detection is sensitive to signal deformations, since it involves derivative calculation. Unfortunately the optical mapping signals suffers from: (1) low sampling rate also low spatial resolution; (2) signal mixing from multiple locations; (3) motion and other artifacts; (4) potential errors caused by unrobustness of measurement method. Therefore, the validity of the manual labeled data is not negated by this negative evidence.

On the other hand, the slope value is generally lower in atria region and higher in nodal region as we expected. On the 'high quality' and some 'low quality' data, slope value also increases with the increasing of temperature, which is an expected property of pacemaking cells. This is also a strong evidence supporting the validity of manual labeled data functionally. On data that is either corrupted heavily or having relatively small nodal region area, this slope value-temperature correlation is not obvious, which may result from multiple reasons, including motion artifacts, observation angle and individual physiological problems of zebrafish heart caused in the experiment preparation.

Chapter 5

Automatic Identification

In order to enable high-throughput analysis of optical mapping data in cardiac research, a fast pacemaker identification method is important for physiological researchers. In this chapter, we propose two novel identification methods, show examples and perform quantitative evaluation based on manually labeled data.

5.1 Methods

5.1.1 Thresholding based on slope value

We presented statistical analysis of diastolic depolarization slope value in Chapter 4. As the slope value showed significant difference in nodal region and atrial region, it is natural to use the slope value to distinguish between nodal and atrial region. Here, we used two different type of slopes: one was the slope value of the signal on every pixel; the other was the slope value of small clusters after a first round robust fuzzy c-means (RFCM) clustering as we performed in Section 3.1.1. In addition, since slope values are expected to be different under different temperatures, we acquired an optimal threshold that maximizes classification performance (evaluation method explained in Section 5.2.1) on manually labeled data at a given temperature. We then use this threshold to separate nodal and atrial regions in new, unlabeled data.

5.1.2   Two-step clustering

An expert identifying zebrafish heart OM signals does not simply 'calculate' the slope value in mind. The expert utilizes the entire signal waveform to make this decision. For automatic identification, we also would like to use the signal waveform to provide more information beyond a single slope measurement. We anticipate this approach to provide better performance than classification based on single slope value.

However, according to our observation, the shapes of the cardiac action potential waveforms of different individual zebrafish hearts show considerable variability under different experimental conditions and is corrupted by different type of artifacts. This variability makes logical algorithm design and supervised learning difficult to perform, because 'new' waveforms will be out of view of predefined logic or encountered data. To make the algorithm robust to these sources of variability, we propose to use a two-step unsupervised clustering framework to group pixels into small number of regions, which are easy for experts to label in very short time later.

For the first round clustering, we used the RFCM method introduced in Section 3.1.1 to separate atrium into many different spatially connected small regions. The average signals for these clusters were then calculated. This step is equivalent to further smoothing in signal vector space with spatial consideration. The calculated average signals are expected to be less noisy than the original single pixel signal.

The second stage clustering was done on this average signals. In this stage of clustering, we grouped signals with similar physiological properties. Clustering the signals in their original vector space was expected to be insufficient for this purpose. This is due to the fact that the distance in original vector space does not represent the physiological similarity of signals from different cardiac cells. Instead, having similar rates of change and reaching 'hills' and 'valleys' in similar time, in other words, having similar 'trends', is likely correlated to similarities in physiology. At least, we expect this to be true for separating nodal signal from atrial signal based on the observation of our labeled signals.

To represent the descriptive 'trends' by quantitative values, a novel 'discretization of derivative' (DoD) procedure was proposed. First, the derivative of the original signal is calculated. Derivative is a good representation of 'trends'. However, sometimes signals sharing the same trend can be quite different in the magnitude of derivative values. This makes clustering in the derivative vector space also difficult. To cope with this problem, a

(a) Original   (b) Derivative   (c) After discretization   (d) After mode filter
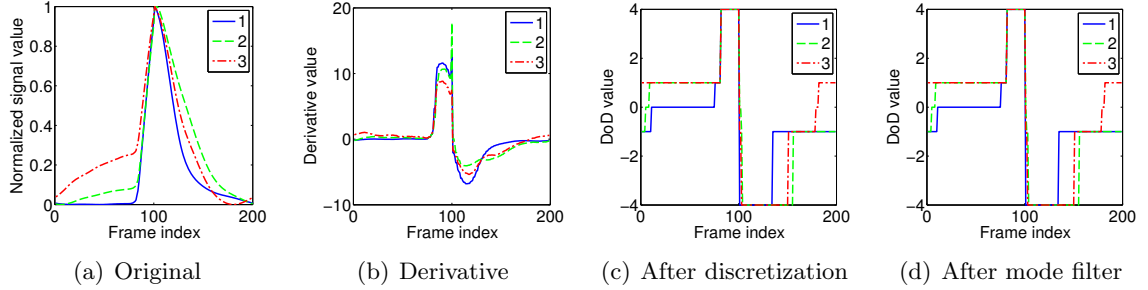
Figure 5.1: Procedure of performing discretization of derivative: (a) Original average signals from different clusters, (b) derivative of these 3 signals, (c) discretized derivative, (d) After mode filter is performed. In (d), 2 and 3 are more similar compared to in the original vector space.

novel discretization step is proposed. In this step, the numerical derivatives are thresholded to different preset values, which represent 'fast increasing', 'slow increasing', 'flat', 'slow decreasing' and 'fast decreasing'. Specifically, we have 3 parameters: slow threshold T_{slow}, fast threshold T_{fast} and fast value V_{fast}. For any derivative value x, we want to calculate a discretized y:

$$y = \begin{cases} -V_{fast} & \text{if } x < -T_{slow}, \\ -1 & \text{if } -T_{fast} \leqslant x < -T_{slow}, \\ 0 & \text{if } -T_{slow} \leqslant x < T_{slow}, \\ 1 & \text{if } T_{slow} \leqslant x < T_{fast}, \\ V_{fast} & \text{if } x \geqslant T_{fast}. \end{cases} \tag{5.1}$$

Here, 0 is the 'flat'; 1 and -1 represent 'slow increasing' and 'slow decreasing' respectively; V_{fast} and $-V_{fast}$ represent 'fast increasing' and 'fast decreasing' respectively. Finally, mode filter was used to smooth the trends acquired from DoD procedure.

The DoD process is illustrated in Figure 5.1. The Euclidean distance between original signal 1 and 2 d_{12} is 1.56, and distance between signal 2 and 3 d_{23} is 1.59, $d_{12} < d_{23}$. However, physiologically 2 and 3 are more similar, because they share the same 'trends'. After the DoD transformation $d_{12} = 17 > 12 = d_{23}$, which means it reflects the actual similarity better than in the original signal space.

Next, k-means clustering with a smaller number of classes was performed on the DoD transformed representation to give the final grouping of functional time series. After the

RFCM-DoD-Kmeans signal grouping process, final grouping results were presented to physiologists for them to make fast decisions.

Besides this final manual decision step, we also proposed an empirical automatic method to select nodal type group from the grouped signals. First, we chose the group with largest average diastolic slope value as nodal region. Second, for other groups, if their average slope values is larger than one third of the largest average slope value, they are labeled as nodal region too. This facilitates the potential nodal signals to be identified automatically.

5.2   Experiments

5.2.1   Evaluation criteria

For evaluation of different identification methods, we use the Dice similarity coefficient (DSC) [11] between automatically identified region R_{auto} and manually labeled region R_{manual}:

$$s = \frac{2|R_{auto} \cap R_{manual}|}{|R_{auto}| + |R_{manual}|} \ , \tag{5.2}$$

where $|R|$ is the area of region. R and $|R_{auto} \cap R_{manual}|$ represent the area of intersection of these two regions. DSC was directly used for evaluating the performance of identification based on slow slope value and automatic identification method based on 2-step clustering.

Besides the automatic identification methods, we also want a measurement for evaluating the 2-step clustering method itself to know whether the bottleneck is 2-step clustering or following nodal region identification step. In this case, DSC is not directly usable, since clustering outputs several groups without nodal or atrial labels.

Therefore, we defined the 'best possible Dice similarity coefficient' (BPDSC) to measure the performance of grouping methods. To calculate BPDSC, we first ranked a the final clustered regions according to their individual DSC with the labeled nodal region. Then we combined the 2nd region with the 1st region, and see if they had higher DSC. If it was higher, we would combine the 2nd region with the 1st region. Next, we added the 3rd one into it and compare DSC. This process stops when adding additional region decreases the DSC. When the algorithm stops, the resulting DSC is taken as the BPDSC.

5.2.2   Experiments

In order to evaluate the proposed methods, we conducted a series of experiments.

Slope value thresholding

For slope value thresholding, we tested it on single pixel signals, signals from RFCM clusters used for manually labeling (referred as 'original clusters' below) and newly generated RFCM clusters (referred as 'other clusters' below). It is expected that the performance on 'other clusters' is lower than on 'original clusters', since the evaluation groundtruth is based on the 'original clusters'. Therefore, performance on 'other clusters' is a more appropriate measure of the true performance.

In terms of selection of threshold, we do 2 different experiments. The 1st experiment is to iterate over all possible thresholds and pick the one with highest DSC. We refer to this threshold as 'best threshold'. The DSC given 'best threshold' reflects the capability of thresholding method. The 2nd experiment is to use the same temperature data from other hearts as training data. The threshold is acquired by iterating over all possible thresholds on the training data, and pick the one with highest DSC and then use it for the data that we would like to identify nodal region. This is one way of predicting the threshold given labeled data.

Clustering based methods

The proposed RFCM-DoD-Kmeans method involves parameter selection. We fixed the number of final K-means clustering to be 6, so that it allows physiologists quickly choose 'pacemaker group' from the grouping result and the extra 1 class besides 5 manually labeling classes also provides robustness when intraclass variation is large.

Besides number of clusters, choice of distance measure affects the K-means algorithm. We considered 4 distances, Euclidean, cityblock, cosine and correlation. The proposed DoD transformation has 4 parameters to choose according to Section 5.1.2. They are slow threshold T_{slow}, fast threshold T_{fast}, fast value V_{fast} and mode filter w_{mode}. We considered 3 different values for each of them: for T_{slow}, we used 0.3, 0.5, 0.7; for T_{fast}, we used 2, 2.5, 4; for V_{fast}, we used 1.5, 2, 4 and for w_{mode}, we used 5, 7, 11.

For RFCM-DoD-Kmeans, we did experiments on 324 combinations of these 5 parameters. At this stage, we used the same RFCM clusters as those used for manual labeling. For K-means clustering, we run it for 10 times and choose the one with 'smallest within-cluster sums of point to centroid distances'. For every experiment, the BPDSC was calculated to represent performance. Under this context, the BPDSC S was a multiple dimensional

matrix and a function of multiple parameters: $S(T_{slow}, T_{fast}, V_{fast}, w_{mode}, dist)$.

Next, we selected the best parameters on all data, high quality data and low quality data by selecting maximum value from their BPDSC function:

$$p_i^{best} = \arg\max_{p_i} S_i(p_i), \ i \in \{all, good, bad\} \ , \tag{5.3}$$

where $p_i \in T_{slow}, T_{fast}, V_{fast}, w_{mode}, dist$ represents the parameters. In the following sections, p_i^{best} is referred as best parameters.

Besides best parameters, we also acquired a set of 'best' parameters in an average sense. For example, to know the best value of T_{slow}, we did summation on every other parameters to acquire the function $S(T_{slow})$ and choose the T_{slow} to maximize it. We refer them as 'best average parameters':

$$p_{ij}^{best\ average} = \arg\max_{p_{ij}} S_{ij}(p_{ij}) \ ,$$
$$i \in \{all, good, bad\}, \ j \in \{T_{slow}, T_{fast}, V_{fast}, w_{mode}, dist\} \ , \tag{5.4}$$

$$S_{ij}(p_{ij}) = \sum_{k_1} \cdots \sum_{k_n} S_i(p_{ik_1}, \cdots, p_{ij}, \cdots, p_{ik_n}) \ ,$$
$$k_1, \cdots, k_n \in \{T_{slow}, T_{fast}, V_{fast}, w_{mode}, dist\} \backslash j \ . \tag{5.5}$$

After acquiring these 'best' and 'best average' parameters, we used them to test the performance of entire RFCM-DoD-Kmeans pipeline. For any single data acquisition, we ran RFCM-DoD-Kmeans 10 times and use the average BPDSC as the performance measurement. Here, the generated RFCM clusters are different from those used for manually labeling, so they are the 'other clusters' defined earlier in this section.

For evaluating proposed RFCM-DoD-Kmeans method, we also performed experiments on 2 other simpler clustering based grouping methods: (1) directly doing K-means clustering on signals for single pixels; (2) doing K-means right after RFCM without doing DoD. In what follows, we refer to them as direct K-means and RFCM-Kmeans.

For these 2 methods, the number of final clusters was also set to be 6. For each run of Direct K-means, we chose the one out of 10 with smallest within-cluster sums of point to centroid distances. This process was run 10 times on each of the data, the average BPDSC indicates the performance.

We run RFCM-Kmeans on both 'original' RFCM clusters and 'other' RFCM clusters. They were experimented in the same way as the proposed RFCM-DoD-Kmeans pipeline, except for that DoD is not involved.

Following each run of any clustering method, the nodal regions were selected using the empirical method proposed in Section 5.1.2.

5.3 Results

In this section, we show examples of automatic identification of using different methods and compare different identification methods quantitatively with manually labeled data.

5.3.1 Examples of different detection result

Slope value based identification

Figure 5.2 shows examples of using slope value to detect pacemaking region. Pacemaking region identification based on thresholding of pixel signals is shown in Figure 5.2 (c) and (d). In this example the DSC between manually and automatically labeled nodal region is 0.75. Identification based on thresholding on cluster signals is shown in Figure 5.2 (e) and (f) and the DSC was calculated to be 0.92.

Clustering based identification

Figure 5.3 shows examples of nodal region identification based on using clustering methods. Example results and corresponding BPDSC's from direct Kmeans, RFCM-Kmeans and RFCM-DoD-Kmeans are presented. Their corresponding BPDSC's are 0.72, 0.77 and 0.87 respectively. Figure 5.4 shows more examples of RFCM-DoD-Kmeans results.

5.3.2 Quantitative performance analysis

Slope value thresholding

Figure 5.5 shows the boxplots of DSCs from experiments of slope value based thresholding. The thresholding was done on pixels, 'original clusters' and 'other clusters'. Performances of all data, high quality data and low quality data are plotted separately. Figure 5.5 (a) shows the DSCs of best thresholds. Figure 5.5 (b) shows the DSCs of trained thresholds using different hearts under same temperature.

(a) Manually labeled atrium regions

(b) Manually labeled nodal region

(c) Pixel level slope value map

(d) Nodal regions by thresholding on pixels

(e) Cluster level slope value map

(f) Nodal regions by thresholding on clusters
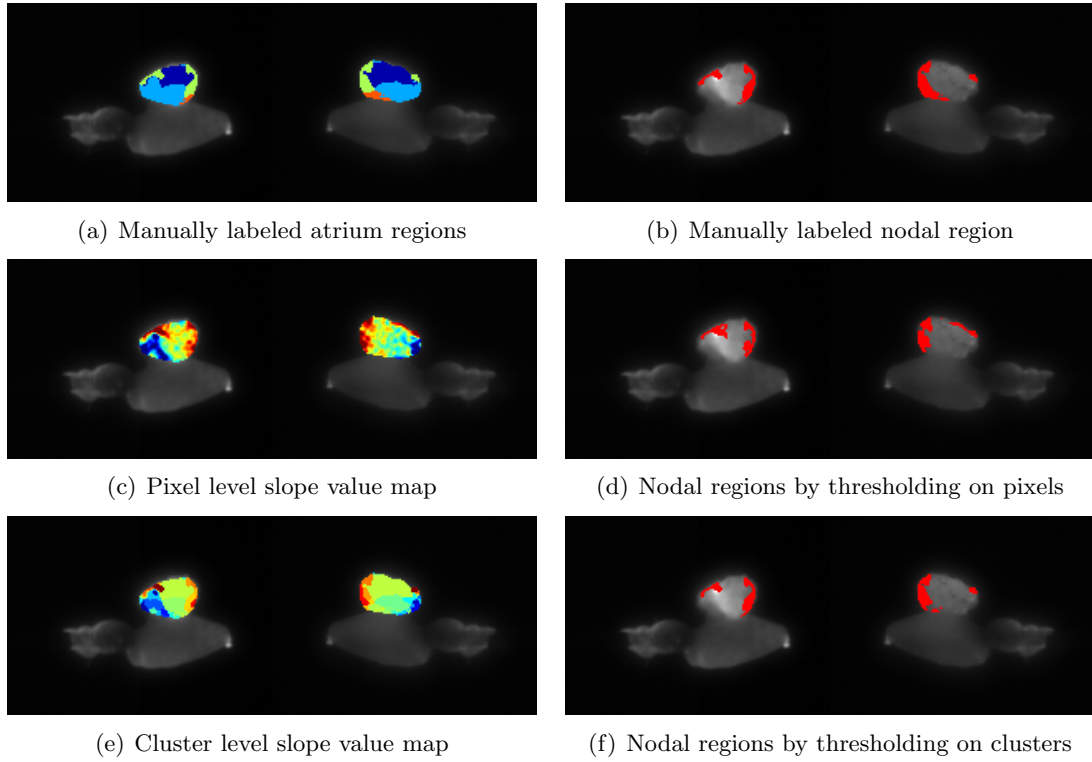
Figure 5.2: Examples of nodal region identification based on slope value: (a) manually labeled atrium with color coded regions, (b) manually labeled nodal region, (c) slope value map calculated on every pixel, (d) identified nodal regions by thresholding on pixels (DSC 0.75), (e) slope value map calculated on average signal of every cluster, (f) identified nodal regions by thresholding on clusters (DSC 0.92)

Clustering based methods

Figure 5.6 shows the the BPDSC boxplots of clustering methods. Figure 5.6 (a) and (b) show the boxplots of running the proposed 2-step clustering result with best parameters and best average parameters. In (a), the DoD-Kmeans process is done on the RFCM clusters with manual labels, namely 'original clusters'. In (b), the DoD-Kmeans process is done on the RFCM clusters without manual labels, namely 'original clusters'. In (c), mean BPDSC's of under different parameters and on different quality data are plotted. Figure 5.6 (d), (e), (f) contain the boxplots of direct K-means and RFCM-Kmeans using different distance measures for comparison.

Figure 5.7 shows the DSC boxplots of clustering method based automatic nodal region

(a) Direct K-means regions/signals

(b) Direct K-means nodal region

(c) RFCM-Kmeans regions/signals

(d) RFCM-Kmeans nodal region

(e) RFCM-DoD-Kmeans regions/signals

(f) RFCM-DoD-Kmeans nodal region
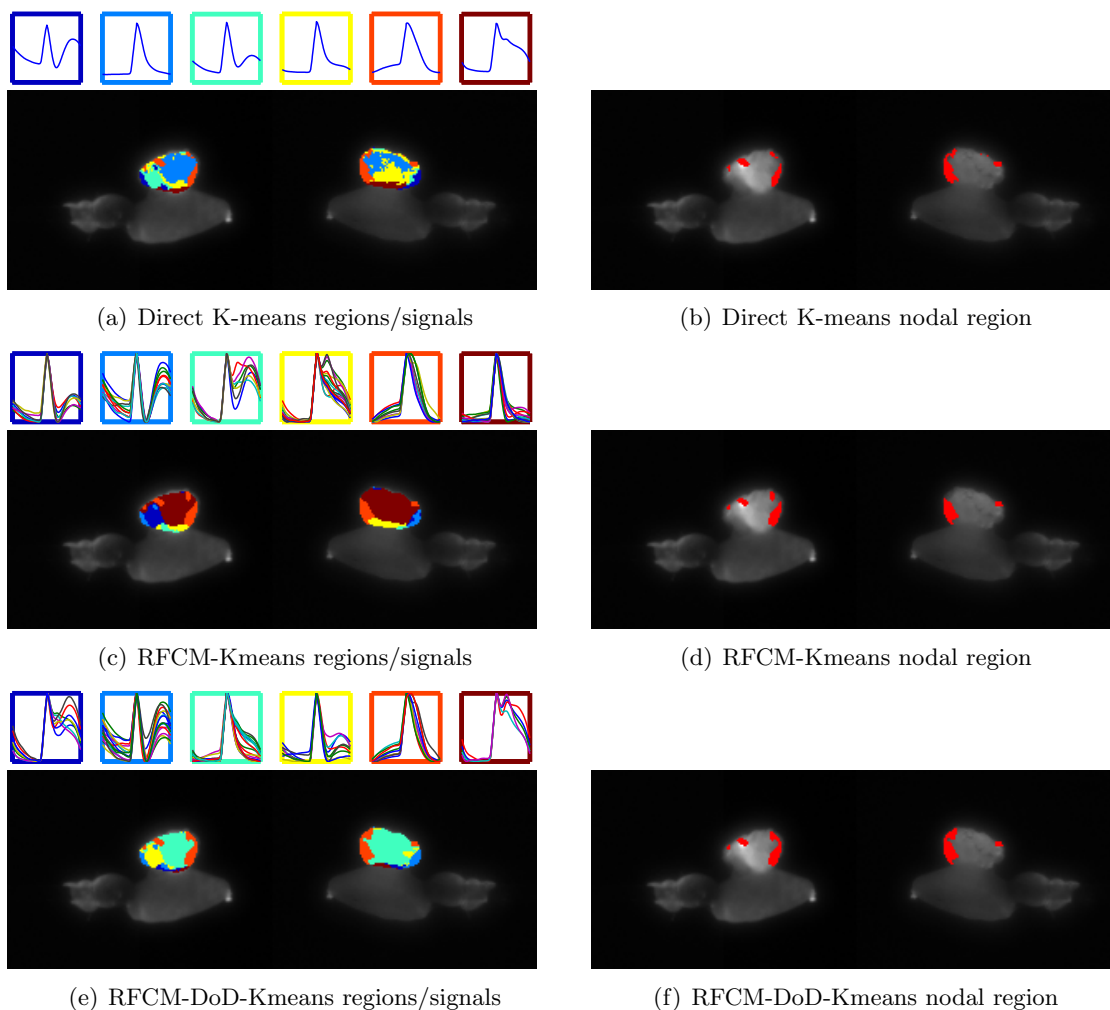
Figure 5.3: Examples of nodal region identification based on clustering (nodal regions/signals are labeled as bright red): (a) Direct K-means clustering color coded regions and signals, (b) Direct K-means identified nodal region (DSC 0.72), (c) RFCM-Kmeans clustering color coded regions and grouped signals, (d) RFCM-Kmeans identified nodal region (DSC 0.78), (e) RFCM-DoD-Kmeans color coded regions and grouped signals, (f) RFCM-DoD-Kmeans identified nodal region (DSC 0.87)
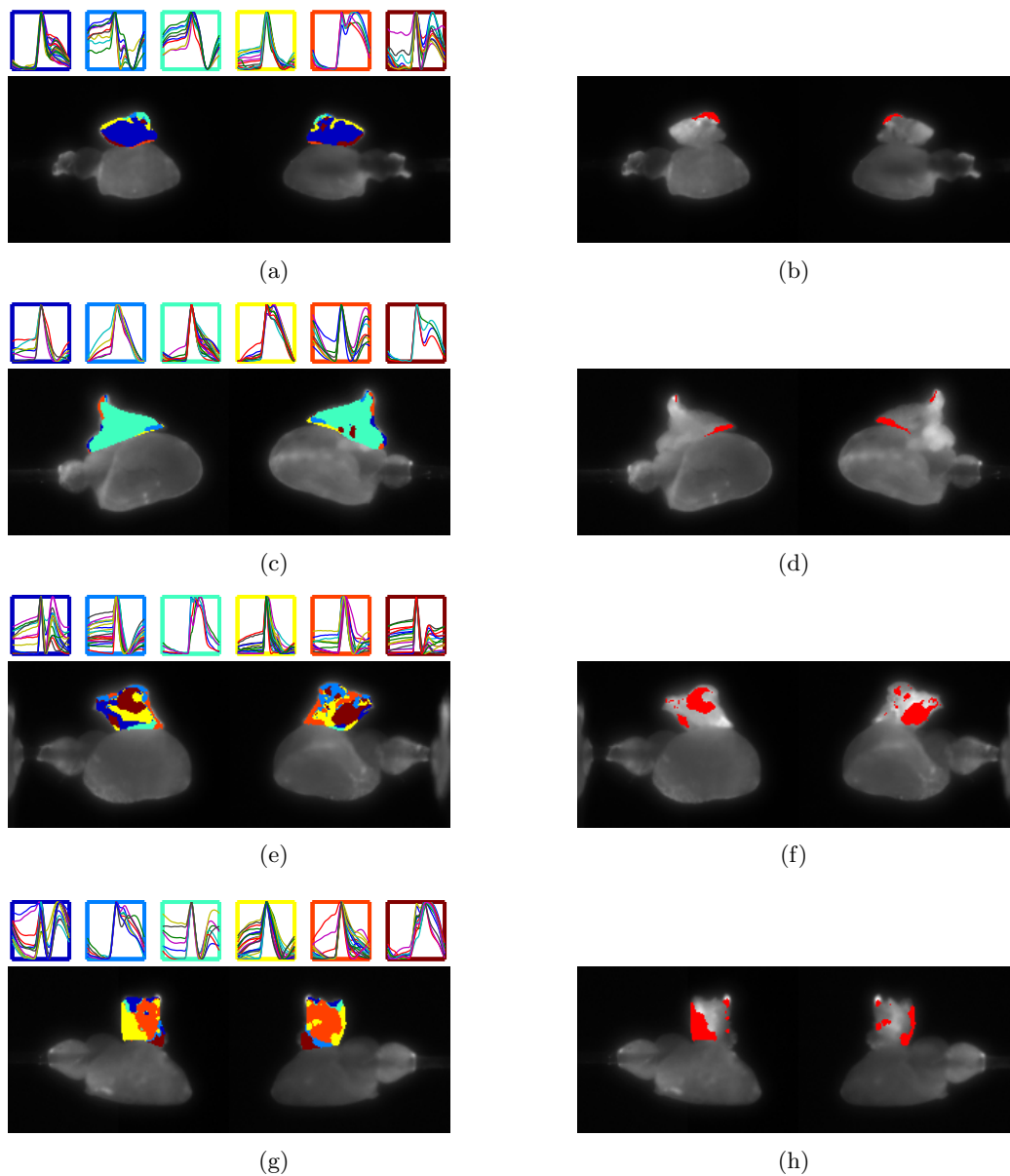
Figure 5.4: More examples of nodal region identification based on RFCM-DoD-Kmeans: (a), (c), (e) and (g) are the grouping results. (b), (d), (f) and (h) are the corresponding results after nodal group selection. The DSC is acquired by comparing the automatic identification with the manual result shown in Figure 3.4 and Figure 3.3. (b) has a DSC of 0.51, (d) has a DSC of 0.58, (f) has a DSC of 0.57, and (h) has a DSC of 0.65.

(a) Best thresholds

(b) Trained thresholds
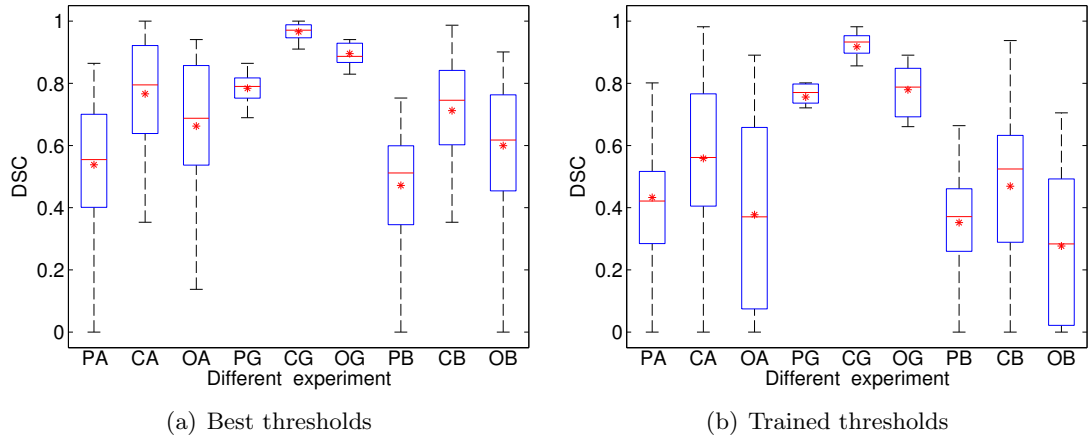
Figure 5.5: Boxplot showing DSC of nodal region detection using slope value: Here, P means pixel, C means 'original clusters', O means 'other clusters', A means all data, G means good data and B means bad data. For example, PA means the thresholding of single pixels on good data. (a) The thresholds used is acquired by searching all possible thresholds on the test data itself to find the threshold with the highest DSC . This represents the highest DSC we can achieve by thresholding on each data. (b) The thresholds used is acquired by searching all possible thresholds on training data with same temperature as test data. This represents the highest DSC we can achieve on unseen data given labeled training data.

identification. Each of the subfigures has same meaning as those shown in Figure 5.6.

Comparison between different methods

Table 5.1 compares the DSCs and BPDSC's of different methods. Slope value based thresholding on clusters outperforms other method given the best threshold. This again validates the fact that slope value is an important factor that the expert considers while labeling. However, slope thresholding given trained threshold has poor performance, which means the slope value are not effective to identify nodal region in new data.

For clustering based methods, average performance of RFCM-DoD-Kmeans methods given different parameters are similar to direct Kmeans and RFCM-Kmeans methods. With selected parameters, the RFCM-DoD-Kmeans outperforms direct Kmeans and RFCM-Kmeans significantly in high quality data. But the performance in all data and low quality data is still similar. This is probably due to the fact that DoD transformation is not capable of dealing with various motion corrupted signals. The high BPDSC of RFCM-DoD-Kmeans represents its identification power when the final group selection is done by a manual rater

(a) RFCM-DoD-Kmeans on 'original clusters'

(b) RFCM-DoD-Kmeans on 'other clusters'

(c) Mean BPDSC with different parameters

(d) Direct Kmeans with different distance

(e) RFCM-Kmeans on 'original clusters'

(f) RFCM-Kmeans on 'other clusters'

Figure 5.6: BPDSC boxplots of pacemaking region detection using clustering methods: for X axis tick labels, the first letter A means on all datasets, G means on good datasets, B means on bad datasets; for (a), (b), the second letter B means best parameters, and A means best average parameters; for (d), (e), (f) the lower case letters represent distances used in Kmeans, sqe is Euclidean, cit is city block, cos is cosine, cor is correlation. (a) BPDSC's on individual recordings processed by RFCM-DoD-Kmeans on 'original clusters'; (b) BPDSC's on individual recordings processed by RFCM-DoD-Kmeans on 'other clusters'; (c) Mean BPDSC of all data with different parameters; (d) BPDSC's on individual recordings processed by direct Kmeans with different distance; (e) BPDSC's on individual recordings processed by RFCM-Kmeans on 'original clusters'; (f) BPDSC's on individual recordings processed by RFCM-Kmeans on 'other clusters'.

(a) RFCM-DoD-Kmeans on 'original clusters'

(b) RFCM-DoD-Kmeans on 'other clusters'

(c) Mean BPDSC with different parameters

(d) Direct Kmeans with different distance

(e) RFCM-Kmeans on 'original clusters'
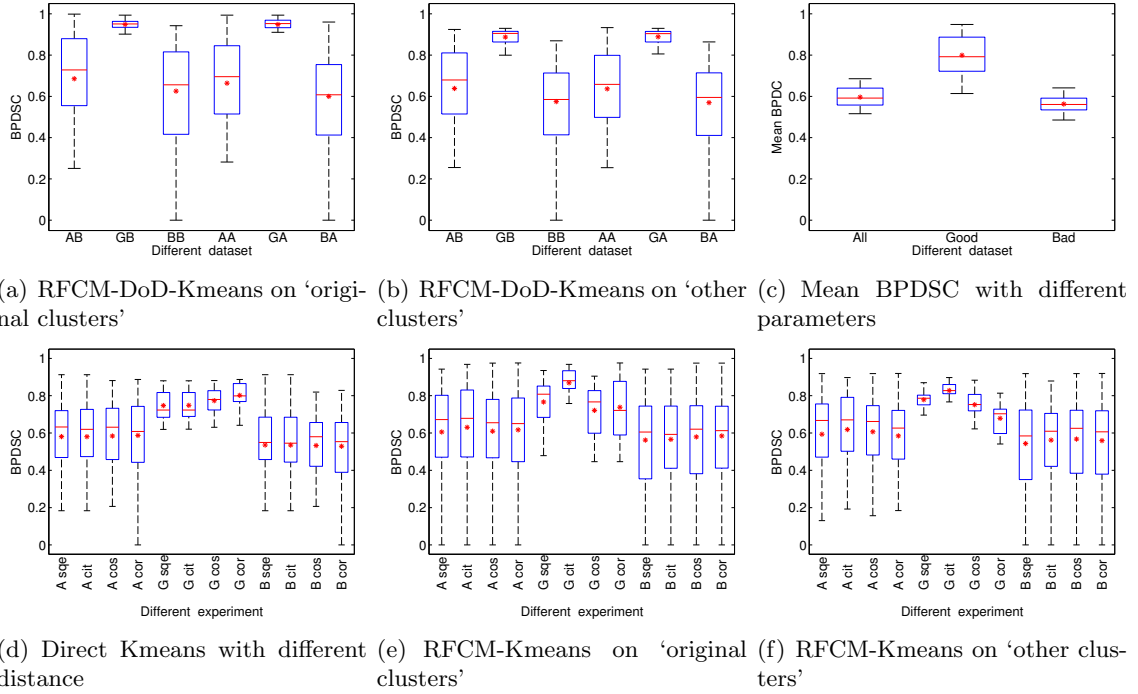
(f) RFCM-Kmeans on 'other clusters'
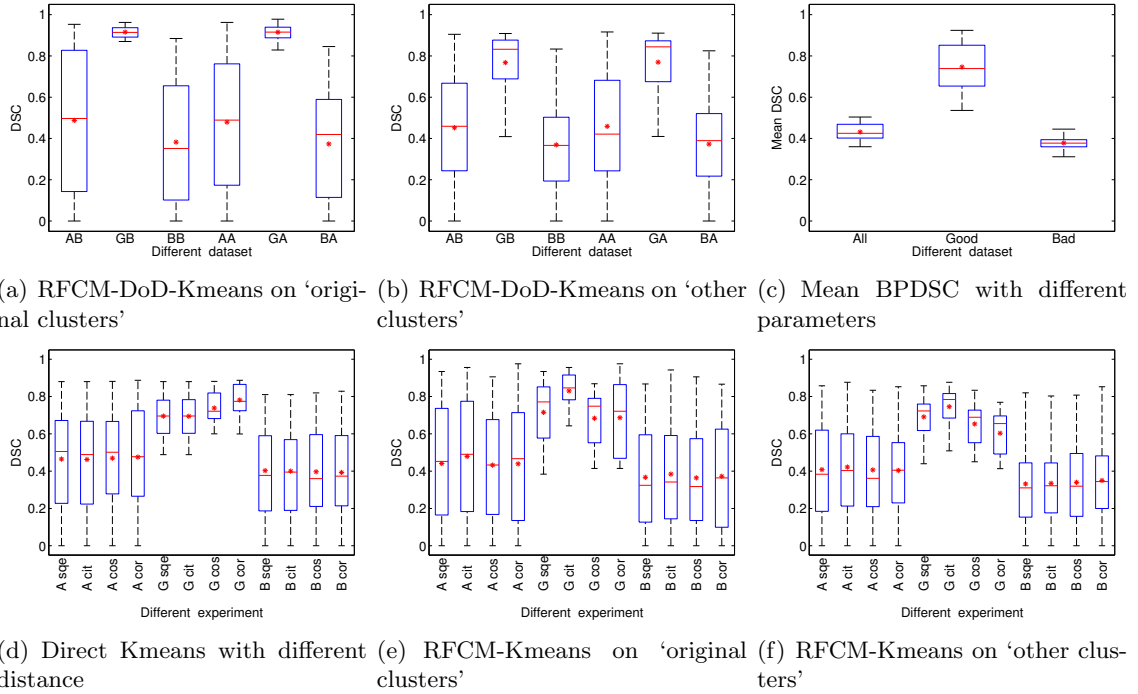
Figure 5.7: DSC boxplots of fully automated pacemaking region detection using clustering methods: for X axis tick labels, the first letter A means on all datasets, G means on good datasets, B means on bad datasets; for (a), (b), the second letter B means best parameters, and A means best average parameters; for (d), (e), (f) the lower case letters represent distances used in Kmeans, sqe is Euclidean, cit is city block, cos is cosine, cor is correlation. (a) DSCs on RFCM-DoD-Kmeans based automatic identification on 'original clusters'; (b) DSCs on RFCM-DoD-Kmeans based automatic identification on 'other clusters'; (c) Mean DSC of all data with different parameters; (d) DSCs on direct Kmeans based automatic identification with different distance; (e) DSCs on RFCM-Kmeans based automatic identification on 'original clusters'; (f) DSCs on RFCM-Kmeans based automatic identification on 'other clusters'.

			All	Good	Bad
Slope value based	Best threshold	pixel	0.54	0.78	0.47
		cluster	0.66	0.90	0.60
	Trained threshold	pixel	0.43	0.76	0.35
		cluster	0.38	0.78	0.28
Clustering BPDSC	Direct K-means (distance with best performance)		0.59	0.80	0.54
	RFCM-Kmeans (distance with best performance)		0.62	0.83	0.57
	RFCM-DoD-Kmeans	Average over parameters	0.60	0.80	0.56
		Best parameters	0.64	0.89	0.57
		Best average parameters	0.64	0.89	0.57
Automatic nodal region identification	Direct K-means (distance with best performance)		0.48	0.78	0.40
	RFCM-Kmeans (distance with best performance)		0.42	0.75	0.35
	RFCM-DoD-Kmeans	Average over parameters	0.43	0.75	0.38
		Best parameters	0.45	0.77	0.37
		Best average parameters	0.46	0.77	0.37

Table 5.1: DSC comparison of different experimented methods: In terms of evaluation involving cluster signals, 'other clusters' is used as it is more representative for the true performance. Multiple experiments has been done on direct K-means and RFCM-Kmeans with the different distance measures. Only the results with the best distance measure is presented in this table.

or a suitable automatic group selection is used.

Additionally in Table 5.1, the average DSC of the fully automated process is lower than the BPDSC of RFCM-DoD-Kmeans, which indicates that our current auto group selection method failed to identify some of the separated nodal groups. Therefore, the combination of RFCM-DoD-Kmeans and expert selection is currently the best method. As we are performing the RFCM-DoD-Kmeans and final expert selection for now, we will develop better auto selection method in the future.

5.4 Conclusion and discussion

This chapter described different pacemaking region identification methods including slope value based thresholding, direct K-means clustering, RFCM-Kmeans and RFCM-DoD-Kmeans. To evaluate performance of signal clustering algorithms in pacemaker detection, we introduced the 'best possible Dice similarity coefficient' (BPDSC). We acquired the BPDSCs of clustering algorithms and Dice similarity coefficients of slope value based thresholding by comparing their results with manually labeled data.

Slope value is a good feature for separating the nodal region and atrial region, if we can find a 'best' threshold. However, it does not work well when we tried to acquire this threshold based on other hearts under same temperature.

Among clustering based methods, RFCM-DoD-Kmeans method outperforms direct K-means and RFCM-Kmeans significantly in high quality datasets given good set of parameters. But the differences of them in all datasets or low quality datasets are only marginal. This is probably due to the large variance of cardiac signal in 'low quality' datasets caused by artifacts. Due to the limited amount of good data, we have not tested the robustness of good parameters over different good datasets.

Additionally, we made an attempt to fully automate the clustering based identification process, although its final result is not satisfactory. A better auto selection method remains a topic for future research.

Chapter 6

Conclusion

6.1 Conclusion

This thesis has introduced a novel preprocessing pipeline for high throughput analysis of zebrafish optical mapping data. The first step in the pipeline which involves spatial-temporal averaging, drifting removal and scaling and cycle averaging. This successfully converted the low SNR original OM data to a set of clear cardiac signals. Subsequently, these enhanced signals were clustered to spatially connected patches using robust fuzzy c-means. Manual labeling was performed on these patches. Next, activation time and resting slope value were measured for signal in each pixel and statistically analyzed with the manual labels. Finally, a novel automatic pacemaker identification method was proposed. Slope value-based method utilized the fact that pacemaking cells have positive diastolic depolarization slope and used thresholding on slope value to detect pacemaking regions. A novel RFCM-DoD-Kmeans approach was used to divide the heart into several physiologically meaningful regions by performing clustering on a transformed signal space.

Manual labeling was validated from three aspects: (1) structurally, regions with same labels connected pieces; (2) functionally, labeled nodal regions have higher resting slope value than labeled atrial regions and increasing slope over temperature is observed in labeled nodal region; (3) part of the labeled pacemaking region coincide with the locations reported in literature.

Automatic identification methods were evaluated based on manually labeled data and performance of the slope value based method evaluated. A 'best possible Dice similarity coefficient' (BPDSC) was proposed to measure the performance of clustering based methods.

The proposed novel RFCM-DoD-Kmeans method was compared with slope value based thresholding, direct K-means and RFCM-Kmeans. On 'high quality' data, RFCM-DoD-Kmeans gives satisfactory identification performance and outperforms other methods given good set of parameters.

6.2 Primary contributions

The primary contributions of this thesis include:

- Introduced a novel cycle averaging method into preprocessing pipeline.

- Used RFCM as a novel pre-clustering step to reduce work load of manual labeling in zebrafish OM data.

- Implemented measurement methods and performed statistical analysis on measured values.

- Proposed a novel unsupervised signal grouping pipeline for pacemaking region identification.

These methods can be termed as "computational instrumentation" and they will find utility in high throughput quantitative analysis of optical mapping signals.

6.3 Future work

As a first attempt on pacemaking region identification from optical mapping data with a signal analysis perspective, concepts and methods introduced in this thesis likely have many possible future directions and applications.

6.3.1 Preprocessing

To reduce the mixing of different type of signals, methods that involves less spatial and temporal averaging should be considered. Longer recording will increase the number of cycles to be averaged and is therefore helpful for reducing spatial and temporal averaging. Also, in preprocessing stage, signal averaging can be performed in a 'bilateral' fashion. This means that averaging will happen if and only if they are close in both spatial and functional

space, which is the space spanned by the signal vector. This is similar to the well known bilateral filter [46] in image processing.

In the 'low quality' data, signals are mostly corrupted by motion artifacts although chemical excitation-contraction uncoupler has already been used. Therefore, quantification of motion artifacts is helpful. If we know the temporal interval that is likely corrupted by motion, we can avoid making measurements in that interval and discard that part when trying to do automatic identification.

6.3.2 Manual labeling

Due to the fact that our manual labeling is done on RFCM clusters, its consistency over multiple runs needs to be validated. Multiple times of same expert labeling and labels from multiple experts will help validating the intra-rater and inter-rater reliability respectively.

In addition, different clustering methods other than RFCM might be more robust across multiple runs for this pre-clustering process.

6.3.3 Measurements

The current methods for measuring activation time and slope value is simple and needs be validated. Activation time measurement might not be robust since it involves calculation of a time derivative of the noisy signal. It may also provide poor temporal resolution since our current camera frame rate is low. Hence, methods for obtaining sub-frame accuracy of activation time are worth investigating. Currently, slope value calculation relies on fast depolarization starting time measurements, and a fixed length interval for fitting the slope, which might be unreliable under motion artifacts. Methods that are robust to the presence of motion artifacts will help improve the measurement accuracy.

6.3.4 Automatic identification

For pacemaker identification, there are several aspects that can be further improved. Better algorithm for threshold determination will improve the identification performance significantly. For the proposed DoD transformation, its robustness under different 'high quality' datasets given same set of parameters needs to be tested when more 'high quality' data is available Also, as the purpose of DoD is to bring the signals into metric space that reflects the physiologically similarity under different deformation, a more suitable transformation for

this purpose will boost the performance of the class of clustering-transformation-clustering approaches. In addition, good automatic group selection method is desired for fully automating the nodal region identification procedure.

In addition, other than unsupervised clustering methods proposed in this thesis, supervised learning methods such as support vector machine [6] , random forests [7] and deep learning [5] can be used and will likely give good identification performance given larger amount of manually labeled data.

Appendix A

Diastolic Slope Ending Point

For a given signal, we want to find the ending point of depolarization slope. In this thesis, the ending point is assumed to be the point furthest from the straight line connecting the first and peak point, as illustrated in Figure A.1.
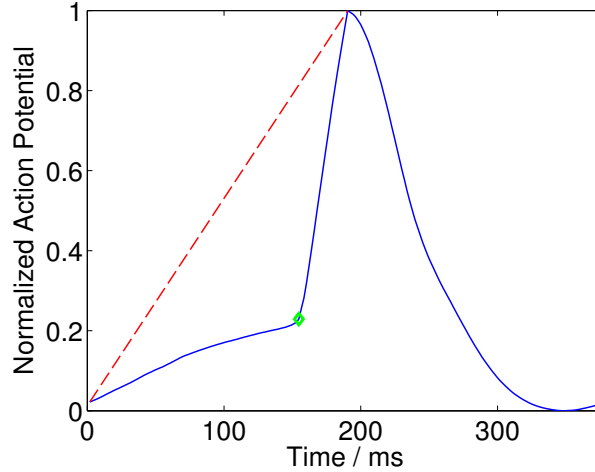
Figure A.1: Illustration of calculating diastolic slope ending point.

We denote the signal as $f(t)$, the first point as $(t_{1st}, f(t_{1st}))$, the peak point as $(t_{peak}, f(t_{peak}))$ and the wanted ending point as $(t_{end}, f(t_{end}))$. Given $(t_{1st}, f(t_{1st}))$ and $(t_{peak}, f(t_{peak}))$, the equation for the straight line connecting them is:

$$(f(t_{peak}) - f(t_{1st}))x + (t_{1st} - t_{peak})y + (t_{peak}f(t_{1st}) - t_{1st}f(t_{peak})) = 0 \qquad \text{(A.1)}$$

We also know that the the distance from a point (x_0, y_0) to the line $Ax + By + C = 0$

61

is $\frac{|Ax_0+By_0+C|}{\sqrt{A^2+B^2}}$, therefore we can calculate t_{end} by equation 4.4:

$$t_{end} = \underset{t_{1st}<t<t_{activation}}{\arg\max} \frac{|(f(t_{peak}) - f(t_{1st}))t - (t_{peak} - t_{1st})f(t) - t_{1st}f(t_{peak}) + t_{peak}f(t_{1st})|}{\sqrt{(f(t_{peak}) - f(t_{1st}))^2 + (t_{peak} - t_{1st})^2}}$$

The ending point we found using this method is invariant to scaling of $f(t)$ and t. To prove this, we show that $t_{end} = \alpha t^0_{end}$, given the scaling of $f(t) = \beta f_0(t)$ and $t = \alpha t^0$.

$$
\begin{aligned}
t_{end} &= \underset{t_{1st}<t<t_{activation}}{\arg\max} \frac{|(f(t_{peak}) - f(t_{1st}))t - (t_{peak} - t_{1st})f(t) - t_{1st}f(t_{peak}) + t_{peak}f(t_{1st})|}{\sqrt{(f(t_{peak}) - f(t_{1st}))^2 + (t_{peak} - t_{1st})^2}} \\
&= \underset{\alpha t^0_1<t<\alpha t^0_a}{\arg\max} \frac{|(\beta f(\alpha t^0_p) - \beta f(\alpha t^0_1))t - (\alpha t^0_p - \alpha t^0_1)\beta f(t) - \alpha t^0_1\beta f(\alpha t^0_p) + \alpha t^0_p\beta f(\alpha t^0_1)|}{\sqrt{(\beta f(\alpha t^0_p) - \beta f(\alpha t^0_1))^2 + (\alpha t^0_p - \alpha t^0_1)^2}} \\
&= \underset{\alpha t^0_1<t<\alpha t^0_a}{\arg\max} |(f(\alpha t^0_p) - f(\alpha t^0_1))t - (\alpha t^0_p - \alpha t^0_1)f(t) - \alpha t^0_1 f(\alpha t^0_p) + \alpha t^0_p f(\alpha t^0_1)| \\
&= \alpha \underset{t^0_1<t<t^0_a}{\arg\max} |(f(\alpha t^0_p) - f(\alpha t^0_1))\alpha t - (\alpha t^0_p - \alpha t^0_1)f(\alpha t) - \alpha t^0_1 f(\alpha t^0_p) + \alpha t^0_p f(\alpha t^0_1)| \\
&= \alpha t^0_{end} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (A.2)
\end{aligned}
$$

where t_1, t_p and t_a represent t_{1st}, t_{peak} and $t_{activation}$, respectively for saving space.

Bibliography

[1] Aristides B Arrenberg, Didier YR Stainier, Herwig Baier, and Jan Huisken. Optogenetic control of cardiac function. *Science Signalling*, 330(6006):971, 2010.

[2] Keith Baker, Kerri S Warren, Gary Yellen, and Mark C Fishman. Defective pacemaker current (ih) in a zebrafish mutant with a slow heart rate. *Proceedings of the National Academy of Sciences*, 94(9):4554–4559, 1997.

[3] Jeroen Bakkers. Zebrafish as a model to study cardiac development and human cardiac disease. *Cardiovascular research*, 91(2):279–88, July 2011.

[4] S Serge Barold, Arzu Ilercil, Fabio Leonelli, and Bengt Herweg. First-degree atrioventricular block. *Journal of Interventional Cardiac Electrophysiology*, 17(2):139–152, 2006.

[5] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009.

[6] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] Neil C Chi, Robin M Shaw, Benno Jungblut, Jan Huisken, Tania Ferrer, Rima Arnaout, Ian Scott, Dimitris Beis, Tong Xiao, Herwig Baier, et al. Genetic and physiologic dissection of the vertebrate cardiac conduction system. *PLoS biology*, 6(5):e109, 2008.

[9] B.R. Choi and G. Salama. Optical mapping of atrioventricular node reveals a conduction barrier between atrial and nodal cells. *American Journal of Physiology-Heart and Circulatory Physiology*, 274(3):H829–H845, 1998.

[10] Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, 2009.

[11] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[12] Igor R Efimov, Vladimir P Nikolski, and Guy Salama. Optical imaging of the heart. *Circulation research*, 95(1):21–33, 2004.

[13] I.R. Efimov and T.N. Mazgalev. High-resolution, three-dimensional fluorescent imaging reveals multilayer conduction pattern in the atrioventricular node. *Circulation*, 98(1):54–57, 1998.

[14] V.G. Fast and R.E. Ideker. Simultaneous optical mapping of transmembrane potential and intracellular calcium in myocyte cultures. *Journal of cardiovascular electrophysiology*, 11(5):547–556, 2007.

[15] E.M. Goldys. *Fluorescence Applications in Biotechnology and Life Sciences.* Wiley, 2009.

[16] Liza Gross. Regenerating zebrafish hearts reveal the molecular agents of repair. *PLoS Biology*, 4(8):e281, 2006.

[17] Todd J Herron, Peter Lee, and José Jalife. Optical imaging of voltage and calcium in cardiac cells & tissues. *Circulation research*, 110(4):609–623, 2012.

[18] William J Hucker, Crystal M Ripplinger, Christine P Fleming, Vadim V Fedorov, Andrew M Rollins, and Igor R Efimov. Bimodal biophotonic imaging of the structure-function relationship in cardiac tissue. *Journal of biomedical optics*, 13(5):054012–054012, 2008.

[19] Kodak Inc. Product summary: Kodak kai-0340 image sensor, 2008.

[20] C.J. Jou, K.W. Spitzer, and M. Tristani-Firouzi. Blebbistatin effectively uncouples the excitation-contraction process in zebrafish embryonic heart. *Cellular Physiology and Biochemistry*, 25(4-5):419–424, 2010.

[21] Richard Klabunde. *Cardiovascular physiology concepts.* LWW, 2011.

[22] Sebastian Kurtek, Anuj Srivastava, and Wei Wu. Signal estimation under random time-warpings and nonlinear signal alignment. NIPS, 2011.

[23] Di Lang, Matthew Sulkin, Qing Lou, and Igor R Efimov. Optical mapping of action potentials and calcium transients in the mouse heart. *Journal of visualized experiments: JoVE*, (55), 2011.

[24] Ulrike Langheinrich, Gabriele Vacun, and Thomas Wagner. Zebrafish embryos express an orthologue of herg and are sensitive toward a range of qt-prolonging drugs inducing severe arrhythmia. *Toxicology and applied pharmacology*, 193(3):370–382, 2003.

[25] Jacob I Laughner, Yuanzheng Gong, Benjamen A Filas, Song Zhang, and Igor R Efimov. Structured light imaging of epicardial mechanics. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5157–5160. IEEE, 2010.

[26] Jacob I Laughner, Fu Siong Ng, Matthew S Sulkin, R Martin Arthur, and Igor R Efimov. Processing and analysis of cardiac optical mapping data obtained with potentiometric dyes. *American journal of physiology. Heart and circulatory physiology*, 303(7):H753–65, October 2012.

[27] Qing Lou, Wenwen Li, and Igor R Efimov. Multiparametric optical mapping of the langendorff-perfused rabbit heart. *Journal of visualized experiments: JoVE*, (55), 2011.

[28] F.L. Meijler and M.J. Janse. Morphology and electrophysiology of the mammalian atrioventricular node. *Physiological reviews*, 68(2):608–647, 1988.

[29] David J Milan, Andrea C Giokas, Fabrizio C Serluca, Randall T Peterson, and Calum A MacRae. Notch1b and neuregulin are required for specification of central cardiac conduction tissue. *Science Signalling*, 133(6):1125, 2006.

[30] David J Milan, Albert M Kim, Jeffrey R Winterfield, Ian L Jones, Arne Pfeufer, Serena Sanna, Dan E Arking, Adam H Amsterdam, Khaled M Sabeh, John D Mably, et al. Drug-sensitized zebrafish screen identifies multiple genes, including gins3, as regulators of myocardial repolarization. *Circulation*, 120(7):553–559, 2009.

[31] David J Milan and Calum A MacRae. Zebrafish genetic models for arrhythmia. *Progress in biophysics and molecular biology*, 98(2):301–308, 2008.

[32] David J Milan, Travis A Peterson, Jeremy N Ruskin, Randall T Peterson, and Calum A MacRae. Drugs that induce repolarization abnormalities cause bradycardia in zebrafish. *Circulation*, 107(10):1355–1358, 2003.

[33] Sergey F Mironov, Frederick J Vetter, and Arkady M Pertsov. Fluorescence imaging of cardiac propagation: spectral properties and filtering of optical action potentials. *American Journal of Physiology-Heart and Circulatory Physiology*, 291(1):H327–H335, 2006.

[34] C.S. Myers and L.R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word. *Bell System Technical Journal*, 1981.

[35] P. Nemtsas, E. Wettwer, T. Christ, G. Weidinger, and U. Ravens. Adult zebrafish heart as a model for human heart? an electrophysiological study. *Journal of molecular and cellular cardiology*, 48(1):161–171, 2010.

[36] Marcello Pagano, Kimberlee Gauvreau, and Marcello Pagano. *Principles of biostatistics*. Duxbury Pacific Grove, CA, 2000.

[37] DL Pham. Spatial Models for Fuzzy Clustering. *Computer Vision and Image Understanding*, 84(2):285–297, November 2001.

[38] Dzung L Pham. Fuzzy clustering with spatial constraints. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 2, pages II–65. IEEE, 2002.

[39] Robert Plonsey and Roger C Barr. *Bioelectricity: a quantitative approach.* Springer, 2007.

[40] M Khaled Sabeh, Hussein Kekhia, and Calum a Macrae. Optical Mapping in the Developing Zebrafish Heart. *Pediatric cardiology*, March 2012.

[41] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[42] Jordan T Shin, Eugene V Pomerantsev, John D Mably, and Calum A MacRae. High-resolution cardiovascular function confirms functional orthology of myocardial contractility pathways in zebrafish. *Physiological genomics*, 42(2):300–309, 2010.

[43] Mark E Silverman, Charles B Upshaw Jr, and Helmut W Lange. Woldemar mobitz and his 1924 classification of second-degree atrioventricular block. *Circulation*, 110(9):1162–1167, 2004.

[44] Anuj Srivastava, Wei Wu, Sebastian Kurtek, Eric Klassen, and JS Marron. Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*, 2011.

[45] Federico Tessadori, Jan Hendrik van Weerd, Silja B Burkhard, Arie O Verkerk, Emma de Pater, Bastiaan J Boukens, Aryan Vink, Vincent M Christoffels, and Jeroen Bakkers. Identification and functional characterization of cardiac pacemaker cells in zebrafish. *PloS one*, 7(10):e47644, October 2012.

[46] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998.

[47] Hidekazu Tsutsui, Shin-ichi Higashijima, Atsushi Miyawaki, and Yasushi Okamura. Visualizing voltage dynamics in zebrafish heart. *The Journal of physiology*, 588(Pt 12):2017–21, June 2010.

[48] A.O. Verkerk, G.S.C. Geuzebroek, M.W. Veldkamp, and R. Wilders. Effects of acetylcholine and noradrenalin on action potentials of isolated rabbit sinoatrial and atrial myocytes. *Frontiers in Physiology*, 3, 2012.

[49] Kongming Wang and Theo Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276, 1997.

[50] Christopher Wren. *Concise guide to pediatric arrhythmias.* Wiley-Blackwell, 2011.

[51] R. Xu and D. Wunsch. *Clustering.* IEEE Press Series on Computational Intelligence. Wiley, 2008.