Robust algorithms for property recovery in motion modeling, medical imaging and biometrics

Yong Zhang

University of South Florida

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the American Studies Commons

Scholar Commons Citation

Zhang, Yong, "Robust algorithms for property recovery in motion modeling, medical imaging and biometrics" (2005). Graduate Theses and Dissertations.

http://scholarcommons.usf.edu/etd/931

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Robust Algorithms for Property Recovery in
Motion Modeling, Medical Imaging and Biometrics

by

Yong Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Co-Major Professor: Dmitry B. Goldgof, Ph.D.
Co-Major Professor: Sudeep Sarkar, Ph.D.
Don R. Hilbelink, Ph.D.
Anthony J. Llewellyn, Ph.D.
Leonid V. Tsap, Ph.D.

Date of Approval:
May 3, 2005

Keywords: Nonrigid Deformation, Elasticity, Computer Vision, Face Recognition.

© Copyright 2005, Yong Zhang
DEDICATION

To my parents
ACKNOWLEDGMENTS

I would like to thank my two major advisors, Dr. Goldgof and Dr. Sarkar, for their academic guidance and years of support. I was strongly influenced by their “vision” of what Computer Vision should be as a relatively new discipline. The one thing that I appreciate the most is that they gave me great freedom to pursue my research interest and provide opportunities so that I can work on a wide range of topics. I couldn’t have chosen better supervisors.

I am very fortunate to have the opportunity to work with Dr. Hall. Without his insight and expertise, the two papers on the constrained genetic approach would not be possible.

I am grateful to Dr. Hilbelink, Dr. Llewellyn and Dr. Tsap for taking the time from their busy schedule to be the committee member and providing very helpful comments on the draft.

I also thank the Department of Computer Science and Engineering at the University of South Florida for providing the lab facility that allowed me to complete my dissertation.

Most important of all, I wish to thank my family for their unfailing support and continually helping me in every way possible.
TABLE OF CONTENTS

LIST OF TABLES iii

LIST OF FIGURES iv

ABSTRACT v

CHAPTER 1 INTRODUCTION 1
1.1 Overview 1
1.2 Fundamental Issues and Related Work 2
1.2.1 Property Recovery and Inverse Problem 2
1.2.2 Computational Issues and Algorithms 3
1.2.3 Nonrigid Motion Modeling 6
1.3 Contributions 9

CHAPTER 2 RECOVERY ALGORITHMS 10
2.1 Forward Model: Linear Elastic Deformation 10
2.2 Recovery Algorithm: Deterministic Approach 12
2.2.1 Tikhonov Regularization 12
2.2.2 Baseline Solution Strategy 13
2.2.3 GCV Solution Strategy 17
2.2.4 Uniqueness of Solution 19
2.3 Recovery Algorithm: Constrained Genetic Approach 22
2.3.1 Genetic Encoding 22
2.3.2 Rank Based Constraint 24
2.3.3 Balancing Fitness and Penalty by Stochastic Ranking 28
2.3.4 Genetic Operators 29
2.4 Deterministic or Stochastic Approach 32
2.4.1 Experiments with GA, CGA and GNM 32
2.4.2 Experiments with CGA and GNM 37

CHAPTER 3 APPLICATION IN NONRIGID MOTION MODELING 45
3.1 Motion Parameters in Physical Model 45
3.1.1 Physical Motion Parameters 45
3.1.2 Physical and Mathematical Justifications 46
3.2 Synthesis with Recovered Parameter 52
3.2.1 Three Perspectives 52
3.2.2 Dirichlet-Type Motion Synthesis with Recovered Property 55
3.3 Experiments 58
3.3.1 Recovery of Physical Motion Parameter
3.3.2 Motion Synthesis with Recovered Parameter

CHAPTER 4 APPLICATION IN BURN SCAR ASSESSMENT
4.1 Model Construction
4.1.1 Extraction of Natural Point Features
4.1.2 Adaptive Meshing
4.2 Estimate Scar Elasticity
4.2.1 Range Scanner Setting
4.2.2 Boundary Condition Specification
4.3 Experimental Results
4.3.1 Data Set
4.3.2 Quantify Scar Elasticity
4.3.3 Relative Elastic Index

CHAPTER 5 APPLICATION IN FACE RECOGNITION
5.1 Hypothesis: Face Anatomy and Muscle Biomechanics
5.2 Computational Method
5.3 Experimental Results

CHAPTER 6 SENSITIVITY ANALYSIS
6.1 Computational Method
6.1.1 Primary Problem
6.1.2 Performance Measure and Dimensional Sensitivity
6.1.3 Adjoint State Method
6.1.4 Derivative Computation
6.1.5 Normalized Sensitivity
6.2 Experiments with Synthetic Model
6.2.1 Model Configuration and Performance Measure
6.2.2 Normalized Sensitivity Distribution
6.2.3 Dimensional Sensitivity Distribution
6.2.4 Relative Error Distribution
6.3 Experiments with Burn Scar
6.3.1 Model Configuration
6.3.2 Normalized Sensitivity
6.3.3 Dimensional Sensitivity
6.4 Experiments with Deformed Face Model
6.4.1 Face Model Configuration
6.4.2 Normalized Sensitivity
6.4.3 Dimensional Sensitivity

CHAPTER 7 CONCLUSIONS AND DISCUSSIONS

REFERENCES

ABOUT THE AUTHOR
LIST OF TABLES

Table 1. General Object Reconstruction Diagram. 1
Table 2. Inversion Algorithms for Elastic Property Reconstruction. 5
Table 3. Link Table for Maintaining Original Spatial Connections. 24
Table 4. CGA Experiments with Different Population Sizes. 39
Table 5. Performance of CGA and GNM with Different Initial Conditions. 40
Table 6. Adaptive Smoothing in Motion Analysis and Surface Reconstruction. 51
Table 7. Three Perspectives for Motion Synthesis. 55
Table 8. Synthesized Motion along Cross-Section Line. 65
Table 9. The Performance of Point Feature Detector. 71
Table 10. The Estimated Relative Scar Elasticity (kPa). 74
Table 11. Physician’s Rating and Relative Elastic Index (REI). 77
Table 12. Face Experiments. 84
# LIST OF FIGURES

| Figure 1 | Genetic Coding of the Young’s Modulus in a Finite Element Model. | 23 |
| Figure 2 | The Rank Based Scheme for Incorporating Qualitative Prior Information. | 27 |
| Figure 3 | Synthetic Model and the Qualitative Prior Knowledge. | 33 |
| Figure 4 | Reconstructed Young’s Modulus by GA, CGA and GNM. | 34 |
| Figure 5 | Reconstructed Young’s Modulus by the CGA with Biased Prior. | 36 |
| Figure 6 | Configuration of the Forward Model. | 37 |
| Figure 7 | Convergence Behavior of CGA with Different Population Sizes. | 39 |
| Figure 8 | Convergence Pathes of CGA and GNM with Different Starting Points. | 44 |
| Figure 9 | Results of GNM with Various Initial Conditions. | 44 |
| Figure 10 | Configurations of Spring Before and After Stretching. | 52 |
| Figure 11 | Adaptive Control of Motion Field by the Young’s Modulus. | 53 |
| Figure 12 | Configuration of a Synthetic Spring. | 58 |
| Figure 13 | Two Forward Simulations with Different but Proportional Elasticities. | 59 |
| Figure 14 | Inverse Solutions with Different Regularization Parameters. | 61 |
| Figure 15 | Norms against Regularization Parameter and Semi-Convergence. | 62 |
| Figure 16 | Performances of Recovery Algorithms. | 62 |
| Figure 17 | Image Sequence of a Bandage for Motion Synthesis. | 64 |
| Figure 18 | Finite Element Model and the Recovered Elasticity. | 64 |
| Figure 19 | Comparison of Two Experiments. | 67 |
| Figure 20 | Position Shift Between the Corresponding Features. | 70 |
| Figure 21 | Adaptive Meshing Using Point Features. | 72 |
Figure 22. The Geometry of K2T Range Scanner for Burn Scar Study.  73
Figure 23. Estimated Elasticity and Strain.  76
Figure 24. Correlation Between Physician’s Rating and REIs.  77
Figure 25. Strain Pattern of a Modified Face.  83
Figure 26. Four Images Acquired for Each Subject.  83
Figure 27. ROCs for Regular and Modified Faces.  85
Figure 28. Synthetic Model Configuration.  96
Figure 29. Deformed Models.  97
Figure 30. Normalized Sensitivity Contour Maps.  98
Figure 31. Dimensional Sensitivity Contour Maps.  98
Figure 32. Relative Error Distributions.  100
Figure 33. Deformation of Burn Scar.  100
Figure 34. Normalized Sensitivity Maps of Scar.  101
Figure 35. Dimensional Sensitivity Maps of Scar.  103
Figure 36. Face Deformation with Expression.  103
Figure 37. Normalized Sensitivity Maps of Face Model.  105
Figure 38. Dimensional Sensitivity Maps of Face Model.  107
ROBUST ALGORITHMS FOR PROPERTY RECOVERY IN
MOTION MODELING, MEDICAL IMAGING AND BIOMETRICS

Yong Zhang

ABSTRACT

The past two decades has witnessed growing interest in physics based techniques in computer vision, computer graphics and medical imaging. The main advantage of a physical model is its mathematical rigor and physical soundness, which makes it an ideal tool to study complex nonrigid motion. However, since a model based on continuum mechanics is computationally demanding, an idealized framework is often adopted where physical motion parameters are significantly simplified, which inevitably affects the accuracy and reliability of modeling results.

In this study, a new modeling approach is developed that features the reconstruction of actual material properties such as the Young’s modulus and the Poisson’s ratio. Justified by the constitutive law and mathematical considerations, the Young’s modulus is identified as a unique physical motion parameter. By imposing an adaptive smoothness constraint, the Young’s modulus helps preserve the local characteristics (discontinuity) of an object’s deformation, a role similar to the weighting coefficient in the study of edge-preserving visual surface reconstruction.

The contribution of this work is fourfold: (1) two recovery algorithms are developed to solve the inverse elastic problem: A deterministic algorithm that is based on the Gauss-Newton method and the general cross validation, and a stochastic algorithm that is based on the constrained genetic evolution; (2) a new modeling approach is proposed that has the ability to recover nonrigid motion in terms of the physical parameters. The use of recovered parameters can be implemented within a boundary-driven motion synthesis scheme; (3) A
sensitivity method is proposed to evaluate the impact of different parameters. The method uses the adjoint state equation and hence is suitable for large scale models. (4) the proposed modeling approach has been applied to burn scar assessment and face recognition.
CHAPTER 1
INTRODUCTION

1.1 Overview

One of the fundamental issues in computer vision and image understanding is to extract basic attributes of an object from 2D images. Those attributes can be either directly utilized or further processed to facilitate higher level tasks such as classification, recognition and synthesis etc. The commonly studied attributes and associated reconstruction methods are listed in Table 1.

Much progress has been made in reconstructing 3D shape and optical properties (surface reflectance) under the common theme of “shape from X”. However, little has been done to recover the material properties that are hidden beneath the visible surface such as the elasticity and the viscosity. It is interesting to note that, even in the famous Marr’s diagram, only the geometrical primitives (surface and shape) are emphasized. On the other hand, human vision can easily detect the subtleties in the way an object moves and deforms, and further relate it to the underlying mechanical properties. For example, simply looking at the expressions of a baby’s face and a wrinkled face can lead us to believe (no matter how implicitly such a perception is formed) that the baby’s skin is more elastic, which is probably

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Examples</th>
<th>Reconstruction Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>geometric</td>
<td>3D shape/structure</td>
<td>shape from X</td>
</tr>
<tr>
<td>surface</td>
<td>reflectance and color</td>
<td>radiometric models</td>
</tr>
<tr>
<td>kinematic</td>
<td>deformation matrix</td>
<td>affine motion models</td>
</tr>
<tr>
<td>dynamic</td>
<td>force and boundary conditions</td>
<td>physical motion models</td>
</tr>
<tr>
<td>mechanical</td>
<td>elasticity and viscosity</td>
<td>property estimation</td>
</tr>
</tbody>
</table>
one of the reasons that we find the baby’s face more appealing. As another example, by observing the way clouds move, sea waves propagate and a flag flaps in the wind, a person can develop a sense that the objects have different material properties, even though he or she may not be able to give a quantitative description using the language of continuum mechanics. One of the objectives of this research is to develop a method that can recover material properties from images, in particular, the Young’s modulus of elastic objects.

In addition to the aforementioned theoretical interest in material perception, reconstructing elastic property is also of interest to researchers in many practical fields. For example, identifying elasticity abnormalities in soft tissue has high potential in early cancer detection [37, 94, 63, 98, 34], because pathological processes often cause property changes in the diseased tissues that amplify themselves as hard inclusions. Knowing elastic properties of deformable objects is also critical for physics-based modeling such as motion tracking, realistic animation, scar assessment, visualization and surgical planning [74, 127, 135, 82, 99, 35], because the quality of elastic properties could have a strong impact on the performance of physical models. Recovering elastic properties has also found applications in structural damage identification, geophysical exploration, wafer engineering, robotics design and composite material characterization [22, 7, 38, 67, 45].

In this study, applications of the elastic property recovery will be demonstrated in three areas: (1) nonrigid motion modeling; (2) quantitative burn scar assessment; (3) face recognition (biometrics).

1.2 Fundamental Issues and Related Work

1.2.1 Property Recovery and Inverse Problem

Like many vision problems, inferring the elastic properties from the observed deformation is an ill-posed inverse problem. A problem can be regarded as either a forward problem or an inverse problem depending on its cause-effect alignment. A problem that describes the effect based on the given cause is a forward problem, while a problem that is aimed at understanding the cause of a consequence is an inverse problem. In many disciplines, the
forward problem is to find out the behavior of a system responding to the external excitation (cause). For example, we can model the nonrigid deformation of an elastic object caused by the surface force. On the other hand, the inverse problem is to infer the quantities that directly or indirectly caused the effect, such as the force and the elastic properties.

A physical law can specify the cause and the corresponding effect with mathematical equations. It has long been found that inverse problems are characterized by their unstable solutions, while the forward problems are usually well behaved numerically. By providing a formal mathematical definition of the well-posedness and the ill-posedness, Hadamard [53] laid the foundation of inverse problem theory and paved the road of finding a meaningful solution of many inverse problems. In the Hadamard sense, a well-posed problem must have a solution that satisfies the following three conditions: existence, uniqueness and stability (continuous dependence of the solution on the input data). A problem is ill-posed if one of those conditions failed. Most of the forward problems are well-posed and the corresponding inverse problems are ill-posed. The difficulty in solving an ill-posed inverse problem is often related to the non-uniqueness and the instability of the solution.

In practice, an inverse problem is often transformed into an optimization problem. Many algorithms have been developed to reduce the solution space and to smooth the solution path in order to restore the uniqueness and the stability of the numerical solution. The basic idea is to utilize a priori knowledge about the physical system. For example, deterministic regularization is a commonly used method for parameter estimation [30]. Statistical methods in the Bayesian framework have also been investigated [33, 124]. In the context of computer vision, Poggio et al. [106] present a comprehensive review of the inverse problem theory and its implication to image understanding and visual perception.

1.2.2 Computational Issues and Algorithms

The regularized functional of a nonlinear ill-posed inverse problem is difficult to solve, because its convexity can only be guaranteed very locally. In the deterministic domain, the use of gradient-based methods is common [30]. The gradient-based methods are sensitive to
the starting point and may stuck in the local extrema. Therefore, solving inverse problems with the stochastic algorithms has received much attention recently. The hybrid algorithm that utilizes the strengths of both the deterministic and the stochastic methods has also been proposed [22].

The related work is summerized in Table 2. The list only includes the papers that deal with the computational aspect of solving an inverse elastic problem. For studies on the direct measurement and the development of special imaging modalities such as the magnetic resonance elastography, the ultrasonic elastography and the optical coherence tomography, readers are refered to [42, 31, 34, 95, 71].

It is apparent that most of the studies chose the deterministic algorithms that use either direct or iterative inversions. Kallel and Bertrand [68] used the Newton-Raphson method combined with a finite element model to estimate the Young’s modulus of synthetic tissues. Doyley et al [27] studied a modified iterative Newton-Raphson method. Van Houten et al [139] proposed a multi-resolution method (subzone) to improve both the efficiency and robustness. Other deterministic algorithms such as the level set method [6], the adjoint state method [90], the steepest descent search [135] and the iterative regularization [29] have also been reported.

Due to the lack of quantitative comparison data for the two approaches (deterministic vs. stochastic), it is often a difficult task to design an optimal solution strategy that is both robust and efficient. In this study, both the deterministic and the stochastic approaches will be investigated. Specifically, a Newton-Raphson method based on the general cross validation and a constrained genetic algorithm will be discussed. Several comparative experiments were also performed to examine the robustness and the convergence rate of the two algorithms.

Both linear and nonlinear constitutive models have been studied. The linear model is computationally attractive and usually a good approximation for many objects. In this study, a linear elastic model will be used.
Table 2. Inversion Algorithms for Elastic Property Reconstruction.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Date</th>
<th>Algorithms</th>
<th>Models</th>
<th>Experiments (Materials)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barbone et al [12]</td>
<td>2004</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Chiroyi et al [21]</td>
<td>2000</td>
<td>Genetic</td>
<td>Nonlinear</td>
<td>Crystal lattice</td>
</tr>
<tr>
<td>Chwiaczowsky et al [22]</td>
<td>2004</td>
<td>Genetic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Doyley et al [27]</td>
<td>2000</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Engl et al [29]</td>
<td>2003</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Polymer</td>
</tr>
<tr>
<td>Eskin et al [32]</td>
<td>2002</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Franca et al [38]</td>
<td>2004</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Silicon wafer</td>
</tr>
<tr>
<td>Furukawa et al [43]</td>
<td>1997</td>
<td>Genetic</td>
<td>Nonlinear</td>
<td>Cr-Mo steel</td>
</tr>
<tr>
<td>Garboczi et al [45]</td>
<td>1995</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Hori et al [60]</td>
<td>2003</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Huang et al [61]</td>
<td>2001</td>
<td>Deterministic</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Ji et al [65]</td>
<td>2004</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>John et al [66]</td>
<td>2003</td>
<td>Genetic</td>
<td>Linear</td>
<td>Pelvis bone</td>
</tr>
<tr>
<td>Joukhadar et al [67]</td>
<td>1997</td>
<td>Genetic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Kallel et al [68]</td>
<td>1996</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Kirkpatrick et al [71]</td>
<td>2002</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Tissues</td>
</tr>
<tr>
<td>Maurice et al [79]</td>
<td>2004</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Vessel walls</td>
</tr>
<tr>
<td>Miga [83]</td>
<td>2003</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Muthupillai et al [87]</td>
<td>1995</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Gel</td>
</tr>
<tr>
<td>Nakamura et al [88]</td>
<td>1993</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Oberai et al [90]</td>
<td>2003</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>O’Donnell et al [91]</td>
<td>2004</td>
<td>Deterministic</td>
<td>Nonlinear</td>
<td>Heart tissue</td>
</tr>
<tr>
<td>Oliphant et al [92]</td>
<td>2001</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Phantom</td>
</tr>
<tr>
<td>Pellot-Barakat et al [100]</td>
<td>2004</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Phantom</td>
</tr>
<tr>
<td>Plewes et al [105]</td>
<td>2000</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Samani et al [113]</td>
<td>2001</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Breast tissues</td>
</tr>
<tr>
<td>Sumi et al [122]</td>
<td>1998</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Tsap et al [135]</td>
<td>1998</td>
<td>Deterministic</td>
<td>Nonlinear</td>
<td>Skin</td>
</tr>
<tr>
<td>Van Houten et al [139]</td>
<td>1999</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Zhu et al [149]</td>
<td>2003</td>
<td>Deterministic</td>
<td>Linear</td>
<td>Numerical</td>
</tr>
</tbody>
</table>

Papers that use the hybrid approach combining the deterministic and genetic algorithms are classified as Genetic. Papers that considered geometrical or material linearities to simplify the forward model are classified as Linear. No more than two papers are selected from the same research group (authors) unless the algorithms used are significantly different.
1.2.3 Nonrigid Motion Modeling

Much of this study is conducted in the context of nonrigid motion modeling. Nonrigid motion is more difficult to analyze than its rigid counterpart, because it deals with a much wider range of objects and more complex deformation patterns. It is unlikely that a single solution scheme can be found that is applicable to all types of nonrigid motion. Therefore, the following strategy has been adopted: (1) classify nonrigid motion into different types and deal with each type individually; (2) use a motion model to facilitate the analysis because a good model can significantly reduce the recovery and tracking complexities. Huang [62] classified nonrigid motion into three basic types: Articulated motion, elastic motion and fluid motion. Kambhamettu et al. [69] further extended the classification into more detailed subgroups based on the geometrical characteristics of deformation. Aggarwal et al [2] surveyed the existing nonrigid motion modeling techniques, particularly those based on geometrical and physical models.

A key issue in the model-based approach is to define and recover motion parameters. Motion parameters play an important role of carrying various constraints that we wish to impose on the motion. The governing equations of a motion model are characterized by local descriptions of the degrees of freedom of deforming objects. Theoretically, this locality property enables us to achieve a modeling accuracy close to an analytical solution with an infinitely small discretization unit, such as a finite element or a finite difference patch. Unfortunately, such a solution strategy quickly becomes intractable because the computational cost associated with very fine mesh is prohibitively expensive. Therefore, a model that can describe the motion on both local and global scales is strongly desirable. Posing global constraints on motion model through appropriate motion parameters is a commonly used strategy. An extreme example is the motion model of a rigid body where local motion is automatically global because of the rigidity constraint.

The ideal motion parameters that can pose global constraints should possess following attributes: (1) the number of parameters is manageable to reduce the computational com-
plexity; (2) parameters remain constant during deformation; (3) parameters are well defined (either physically or kinematically) so that their values can be verified.

Because of the intimate relationship between an object’s shape and motion, global constraints are often formulated as various shape functions. A constrained deformable superquadric function has been used for simultaneous shape reconstruction and object recognition [126, 80]. Along the same line, a hyperquadric function combined with a kinematic model was proposed to recover the nonrigid motion without the need of correspondence [75]. The drawback of using a global parametric shape function is that the parameters are temporal functions and have to be continuously updated during tracking.

A truncated series of vibration modes has been used to break down the nonrigid motion into different components [101]. The small set of low frequency modal parameters allows for a stable estimation and representation of motion field. A similar frequency-based method was employed to analyze the nonrigid motion in 4D volume images [89]. The main limitation of those approaches is that the modal parameters do not have the clear physical meanings and their values are difficulty to verify.

The aforementioned parameters are based primarily on the kinematic characteristics of motion. The advantage of kinematic parameters is their geometrical intuitiveness and generality. The main disadvantage is the lack of physical descriptions of motion. For instance, forces that cause the motion as well as the object’s internal response are not accounted in the kinematic parameters.

Objects behave differently under the same loading condition. An intuitive explanation is that an object’s capability to resist being deformed is related to its unique material constitution. An object’s capability to resist external forces can be viewed as a special internal constraint that is automatically imposed on its motion equations. This type of constraint can be quantified through a constitutive equation, which is in turn determined by an object’s intrinsic material properties. Imposing motion constraints based on material characteristics through the constitutive equation is physically plausible and is consistent with the practice of continuum mechanics.
Tsap et al. [136] have utilized the recovered elasticity to assist nonrigid motion modeling. The recovery algorithm is based on a one-dimensional gradient decent method. The motion analysis is carried out within a multi-layer mesh refinement framework. The method has also been applied to burn scar study [135]. Modeling nonrigid motion using a full scale physical model has been reported in the study of fluid motion [82], although no attempt has been made to recover the material property of fluid such as the viscosity.

In this study, based on thorough analysis of the characteristics of motion equation and mathematical justification of property controlled adaptive motion smoothing, the Young’s modulus is identified as an unique physical motion parameter to pose continuity constraints on the deformation of elastic objects. As a physical motion parameter, the Young’s modulus has the following desirable attributes:

1. As a distributed parameter, the Young’s modulus can be assigned to each element of a finite element model. At the first glance, it seems an expensive option for a model of many elements. However, biological objects usually have property abnormalities in a few isolated areas, and the background can be regarded as homogeneous. Therefore, an object’s elastic property can be adequately represented by a few distinct values.

2. The Young’s modulus is a function of temperature in thermodynamics. Fortunately, isothermal condition is valid in most cases. The variation of Young’s modulus of biological materials within the normal range of body temperature is negligible and thus can be viewed as a constant. This feature distinguishes the Young’s modulus from kinematic parameters. In long sequence tracking and medical image registration, the Young’s modulus only needs to be recovered once.

3. The Young’s modulus has a clear physical interpretation in elasticity theory and its value can be verified. The ground truth Young’s modulus of biological tissues can be gathered from standard tensile test [41].
1.3 Contributions

In summary, the main contributions of this work are:

1. In the context of nonrigid motion analysis, the Young’s modulus is identified as a unique physical motion parameter, rather than a regular variable in a forward finite element model. The uniqueness of using the Young’s modulus as a motion parameter lies in the fact that it essentially controls the distribution of motion field with a mechanism that is similar to the regularized adaptive smoothing. More importantly, the Young’s modulus poses an implicit continuity constraint on the motion inside an object, which allows us to synthesize the interior motion from boundary observations by solving a well-posed Dirichlet type boundary value problem.

2. Two new recovery algorithms are developed to solve the ill-posed nonlinear inverse elastic problem, one is a deterministic approach based on the Gauss-Newton minimization and the general cross validation selection (GCV), and the other is a stochastic approach based on the constrained evolutionary computation. The coupling of Gauss-Newton iteration and GCV search improves both the efficiency and the robustness of the algorithm. The constrained genetic algorithm proves to be valuable in the case where a good initial condition is not available, especially for large scale models.

3. A sensitivity method is developed to improve the modeling accuracy. The method allows us to quantify the impact of different parameters using the sensitivity maps. The most vulnerable areas in a model can be identified with the aid of sensitivity analysis. The algorithm uses the adjoint state method and therefore is suitable for large scale finite element models.

4. The proposed modeling approach has been applied to burn scar assessment and face recognition. The use of natural features enhances the applicability of the current scar rating method. The experiment using elastic strain pattern of profile images indicates that the proposed new biometrics holds great potential in face recognition.
CHAPTER 2
RECOVERY ALGORITHMS

2.1 Forward Model: Linear Elastic Deformation

The dynamics of a deformable body defined in a bounded domain \( (\Omega \subset \mathbb{R}^n, \partial \Omega = \Gamma_1 + \Gamma_2) \) under the isothermal condition is governed by the following partial differential equation with the Cauchy data \( C_d \) [41, 132, 24]:

\[
\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \sigma^T + \mathbf{F}, \quad \text{in } \Omega,
\]

\[
C_d = \{ \mathbf{u}|_{\Gamma_1} = \bar{\mathbf{u}}, \quad \frac{\partial \mathbf{u}}{\partial n}|_{\Gamma_2} = \mathbf{g} \}, \quad \text{on } \partial \Omega = \Gamma_1 + \Gamma_2.
\]

where \( \mathbf{u}(u,v,w) \) denotes the displacement vector of a point at \( (x,y,z) \), \( \mathbf{F}(f_x, f_y, f_z) \) is the body force, \( \rho \) is the mass density, \( \sigma \) is the stress tensor, \( T \) denotes transpose, \( \nabla \cdot \) is the divergence operator with respect to a tensor, \( n \) is the outward unit normal on the boundary, and \( (\bar{\mathbf{u}}, \mathbf{g}) \) are the Dirichlet and Neumann data on the boundary \( (\Gamma_1, \Gamma_2) \), respectively.

The motion equation of elastic body is simplified with two linear conditions. A Cauchy strain tensor (geometrically linear) is used,

\[
\varepsilon = \frac{1}{2} [\nabla \mathbf{u} + (\nabla \mathbf{u})^T],
\]

where the gradient operator \( \nabla \) is defined with respect to the displacement vector:

\[
\nabla \mathbf{u} = \begin{bmatrix}
\frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} \\
\frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} \\
\frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} & \frac{\partial w}{\partial z}
\end{bmatrix},
\]
and consequently strain tensor can also be expressed in 3D Cartesian coordinate as:

$$
\varepsilon = \begin{bmatrix}
\frac{\partial u}{\partial x} & \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) & \frac{1}{2}(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}) \\
\frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \frac{\partial v}{\partial y} & \frac{1}{2}(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}) \\
\frac{1}{2}(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}) & \frac{1}{2}(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}) & \frac{\partial w}{\partial z}
\end{bmatrix}.
$$

(5)

Using the linear constitutive law (material linear, the generalized Hooke’s law) and isotropic property, the following stress-strain relationship can be derived:

$$
\sigma = \lambda (\text{tr}\varepsilon) I + 2\mu\varepsilon = \lambda (\nabla \cdot \mathbf{u}) I + \mu \nabla \mathbf{u} + \mu (\nabla \mathbf{u})^T,
$$

(6)

where $I$ is the identity matrix, $\text{tr}$ denotes trace, $\lambda$ and $\mu$ are the Lamé constants. The symmetry of stress tensor ($\sigma = \sigma^T$) is automatically derived from the symmetry of strain tensor ($\varepsilon = \varepsilon^T$).

With above linear conditions, the governing motion equations can rewritten in terms of displacement that describes the deformation of an inhomogeneous, isotropic and linear elastic object:

$$
\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot [\lambda (\nabla \cdot \mathbf{u}) I + \mu \nabla \mathbf{u} + \mu (\nabla \mathbf{u})^T] + \mathbf{F}, \quad \text{in } \Omega,
$$

(7)

$$
C_d = \{ \mathbf{u}|_{\Gamma_1} = \bar{\mathbf{u}}, \quad \frac{\partial \mathbf{u}}{\partial n}|_{\Gamma_2} = \mathbf{g} \}, \quad \text{on } \partial\Omega = \Gamma_1 + \Gamma_2.
$$

(8)

Material properties more commonly known in the engineering literature such as the Young’s modulus ($E$) and the Poisson’s ratio ($\nu$) are related to the Lamé constants through the following transformations and can be directly substituted into the motion equation:

$$
\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}.
$$

(9)

Since both the Young’s modulus and Poisson’s ratio are considered as spatial functions, the motion equations (7, 8) will not be further simplified into the Navier’s equation.
2.2 Recovery Algorithm: Deterministic Approach

The inversion of a forward equation is often ill-posed and situation gets worse with a highly nonlinear system of elasticity reconstruction. Even if the forward problem can be simplified through linearization, the corresponding inverse problem is often nonlinear [30]. Various regularization methods have to be used to alleviate the ill-posedness, among which the Tikhonov functional is a popular one [131].

2.2.1 Tikhonov Regularization

Considering the static case of elastic deformation:

\[
\nabla \cdot \left[ \lambda (\nabla \cdot \mathbf{u}) \mathbf{I} + \mu \nabla \mathbf{u} + \mu (\nabla \mathbf{u})^T \right] + \mathbf{F} = 0,
\]

(10)

a differential operator \(B\) for displacement \(\mathbf{u}\) with respect to \(E\) can be defined:

\[
B(E) = \nabla \cdot [\lambda (\nabla \cdot \cdot) \mathbf{I} + \mu \nabla \cdot + \mu (\nabla \cdot \cdot) \mathbf{I}].
\]

(11)

Rearranging equation (10) using \(B(E)\), a nonlinear operator equation can be derived:

\[
B(E)\mathbf{u} + \mathbf{F} = 0,
\]

(12)

\[
A(E) = -\mathbf{F}B(E)^{-1} = \mathbf{u}.
\]

(13)

Assuming that the nonlinear operator \(A : X \rightarrow Y\) is continuously Fréchet differentiable in the Hilbert space \((X, Y)\), an inverse problem of estimating the Young’s modulus from noisy data is obtained:

\[
A(E) = \mathbf{u},
\]

(14)

\[
\|\mathbf{u}_t - \bar{\mathbf{u}}\| \leq \delta,
\]

(15)

where \(\mathbf{u}_t\) is the true displacement, and \(\bar{\mathbf{u}}\) is the corrupted data with a noise level of \(\delta\).

This inverse problem is very likely ill-posed in the Hadamard sense [53]. The existence of an inverse solution is the most difficult to prove. Fortunately, for practical problems,
although the existence of an inverse solution cannot be guaranteed with noisy data, at least the physical reality itself offers an solution that can be approximated. In the case of the Young’s modulus recovery, we know that there must be specific Young’s modulus values associated with each real object. If an inverse system is insolvable, it usually can be found out that either the model configuration or boundary specification does not make physical sense. Therefore, we are more concerned about obtaining a numerical solution that is unique and stable, assuming that such a solution does exist.

The recovery problem is formulated as a Tikhonov functional in its variational form:

\[ E = \min_{E \subset \mathbb{R}^n} T(E), \]

\[ T(E) = \|DA(E) - \bar{u}\|^2 + \beta \|S(E - E^*)\|^2, \]

where \( D \) is a projection operator that maps the continuous degrees of freedom to the discrete measurement coordinate, \( E^* \) denotes the prior knowledge, \( \beta \) is the regularization parameter and \( S \) is the smoothness matrix (a discretized version of the gradient or Laplacian operators). The solution of Tikhonov functional represents a compromise between the fidelities to observation data and prior knowledge. It can also be viewed as a regular least-square solution, subject to the constraint of a small norm.

### 2.2.2 Baseline Solution Strategy

The minimizer of the Tikhonov functional can be found by either a deterministic gradient method or a stochastic global search. The Gauss-Newton method is used to solve the nonlinear ill-posed inverse problem of (16). Since the convexity of a nonlinear Tikhonov functional can only be guaranteed locally, it is expected that the algorithm starts with a point not very far from the true solution.

In the following discussions, \( P \) is used to denote the generic parameter to be estimated, which could be the Young’s modulus, the Poisson’s ratio or any other quantities of interest. \( \bar{u} \) is the data vector that consists of both measured displacement and stress. The inter-
polation matrix $D$ is dropped, assuming that the outputs from forward model have been computed on the measurement coordinate.

The Gauss-Newton strategy is to linearize the nonlinear system and then construct an iterative gradient-based search algorithm. Considering the Tikhonov functional of a nonlinear operator $A(P)$:

$$T(P) = \|A(P) - \bar{u}\|^2 + \beta \|S(P - P^*)\|^2,$$

(18) is rewritten in terms of the linearized expression around a local point $P_k$:

$$L(P) = \|A(P_k) + J(P_k)(P - P_k) - \bar{u}\|^2 + \beta \|S(P - P^*)\|^2,$$

(19)

where $J(P)$ is the Fréchet derivative and $L$ denotes the linearized functional.

$L(P)$ is then minimized through the vanishing gradient condition:

$$g(P) = \nabla L(P) = 0,$$

(20)

$$J^T(P_k)(A(P_k) + J(P_k)(P - P_k) - \bar{u}) + \beta S^T S(P - P^*) = 0.$$  

(21)

Replacing $P$ with $P_{k+1}$ in (21), the iterative fixed point formula of estimating $P_k+1$ is obtained:

$$P_{k+1} = P_k + \Delta P,$$

(22)

$$\Delta P = [J^T(P_k)J(P_k) + \beta S^T S]^{-1}[J^T(P_k)(\bar{u} - A(P_k)) + \beta S^T S(P_k - P^*)].$$

(23)

It can be seen that (23) is the normal equation of the following linear system:

$$\begin{bmatrix} J(P_k) \\ \sqrt{\beta} S \end{bmatrix} \Delta P = \begin{bmatrix} \bar{u} - A(P_k) \\ -\sqrt{\beta} S(P_k - P^*) \end{bmatrix}.$$  

(24)

The Conjugate Gradient method is an efficient solver for the large scale linear system. During each Gauss-Newton iteration for minimizing the Tikhonov functional, the Conjugate Gradient method is used to solve (24). The computed increment $\Delta P$ will then be used to update the Tikhonov solution $P_{k+1}$.
The baseline algorithm for solving an ill-posed elastic problem is given as follows:

Baseline Algorithm:

1. select the initial solution
2. For i = 1 to maximum tests of regularization parameter
3. compute the initial Tikhonov functional with P(i) and beta(i)
4. For k = 1 to maximum Gauss_Newton iterations
5. compute the Frechet derivative J(P)
6. compute the gradient g(P)
7. if converge criteria is satisfied
8. P(i) = P(k)
9. exit Gauss_Newton iteration
10. else
11. compute the increment with the Conjugate Gradient: dP
12. update solution: P(k) = P(k-1) + gamma * dP
13. compute new Tikhonov functional with P(k)
14. if (new Tikhonov < old Tikhonov)
15. accept new solution: P(k)
16. update the Tikhonov functional with P(k)
17. else
18. choose a smaller step size: gamma
19. go to step 12
20. End for
21. if P(i) satisfies the discrepancy principle
22. exit
23. else
24. choose beta(i+1) using the norms and interpolation of beta(i), beta(i-1), ...
25. End for

The baseline algorithm employs a strategy that tries out many possible values of regularization parameter ($\beta_i$) and chooses the best one that satisfies certain criteria such as the discrepancy principle [86]:

$$\|A(P(\beta_i)) - \bar{u}\| \leq \delta. \quad (25)$$
An intuitive interpretation of the discrepancy principle is that it is meaningless to pursue a solution whose residual is below the bound of noise level. If the Gaussian data noise is assumed with $\eta \sim N(0, \sigma)$:

$$\bar{u} = A(P_{\text{true}}) + \eta,$$

and solution $P(\beta)$ is close to the true solution $P_{\text{true}}$, then the expectation of residual norm can be approximated as follow:

$$\varepsilon\{\|A(P(\beta)) - \bar{u}\|^2\} \approx \varepsilon\{\|\eta\|^2\} = N\sigma^2,$$

with $N$ as the dimension of data vector. So, the outer loop of the baseline algorithm should be stopped as soon as the discrepancy rule is satisfied:

$$|\|A(P(\beta)) - \bar{u}\|^2 - N\sigma^2| < \text{tolerance}$$

If the Gauss-Newton solution with a given $\beta$ does not satisfy the discrepancy criteria, a new regularization parameter has to be chosen and tested. This implementation scheme is relatively expensive, because potentially many different $\beta$ values have to be tested, and for each $\beta$, the nonlinear Tikhonov functional (18) has to be solved. If the algorithm can start with a $\beta_i$ that is close to the optimal value and the residual norm is monotonic around $\beta_i$, the interpolation and bracketing techniques can be applied to search for the next $\beta_{i+1}$ using the residual and solution norms of previous values ($\beta_i, \beta_{i-1}, ...$), until the optimal one is found [9]. The main problem with the methods based on the discrepancy rule is that information about data noise may not be available in practice. Therefore, we have to look into more heuristically-based methods such as the generalized cross validation (GCV) [140, 50].
2.2.3 GCV Solution Strategy

GCV is designed based on the leave-one-out principle and has been successfully applied in linear settings. Considering a linear system and its Tikhonov functional in standard form:

\[ Gx = y, \]  

\[ T(x) = \|Gx - y\|^2 + \beta \|x\|^2, \]  

the above functional can be rewritt with the \( k \)th element of data missing:

\[ T(x_k) = \sum_{i \neq k} \|(Gx)_i - y_i\|^2 + \beta \|x\|^2, \]

where \( x_k \) is the solution of minimizing \( T(x_k) \), and \( N \) is the dimension of \( y \).

The philosophy behind the cross validation is that if an element of data is left out, the corresponding solution of minimizing (31) should be able to predict the missing element reasonably well. In other words, if the \( \beta \) used in minimization of (31) is a good choice, the difference between the predicted and observed \( k \)th datum should be small. So, a cross validation function can be constructed by summing up the squared distance for all elements:

\[ CV(\beta) = \sum_{k=1}^{N} \frac{|(Gx_k)_k - y_k|^2}{|1 - Q_{kk}(\beta)|^2}, \]

and the optimal regularization parameter can be found by minimizing the CV function: \( \beta = \min CV(\beta) \). To avoid solving \( N \) equations involved in \( CV(\beta) \), the following formula can be used that has been thoroughly studied by Wahaba in the context of fitting spline model [140]:

\[ CV(\beta) = \sum_{k=1}^{N} \frac{|(Gx_k)_k - y_k|^2}{1 - Q_{kk}(\beta)}, \]

where \( Q(\beta) = G(G^TG + \beta I)^{-1}G^T \), and \( x(\beta) = (G^TG + \beta I)^{-1}G^Ty \), the solution of \( T(x) = 0 \).
To remove its undesired property of dependence on orthogonal data transformation, (33) was extended to the generalized cross validation (GCV) [50]:

$$GCV(\beta) = \frac{\|Gx(\beta) - y\|^2}{(\text{trace}[I - Q(\beta)])^2},$$

(34)

and $GCV(\beta)$ can be minimized through either direct inversion or iterative procedures [51].

For the nonlinear inverse problem (14), it is noticed that its linearized Tikhonov functional:

$$L(P) = \|A(P_k) + J(P_k)(P_{k+1} - P_k) - \bar{u}\|^2 + \beta\|S(P_{k+1} - P^*)\|^2,$$

(35)

is a regularization of the following linear system:

$$JV = d$$

(36)

with $J = J(P_k)$, $V = P_{k+1} - P^*$, and $d = JP_k - JP^* - A(P_k) + \bar{u}$. So, $L(P)$ is rewritten as:

$$L(P) = \|JV - d\|^2 + \beta\|SV\|^2.$$  

(37)

Using the same formula as in (34), the GCV for $L(P)$ is obtained as:

$$GCV(\beta) = \frac{\|JV(\beta) - d\|^2}{(\text{trace}[I - Q'(\beta)])^2},$$

(38)

with $Q'(\beta) = J(J^T J + \beta S^T S)^{-1}J^T$, and $V(\beta) = (J^T J + \beta S^T S)^{-1}J^T d$.

Now that the regularization parameter is estimated for a linearized functional (35) without the requirement of prior knowledge of data noise, a scheme is needed to solve the original nonlinear Tikhonov functional (18) utilizing the GCV method. If we simply place the GCV procedure at the end of the outer loop of baseline algorithm, we would end up with the same situation of solving a nonlinear function many times. A more efficient approach is to incorporate the GCV inside the Gauss-Newton iteration so that $\beta$ and $P$ can be solved simultaneously [52]. The basic idea is that, for each Gauss-Newton iteration, a new reg-
ularization parameter corresponding to $L(P)$ in (35) will be estimated using the GCV, and eventually both the regularization parameter and solution will converge. The modified algorithm is listed in the following pseudo code:

Gauss-Newton with GCV:

1. select the initial solution
2. For k = 1 to maximum Gauss_Newton iterations
   3. compute the Frechet derivative $J(P)$
   4. compute the gradient $g(P)$
   5. compute $\beta(k)$ using the GCV
   6. compute the increment with the Conjugate Gradient: $dP$
   7. update solution: $P(k) = P(k-1) + \gamma \cdot dP$
   8. compute new Tikhonov functional with $P(k)$, $\beta(k)$
   9. compute old Tikhonov functional with $P(k-1)$, $\beta(k)$
   10. if (new Tikhonov < old Tikhonov)
       11. accept new solution: $P(k)$
   12. else
       13. choose a smaller step size: $\gamma$
       14. go to step 7
   15. if both $\beta(k)$ and and $P(k)$ converge
   16. exit
17. End for

2.2.4 Uniqueness of Solution

So far, the stability of inverse solution has been the primary concern. Situation becomes more complex with the nonuniqueness issue, because even in the ideal case with zero data error, certain formulation of inverse problem can lead to the existence of multiple solutions.

Let us consider a one dimensional linear elastic problem with the Dirichlet condition:

$$\frac{d}{dx} \left[ E(x) \frac{du}{dx} \right] = \frac{d}{dx} \left[ E(x) \epsilon(x) \right] = 0, \quad x \in [x_1, x_2].$$  (39)
\[ u(x_1) = u_1, \quad u(x_2) = u_2. \quad (40) \]

where \( \epsilon = \epsilon_{xx} = \frac{du}{dx} \), \( \sigma = \sigma_{xx} = E(x) \epsilon = E(x) \frac{du}{dx} \). \( u(x) \) is known in the domain, implying that \( \epsilon(x) \) is also known. This is a typical heat diffusion problem and \( E(x) \) cannot be determined uniquely from \( u(x) \) or \( \epsilon(x) \), because integration of (39) gives a solution of \( E(x) \) with an arbitrary constant \( C \):

\[ E(x) = \frac{C}{\epsilon(x)}. \quad (41) \]

A unique solution of \( E(x) \) can be obtained if we have either the Neumann data or a direct measurement of \( E(x) \) in the domain. For example, given the following Neumann condition:

\[ E(x) \frac{du}{dx} \bigg|_{x_1} = g, \quad (42) \]

\( C = g \) and a solution can be obtained everywhere from \( \epsilon(x) \):

\[ E(x) = \frac{g}{\epsilon(x)}. \quad (43) \]

Similarly, given a measurement \( E(x_a) \), \( x_a \in [x_1, x_2] \), a unique solution with \( C = E(x_a) \epsilon(x_a) \) is obtained,

\[ E(x) = \frac{E(x_a) \epsilon(x_a)}{\epsilon(x)}. \quad (44) \]

For a 2D elastic object, the equilibrium and constitutive equations are as follows:

\[
\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} = 0, \quad \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{xy}}{\partial x} = 0,
\]

\[
\sigma_{xx} = 2\mu \epsilon_{xx} + \lambda (\epsilon_{xx} + \epsilon_{yy}), \quad \sigma_{yy} = 2\mu \epsilon_{yy} + \lambda (\epsilon_{xx} + \epsilon_{yy}), \quad \sigma_{xy} = 2\mu \epsilon_{xy}. \quad (45)
\]

Combining two equations in (45) and substitute \( \sigma_{ij} \) by \( \epsilon_{ij} \), a single governing equation is obtained:

\[
\frac{\partial}{\partial x} [2\mu (\epsilon_{xx} + \epsilon_{xy}) + \lambda (\epsilon_{xx} + \epsilon_{yy})] + \frac{\partial}{\partial y} [2\mu (\epsilon_{yy} + \epsilon_{xy}) + \lambda (\epsilon_{xx} + \epsilon_{yy})] = 0. \quad (47)
\]
(47) is rewritten by replacing \((\mu, \lambda)\) with \((E, \upsilon)\) using (9) and obtain an equation with \(E\) as the only unknown (a constant Poisson ratio is used in the range of 0.490 - 0.495):

\[
\frac{\partial}{\partial x} [Ef_1(x,y)] + \frac{\partial}{\partial y} [Ef_2(x,y)] = 0,
\]

(48)

\[
f_1(x,y) = \frac{\epsilon_{xx} + \epsilon_{xy}}{1 + \upsilon} + \frac{\upsilon(\epsilon_{xx} + \epsilon_{xy})}{(1 + \upsilon)(1 - 2\upsilon)}, \quad f_2(x,y) = \frac{\epsilon_{yy} + \epsilon_{xy}}{1 + \upsilon} + \frac{\upsilon(\epsilon_{xx} + \epsilon_{xy})}{(1 + \upsilon)(1 - 2\upsilon)}.
\]

(49)

After differentiation, (48) is further transformed to:

\[
f_1(x,y) \frac{\partial E}{\partial x} + f_2(x,y) \frac{\partial E}{\partial y} = f_3(x,y)E.
\]

(50)

\[
f_3(x,y) = \frac{\partial f_1(x,y)}{\partial x} + \frac{\partial f_2(x,y)}{\partial y}.
\]

(51)

(50) is a typical first order linear partial differential equation with respect to \(E\). To obtain a unique solution of (50), the following Cauchy problem [109] has to be solved:

\[
f_1(x,y) \frac{\partial E}{\partial x} + f_2(x,y) \frac{\partial E}{\partial y} = f_3(x,y)E, \quad \Omega \in \mathbb{R}^2.
\]

(52)

\[C = \{x = h_1(t), \quad y = h_2(t)\}, \quad \Omega.
\]

(53)

\[E = h_3(t), \quad \text{on} \quad C.
\]

(54)

where \(C\) is the initial curve in \(\Omega\), defined by two given functions \((h_1, h_2)\). Function \(h_3\) on curve \(C\) defines the Cauchy data, and \(t\) is a real-valued parameter. The Cauchy theorem requires that \(C\) is not characteristic with respect to (52), i.e. \(C\) should not be tangent to any characteristic curve of (52).

The Cauchy theorem ensures the uniqueness of solution only in a local neighborhood that contains \(C\) with the Cauchy data defined on it. This local area may or may not correspond to the whole modeling domain, depending upon the partial differential equation, initial curve and the Cauchy data. If the Cauchy data can be specified along the domain boundary \((\partial \Omega)\), \(E(x,y)\) may be determined uniquely, provided that the specified boundary is not
characteristic and all integral curves of (52) pass through the modeling domain. A recent study by Barbone and Bamber [12] shows interesting results on special configurations of 2D elastic models. Readers are referred to [138, 88] for in-depth discussions on the uniqueness of inverse solution through Dirichlet-to-Neumann mapping. The key point is that, to obtain a unique elasticity estimation, the knowledge of either stress or elasticity itself is needed in the modeling domain (likely along the boundary). Displacement data alone will not be sufficient to guarantee a unique solution. If displacement and material properties are measurable on the boundary, stress/traction data can be derived and added to the data vector ($\vec{u}$) in the recovery algorithm.

2.3 Recovery Algorithm: Constrained Genetic Approach

Since the deterministic methods are sensitive to the initial conditions, the stochastic methods have also been studied recently. In this section, a constrained genetic algorithm (CGA) will be discussed that focus on the following issues: (1) encoding a 2D/3D spatial variable such as $E(x)$ in a 1D genetic computational unit; (2) expressing and incorporating qualitative prior knowledge in the genetic algorithm as a rank based constraint; (3) balancing contributions from the measurement data which has uncertainties and the prior knowledge through stochastic ranking.

2.3.1 Genetic Encoding

Given the finite element model of a deformed object, its Young’s modulus can be interpreted as a chromosome in a CGA. The Young’s modulus value of each element is encoded as a gene in the chromosome through a one-to-one mapping function (Figure 1). As a result, if the finite element model has $N$ elements, the corresponding chromosome would have $N$ genes. Each chromosome in the population pool represents a possible Young’ modulus distribution. If dynamic meshing (multigrid) is used in the finite element model, more sophisticated encoding schemes have to be considered that allow the size and shape of chromosomes to change adaptively during the evolution [48, 49, 115].
Finite Element Mesh

\[ E(i) = \text{Young’s modulus of the } i\text{th element} \]

<table>
<thead>
<tr>
<th>E1</th>
<th>E2</th>
<th>E3</th>
</tr>
</thead>
<tbody>
<tr>
<td>E4</td>
<td>E5</td>
<td>E6</td>
</tr>
<tr>
<td>E7</td>
<td>E8</td>
<td>E9</td>
</tr>
</tbody>
</table>

\[ \begin{align*}
  g1 &= E1 \\
  \vdots & \\
  g9 &= E9 \\
\end{align*} \]

Chromosome

Figure 1. Genetic Coding of the Young’s Modulus in a Finite Element Model.

This one-to-one mapping function is straightforward to implement and works well for a 1D finite element model. However, information about the spatial connections among the neighboring elements in a 2D/3D finite element mesh is completely lost in this encoding scheme. Computations that rely on spatial information cannot be performed properly. For example, in many studies, our interest is to identify only a few isolated areas of abnormal Young’s modulus values from the background of relatively uniform Young’s modulus distribution, and the information about genes’ spatial distribution is needed in a CGA to accomplish the task efficiently. More importantly, the smoothing effect represented by the derivative operators in the deterministic regularization cannot be realized in a rank-based CGA without such spatial information. To overcome those shortcomings of one-to-one mapping, a mechanism is designed to remember the original spatial connections among neighboring elements. An auxiliary link table is created for all chromosomes. Table 3 shows an example that uses a 4-neighbor connection for the 2D quadrilateral mesh in Figure 1. Similarly, a 3-neighbor connection can be considered for a triangle mesh. This link table will be used in the constrained mutation operation. Because the Young’s modulus has
Table 3. Link Table for Maintaining Original Spatial Connections.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Values</th>
<th>4-Neighbor Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>E1</td>
<td>g2 = E2, g4 = E4</td>
</tr>
<tr>
<td>g2</td>
<td>E2</td>
<td>g1 = E1, g3 = E3, g5 = E5</td>
</tr>
<tr>
<td>g3</td>
<td>E3</td>
<td>g2 = E2, g6 = E6</td>
</tr>
<tr>
<td>g4</td>
<td>E4</td>
<td>g1 = E1, g5 = E5, g7 = E7</td>
</tr>
<tr>
<td>g5</td>
<td>E5</td>
<td>g2 = E2, g4 = E4, g6 = E6, g8 = E8</td>
</tr>
<tr>
<td>g6</td>
<td>E6</td>
<td>g3 = E3, g5 = E5, g9 = E9</td>
</tr>
<tr>
<td>g7</td>
<td>E7</td>
<td>g4 = E4, g8 = E8</td>
</tr>
<tr>
<td>g8</td>
<td>E8</td>
<td>g5 = E5, g7 = E7, g9 = E9</td>
</tr>
<tr>
<td>g9</td>
<td>E9</td>
<td>g6 = E6, g8 = E8</td>
</tr>
</tbody>
</table>

continuous values that can vary in the range of several orders of magnitude, a real-valued (double) encoding approach is used [36].

2.3.2 Rank Based Constraint

In the stochastic framework of genetic computation, the ill-posedness of a nonlinear inverse problem implies many local plateaus in the landscape of the admissible solution space. By having a diverse population pool, genetic algorithms can explore a much wider solution space than the gradient methods and thus have a better chance to find the optimal solution by escaping the local minima. The ill-posed nature of the inverse problem also shows up as highly unstable solutions that are physically meaningless. To overcome those numerical difficulties associated with the ill-posedness, various constraints that represent prior knowledge must be imposed. In a CGA, constraints can take the form of penalty functions [23], which are equivalent to the regularization stabilizers or the preconditioners in the deterministic methods [30].

The objective function to be minimized by a CGA is formulated as a combination of the fitness function and the penalty function:

\[
obj(E) = \|Q(E) - b\|_n^2 + \theta P(E) \tag{55}
\]
where $Q(E)$ denotes the output from the forward model, $\mathbf{b}$ is the data vector, $n$ is the number of nodes on which the measurement is made, $P(E)$ is the penalty function, and $\theta$ is the weight coefficient.

In studies of spline and surface fitting, a smoothness constraint imposed on the spline and the surface is often formulated as a quadratic integral functional such as the Sobolev norm. Similarly, the Young’s modulus as a distributed parameter can be regarded as a spatial function that possesses certain degrees of continuity and the smoothness constraints can be realized as:

$$P(E) = \int_{R} |\nabla E(x)|^2 dx$$

where $\nabla$ could be any derivative operator such as the gradient or the Laplacian.

Knowledge about the relative Young’s modulus distribution can be obtained from either an expert’s visual assessment or from low level image cues such as intensity, color and texture. This type of prior knowledge is often expressed qualitatively and is not suited for penalty functions that are composed of continuous differential operators. To utilize the qualitative prior knowledge in a genetic algorithm, an alternative rank-based method is proposed to compute the penalty function. In each possible solution (chromosome), elements (genes) will be ranked based on their relative Young’s modulus values and their positions will be recorded in a sorted rank table. Similarly, the qualitative prior knowledge is transformed into another rank table. For each element, the difference of its ranked positions in the two rank tables is computed. The rank discrepancy of all elements is then summed up to represent the distance between the solution and the prior knowledge:

$$P(E) = \sum_{i=1}^{m} \|r_i - R_i\|^2$$

where $r_i$ is the rank position of element $i$ in the rank table for the solution, $R_i$ is the rank position of element $i$ in the rank table based on prior knowledge, and $m$ is the number of elements.
Figure 2 illustrates the ranking scheme with a simple two-dimensional model of four elements. As the qualitative prior knowledge, the Young's modulus value of each element is labeled as “high”, “mid” and “low”. This qualitative information is then transformed into a rank table where elements are sorted in descending order from “high” to “low”. If \( n \) elements \( (n > 1) \) have the same label, they all can have \( n \) potential rank positions, which will be determined by their counterparts in the solution rank table. For instance, both element (1) and element (4) are labeled as “low”, therefore their rank position can be either 3 or 4. In Solution 1, the ranks of element (2) and element (3) in the solution table match exactly with their ranks in the table of prior knowledge. However, for element (1), its rank in the solution table is 3, while its rank in prior knowledge table is 3 or 4. In the case of multiple ranks, the value that is closest to its counterpart in the solution table is selected, which is 3 for element (1). Similarly, for element (4), 4 is selected from its multiple prior rank of 3 or 4. The final penalty value for Solution 1 is zero \( (P(E) = 0) \). In contrast, Solution 2 has a Young’s modulus distribution that is quite different from the specified prior knowledge and thus a shuffled rank table, which leads to a higher penalty value of 6.

One potential problem with this rank-based penalty function is that two solutions of the same rank table are penalized equally, even if their absolute Young’s modulus values are quite different. As demonstrated in Figure 2, Solution 3 has a Young’s modulus distribution that is 10 times higher than that of Solution 2. But their solution rank tables are exactly the same and therefore receive the same amount of penalty of 6. This problem is related to the non-uniqueness of the inverse solution and can be resolved by incorporating both displacement and force in the data vector \( b \). In other words, Solution 2 and Solution 3 will have different fitness functions because of their different \( b \) and \( Q(E) \) values. This ambiguity can also be partially resolved by introducing another penalty function that specifies a range (both the upper bound and the lower bound) within which the Young’s modulus of an element is allowed to vary.

This rank-based approach has the advantage that it is intrinsically piecewise and helps preserve the parameter discontinuity (although a strong smoothness constraint can still be
### Prior Knowledge

<table>
<thead>
<tr>
<th>(1)</th>
<th>low</th>
<th>(2)</th>
<th>(3)</th>
<th>mid</th>
<th>(4)</th>
<th>low</th>
</tr>
</thead>
</table>

### Prior Rank Table

<table>
<thead>
<tr>
<th>Element ID</th>
<th>Prior Rank</th>
<th>Prior Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td>high</td>
<td>(2)</td>
</tr>
<tr>
<td>(3)</td>
<td>mid</td>
<td>(3)</td>
</tr>
<tr>
<td>(1)</td>
<td>low</td>
<td>(1)</td>
</tr>
<tr>
<td>(4)</td>
<td>low</td>
<td>(4)</td>
</tr>
</tbody>
</table>

### Solution 1

<table>
<thead>
<tr>
<th>Element ID</th>
<th>E (kPa)</th>
<th>Solu. Rank</th>
<th>Prior Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td>40</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(3)</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>(1)</td>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>(4)</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

**Penalty of Solution 1**

\[
p(E) = |3-3|^2 + |1-1|^2 + |2-2|^2 + |4-4|^2 = 0
\]

### Solution 2

<table>
<thead>
<tr>
<th>Element ID</th>
<th>E (kPa)</th>
<th>Solu. Rank</th>
<th>Prior Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4)</td>
<td>35</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>(2)</td>
<td>22</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>(3)</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>(1)</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

**Penalty of Solution 2**

\[
p(E) = |4-4|^2 + |2-1|^2 + |3-2|^2 + |1-3|^2 = 6
\]

### Solution 3

<table>
<thead>
<tr>
<th>Element ID</th>
<th>E (kPa)</th>
<th>Solu. Rank</th>
<th>Prior Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4)</td>
<td>350</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>(2)</td>
<td>220</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>(3)</td>
<td>60</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>(1)</td>
<td>40</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

**Penalty of Solution 3**

\[
p(E) = |4-4|^2 + |2-1|^2 + |3-2|^2 + |1-3|^2 = 6
\]

Figure 2. The Rank Based Scheme for Incorporating Qualitative Prior Information.
The numbers in the parentheses denote element ID and the numbers at the lower-right corner of elements represent the Young’s modulus.
imposed in the areas of little Young’s modulus variation). This approach is particularly suitable for studies that aim at identifying and quantifying property abnormalities. In our previous studies on burn scar assessment [135], qualitative prior knowledge was collected from physicians who isolate and rate the scars based on a relative rating scale. Automatic methods for extracting the information directly from image intensity, texture and color can also be considered.

2.3.3 Balancing Fitness and Penalty by Stochastic Ranking

In the constrained objective function (55), the fitness and penalty terms are computed on different quantities. The fitness is measured as difference of displacement (meter), while the penalty is based on the difference of rank orders (unitless). An optimal weight coefficient \( \theta \) is needed to balance their contributions. If \( \theta \) is too small, the data noise will not be penalized enough and the resulting solution becomes unstable. If \( \theta \) is too large, solution will be forced into a smoothed prior space and most of the data signals, i.e. the information used to infer the Young’s modulus, will be lost due to over-penalization. In the deterministic domain, several choices are available for determining the optimal regularization parameter. If the noise level of observation data is known, methods based on the discrepancy principle [86] such as the Miller method [84] can be considered. In case that information about the data noise is not available, heuristic methods such as the generalized cross validation [140] or the L-curve [55] method are commonly used. However, those methods are deterministic and not suited for handling rank-based qualitative data.

In a recent study on the constrained evolutionary optimization, Runarsson and Yao [110] presented a stochastic method to strike a balance between the fitness and penalty functions, without the need of computing the weight coefficient explicitly. Determination of an optimal \( \theta \) is related to the dynamic and adaptive ranking of individual chromosome in a population. Ranking is based on the relative dominance of either the fitness function or the penalty function between two adjacent individuals. The balance of dominance is achieved by introducing a bubble-sort-like dynamic ranking procedure for an individual
to win a comparison. It is found that this stochastic method is well suited for handling qualitative prior knowledge in elasticity reconstruction. Readers are referred to [110] for detailed discussions on the method and related implementation issues.

2.3.4 Genetic Operators

To minimize the objective function (55), Several important genetic operators are specified.

Mutation: A standard Gaussian mutation operator is used,

\[ g_j^* = g_j + \sigma N(0, 1) g_j \]  \hspace{1cm} (58)

where \( g_j \) is the value of gene \( j \) before the mutation, \( g_j^* \) is the value of gene \( j \) after the mutation. \( N(0,1) \) is a random Gaussian number (mean = 0, standard deviation = 1) and \( \sigma \) is the mutation step size. The dynamic control of mutation step size is determined by a predefined decay rate \( \tau \) as:

\[ \sigma(k+1) = \tau \sigma(k), \]  \hspace{1cm} (59)

where \( k \) is the generation counter. \( \tau \) is usually set in the range of 0.99 - 1.0. A smaller value of \( \tau \) may help speed up the convergence rate but has the risk of premature convergence. Other than Gaussian mutation, other types of real-coded mutation operators are also considered such as uniform mutation and step mutation. Those mutation approaches did not show any advantage over Gaussian mutation in terms of both solution accuracy and computation efficiency.

Experiments were conducted with the method of dynamically setting the mutation probability (\( P_m \)) based on population statistics and no significant improvement was found in terms of the solution accuracy. One of the purposes of having a relatively high mutation rate is to maintain the population diversity and prevent premature convergence. To utilize the spatial connectivity among finite elements as recorded in the link table (Table 3), each gene will be compared with its neighboring genes after the mutation. If the gene has a Young’s modulus value that is much higher/lower than the maximum/minimum of its neighboring
genes, the gene will be assigned a mean value of its neighbors. This process has the effect of partially restoring the smoothness property of the quantitative regularization functional that is missed in the rank-based penalty function (57).

**Crossover:** It is also found that the one-point crossover operator and the multiple-point crossover operator performed equally well, at least for this particular inverse problem setting. The crossover probability ($P_c$) is fixed to 0.7. The one-point crossover function is implemented in a traditional fashion: children are generated by joining two parents at a randomly selected crossover position and then swapping each sides.

**Parent Selection and Replacement:** The parent selection operator is implemented as tournament selection ($k=2$). Experiments were conducted with a wide range of replacement ratios (0.05 - 1.0), i.e., the percentage of parent population to be replaced by new chromosomes. Given a population of size $S$ and a replacement ratio of $r$, the number of parents to be replaced is: $N = rS$. Smaller replacement ratios ($<0.3$) did not yield satisfying results because of the lack of contributions from new chromosomes to the population diversity. However, no significant difference was observed with replacement ratios ranging from 0.5 to 1.0. On the other hand, the higher the replacement ratio, the longer the simulation time.

In [110], it was shown that the stochastic ranking probability ($P_f$) in the range of (0.45 - 0.475) gave the best result for 13 benchmark functions. Two $P_f$ values (0.45 and 0.475) were tested in experiments, and no noticeable improvement or degradation was observed in the algorithm’s performance. So a $P_f$ value of 0.45 is used in all of the following comparison studies.

An elitism strategy is also enabled during the replacement operation, in which some elite members of the old generation are chosen to survive to the next generation without competition (potentially being replaced by a better offspring). Given the genetic parameters specified above, trial tests with different elite ratios (1% - 15%) showed that 3% gave a slightly better result on average (although very marginal). So an elite ratio of 3% is chosen in all the experiments.

The pseudo code for the constrained genetic algorithm is as follows:
/* population size */
S = 0.8 /* replacement ratio */
τ = 0.997 /* mutation decay rate */
σ = 0.2 /* initial mutation step size */
Pc = 0.2 /* one-point crossover probability */
Pf = 0.45 /* stochastic ranking probability */
for i = 1 to population S(0)
    chromosome (i) = a value from a Gaussian distribution
end for
for k = 1 to maximum generation
    for i = 1 to population S(k)
        E (i) = chromosome (i)
        call forward finite element model with E (i)
        compute fitness and penalty functions for chromosome (i)
    end
    rank the whole population S(k) with stochastic ranking (Pf)
    pick a set of parents from population S(k) through tournament selection
    use one-point crossover to generate N offspring with a probability of Pc
    for i = 1 to N offspring
        Mutate chromosome(i): g = g + σN(0, 1) g
        E (i) = chromosome (i)
        call forward finite element model with E (i)
        compute fitness and penalty functions for chromosome (i)
    end
    pick N parents from population S(k) through tournament selection
    rank N picked parents and N offspring with stochastic ranking (Pf)
    for i = 1 to N
        for j = 1 to N
            if rank of parent (i) ≤ rank of offspring (j)
                then
                    replace parent (i) with offspring (j)
                    remove offspring (j)
                end
            end
    end
    a new population S(k+1)
    update mutation step size: σ(k+1) = τσ(k)
end
2.4 Deterministic or Stochastic Approach

2.4.1 Experiments with GA, CGA and GNM

A synthetic numerical model is used to compare the efficacy of three recovery algorithms: a regular genetic algorithm (GA), the proposed constrained genetic algorithm (CGA), and the deterministic Gauss-Newton methods (GNM). The forward model is a two dimensional thin shell (5 cm by 5 cm) and is discretized to 61 nodes and 100 triangle elements (Figure 3). A small square at the upper-left corner of the model is the abnormal area and has a higher Young’s modulus value of 250 kPa. The rest of background elements have a uniform Young’s modulus of 50 kPa. The element type is 3-node triangular shell without out-of-plane deformation ($w$ component in $z$ direction). A linear scheme is used in the interpolation (shape) functions:

\[
    u(x,y) = H_i u_i + H_j u_j + H_k u_k, \quad (60)
\]

\[
    v(x,y) = H_i v_i + H_j v_j + H_k v_k, \quad (61)
\]

\[
    H_i(x,y) = \frac{1}{2A} \left[ (x_j y_k - x_k y_j) + (y_j - y_k)x + (x_k - x_j)y \right], \quad (62)
\]

\[
    H_j(x,y) = \frac{1}{2A} \left[ (x_k y_i - x_i y_k) + (y_k - y_i)x + (x_i - x_k)y \right], \quad (63)
\]

\[
    H_k(x,y) = \frac{1}{2A} \left[ (x_i y_j - x_j y_i) + (y_i - y_j)x + (x_j - x_i)y \right], \quad (64)
\]

where $H$ denotes the interpolation function, $A$ is the area of triangle element (see Figure 3 (c)), $(i,j,k)$ are used to index three nodal positions and $(x,y)$ represent the location in a global 2D coordinate.

Boundary conditions are specified as follows: Forces (concentrated loads) are applied to the nodes of top boundary (1.2N on each node). Displacement constraints (Dirichlet type) are fixed to zero on the nodes of bottom boundary (see Figure 3 (a)). Those boundary conditions are used to generate noise-free displacement data on each node. As a result, the data vector $b$ includes forces of top boundary nodes as well as displacements of all nodes other than those on the bottom boundary.
Three algorithms (GA, CGA and GNM) were tested with the noise-free data to study to what degree the Young’s modulus in the abnormal area can be reproduced. The data vector includes both displacement and force to ensure the uniqueness of the inverse solution. As prior knowledge, four elements in the upper-left square were identified as a potential abnormal area. These four elements were labeled as “high” and other elements in the background were labeled as “low” (Figure 3).

3% white noise was then added to the noise-free data. Again, the GA, CGA and GNM were studied using the noisy data to assess the performance. For this small 2D model, we found that a population size of 50 is large enough for the GA and CGA to converge to a good solution with less than 300 generations. The performance of the three algorithms was evaluated based on the shape and absolute Young’s modulus value of the reconstructed abnormal area. For the GA and CGA experiments, we ran 30 simulations for each case and therefore the conclusions are drawn from the average of 30 simulations.

The reconstructed Young’s modulus using the three algorithms are shown in Figure 4. Given the noise-free data, all algorithms were able to find a near-optimal solution with the abnormal area well defined in terms of both the geometry and absolute values. The results from the CGA and GNM are almost identical. Given the noise corrupted data, both

Figure 3. Synthetic Model and the Qualitative Prior Knowledge.
In (a), the Young’s modulus of the background area is 50 kPa. In (b), the background area is labeled as “low”. (c) illustrates the linear interpolation scheme of triangle element used in equations (31) to (35). \((u, v)\) are displacement components in a 2D coordinate, \((x, y)\) denotes the location of a point and \((i, j, k)\) are used to index three nodes.
Figure 4. Reconstructed Young’s Modulus by GA, CGA and GNM.

The figures are plotted using greyscale, with black and white indicating high and low Young’s modulus values, respectively. The figures were generated using the GMV (the General Mesh Viewer) visualization software [47] that may add certain degrees of smoothing effect on the Young’s modulus distribution (the Young’s modulus value inside an element is constant).
GNM and CGA still successfully identified the abnormal area, although the boundaries were blurred due to the smoothing effect. However, the GA failed to find an acceptable solution as evidenced by the multiple abnormal areas (Figure 4 (f)) that do not exist in the forward model.

A Gaussian distribution with a mean of 50 kPa was used as the initial condition for all the experiments. In the experiments of GNM, the regularization parameter was set to 1.93E-8, and the smoothness matrix was the zero order finite difference matrix, i.e. the identity matrix. If necessary, higher order matrix can be used. For example, for an evenly spaced solution vector of 4 elements, the first order finite difference matrix is:

\[
\begin{bmatrix}
1 & -1 & 0 & 0 \\
-1 & 2 & -1 & 0 \\
0 & -1 & 2 & -1 \\
0 & 0 & -1 & 1
\end{bmatrix}
\]

(65)

Since the higher order matrices have a strong tendency to over-smooth the vector, we used the identity matrix in all the experiments. The simulation was stopped when either a maximum Gauss-Newton iteration number is reached (100 for this experiment) or a stopping criteria is met. The stopping rule is defined as follow:

\[
convergence = \|\text{grad}(E)\|/\|\psi(E_k)\| < \text{tolerance}
\]

(66)

where \(\psi(E_k)\) is the linearized Tikhonov functional with respect to the current solution \(E_k\) and \(\text{grad}(E)\) is the gradient operator of Tikhonov functional:

\[
\text{grad}(E) = D^T(E_k)(Q(E_k) + D(E_k)(E - E_k) - b) + \alpha W^T W (E - E_{prior}).
\]

(67)

The convergence tolerance was set to 1.9e-10 for all the experiments with GNM, including those in the next section. This tolerance value is relatively stringent. A good initial
solution guess usually allows the algorithm to converge with less than 20 iteration steps, while a poor initial point results in the maximum iteration number to be reached (diverge).

One important issue in using the rank-based CGA for the Young’s modulus reconstruction is the quality of prior knowledge. For example, if the prior knowledge used in the algorithm does not correspond to the true Young’s modulus distribution very well, how much will the final solution be affected? An experiment was carried out with slightly biased prior knowledge. The model configuration, initial condition and noisy data were the same as those in the previous experiments. The biased prior and the reconstructed Young’s modulus are shown in Figure 5. A rectangle of four elements was identified as the potential abnormal area and was labeled “high” incorrectly, but still close to the true abnormality square. It can be seen that the shape of the reconstructed abnormal area is influenced by the biased prior, although the CGA still managed to find the true abnormality square. The prior knowledge has a tendency to attract the inverse solution towards itself. In applications, we do not expect the prior knowledge to be severely biased, but its side-effect should not be ignored and must be corrected as much as possible.

It is worth noting that there are two basic types of prior involved in the Young’s modulus reconstruction with the CGA. One is the smoothness prior that is implicit in the quadratic
number of elements = 900, number of nodes = 961, number of elements in the abnormal area = 60. (b) illustrates the interpolation scheme of quadrilateral element used in the equations (39) to (44). \((u,v)\) are displacement components, \((x,y)\) denotes the location in global coordinate system, \((s,t)\) represents the local coordinates within an element, and \((i,j,k,l)\) are used to index four nodes.

penalty term, and the other one is explicitly specified as the ranked prior table. The former is a global constraint while the later is local in nature. Therefore, the influence of a bad local prior cannot be remedied by adjusting the global regularization parameter or stochastic ranking in CGA. More study is needed to find more sophisticated and effective methods so that the biased prior knowledge can be handled properly.

2.4.2 Experiments with CGA and GNM

There have been debates on whether the genetic algorithms have any advantage over the traditional gradient decent methods, because the computational efficiency of the genetic algorithms is often a concern, especially for large-scale problems. To understand how various factors might affect the performance of a CGA in solving inverse problems, we carried out another set of experiments using both the CGA and GNM.
The forward model is similar to the one that we used in the previous experiments, except a much larger parameter space of 900 elements (Figure 6), and hence it is more difficult to find an optimal solution of the corresponding inverse problem. The model has a size of 10 by 10 (cm) and is meshed with quadrilateral elements. 60 elements in the center of the model are the abnormal area with a Young’s modulus value of 400 kPa, and the background elements have a lower Young’s modulus value of 50 kPa. The element type used is 4-nodes quadrilateral shell. Isoparametric mapping and the Lagrange formula are used in the interpolation (shape) functions:

\[
\begin{align*}
    u(x, y) &= H_i u_i + H_j u_j + H_k u_k + H_l u_l, \\
    v(x, y) &= H_i v_i + H_j v_j + H_k v_k + H_l v_l,
\end{align*}
\]

\[
\begin{align*}
    H_i(s, t) &= \frac{1}{4}(1 - s)(1 - t), \\
    H_j(s, t) &= \frac{1}{4}(1 + s)(1 - t), \\
    H_k(s, t) &= \frac{1}{4}(1 + s)(1 + t), \\
    H_l(s, t) &= \frac{1}{4}(1 - s)(1 + t),
\end{align*}
\]

where \( H \) denotes the interpolation function, \((i, j, k, l)\) are used to index four nodal positions, \((s, t)\) denote the local coordinates of quadrilateral element, and \((x, y)\) represent the global coordinates (see Figure 6 (b)).

The boundary conditions were specified as follows: The model was compressed by forces exerted on the top boundary (0.5N on each node) while the bottom boundary was fixed (Dirichlet type with all displacement components set to zero).

First, an appropriate population size needs to be determined. A series of experiments was conducted with different population sizes (10, 20, 50, 100, 500, 1000). The CGA simulation was stopped whenever the objective function reached a preset minimum (0.03 in all experiments). Each simulation was started with a good initial guess to ensure that the solutions obtained are equally mechanically meaningful (close to the true solution). For
Table 4. CGA Experiments with Different Population Sizes.

<table>
<thead>
<tr>
<th>Population Size</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generations</td>
<td>1563</td>
<td>825</td>
<td>637</td>
<td>559</td>
<td>430</td>
<td>195</td>
</tr>
<tr>
<td>Time (min)</td>
<td>17.6</td>
<td>24.5</td>
<td>44.2</td>
<td>82.3</td>
<td>286.7</td>
<td>339.5</td>
</tr>
</tbody>
</table>

Results are reported as the average of 30 runs for each population size. All simulations were started with a good initial point and stopped when the objective function reached a preset value.

It is clear that the CGA with a larger population size can reach a specified minimum with many less generations than the CGA with a smaller population size, but requires much more simulation time. In other words, for the CGA with a larger population size, each generation takes more time due to the computational cost associated with its large pool. Although the CGA with a smaller population size uses much less execution time, this benefit is often offset by the concern of premature convergence due to a less diversified pool that is common to smaller populations. Taking the above factors into account, a population size of 100 was chosen that can maintain enough population diversity, yet still has a “reasonable” computational cost.
Table 5. Performance of CGA and GNM with Different Initial Conditions.

<table>
<thead>
<tr>
<th>Initial Condition $N(\mu, \sigma)$</th>
<th>Num. of Simul.</th>
<th>Converg. Ratio</th>
<th>Average Iterat. Steps</th>
<th>Average Converg. Time (min)</th>
<th>Num. of Simul.</th>
<th>Converg. Ratio</th>
<th>Average Generat. Count</th>
<th>Average Converg. Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(50, 10)$</td>
<td>100</td>
<td>95%</td>
<td>8</td>
<td>9</td>
<td>30</td>
<td>100%</td>
<td>315</td>
<td>45</td>
</tr>
<tr>
<td>$N(100, 50)$</td>
<td>100</td>
<td>82%</td>
<td>17</td>
<td>20</td>
<td>30</td>
<td>97%</td>
<td>549</td>
<td>76</td>
</tr>
<tr>
<td>$N(200, 50)$</td>
<td>100</td>
<td>44%</td>
<td>25</td>
<td>33</td>
<td>30</td>
<td>93%</td>
<td>834</td>
<td>137</td>
</tr>
<tr>
<td>$N(300, 50)$</td>
<td>100</td>
<td>17%</td>
<td>34</td>
<td>39</td>
<td>30</td>
<td>91%</td>
<td>1494</td>
<td>251</td>
</tr>
<tr>
<td>$N(400, 50)$</td>
<td>100</td>
<td>3.8%</td>
<td>52</td>
<td>61</td>
<td>30</td>
<td>88%</td>
<td>3483</td>
<td>582</td>
</tr>
<tr>
<td>$N(500, 50)$</td>
<td>100</td>
<td>4.1%</td>
<td>48</td>
<td>56</td>
<td>30</td>
<td>84%</td>
<td>7347</td>
<td>1249</td>
</tr>
</tbody>
</table>

For each of the 6 initial condition distributions (row), we ran 100 GNM simulations and 30 CGA simulations from which the convergence ratio, the average iteration steps (generation count) and the average convergence time were computed. The convergence ratio was computed as the number of converged simulations divided by the total number of simulations. Only the results of converged simulations were used in the computation of the average iteration steps (generation count) and the average convergence time. Note that the true solution is 400 kPa in the abnormal area, and 50 kPa in the background, with a mean of 73 kPa.

The behavior of a CGA that has $M$ chromosomes in its pool can be roughly viewed as being equivalent to $M$ local gradient methods that are run simultaneously. The difference is that the chromosomes in a CGA have certain degrees of information sharing among themselves through the crossover operation, while the gradient methods running in parallel are completely independent of each other. It would be interesting to compare the performance of a CGA of $M$ chromosomes with that of $M$ independent gradient methods.

The gradient methods are known to be sensitive to the initial conditions. A guess is used that obeys the normal distribution law, $N(\mu, \sigma)$. In other words, the initial Young’s modulus values assigned to 900 elements should have a sample mean of $\mu$ and a standard deviation of $\sigma$. The experiments used 6 such initial conditions (unit = kPa): $N(50, 10)$, $N(100, 50)$, $N(200, 50)$, $N(300, 50)$, $N(400, 50)$, $N(500, 50)$, in the order of gradually deviating from the true solution (Table 5). All of the experiments were performed on a Sun Sparc ultra 5 machine (248 MHz, 2560 Mb).
The experiments are designed as follows: for each of the 6 initial condition distributions, we ran 30 CGA simulations and 100 GNM simulations and report the results in Table 5. So, each row in Table 5 represents the average performance of 30 CGA simulations and 100 GNM simulations, all with their initial guesses generated by the same normal distribution function. A GNM simulation was stopped if the convergence criteria (37) was satisfied with a specified tolerance value (1.9e-10) or a preset maximum iteration number was reached (400 in this experiment). Similarly, a CGA simulation was stopped if the objective function value was below a threshold (0.05) or the change of objective function between two consecutive generations was below a threshold (0.0005). The two algorithms were then evaluated based on three performance indexes: Convergence ratio, iteration steps (generation count) and convergence time.

The convergence ratio is defined as the number of converged simulations divided by the total number of simulations. For the GNM, the convergence ratio dropped rapidly as the initial condition moved away from the true solution (the third column in Table 5). For example, starting with two relatively bad initial conditions \( N(400, 50) \) and \( N(500, 50) \), most of the GNM simulations failed to converge (or converged to a local minimum far away from the true solution). In contrast, the convergence ratio of the CGA is more stable. The performance deterioration of the CGA caused by a poor initialization is much less drastic (the seventh column in Table 5).

The gain in the convergence ratio of a CGA comes with a price of longer convergence time. A single CGA simulation took much more time to converge than a single GNM simulation did (the fifth and ninth columns in Table 5). On the other hand, if we view the outcome of a single CGA of 100 chromosomes as being equal to that of 100 independent GNMs, then running 100 independent GNMs seems more expensive than running a single CGA. However, this comparison could be misleading. It should be stressed that the bottleneck of a GNM simulation is the computation of the Jacobian matrix. In our current implementation, a finite difference approximation method is used, which implies that for a problem of \( m \) parameters, at least \( m^2 \) calls of the forward model were needed for each
iteration step. This accounts for about 92% of the total GNM simulation time. If a more efficient method such as the adjoint state method is used [90], the simulation time of a GNM can be significantly reduced, and the advantage of a single CGA against 100 independent GNMs will likely disappear. Furthermore, if we are dealing with a complex 3D problem, a larger population size might be needed, which will also increase the computational cost of the CGA.

The iteration steps and generation count show similar trends in that they increased gradually as the initial condition got worse. But they are often implementation-dependent and thus are less informative about the algorithm’s behavior than the convergence ratio and convergence time.

From the above experimental results and analysis, several observations can be made:

1. The CGA is more robust than the GNM in the sense that the CGA is less sensitive to the initial conditions.

2. A single GNM is more efficient than a single CGA in terms of the convergence time. If a good initialization is available, the GNM is a more attractive choice.

3. A better solution strategy is to combine the GNM and the CGA in a way that both their strengths can be fully utilized. For instance, the CGA can be used to find a relatively good initial approximation, upon which the GNM can be launched to speed up the search.

To illustrate the convergence behaviors of the CGA and the GNM, the minimization steps of four example runs were plotted in Figure 8. Since it is difficult to visualize the solution landscape in high dimensions, the space was simplified in the following way: For each solution vector, two averaged values were computed, one for the abnormal area and one for the background:

\[ E_a = \sum_{i=1}^{I} E_i, \quad E_b = \sum_{j=1}^{J} E_i, \]

where \( I \) and \( J \) are the number of elements in the abnormal area and the background, respectively. Using the two averaged values as new parameters, each solution and the
The corresponding objective function can be plotted in a 2D contour map. Each point in the 2D map represents a possible solution. The true solution shown as a black circle is located at the global minimum of solution space (objective function = 0).

As shown in Figure 8 (a,c), if starting with a point close to the true solution, the GNM took only a few steps to reach the true solution while the CGA could take a couple of hundred generations. Given a poor initial guess, Figure 8 (d) depicts the convergence path of a successful CGA simulation, while Figure 8 (b) shows a GNM simulation that failed to converge to the true solution (stuck in a local minimum).

It should be emphasized that Figure 8 is only used for illustration. The actual solution space of 900 parameters is far more complex than the 2D contour map. There are numerous local extrema that cannot be seen in this extremely simplified two parameter space.

The results of 100 GNM simulations are plotted in Figure 9. The 100 examples are randomly picked from the total of 600 simulations (we had 100 simulations for each of the six initial conditions). The plot helps us visualize the influence of initialization on the behavior of local gradient methods. It is apparent that as the initial condition deviated away from the true solution, more and more GNM simulations failed to converge (trapped in local minimum).

Based on the above comparison results, the deterministic Gauss-Newton method is used in all of the following application experiments, due to the consideration of its computational efficiency.
Figure 8. Convergence Paths of CGA and GNM with Different Starting Points.

The contours (dashed lines) represent the values of objective function. The black circle indicates the position of the true solution. Note that this is a VERY simplified view of the solution space. The actual solution landscape in higher dimensions is much more complex and is characterized by numerous local minimums.

Figure 9. Results of GNM with Various Initial Conditions.

The circles and filled triangles represent the cases that converged and diverged, respectively.
3.1 Motion Parameters in Physical Model

Let's first examine the motion equations and identify the physical motion parameters that can be recovered and used for motion analysis. Justifications will then be provided for choosing certain parameters based on both physical and mathematical considerations.

3.1.1 Physical Motion Parameters

In motion equations (7, 8), four types of quantity can be identified: Displacement (u), material properties (ρ, E, ν), boundary conditions (¯u, g) and body force (F). Displacement vector represents the solution of motion equations and certainly is not a candidate of motion parameter. Body forces (mainly the gravity) can be considered as known constants in most applications and will not be treated as motion parameters. By the classical elasticity theory [120, 76], we can seek solutions of u throughout the modeling domain by specifying appropriate material properties and boundary conditions. Therefore, material properties and boundary conditions are the possible choices of motion parameter.

Direct measurement of actual forces (Neumann data g) is usually not available. But displacements (Dirichlet data ¯u) can be estimated from image sequences and robust algorithms have been developed using either feature correspondence or optical flow. With the displacement data, physical model-based motion analysis can be characterized as a Dirichlet boundary value problem. Material properties are therefore chosen as physical motion parameters.

Although all motion parameters can be recovered simultaneously, a sequential approach seems more practical in which all parameters are set to constant, except the one to be esti-
mated. For many biological materials, Poisson’s ratio $\nu$ can be approximated as a constant. For example, soft tissues can be viewed as near-incompressible, and a $\nu$ value in the range of 0.490 - 0.495 is computationally safe. Mass density can be acquired from literature or direct measuring. Note that mass density is only needed in a transient simulation, and is canceled out with the temporal derivative in static case. The focus is on recovering and using the Young’s modulus as the only physical parameter to facilitate motion modeling.

3.1.2 Physical and Mathematical Justifications

Motion lies in the core of mechanics and has been studied from both kinematic and dynamic perspectives. In computer vision and graphics, motion analysis is pursued along the same routes: both kinematic and physical models are used. However, the parameters (constraints) introduced to vision and graphic models are often based on geometrical considerations and may not be physically realistic. the choice of physical motion parameters is motivated by both mathematical convenience and physical soundness.

*Constraint, Constitutive Equation and Material Property:* Field equations (including motion equation) derived from the law of conservation (mass, momentum, energy and entropy) have two important features: (1) they are local, and (2) they are valid for all types of medium irrespective of their material constitutions. Mathematically, this results in an underdetermined system which requires extra constraints to ensure a unique solution. Although any form of constraint can be applied, almost all constitutive equations in continuum mechanics are postulated based on experimental observations and mathematical idealization. From a physical point of view, deformation is a result of object’s response to external forces, so it is natural to formulate the constitutive equation as a relationship between stress and deformation. The Hooke’s law for elastic solids and the Stoke’s law for viscous fluids are well known examples. Constitutive equations are derived from macroscopic observations, and have supports from fundamental constitutive axioms as well as microscopic statistical dynamics. Intuitively, the strong internal constraint (resistance to being deformed) of solids is related to the strong inter-atomic bounds, while the fact that gases exhibit little resis-
tance to external pressure is a consequence of their weak intermolecular bounds. Rigorous treatment of the constitutive theory is not in the scope of this paper. What is emphasized is that imposing motion constraints through the constitutive equation has a firm root in fundamental physics and is well justified [133, 24].

The stronger the constraint the simpler the motion equation. For example, equation (7) derived with two linear constraints (3, 6) is much easier to handle than the original field equation (1). Note that (3) is a kinematic description while (6) is a constitutive equation. With an even stronger constraint of \( \varepsilon = 0 \), (7) becomes the motion equation of rigid solids where the divergence of stress disappears (more precisely, the rigidity condition is \( G = 0 \), with \( G \) as the Green strain tensor: \( G = \frac{1}{2} [\nabla \mathbf{u} + (\nabla \mathbf{u})^T + \nabla \mathbf{u} (\nabla \mathbf{u})^T] \)). On the other hand, constraint must be significantly relaxed for fluid motion. Unlike elastic solids that can remember and restore the original configuration after stress is removed, most fluids are memoryless with respect to their initial configuration and thus fluid motion is often studied with Euler descriptions (velocity and strain-rate tensor) rather than Lagrangian descriptions (displacement and Cauchy-Green strain tensor). Because of the weak constraint, fluid motion equations are more difficult to handle.

**Adaptive Smoothing of Motion Field:** Another motivation for studying material properties as physical motion parameters is that they play an important role in the regularized motion synthesis. If motion field is viewed as an unknown function of space, \( \mathbf{u}(\mathbf{x}), \mathbf{x} \subset \mathbb{R}^n \), one of the goals of motion synthesis is to estimate a relatively dense distribution of \( \mathbf{u} \) from a sparse observation \( \bar{\mathbf{u}} \). Like many vision problems [106], motion synthesis is likely ill-posed and must be regularized:

\[
\mathbf{u} = \arg\min \{ \| \mathbf{u} - \bar{\mathbf{u}} \|^2 + \beta R(\mathbf{u}) \},
\]

where \( \mathbf{u} \) is the motion field to be synthesized, \( \bar{\mathbf{u}} \) is the observed sparse and noisy motion data, \( \beta \) is the regularization parameter, and \( R(\mathbf{u}) \) is a regularization (smoothness) function.

If discontinuity exists in the motion field, an adaptive regularization scheme is needed to avoid oversmoothing. Adaptive smoothing has been intensively studied in surface recon-
struction [46, 16, 128, 129], and the basic algorithm can be represented as an optimization of the following energy function:

\[
W(f) = \|Hf - b\|^2 + \beta \int_{\Omega} w(x) \phi(Df(x)) \, dx,
\]

(75)

where \(f(x), x \subset R^2\), is a piecewise smooth surface, \(H\) is a linear operator, \(b\) is the observation data, \(w(x)\) is a weighting function, \(Df(x)\) is a derivative of \(f\), and \(\phi\) is an energy function defined on \(Df(x)\). Assuming that motion field can also be approximated as a piecewise smooth function, motion synthesis problem (74) can be formulated as a minimization of the following functional:

\[
W(u) = \|u - \bar{u}\|^2 + \beta \int_{\Omega} w(x) \phi(Du(x)) \, dx.
\]

(76)

On the other hand, motion of a continuum can also be studied with the variational approach. Without losing generality, a static case of linear elastic body in the integral form is considered:

\[
W(u) = \int_{\Gamma_1} \beta_1(u - \bar{u}) + \int_{\Omega} \beta_2(\varepsilon - \frac{1}{2}(\nabla u + (\nabla u)^T) + \int_{\Omega} \left[ \frac{1}{2} \lambda (tr(\nabla u + (\nabla u)^T))^2 + \frac{1}{4} \mu (\nabla u + (\nabla u)^T) \cdot (\nabla u + (\nabla u)^T) \right] + \int_{\Omega} Fu + \int_{\Gamma_2} F_s u,
\]

(13b)

where \(F_s\) is surface force, \(\beta_1\) and \(\beta_2\) are Lagrangian multipliers, \(\Omega\) denotes object’s configuration with a boundary split into two sections: \(\Gamma_1\) for Dirichlet condition and \(\Gamma_2\) for Neumann condition. Using variational calculus (\(\delta W = 0\)), motion equations (7, 8) can be obtained as the Euler equations of (13b). This derivation is also known as the Hellinger-Reissner variational principle in elasticity theory [58].

Comparing (76) and (13b), it can be seen that the residual norm \(\|u - \bar{u}\|^2\) is equivalent to the Dirichlet condition \(\int_{\Gamma_1} \beta_1(u - \bar{u}),\) and the smoothness norm is related to the strain...
energy, the third integral in (13b). It is interesting to note that material properties, \( \mu(x) \) and \( \lambda(x) \), play a role similar to \( w(x) \) in (76) as the adaptive regularization controller, although the strain energy is more involved than many energy functions \( \phi \) used in surface reconstruction.

Material properties being adaptive regularization controller is more obvious in differential equations. In the study of multi-scale image description, Perona and Malik [102] proposed an adaptive smoothing scheme that utilizes the anisotropic diffusion equation:

\[
\frac{\partial f(x, t)}{\partial t} = \nabla \cdot (c \nabla f), \tag{77}
\]

\[
c(x, t) = R(\nabla f). \tag{78}
\]

where \( f(x, t) \) is image intensity, \( x \) is image coordinate, \( t \) is a scale variable, and \( c \) is the diffusion coefficient that controls the degree of smoothness. Note that \( c \) itself is a function of intensity gradient.

For comparison, the motion equation (7) is rewritten in terms of the Young’s modulus \( E \):

\[
\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \left[ \frac{\nu E}{(1 + \nu)(1 - 2\nu)} (\nabla \cdot \mathbf{u}) \mathbf{I} + \frac{E}{2(1 + \nu)} (\nabla \mathbf{u} + (\nabla \mathbf{u})^T) \right] + \mathbf{F}. \tag{79}
\]

It is not difficulty to see the similarity between (77) and (79). The Hooke’s law (6) implies that, under the same loading condition, \( E \) is a monotonically decreasing function of deformation, which is exactly the desired property of coefficient \( c \) in diffusive adaptive smoothing (see Figure 4 in [102] and related discussions). In other words, a spatially varying elastic property \( E(x) \) in motion equation (79) has the effect of adaptively smoothing the motion field \( (\mathbf{u}) \) with a mechanism that characterizes the nonlinear diffusion equation (77). It should be pointed out, though, unlike the motion equation that has been well established in the elasticity theory, the adoption of nonlinear diffusion equation in image smoothing is motivated by its mathematical convenience, albeit its marked success. For in-depth discussions on the nonlinear diffusion equation with respect to adaptive smoothing, readers are refered to [5, 141].
Several observations can be made from the above comparisons:

1. Both $E$ and $c$ are monotonically decreasing functions of the derivatives of their corresponding unknown variables $(\nabla u, \nabla f)$, a necessary condition to avoid oversmoothing.

2. Similar to the surface reconstruction where small $c(x)$ or $w(x)$ indicate the existence of edges (for example, $c(x) = 0$ or $w(x) = 0$), in motion analysis, small $E(x)$ in an elastic body corresponds to the discontinuity in motion field $u$ (see the following spring example).

3. The Young’s modulus can be viewed as a regularization controller in motion synthesis.

4. The Young’s modulus is an actual physical quantity well-defined in elastic dynamics. In surface reconstruction or image restoration, $c$ is an abstract coefficient that is postulated for computational convenience, and is not considered or intended to be used as a "true" physical quantity, even though $c$ does have a physical definition of conductivity in thermodynamics and hydrodynamics.

5. As a consequence of being a physical quantity, the Young’s modulus is unique for each object and can be used as a constant (under the isothermal condition). On the other hand, many different $c$ functions have been proposed to achieve adaptive smoothing.

The comparative study is summarized in Table 1. Adaptive smoothing will be revisited in the discussion of motion synthesis.

To further illustrate the adaptive smoothing effect of $E$ on motion field, a tension analysis of a simple spring is presented. The spring is 19 cm long and consists of 3 sections of different materials, with the middle section of a much smaller Young’s modulus (Figure 10). Numerical computation was performed with a finite element model that discretizes the spring into 19 elements (1 cm each) and 20 nodes. The spring was fixed at one end and stretched by 5 cm at the other end. The deformation of each node is measured with respect to its pre-stretched position. The distributions of displacement $u(x)$ and the Young’s modulus $E(x)$ are plotted in Figure 11. The middle section of small $E$ causes a large "jump"
Table 6. Adaptive Smoothing in Motion Analysis and Surface Reconstruction.

<table>
<thead>
<tr>
<th>Unknown</th>
<th>Displacement, ( \mathbf{u}(\mathbf{x}) )</th>
<th>Surface/image Reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variational (Energy)</td>
<td>( \int_{\Gamma_1} \beta_1 (\mathbf{u} - \bar{\mathbf{u}}) + \int_{\Omega} \frac{1}{8} \frac{\nu E}{(1 + \nu)(1 - 2\nu)} (\mathbf{u})^T (\mathbf{u})^T ) + ( \int_{\Omega} \beta_2 (\mathbf{u} - \frac{1}{2} (\mathbf{u} + (\mathbf{u})^T) + \int_{\Omega} \mathbf{F} \mathbf{u} + \int_{\Gamma} s \mathbf{u}</td>
<td></td>
</tr>
<tr>
<td>Smoothness Controller</td>
<td>( E(\mathbf{x}) )</td>
<td>( w(\mathbf{x}) )</td>
</tr>
<tr>
<td>PDE</td>
<td>( \frac{\rho}{\partial t^2} = \mathbf{F} + \frac{\nu E}{(1 + \nu)(1 - 2\nu)} (\mathbf{u})^T (\mathbf{u})^T ) + ( C \cdot (\mathbf{F} \cdot (\mathbf{u})^T) )</td>
<td>( \frac{\partial f(\mathbf{x}, t)}{\partial t} = \nabla \cdot (c \nabla f) )</td>
</tr>
<tr>
<td>Smoothness Controller</td>
<td>( E(\mathbf{x}) \propto \frac{1}{\mathbf{u}} )</td>
<td>( c = g(\nabla f), \ c \propto \frac{1}{\sqrt{f}} )</td>
</tr>
<tr>
<td>Comparison ( (E, c) )</td>
<td>(1) ( E ) is a measurable physical quantity.</td>
<td>(1) ( c ) is a coefficient without the corresponding physical definition in surface/image reconstruction.</td>
</tr>
<tr>
<td></td>
<td>(2) As an intrinsic material property, ( E ) can be used as a fixed constant.</td>
<td>(2) Various forms of ( c ) can be constructed as functions of the derivatives of ( f(\mathbf{x}) ).</td>
</tr>
</tbody>
</table>
in the motion field, i.e. a motion discontinuity, while two sections of large $E$ have relatively small and smooth displacements. This has the same effect of setting $c(x)$ or $w(x)$ to zero at edges of a surface/image. It should be emphasized again that this adaptive motion smoothing phenomenon exhibited by an elastic body is determined by its physical property that cannot be changed arbitrarily. For example, setting $E = 0$ does not make physical sense in the context of solid mechanics. Instead, adaptive smoothing of motion field is determined by the relative property difference (heterogeneity) that could be in the range of several orders of magnitude [28]. Based on biomechanical tests [42], the Young’s modulus of typical materials found in biological tissues are in the following ranges (kPa): bone ($\sim 10^7$), collegian ($\sim 10^6$), elastin ($\sim 10^3$) and blood vessel ($\sim 10^2$).

3.2 Synthesis with Recovered Parameter

3.2.1 Three Perspectives

Motion synthesis can be studied from several perspectives: (1) regularized motion reconstruction, (2) Bayesian inference, and (3) Dirichlet type physical modeling.

From a data interpolation point of view, motion synthesis can be viewed as a regularized reconstruction problem where a smooth motion field $u(x)$ is estimated from sparse and noisy data $\bar{u}(x)$ by minimizing function (76). Various $w(x)$ and $\phi$ functions can be designed to

Figure 10. Configurations of Spring Before and After Stretching.
Figure 11. Adaptive Control of Motion Field by the Young’s Modulus. Note the spatial correspondence between the section of small $E$ and the discontinuity in motion field. $E$ is plotted on logarithm scale. Displacement of each node is computed with respect to its original position: $u = x - x_0$.

avoid oversmoothing of motion field. Shulman and Aloimonos [118] proposed a regularized approach to interpolate a smooth motion field with discontinuity preserved. A unique feature of their approach is that the coefficient $w(x)$ is considered as approximately a constant, although it is not a physical quantity.

Motion synthesis can also be viewed as an inference process in the Bayesian framework. With motion observation being approximated by a measurement operator $G$:

$$\bar{u} = Gu + \eta,$$

the posterior probability of $u$, given the noisy data $\bar{u}$, can be computed as:

$$p(u/\bar{u}) = \frac{p(\bar{u}/u)p(u)}{p(\bar{u})}.$$  

Regularized motion synthesis can be treated as a special case of Bayesian inference. Assuming that the likelihood and prior variables in (81) are Markov random fields (MRF), by the Hammersley-Clarkford theorem [13], their probabilities have a Gibbs distribution:

$$p(\bar{u}/u) = \frac{1}{Z_1}e^{-W_1(\bar{u}/u)}, \quad p(u) = \frac{1}{Z_2}e^{-W_2(u)},$$  

53
where $Z_i$ is a normalizing constant known as the partition function, and $W_i$ is the energy function.

By treating the evidence $p(\bar{u})$ as a constant, (81) is rewritten as:

$$p(u/\bar{u}) = \frac{1}{p(\bar{u})Z_1Z_2} \exp[-W_1(\bar{u}/u) - W_2(u)].$$

(83)

So, the posterior variable is also a MRF. Since exponential function is monotonic, Maximum A Posterior (MAP) solution of a MRF is equivalent to the minimizer of its energy function:

$$\arg\max_u [p(u/\bar{u})] = \arg\min_u [W_1(\bar{u}/u) + W_2(u)].$$

(84)

If data is $k$-variant and noise is i.i.d additive normal, $\eta \sim N(\mu = 0, \Sigma = \sigma_1^2 I)$, then:

$$p(\eta) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\eta - \mu)^T\Sigma^{-1}(\eta - \mu)\right] = \frac{1}{(2\pi)^{k/2}\sigma_1^{k}} \exp\left[-\frac{1}{2\sigma_1^2}\eta^2\right],$$

(85)

which implies that the likelihood density is proportional to the exponential function of residual norm:

$$p(\bar{u}/u) \propto \exp\left[-\frac{1}{2\sigma_1^2}\|\bar{u} - Gu\|^2\right], \quad W_1(\bar{u}/u) = \frac{1}{2\sigma_1^2}\|\bar{u} - Gu\|^2.$$  

(86)

Similarly, with a hypothesis of motion field as: $u \sim N(\mu = u_m, \Sigma = \sigma_2^2 I)$, the prior density is proportional to the exponential function of smoothness term in the regularization approach:

$$p(u) \propto \exp\left[-\frac{1}{2\sigma_2^2}\|u - u_m\|^2\right], \quad W_2(u) = \frac{1}{2\sigma_2^2}\|u - u_m\|^2.$$  

(87)

With $W_1$ and $W_2$ in (86) and (87), (84) becomes:

$$\arg\max_u [p(u/\bar{u})] = \arg\min_u \left[\frac{1}{2\sigma_1^2}\|\bar{u} - Gu\|^2 + \frac{1}{2\sigma_2^2}\|u - u_m\|^2\right].$$

(88)

Therefore, Bayesian inference is reduced to the classical quadratic regularization form, as a consequence of MRF and Gaussian condition on both prior distribution and data noise (likelihood). Readers are refered to [124] for a thorough coverage on the statistical techniques for solving inverse problems.
Table 7. Three Perspectives for Motion Synthesis.

<table>
<thead>
<tr>
<th>Perspectives</th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regularized Reconstruction</td>
<td>Residual Norm</td>
<td>Regularization Norm</td>
</tr>
<tr>
<td></td>
<td>$|u - \bar{u}|^2$</td>
<td>$\int_\Omega w(x) \phi(Du(x)) , dx$</td>
</tr>
<tr>
<td>Bayesian Inference</td>
<td>Likelihood</td>
<td>Solution Prior</td>
</tr>
<tr>
<td></td>
<td>$\eta \sim N(0, \sigma_1^2 \mathbf{I})$, $p(\bar{u}/u) \propto \exp[-\frac{1}{2\sigma_1^2} |\bar{u} - Gu|^2]$</td>
<td>$\mathbf{u} \sim N(\mathbf{u}_m, \sigma_2 \mathbf{I})$, $p(\mathbf{u}) \propto \exp[-\frac{1}{2\sigma_2} |\mathbf{u} - \mathbf{u}_m|^2]$</td>
</tr>
<tr>
<td>Physical Motion Equation (Elasticity Theory)</td>
<td>Dirichlet Condition</td>
<td>Constitutive Constraint by Material Properties</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{u} = \bar{u}$, on $\partial \Omega$</td>
<td>$\nabla \cdot \mathbf{F} + \sigma = 0$, $\sigma = \lambda(\nabla \cdot \mathbf{u}) \mathbf{I} + \mu \nabla \mathbf{u} + \mu(\nabla \mathbf{u})^T$</td>
</tr>
</tbody>
</table>

In previous sections, the analogy between the regularized motion reconstruction (13) and elastic dynamics in variational form (13b) has been discussed, particularly the role of Young’s modulus as an adaptive smoothness controller of motion field, the equivalence of residual norm to Dirichlet condition, and the correspondence between regularization term and strain energy. Motion equations in PDE forms (7,8) can be derived from the variational form by the Hellinger-Reissner principle [58]. Three perspectives are summarized in Table 2, with an observation of following equivalencies: (1) residual norm $\rightarrow$ likelihood $\rightarrow$ Dirichlet condition; (2) regularization norm $\rightarrow$ prior $\rightarrow$ strain energy (constitutive constraint).

### 3.2.2 Dirichlet-Type Motion Synthesis with Recovered Property

To be consistent with our original motivation of analyzing nonrigid motion using physical motion parameters, motion synthesis is approached by solving a Dirichlet boundary value problem. This approach has the following advantages: (1) Dirichlet problem is well-posed and proofs can be readily found in computational physics and applied mathematics [24, 73]; (2) given material properties and boundary measurement, motion can be synthesized throughout the modeling domain; (3) synthesis algorithm can be implemented with finite element method. The coupling of tracking and synthesis can be realized in a model whose
tracking module is responsible for monitoring and measuring boundary deformation while the synthesis module computes motions inside the object.

Reconstructing a smooth function from noisy data with discontinuity preserved is a difficulty problem, because the location of discontinuity is unknown \textit{a priori}. In surface reconstruction, Geman and Geman [46] presented a “line process” model by treating both surface function and discontinuity as dual Markov random fields. Blake and Zisserman [16] proposed a physics-based approach by formulating surface reconstruction as “weak spring” or “weak thin plate” models. Since then, two general solution strategies have been adopted: (1) reconstruct surface and discontinuity simultaneously; (2) reconstruct surface with the discontinuities predefined.

The main difficulty of reconstructing surface and discontinuity simultaneously is that the energy function to be minimized is nonconvex. Various deterministic and stochastic algorithms have been proposed, such as graduated nonconvexity [16], simulated annealing [46], local validation [129] and others.

If discontinuity can be predetermined, as suggested by Terzopoulos [129], discontinuity-preserving reconstruction problem would be much easier to handle numerically. In other words, the prior knowledge about discontinuity would be obtained from other cues, rather than being inferred directly from noisy data. Unfortunately, in surface reconstruction, such prior knowledge of discontinuity is rarely available, although recent progress in Gibbs prior learning [150] sheds some lights on the possibility of obtaining generic structural information from natural scenes.

In physics-based motion synthesis, if the Young’s modulus is known, the second reconstruction strategy becomes a feasible option, because the Young’s modulus is an adaptive smoothing controller with its low values indicating the presence of motion discontinuity. Homogeneous property is often assumed in motion modeling. The validity of using homogeneous property is task-dependent. In applications that are dominated by global motion tracking, errors in local motion caused by the use of homogeneous property might be acceptable. On the other hand, in applications where detailed local deformation has significant
implications, property heterogeneity must be taken into account. A simple homogeneous property could lead to large modeling errors (see examples in Section 5). This is especially true in medical imaging, where local deformations could provide invaluable diagnostic information.

Implementation of the Dirichlet-type motion synthesis can be realized by discretizing the motion equation with finite element method. Derivation of finite element formulation using the Rayleigh-Ritz or Galerkin methods can be found in many textbooks. The algebraic equation that assembles all elements is:

\[ K u = F_g \] (89)

where \( K \) is the stiffness matrix that includes material properties, \( u \) is nodal displacement solution, and \( F_g \) is the generalized force. In practice, Dirichlet condition can be implemented using the penalty method. For example, to specify a displacement at node \( i \), \( u_i = y \), a penalty term can be added to (89), \( k u_i = ky \). By setting \( k \) to a very large number with respect to the corresponding diagonal element in the stiffness matrix, \( k \gg k_{ii} \), Dirichlet condition can be satisfied numerically [151].

Motion synthesis is approached with Dirichlet data because it is more accessible than Neumann data, albeit the later is more intuitive in terms of driving the deformation. If necessary, force can be derived from displacement with appropriate mapping functions. By constructing a potential function of intensity gradient and using a suitable projection, a virtual force can be derived from image measurement [126]. To obtain actual boundary force, transformation similar to Dirichlet-to-Neumann mapping can be used:

\[ \Lambda_{E,\nu}(\bar{u}) = n \left\{ \frac{\nu E}{(1 + \nu)(1 - 2\nu)} \text{tr}\varepsilon(u) \right\} I + \frac{E}{(1 + \nu)} \varepsilon(u) \right\} \bigg|_{\partial\Omega} \] (90)

where \( \Lambda_{E,\nu}(\bar{u}) \) is the Dirichlet-to-Neumann map, \( n \) is the unit outer normal to \( \partial\Omega \), and \( \bar{u} \) is the specified boundary displacement. The physical interpretation of \( \Lambda(\bar{u}) \) is that the boundary displacement is transformed into boundary traction. This derivation of actual force requires the knowledge of material properties, at least on the boundary.
The synthesis scheme can be extended to transient cases by solving an initial-boundary value problem. In motion analysis, a model with recovered physical parameter has the advantage that the parameter can be used as a constant without the need of update in the subsequent tracking and synthesis. Only feature that need to be tracked is object’s boundary, which is usually the most salient image cue.

3.3 Experiments

Two experiments are presented to demonstrate the proposed motion analysis method: (1) a 1D synthetic spring is used to illustrate the recovery algorithms; (2) an image sequence of a 2D elastic object is used to demonstrate how the recovered heterogeneous Young’s modulus can be utilized to synthesize interior motion from boundary observations.

3.3.1 Recovery of Physical Motion Parameter

The advantage of experimenting with synthetic model is that its well controlled nature (data error and true solution are known and adjustable) allows us to examine all aspects of the recovery algorithm with various test scenario. The information contained in a simple spring experiment is surprisingly rich, which helps us to gain insights on the behavior of algorithms and to optimize the implementation.

*Forward Model:* The model configuration is as follows (Figure 12): (1) spring has a length of 19 cm, discretized into 19 elements of equal length; (2) element is of uniaxial tension-compression type without bending effect. Each element has a cross-section area of 0.1 cm² and each node has one degree of freedom; (3) material properties are isotropic and linear. A Poisson ratio of 0.495 is assigned to all elements. Property heterogeneity in terms of Young’s modulus is as follow: all elements have a value of 500 kPa, except the five
elements in the middle that have a higher value of 2500 kPa. This type of configuration is of interest in medical imaging because diseased tissues often show up as hard inclusions surrounded by soft tissues; (4) one end of spring is fixed and the other end is pulled by 5 cm, which is equivalent to a force of 1.6667 NT; (5) spring is in equilibrium after deformation (static simulation).

**Inverse Problem:** Two types of data are considered for recovering spring’s Young’s modulus: (1) data that includes only displacements, and (2) data that includes both displacements and boundary forces. Ideal displacement data is generated by running the forward model with properties and boundary conditions as specified above. Noisy data is obtained by adding 1% white noise to the ideal data.

**Solution Uniqueness and Data Types:** A few examples are used to demonstrate numerically that the uniqueness of inverse solution is determined by the *data type* (Dirichlet or Neumann). Two forward simulations are conducted using models that are identical except that the second model has an elasticity 10 times higher than the first one. Given the same Dirichlet conditions (fixed at node 1 and pulled at node 20 by 5 cm), two models generated the same deformation on each node (Figure 13). This exercise indicates that each displacement data set is related to an infinite number of possible elasticity distributions as long as they are proportional to each other: $E_i = bE_j$, with $b$ being a real constant. Therefore, a unique elasticity solution cannot be obtained from displacement data only.
Two inverse experiments are used to show the nonuniqueness of elasticity solution associated with displacement data. For each experiment, the same ideal data was used, \( \mathbf{u} = (u_1, u_2, ..., u_{20}) \), but different initial guesses (800 kPa and 1600 kPa, both uniform). The recovered elasticity is proportional to the true solution, but with different absolute values. The experiments imply that the solution space of inverse problem formulated with only displacement data has many minima of equal magnitude. The experiments also indicate that only relative elasticity can be recovered with displacement data. For applications where the primary interest is to identify local property abnormalities, recovering a relative elasticity distribution is certainly an attractive alternative.

With a data set that includes both displacement and force, \( \mathbf{u} = (u_1, u_2, ..., u_{20}, F_{20}) \), a unique elasticity solution was obtained regardless of the initial values being used (not too far away from the true solution). Since actual force measurement is rarely available in practice, surface traction can be derived using (90), provided that the Young’s modulus on the boundary is known. For the spring model, 5 cm displacement at node 20 (\( E = 500 \) kPa) is equivalent to a pulling force of 1.6667 Newton.

Solution Stability and Data Quality: In addition to the uniqueness issue, data noise causes more serious problem of solution instability. In the next experiment, a data set that was corrupted by 1% white noise was used, \( \bar{\mathbf{u}} = (\bar{u}_1, \bar{u}_2, ..., \bar{u}_{20}, \bar{F}_{20}) \). The results are plotted in Figure 14. The elasticity solution obtained without regularization (\( \beta = 0 \)) shows a typical “zigzag” pattern, indicating high instability of inverse solution. Small perturbation errors in the data are dramatically amplified in solutions. The ill-posed nature of inverse problems is often overlooked and solution is computed by naive least square approach, i.e. \( \beta = 0 \). This example demonstrates that the solution obtained with a minimal residual norm is physically meaningless. In other words, the regular least square method would “overfit” solution to the noisy data to a degree that noise propagation completely dominates the solution profile.

On the other hand, if the residual norm is penalized too much with a large \( \beta \), inevitably the solution will be oversmoothed (Figure 14 (b)). The key is therefore to balance the
contributions from data and prior with an optimal regularization parameter (Figure 14 (d)). I plot the solution norm $\| E \|$, residual norm $\| A(E) - \bar{u} \|$ and relative error norm $\| E - E_{true} \| / \| E_{true} \|$ against regularization parameter in Figure 15. The typical semi-convergence property of regularization algorithm is clearly visible and the optimal $\beta$ is identified to be $1.6E-15$. The small value of $\beta$ is mainly due to the large magnitude disparity between solution and data, which can be scaled properly, if numerical underflow is a concern.

**Performance:** The baseline and GCV-based algorithms are applied to the same inverse problem to assess their performance. The model configuration of spring and boundary condition are the same as those in previous experiments. To ensure convergence of Gauss-Newton iteration in GCV-based method, the iteration was started with a larger $\beta$ and enforcing a cooling sequence using $\beta_{i+1} = max(\beta^* + \gamma \beta_i)$, with $\beta^*$ being the currently estimated value from GCV function and $\gamma < 1.0$. The initial solution is a uniform value of 800 (kPa).
The results are shown in Figure 16. Both algorithms are successful in terms of the quality of recovered elasticity, with the heterogeneity zone in the middle of spring clearly identified. Both algorithms converged within 15 iterations, but the baseline approach took considerable more time because of the large number of nonlinear equation (18) that has to be solved. The converged $\beta$ from the GCV-based algorithm is slightly smaller (0.47E-15) than the value chosen by the baseline algorithm (1.57E-15).
3.3.2 Motion Synthesis with Recovered Parameter

Data Acquisition and Parameter Recovery: The object used in the experiment is an elastic bandage with a dimension of 11.6 by 7.4 cm (Figure 17). The bandage is designed for wound care and made of material that has a Young’s modulus close to that of artificial skin (polytetrafluoroethylene). To create property heterogeneity, less elastic tapes were attached at both ends of bandage. Measured Young’s modulus of bandage and tape from tensile test are 160 and 7900 (kPa), respectively. As a result, the final elastic object has three sections: an easy-to-deform section in the middle and two less-deformable sections on both ends.

A 7x10 rectangle grid was marked on the bandage to facilitate correspondence matching. An image sequence was acquired with Minolta range camera while the bandage was gradually stretched laterally. Frame 6 and frame 7 are the same except that, in frame 7, the central portion of bandage was temporally blocked by another object (Figure 17 (d)). Frame 7 will be used to demonstrate the boundary-based motion synthesis scheme in the presence of occlusion.

A finite element model was built directly on top of the grid, with 70 nodes and 54 quadrilateral elements (Figure 18 (a)). Frame 1 and frame 4 were used to recover the Young’s modulus. Displacements were computed with the aid of grid correspondence and assigned to each node of model. To recover the Young’s modulus with a higher resolution, the mesh was refined through bilinear interpolation (Figure 18 (b)). The Young’s modulus on boundary elements were known and used to derive boundary stress by (90). The results are plotted in Figure 18 (d). The overall distribution of recovered elasticity matches very well with measured elasticity, with the central heterogeneity section clearly identified.

Significance of Recovered Heterogeneous Property: Motions in the interior of a deformable object can be synthesized from boundary observations by solving a Dirichlet-type forward problem, provided that material properties are known. In this case, the use of homogeneous property is not valid, because it automatically leads to a simple smooth motion field.
Figure 17. Image Sequence of a Bandage for Motion Synthesis.

Figure 18. Finite Element Model and the Recovered Elasticity.
Table 8. Synthesized Motion along Cross-Section Line.

<table>
<thead>
<tr>
<th>Node ID</th>
<th>Measured motion (m)</th>
<th>Synthesized Motion (homog. elasticity)</th>
<th>Error (%)</th>
<th>Synthesized Motion (recovered elasticity)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>-0.0121</td>
<td>-0.0121</td>
<td>0</td>
<td>-0.0121</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>-0.0118</td>
<td>-0.0089</td>
<td>25</td>
<td>-0.0120</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>-0.0108</td>
<td>-0.0058</td>
<td>46</td>
<td>-0.0123</td>
<td>14</td>
</tr>
<tr>
<td>34</td>
<td>-0.0099</td>
<td>-0.0021</td>
<td>79</td>
<td>-0.0080</td>
<td>19</td>
</tr>
<tr>
<td>35</td>
<td>-0.0009</td>
<td>0.0010</td>
<td>211</td>
<td>-0.0012</td>
<td>33</td>
</tr>
<tr>
<td>36</td>
<td>0.0056</td>
<td>0.0042</td>
<td>25</td>
<td>0.0065</td>
<td>16</td>
</tr>
<tr>
<td>37</td>
<td>0.0132</td>
<td>0.0076</td>
<td>42</td>
<td>0.0155</td>
<td>17</td>
</tr>
<tr>
<td>38</td>
<td>0.0189</td>
<td>0.0109</td>
<td>42</td>
<td>0.0181</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>0.0191</td>
<td>0.0146</td>
<td>24</td>
<td>0.0180</td>
<td>6</td>
</tr>
<tr>
<td>40</td>
<td>0.0186</td>
<td>0.0186</td>
<td>0</td>
<td>0.0186</td>
<td>0</td>
</tr>
</tbody>
</table>

Finite element model is used to synthesize interior motion of bandage in frame 7 that is partially blocked. Displacements on bandage boundary were measured between frame 1 and frame 7 and then assigned to the boundary nodes of model. For comparison, two experiments were performed, one with an assumed homogeneous Young’s modulus of 7900 (kPa) and another one with the recovered Young’s modulus (Figure 18 (d)). Synthesized motion at the center of each element was plotted as small squares in frame 6 (Figure 19 (a,b)). Note that Frame 6 is the same as frame 7 except the occlusion.

There exist significant mismatches between true positions and synthesized positions using homogeneous Young’s modulus (Figure 19 (a)), especially in the middle section. Model with recovered Young’s modulus reduces synthesis error significantly (Figure 19 (b)). To visualize the adaptive control of Young’s modulus on motion field, a cross-section is drawn in the center of bandage along the direction of stretching for both displacement and elasticity distributions (Figure 19 (c,d,e,f)). As expected, homogenous elasticity causes an unrealistic smooth transition in motion field while recovered heterogeneous elasticity helps preserve the observed motion discontinuity. The results are also listed in Table 3. This example also demonstrates the advantage of the proposed synthesis scheme over other approaches that require tracking all features explicitly, which could be defeated by the existence of occlusion.
This motion synthesis scheme using the recovered object-specific heterogeneous property holds great potentials in medical imaging. Bajcsy et al. [11] presented an elastic registration method based on a simplified Navier’s equation that was widely adopted in medical image registration and surgery planning [99, 54]. Method based on boundary mapping was also proposed for nonrigid registration [25]. Many researchers recognize that it is essential to incorporate anatomical structure with spatially variable properties into biomechanical models.

This synthesis scheme can also be used in realistic visualization and animation, especially those use physics-based modeling techniques [74]. Vivid simulation of live objects that involves subtle deformation of soft tissues, such as facial expression, is only possible if detailed knowledge of tissue properties is incorporated into the model.
Figure 19. Comparison of Two Experiments.
Synthesized motion using the assumed homogeneous elasticity (a,c,e) and the recovered heterogeneous elasticity (b,d,f). Measured displacement is shown as solid line in (e,f). Synthesized displacements are depicted as circles and triangles in (e,f).
CHAPTER 4
APPLICATION IN BURN SCAR ASSESSMENT

Each year more than one million people suffer burn injuries in Canada and the US [111]. Accurate rating of scar condition is needed in order to design an effective treatment plan. Scar rating in clinical settings is done using the Vancouver rating scale or its variants by which experts assess the vascularity, pigmentation, pliability and size of scars [121, 145]. These rating methods suffer from their subjective nature and low consistency among rates. There is a strong need for a quantitative and objective scar rating method based on the biophysical properties of skin tissue [107].

In this application, a physical model-based scar rating method is presented which focuses on estimating the relative elasticity of scars. The contributions of the proposed method are: (1) noninvasive natural image features are used to generate an adaptive mesh and to build a finite element model. In previous works, artificial markers (ink stamps) on the skin to facilitate model construction had been used. The use of natural features enables us to quantify the elastic property without the need of any markers, and hence greatly enhances the applicability of the proposed rating method; (2) a robust procedure of quantifying the scar elasticity is established using the regularization method so that the noisy data can be handled properly.

Two important issues in scar elasticity estimation are the ill-posedness and the computational cost. The stability of inverse solution can be partially restored through the regularization methods. The following approaches are experimented in the context of burn scar assessment: (1) reduce the parameter space by posing stronger constraints; (2) reduce the parameter space by using an adaptive mesh.
4.1 Model Construction

The mechanical behavior of skin is determined by the presence of collagen fibers, elastin fibers and lubricating ground substance [77, 42, 130]. Burn scars tend to have random organization of collagen fibers during tissue regeneration. In the previous studies of burn scar evaluation, it has been shown that the use of elastic and isotropic model is adequate for this particular task [135, 107].

4.1.1 Extraction of Natural Point Features

The Shi-Tomasi detector is used to extract salient points in scar images [117, 114]. For each pixel $p$ in the image, a $2 \times 2$ matrix ($g$) is computed within a window ($w$):

$$g = \begin{bmatrix} \sum_w \frac{\partial I}{\partial x} \frac{\partial I}{\partial x} & \sum_w \frac{\partial I}{\partial x} \frac{\partial I}{\partial y} \\ \sum_w \frac{\partial I}{\partial y} \frac{\partial I}{\partial x} & \sum_w \frac{\partial I}{\partial y} \frac{\partial I}{\partial y} \end{bmatrix}$$

(91)

where $I$ is the image intensity and $(x, y)$ represent row and column. The first derivatives is computed by convolving the intensity with the derivative of Gaussian filter ($G$): $\frac{\partial I}{\partial x} = \frac{\partial G}{\partial x} * I$, $\frac{\partial I}{\partial y} = \frac{\partial G}{\partial y} * I$. The coefficients of gradient matrix is then computed: $\frac{\partial I}{\partial x} \frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y} \frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial x} \frac{\partial I}{\partial y}$, $\frac{\partial I}{\partial y} \frac{\partial I}{\partial y}$. The coefficients of all pixels inside $w$ are summed up to produce the matrix $g$ of pixel $p$.

The eigenvalues of matrix $g$ is then computed: $(\lambda_1, \lambda_2)$. Given a threshold $T$, satisfaction of the condition: $\min(\lambda_1, \lambda_2) > T$, indicates that the window contains a corner/point with two strong edges along the eigenvector directions. A value of $T = 1$ is used for all of the scar experiments. With this value, 400-800 features can be ensured in the scar images, which suffices the need of building a relatively dense finite element mesh.

Another threshold $h$ is used, the minimum distance between two adjacent points, to control feature distribution. Since scars have lower intensity than normal skins, the value of $h$ is changed adaptively based on the intensity variation. An area is first selected that contains only scars and skins. The selected area is then equalized to highlight the intensity contrast between scars and skins. For each feature point, its $h$ is computed based on the average intensity in $w$, scaled linearly by the overall intensity variation in the modeling.
area: \[ h = H_{\text{min}} + (I_w - I_{\text{min}}) \frac{H_{\text{max}} - H_{\text{min}}}{I_{\text{max}} - I_{\text{min}}}, \]

where \( h \) is the minimum distance between a feature and its neighbors, \( I_w \) is the average intensity in \( w \), \((I_{\text{max}}, I_{\text{min}})\) are the maximum and minimum intensities in modeling area after equalization, and \((H_{\text{max}}, H_{\text{min}})\) are user-specified maximum and minimum \( h \) that correspond to \((I_{\text{max}}, I_{\text{min}})\). After computing \( h \) for each feature, all features are sorted based on the number of their neighbors that are in conflict with their \( h \) value. By ”conflict” we mean that the distance between a feature and its neighbors is less than its \( h \). The feature is then iteratively removed that has the most conflicts until no conflict exists. As a result, the final distribution has more features in scars than in skins.

All images have the resolution of 640×480 pixels. Features are extracted using a window size of 9×9. The minimum distance \((H_{\text{min}})\) is in the range of 15-20 pixels, and \( H_{\text{max}} \) is set to be 1.3 - 2.5 times larger than \( H_{\text{min}} \). The choice of \( H_{\text{min}} \) depends on the size of scars. If \( H_{\text{min}} \) is too large, there may not be enough features in scar areas and vice versa. The same argument applies to the choice of \( H_{\text{max}} \). Those heuristic parameters must be tuned for a specific setting depending upon the quality of images and size of scars.

To quantify the precision of feature extraction, 218 corresponding features were examined in scar images. For each patient, two frames were taken (before and after deformation). Features were then extracted from two frames. There exist small shifts between the com-
Table 9. The Performance of Point Feature Detector.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>Max.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature position shift (pixel)</td>
<td>0</td>
<td>4.5</td>
<td>2.1</td>
</tr>
<tr>
<td>actual displacement (pixel)</td>
<td>11.9</td>
<td>36.2</td>
<td>24.8</td>
</tr>
<tr>
<td>error (shift/displacement)</td>
<td>0.0%</td>
<td>13.3%</td>
<td>7.2%</td>
</tr>
</tbody>
</table>

Computed positions of corresponding features in two frames. An example is shown in Figure 20. The white square indicates the computed position by the feature detector. The correspondence pair has a shift of 3 pixels. Note that the feature patterns changed slightly between two frames due to the variations in lighting, projection and skin deformation. The actual displacements due to skin deformation between two correspondences are 26 pixels. Since our interest is to obtain good displacement data, the extraction error is defined as the ratio of the computed position shift to the actual displacement. For the example shown in Figure 20, the extraction error is 11.5%. For 218 feature pairs examined, the results are summarized in Table 9. With current imaging setting and feature extraction method, the average error introduced in feature extraction is less than 10%.

4.1.2 Adaptive Meshing

Given a set of points distributed on object’s surface, an adaptive mesh based on the Delaunay principle [19] can be generated. The following meshing procedure is used: (1) extract point features from images; (2) select the points inside the region of interest (ROI) for which a physical model will be built; (3) link the points on the boundary of ROI to form a closed polygon or surface; (4) generate an adaptive Delaunay mesh using feature points as its nodes; (5) refine the mesh.

Refinement is performed to insert new nodes into the area of interest as a quality assurance procedure. These guidelines are followed in the node insertion: (1) new nodes are added at the region of high curvature to ensure accurate surface representation; (2) new nodes should subdivide the thin element into regular elements to reduce modeling errors; (3) new nodes are added at regions of property discontinuities which could cause modeling errors. Figure 21 shows an example of adaptive triangle mesh.
Figure 21. Adaptive Meshing Using Point Features.
(a) Scar image. (b) Point features and ROI. (c) Adaptive triangle mesh with refinement at the boundary.

4.2 Estimate Scar Elasticity

4.2.1 Range Scanner Setting

A K2T range scanner is used to obtain 3D displacement between corresponding features. The setting of K2T scanner is illustrated in Figure 22 (a). The K2T system consists of a CCD camera and a structured light projector, and computes the depth from images of striped light patterns. An example of structured light patterns is shown in Figure 22 (b). In scar study, the effective imaging area is about 30-35 cm cube. The distance between patients and the range camera is about 100-130 cm.

A set of images (2-4 frames) was taken while patient’s skin was stretched. Since the scar condition of most patients did not allow the use of contact devices, patient’s skin was pulled gently by hands to avoid pain and further damage. Therefore, it was not possible to measure the forces applied to patient’s skin. 3D displacement ($u$) obtained from intensity and range images is used to compute the objective function in an “output-least-squared” form: $\|F(E) - u\|^2$, where $F(E)$ denotes the forward model.

Our interest is to determine the relative elasticity of scars. Good information can be obtained about the geometry of scars from image cues (intensity, color and texture). To reduce the computational complexity, several simplifications are made: (1) the geometry of scars is known; (2) the elasticity of scar is higher than that of normal skin; (3) the elasticity
of background normal skin is known. The minimization of Tikhonov functional is then transformed into a one-dimensional search problem with two steps:

1. Determine the regularization parameter using the L-curve method [55].

2. Change scar elasticity \(E_s\) until the minimizer of the simplified functional is found:

\[
T(E_s) = \|F(E_s) - \mathbf{u}\|^2 + \alpha \|W(E_s - E_s^*)\|^2.
\] (92)

In L-curve method, the log of the regularized solution norm \(\log|E|\) is plotted against the log of the residual norm \(\log|F(E) - \mathbf{u}|\) for a range of \(\alpha\). The optimal \(\alpha\) is chosen to be the one that corresponds to the corner of the L-curve. The intuitive interpretation of the L-curve method is that the solution norm will go to the maximum and the residual norm will go to the minimum as \(\alpha\) approaches zero, and vice versa.

### 4.2.2 Boundary Condition Specification

Displacements are specified at the boundary nodes that are also feature points. The feature points inside the modeling domain are designated as the controlling nodes, on which the measured displacements and the simulated displacements by forward model are used to calculate the residual norm. Since only displacement is available, the elasticity of normal skins is used as a reference to determine the relative elasticity of scars. The reported Young’s
Table 10. The Estimated Relative Scar Elasticity (kPa).

<table>
<thead>
<tr>
<th>Patients</th>
<th>Scar elasticity (Natural features)</th>
<th>Scar elasticity (Artifi. markers)</th>
<th>Absolute error</th>
<th>Relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>970407</td>
<td>53</td>
<td>46</td>
<td>7</td>
<td>13.2%</td>
</tr>
<tr>
<td>970416</td>
<td>12</td>
<td>14</td>
<td>2</td>
<td>16.6%</td>
</tr>
<tr>
<td>970425</td>
<td>41</td>
<td>37</td>
<td>4</td>
<td>9.7%</td>
</tr>
<tr>
<td>970922</td>
<td>19</td>
<td>18</td>
<td>1</td>
<td>5.3%</td>
</tr>
</tbody>
</table>

*Errors are computed between the results using natural features and artificial markers. The elasticity of normal skin used in experiments is 5 kPa.*

...modulus values of soft tissues are in the range of 1-100 kPa [42, 130]. In scar assessment, a Young’s modulus of 5 kPa is used for normal skin.

It is worth noting that the body force should not be confused with the boundary conditions in normal sense such as surface force/traction. The variation of body force (mainly gravity) is a significant factor in the simulation of large-sized objects such as the sea water in a bay or the planet earth. But for small objects such as scars, body force can be treated as a constant.

### 4.3 Experimental Results

#### 4.3.1 Data Set

The data set includes images of four patients that were taken at different healing stages. The boundaries between scars and normal skins can be clearly identified in those images.

#### 4.3.2 Quantify Scar Elasticity

Two independent experiments are carried out, one using artificial markers and one using natural features, to determine if the natural feature is a viable option for replacing the artificial marker. The results for all patients are listed in Table 10. Although direct measurements of scar’s Young’s modulus are not available for those patients, the fact that the differences between the estimated Young’s modulus using artificial markers and using natural features are very small suggests a high level of consistency. The results of patient-
970922 are plotted in Figure 23. As expected, experiment using artificial markers shows a slightly better performance than that using natural features, as indicated by the minimization curves (Figure 23 (f)). But the disparity between two curves is very small. More importantly, the minimum points of two curves are almost identical (18 kPa and 19 kPa). The strain distributions also match well with scar distribution.

The errors in measured displacements mainly came from two sources: (1) feature extraction caused by illumination and texture variations; (2) correspondence mismatch. Since the correspondence was manually established, the disparity between the modeling results of using natural features and using artificial markers is mainly caused by the feature extraction error. As analyzed before, the feature extraction error is less than 10% on average, which is an acceptable value for burn scar assessment.

4.3.3 Relative Elastic Index

Current clinical scar assessment methods are based on a relative rating scale. This study will investigate if there exists a positive correlation between the clinical rating and the estimated elasticity ratio of scars to normal skins, so that a standard can be established that is quantitative and comparable to the one used by physicians. The Relative Elastic Index (REI) is defined as the ratio of average Young’s modulus of scars to that of normal skins:

\[
REI = \frac{\sum_{i=1}^{n} E_i / n}{\sum_{j=1}^{m} E_j / m},
\]

(93)

where \(n\) and \(m\) are the numbers of elements inside scars and normal skins, respectively.

Table 11 lists physician’s ratings, as well as REIs using both artificial markers and natural features. In all experiments, patient’s skin was stretched along two directions (horizontal and vertical). So, each patient was studied twice. In Figure 24, the REIs using natural features and artificial markers are plotted against the physician’s ratings. Two important observations can be made: (1) there exists a good correlation between the physician’s ratings and the REIs using either natural features or artificial markers; (2) the REIs computed using natural feature are consistent with those using artificial markers, especially for less
Figure 23. Estimated Elasticity and Strain.
(a) Scar image. (b) After being stretched. (c) Range image. (d) Strain using natural features. (e) Strain using marker. (f) Minimization curves. In (d) and (e), dark color indicates low strain value of scars.
Table 11. Physician’s Rating and Relative Elastic Index (REI).

<table>
<thead>
<tr>
<th></th>
<th>Physician’s rating</th>
<th>REI using natural features</th>
<th>REI using artificial markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>scar-970407(hori)</td>
<td>4.5</td>
<td>9.5</td>
<td>8.0</td>
</tr>
<tr>
<td>scar-970407(vert)</td>
<td>4.5</td>
<td>9.1</td>
<td>7.3</td>
</tr>
<tr>
<td>scar-970416(hori)</td>
<td>3.0</td>
<td>2.8</td>
<td>3.3</td>
</tr>
<tr>
<td>scar-970416(vert)</td>
<td>3.0</td>
<td>4.1</td>
<td>3.8</td>
</tr>
<tr>
<td>scar-970425(hori)</td>
<td>3.3</td>
<td>6.3</td>
<td>5.0</td>
</tr>
<tr>
<td>scar-970425(vert)</td>
<td>3.3</td>
<td>7.4</td>
<td>5.6</td>
</tr>
<tr>
<td>scar-970922(hori)</td>
<td>2.0</td>
<td>2.2</td>
<td>2.3</td>
</tr>
<tr>
<td>scar-970922(vert)</td>
<td>2.0</td>
<td>2.8</td>
<td>2.0</td>
</tr>
<tr>
<td>normal-skin</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

For each patient, the skin was stretched in two directions, horizontally and vertically. The REI of normal skin (1.0) is defined as the baseline.

Figure 24. Correlation Between Physician’s Rating and REIs.

damaged scars. The slight deviation from the linear monotonic relationship is probably caused by the complex scar patterns, which affect both feature extraction and matching. This problem can be solved by calibrating the model against the ground truth, a possibility that is currently under investigation. It should be pointed out that, due to the limited number of patients, no statistical conclusion was drawn at this point about the correlation between the physicians’ rating and the REI.
CHAPTER 5
APPLICATION IN FACE RECOGNITION

During the past a couple of years, biometrics research has received considerable attention due to its high potential in security related applications. There exists a wide variety of biometrics techniques, some are relatively mature while others are still in their infancy. Each biometrics technique has its pros and cons, and it is not possible to find a single one that can solve all practical problems. Therefore, there is always a need for new biometrics.

Other than fingerprint, face recognition is probably the most natural (and hence a popular) biometrics because we have developed the ability to recognize faces automatically with no conscious effort. Current face recognition methods rely on visible photometric or geometric attributes that are present in intensity images. Based on large amount of research and benchmark studies [20, 40, 103], it has been recognized that those methods suffer from problems associated with following factors: (1) illumination and pose variation; (2) make-up, hairs and glasses; (3) plastic surgery; (4) face deformation during expression (dynamic face analysis in video sequence). Future face recognition methods must address those difficult issues.

A new class of features (or biometrics) is proposed that is derived from the computed strain pattern exhibited during face expression. The method has several advantages:

1. Elastic strain pattern is directly related to the material property of underlying facial muscles. Our hypothesis is that, if the anatomical structure of an individual face (geometry, distribution and strength of bones and muscles) is unique, then this anatomical uniqueness should be reflected in the elastic strain pattern. Strain pattern could be less sensitive to the illumination and pose changes as well as camouflage using makeup.
2. The computation of elastic strain map requires at least two frames that capture the face deformation during expression. Most face recognition methods use static images only and dynamic face expression has been considered as an adverse factor that may cause performance degradation [142]. However, a recent study by Yacoob and Davis [143] indicates that deformed face (smiling face) is more recognizable and actually increases the identification rate. This study will go one step further beyond the visible cues in face expression to recover the elastic strain pattern that might help reveal the underlying anatomical individuality.

3. Anatomy-based physical model has been widely used in realistic facial animation and surgery simulation [97, 127, 74]. Essa and Pentland [116] proposed a physical model to represent facial motion and distinguish face expressions. They used the model to estimate visual muscle activation and to generate motion-energy templates for each expression. However, we are not aware of any study that uses physical model to compute elastic strain pattern for the purpose of face recognition. A finite element model based on continuum mechanics enables us to compute the strain map accurately.

5.1 Hypothesis: Face Anatomy and Muscle Biomechanics

Biometrics refers to the identification of an individual based on distinctive physiological or behavioral characteristics. In face recognition, the hypothesis is that each individual has a unique visible face pattern in terms of shape, color or texture that can be utilized for recognition. This uniqueness in visible pattern is certainly related to, and probably determined by, the uniqueness of the underlying anatomical structure and biomechanical compositions, which could also be useful for recognition, if they are measurable.

Major anatomical units of human face are: bones (skull), muscles, skin, blood vessels and nerves [39]. The use of skull measurement in identification (craniofacial analysis) and its forensic implications is well documented [64]. But it is doubtful that a practical biometrics can be derived due to the limitation of special imaging modality (X-ray is needed for skull measurement). It is also difficult to capture facial nerve patterns with current
imaging technologies. The blood vessel patterns, however, has been utilized in both facial thermography and iris-retina scans with the aid of infrared camera.

Face expression is controlled by muscles and emotional states that trigger and change the muscle movements, which could be quantified by the elastic strain pattern. Elastic strain pattern computed from the observed face deformation not only reflects an individual’s emotional state, but, more importantly, also reveals the intrinsic muscle properties, and therefore can be utilized for recognition. This unique strain pattern associated with an individual can remain unchanged for a long period of time, although it is reasonable to expect some variations caused by aging (loss of elastin fibers and muscle elasticity), injuries and plastic operations.

An ideal face model to compute strain pattern should incorporate all anatomical details. However, such a full-scale model would be too expensive to be realistic for face recognition. To strike a balance between modeling accuracy and computational efficiency, we have to make a compromise on what to model and how to model.

Although a whole face contains more information about an individual, it is more practical to model a portion of face whose deformation is dominated by one or a few major muscles. A section is chosen that is between the cheek bone and jaw line (side view) and covered by the masseter muscle. The masseter muscle is a large, thick and roughly rectangular plate that is responsible for jaw action. Its deformation will be modeled between two positions, namely the closed mouth and the open mouth.

Masseter muscle is of striated type. Due to the location just beneath the skin tissue, its contraction has an immediate effect on skin motion. Therefore, we do not treat skin as a separate functional layer. Instead, we model the deformation of muscle and skin together as an integrated mechanical unit. This single-layer model satisfies requirements of face recognition. But two-layer model may be considered, especially for older people where the motion of muscle and aged skin are less synchronized (wrinkles).
5.2 Computational Method

There are four major steps in computing elastic strain pattern: (1) feature extraction and motion measurement; (2) finite element construction; (3) strain computation; (4) conversion of strain map to intensity image.

*Feature Extraction and Motion Measurement:* First, salient points will be extracted from intensity images and then construct a polygonal surface using the Delaunay principle. Feature extraction is performed using the Shi-Tomasi detector [117].

Due to the nonrigid nature of face deformation, we presently establish correspondence between two frames manually to ensure the quality of displacement vector. The displacement data will be used in the finite element model to specify the Dirichlet boundary condition. This manual processing is not new in early face recognition research where eyes and other facial features are manually located (FERET test [103]). Future work will use an automated correspondence matching method.

*Delaunay Meshing and Model Construction:* The Delaunay triangulation is used to generate an adaptive triangle mesh with a set of points that are randomly distributed on face images. A simple meshing strategy is designed: (1) select points that defines the boundary of the region to be modeled; (2) link those boundary points to form a polygon; (3) select the points that are inside the polygon; (4) generate a triangle mesh that is adaptive to all the selected points.

To ensure the mesh quality, a local mesh refinement procedure is used that detects bad-shaped elements and improves the initial mesh accordingly: (1) node will be added at the region of high curvature to increase the accuracy of surface representation; (2) new node will subdivide the long and thin element into more regular-shaped element.

*Forward Modeling and Strain Computation:* The deformation of face tissues can be described by a motion equation:

\[
\nabla \cdot \left[ \lambda (\nabla \cdot \mathbf{u}) \mathbf{I} + G \nabla \mathbf{u} + G (\nabla \mathbf{u})^T \right] + \rho \mathbf{f}_b = \rho \frac{\partial^2 \mathbf{u}}{\partial t^2},
\]

(94)
where $G$ and $\lambda$ are the Lamé constants.

Equation (94) is solved numerically after being discretized over the Delaunay triangle mesh using the finite element method in variational formulation [151]. Since the primary interest is in static deformation only, the final forward model in the discrete matrix form becomes:

$$\mathbf{Ku} = \mathbf{F}$$

(95)

where $\mathbf{K}$ is stiffness matrix and $\mathbf{F}$ is the generalized force.

To compute the strain pattern of a deformed face, we have to supplement appropriate boundary conditions. The Dirichlet condition is specified using the measured displacement data on the feature points. As a result, it becomes an over-specified boundary value problem of the first kind and can be readily solved using an iterative numerical solver.

**Strain Conversion and PCA Analysis:** The standard principle component analysis (PCA) [137] is used for similarity computation. The strain maps is normalized into intensity images using a simple linear transformation:

$$\frac{e_x - e_{\text{min}}}{e_{\text{max}} - e_{\text{min}}} = \frac{I_x - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}}$$

(96)

where $(e_{\text{max}}, e_{\text{min}})$ are the maximum and minimum strain values for all subjects, $e_x$ is the strain value to be converted, $I_x$ is the converted intensity value. $(I_{\text{max}}, I_{\text{min}})$ are set to 255 and 0, respectively. Because strain value spans over a wide range, information may be lost with this linear conversion. More sophisticated conversion method will be considered in future investigations.

Before PCA analysis, all converted strain intensity images are scaled based on the geometry of facial landmarks (nose, ear, mouth and jaw line). A rectangle region of the size of 150 by 160 pixels (original face images are 640 by 480 pixels) in the center of strain intensity image was then chopped out. This rectangle strain image was then used in PCA analysis. An example strain pattern is shown in Figure 25. The implementation of PCA algorithm used in this study can be found in [14].
The face is attached with a less stretchable tape, which corresponds to small strain (low intensity).

(a) Close mouth + bright light, (b) open mouth + bright light, (c) close mouth + low light, (d) open mouth + low light.
Table 12. Face Experiments.

<table>
<thead>
<tr>
<th></th>
<th>Gallery</th>
<th>Probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1</td>
<td>27 (RF,BL)</td>
<td>27 (RF,LL)</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>20 (RF,BL) + 7 (MF,BL)</td>
<td>20 (RF,LL) + 7 (MF,LL)</td>
</tr>
</tbody>
</table>

RF: Regular Face. MF: Modified Face. BL: Bright Light. LL: Low Light.

5.3 Experimental Results

Two side-view images (intensity + range) were acquired for each subject (open-mouth, close-mouth) with a Minolta Vivid-900 range scanner. Each subject was imaged under two illumination conditions: bright and low light conditions. As a result, each subject has 4 images: open-mouth + bright-light, close-mouth + bright-light, open-mouth + low-light and close-mouth + low-light (see Figure 26). The complete data set contains 27 subjects (108 images). To study the efficacy of the proposed method in the presence of modified faces (tissue properties are changed due to surgery, trauma or burn, either intentionally or accidentally), a transparent rectangular tape was attached on the face of 7 subjects. The tape is less stretchable and thus has the same effect of modified faces and results in abnormal strain patterns. Two experiments were performed with different gallery and probe compositions (see Table 1). The results are presented with the standard receiver operating characteristic (ROC) curves that is used to quantify identification performance. The normalization procedure as reported in FRVT2002 [40] was used in the ROC computation.

The purpose of Experiment-1 is to investigate whether elastic strain pattern has enough discrimination power for recognition of regular faces. Experiment-2 is to understand whether the modified faces are more difficult to recognize or more distinguishable due to their unusual strain pattern.

The ROC curves of two experiments are shown in Figure 27. For the gallery and probe sets that include only regular faces, a verification rate of 67.4% was observed at a false alarm rate of 5%. Although the results were obtained with a relatively small data set, the
Figure 27. ROCs for Regular and Modified Faces.

performance is still quite promising. The data sets do not contain frontal views that are commonly used in face recognition test.

As expected, the experiment with gallery and probe sets that contain modified faces shows a better performance. At a false alarm rate of 5%, we were able to achieve a verification rate of 78.2%. 6 of 7 modified faces were correctly identified in Experiment-2 while only 3 of 7 regular faces of the same subjects were identified in Experiment-1. The two experiments suggests that a person who changed his/her appearance by plastic surgery or other approaches actually has a better chance to be detected because medical surgery cause property changes of facial tissues, which is hard to detect using the methods that rely on visible cues only.

The rationale for the new biometrics is that the underlying anatomical structure is unique for each individual, and this physiological invariant can be explored for face recognition. The proposed method has the advantage that recognition reaches beyond the visible faces that are usually seen and used by other face recognition methods, which can be confused by various camouflage using plastic surgery or makeup.

Several issues that need to be addressed in future investigations are: (1) a larger data set is needed to thoroughly evaluate the performance; (2) correspondence matching needs
to be automated. Optical flow and motion measurement in video sequence also can be considered; (4) using strain pattern on the whole face rather than a portion of face; (5) a comparative study with other methods is necessary.
CHAPTER 6
SENSITIVITY ANALYSIS

In comparison to geometrical and mass-spring models, physical models that are based on continuum mechanics have high computational complexity. As a result, various assumptions are often made to simplify the model and its parameters. For example, a commonly used assumption is that the material properties of an object are isotropic and homogeneous. However, results from large amount of biomechanical tests [42] indicate that the mechanical behavior of many biological materials, especially soft tissues, cannot be accurately described by such a simplified model. Certain types of muscles (such as skeletal muscle) are characterized by strong anisotropic behaviors. More importantly, the property heterogeneity of several orders of magnitude is also common in human organs [28].

Recognizing the inadequacy of simplified physical models, researchers have started to investigate to what degree the various assumptions, especially those about the material properties and the boundary conditions, may affect the model’s performance by means of sensitivity analysis. For example, Alterovitz et al [4] studied the influence of both the physician-controlled parameters and the intrinsic material parameters on the accuracy of needle insertion simulation. In the study of model-based breast cancer diagnosis, Tanner et al [123] compared the results of biomechanical models with different settings of boundary condition and material properties. However, those studies were done on the case-by-case basis using ad hoc comparison methods, and therefore the conclusions cannot be readily generalized to other domains. Moreover, the experiments were performed based on the assumption that the model has homogeneous material properties, which implies that the solution of a Dirichlet type problem could be independent of the internal property variation, and hence the subsequent sensitivity analysis results may not be valid. Another limitation
of their methods is that the sensitivity data is incapable of providing a complete picture of the model’s spatial response to the parameter variation on each point of the model.

In this section, a local gradient-based computation method is discussed that can be used to conduct a systematic and comprehensive sensitivity analysis of any motion model. Specifically, physics-based nonrigid motion modeling will benefit from such a sensitivity analysis in the following aspects:

1. the proposed method allows us to compare the relative importance of different parameters using the normalized sensitivity data and quantify the impact of various assumptions on model’s performance using the dimensional sensitivity data;

2. the algorithm is designed based on the adjoint state method, which significantly reduces the computational cost and is suited for handling large scale numerical models;

3. the sensitivity contour map enables us to identify the vulnerable areas of a model that are most affected by a poor assumption, so that further improvement can be made.

4. as discussed in the previous sections, the parameter value can be obtained by either the direct measurement or the indirect inference. But those acquisition procedures are time-consuming and expensive. It would be economic to first conduct a sensitivity analysis to identify the primary parameters and then to concentrate our effort on the acquisition of those parameters.

6.1 Computational Method

Sensitivity analysis is closely related to optimization problems often encountered in the traditional model calibration tasks, such as optimal shape design, boundary condition specification, discretization strategy investigation, as well as material property assignment [72]. In this study, the focus is on assessing the impact of two elastic parameters (the Young’s modulus and the Poisson’s ratio) on model’s performance, based on the computed sensitivity information. Without loss of generality, we will discuss the deformation of a linear elastic
body and its response to parameter perturbation. However, the methodology developed here can be readily applied to nonlinear systems.

6.1.1 Primary Problem

Given the governing equations of elastic dynamics, a variational method can be used to derive the finite element formulation of the partial differential equations so that a numerical solution of the forward model can be obtained. For example, using the Galerkin method [151], the primary problem in the static equilibrium can be discretized into a linear finite element system:

$$ K u = b, $$

(97)

where $K$ is the stiffness matrix, $u$ denotes the nodal solution, and $b$ is the load vector.

For each finite element, an interpolation matrix ($R_i$) (the basis function) is used to relate the displacements in the local element coordinate ($d_i$) and the global coordinate ($u$), which ensures that the requirement of strain compatibility is satisfied:

$$ d_i = R_i u. $$

(98)

After the discretization, the corresponding constitutive equation (2) and the strain-displacement equation (3) in the discrete matrix form are:

$$ \sigma_i = K_i \varepsilon_i, $$

(99)

$$ \varepsilon_i = D_i u, $$

(100)

where $K_i$ is the coefficient of stiffness matrix and $D_i$ is a matrix derived from the spatial differentiation of $R_i$ and the combination of row items. Note that $K_i$ is comprised of the material properties ($E, \nu$) that will be studied in the sensitivity analysis.
With the above notations, the stiffness matrix and the load vector in (8) that assemble all the finite elements become:

\[
K = \sum_{i=1}^{M} \int_{V} D_i^T K_i D_i dV, \tag{101}
\]

\[
b = \sum_{i=1}^{M} \int_{V} R_i^T q_{1i} dV + \sum_{i=1}^{M} \int_{S} R_i^T q_{2i} dS + \sum_{j=1}^{L} q_{3j}, \tag{102}
\]

where \( M \) is the total number of elements, \( V \) denotes the size of element, \( q_{1i} \) is the body force within an element, \( q_{2i} \) is the surface traction term, \( q_{3j} \) is the point load, and \( L \) is the number of point loads exerted on the body.

Since the interest is in the sensitivity assessment of a mechanical system governed by (8) to the parameter variations, (8) is rewritten in terms of the parameter vector:

\[
K(p)u = b(p), \tag{103}
\]

where \( p \) denotes the generic parameter to be investigated, which could be the material properties, the stress or strain distributions, the loading conditions (either the Dirichlet type or the Neumann type), object’s geometrical configuration, and even the meshing scheme itself. Here we concentrate on the sensitivity analysis of two spatial parameters, the Young’s modulus and the Poisson’s ratio: \( p = \{ p_j \} = \{ E_j, \nu_j \}, j = 1, 2, \ldots M. \)

6.1.2 Performance Measure and Dimensional Sensitivity

The simplest sensitivity analysis method that can be done in nonrigid motion modeling is based on a Taylor series describing the relationship between the state variable \( u \) and the parameter \( p \):

\[
u(p + \Delta p) = u(p) + \sum_{j=1}^{M} \frac{\partial u}{\partial p_j} \Delta p_j + \frac{1}{2} \sum_{j=1}^{M} \sum_{l=1}^{M} \frac{\partial^2 u}{\partial p_j \partial p_l} \Delta p_j \Delta p_l + \ldots \tag{104}\]
where the first-order derivatives \( \frac{\partial u}{\partial p_j} \) are the coefficients of the Jacobian matrix: \( J = [\frac{\partial u}{\partial p_j}] (i = 1, 2, \ldots N; j = 1, 2, \ldots M) \), with \( N \) and \( M \) as the total number of nodes and the total number of elements in a model, respectively.

The Jacobian matrix provides a local description of the response of the state variable to the parameter change and is adequate for analyzing a simple system such as the mass-spring model. However, for most practical engineering problems, more sophisticated sensitivity analysis method is needed to help us understand how a given model depends on certain parameters, so that the most significant ones can be identified and improved. To this end, a performance function (objective function) has to be defined:

\[
H(u, p) = \int_{\Omega} f(u, p) d\Omega.
\] (105)

The function defined in (16) is only a generic integral form. The exact formula of \( f(u, p) \) is often task-dependent, and the use of Euclidean norm is common.

Given a performance function \( H(u, p) \), the dimensional sensitivity (\( S \)) can be defined as its total derivative with respect to the parameter vector:

\[
S = \frac{dH(u, p)}{dp} = \int_{\Omega} \left[ \frac{\partial f(u, p)}{\partial p} + \frac{\partial f(u, p)}{\partial u} \frac{du}{dp} \right] d\Omega,
\] (106)

or in the discrete form:

\[
S = [S_1, S_2, \ldots S_j, \ldots S_M],
\] (107)

\[
S_j = \frac{dH(u, p)}{dp_j} = \frac{\partial H(u, p)}{\partial p_j} + \frac{\partial H(u, p)}{\partial u} \frac{du}{dp_j}.
\] (108)

Without causing any confusion, from now on, (18) and (19) will be represented with a single expression:

\[
S = \frac{dH(u, p)}{dp} = \frac{\partial H(u, p)}{\partial p} + \frac{\partial H(u, p)}{\partial u} \frac{du}{dp}.
\] (109)
6.1.3 Adjoint State Method

The dimensional sensitivity can be computed using the finite difference approximation, the direct differentiation or the adjoint state method [18]. The adjoint state method is chosen because it is more efficient for large scale problems. For a finite element model of modest size that are commonly encountered in physics-based nonrigid motion simulation, the number of elements could easily reach a couple of thousands. Therefore, the direct computation of sensitivity distribution would be too costly.

The adjoint state method is based on the principle of variational calculus [56], and can be derived from the governing partial differential equations in the continuous space (1-5), or from the linear system in the discrete space (14). This study will discuss the adjoint state approach following the second route.

(14) is first differentiated with respect to the parameter $p$:

$$\frac{\partial K(p)}{\partial p}u + K(p) \frac{du}{dp} = \frac{\partial b(p)}{\partial p}. \quad (110)$$

(21) is then multiplied by an arbitrary vector $\phi$ (adjoint state variable):

$$\phi^T \frac{\partial K(p)}{\partial p}u + \phi^T K(p) \frac{du}{dp} = \phi^T \frac{\partial b(p)}{\partial p}. \quad (111)$$

Combining (20) and (22), we have:

$$\frac{dH(u,p)}{dp} = \frac{\partial H(u,p)}{\partial p} + \frac{\partial H(u,p)}{\partial u} \frac{du}{dp} - \phi^T K(p) \frac{du}{dp} - \phi^T \frac{\partial K(p)}{\partial p}u + \phi^T \frac{\partial b(p)}{\partial p}. \quad (112)$$

Since $\phi$ is arbitrary, a $\phi$ is chosen such that the following equation is satisfied (the second and third terms in the right side of (23)):

$$\frac{\partial H(u,p)}{\partial u} \frac{du}{dp} - \phi^T K(p) \frac{du}{dp} = 0. \quad (113)$$
(24) is equivalent to:

\[
\frac{du}{dp}^T K(p)^T \phi = \left( \frac{du}{dp} \right)^T \left( \frac{\partial H(u, p)}{\partial u} \right)^T,
\]

which leads to the adjoint state equation:

\[
K(p) \phi = \left( \frac{\partial H(u, p)}{\partial u} \right)^T
\]

In the above derivation, we took advantage of the fact that the stiffness matrix is symmetric. Also note that the adjoint problem (26) is in the same form as the primary problem (14), except for the load term, \( \left( \frac{\partial H(u, p)}{\partial u} \right)^T \), which is also known as the pseudo-load.

Substituting (24) back into (23), it leads to:

\[
\frac{dH(u, p)}{dp} = \frac{\partial H(u, p)}{\partial p} + \phi^T \left( \frac{\partial b(p)}{\partial p} - \frac{\partial K(p)}{\partial p} u \right).
\]

As a result, we only need to solve the primary problem (14) and the adjoint problem (26) once, in order to obtain the state solution \( \mathbf{u} \) and the adjoint state solution \( \phi \), and then use (27) to compute the dimensional sensitivity. The adjoint state formulation can also be derived using the Lagrange multiplier method by treating (14) as the state equation, (16) as a constraint function and \( \phi \) as the Lagrange multiplier [26].

### 6.1.4 Derivative Computation

The partial derivative terms in (27) are needed in order to calculate the dimensional sensitivity. If the performance function is defined in an explicit form, the second term in (27) can be computed as:

\[
\frac{\partial H(u, p)}{\partial p} = \int_\Omega \frac{\partial f(u, p)}{\partial p} d\Omega.
\]

In most applications, \( f(u, p) \) is defined as a function of the state variable only, \( f(u) \), and therefore its partial derivative with respect to the parameter becomes zero.
With a finite element formulation, the partial derivatives of the stiffness matrix \( (K) \) and the load vector \( (b) \) with respect to the parameter \( (p) \) can be expressed as:

\[
\frac{\partial K(p)}{\partial p} = \sum_{i=1}^{M} \int_{V} D_i^T \frac{\partial K_i(p)}{\partial p} D_i dV \tag{118}
\]

\[
\frac{\partial b(p)}{\partial p} = \sum_{i=1}^{M} \int_{V} R_i^T \frac{\partial q_{1i}(p)}{\partial p} dV + \sum_{i=1}^{M} \int_{S} R_i^T \frac{\partial q_{2i}(p)}{\partial p} dS + \sum_{j=1}^{L} \frac{\partial q_{3j}(p)}{\partial p}. \tag{119}
\]

Since the interest is in the sensitivity values with respect to the material properties only \( (p = \{E, v\}) \), the derivative computations involved in (29) and (30) will only cause a slight increase of the computational cost.

### 6.1.5 Normalized Sensitivity

Dimensional sensitivity as defined in (20) is suitable for studying the dependence of a model on a single parameter (note that the vector \( p = [p_1, p_2, \ldots, p_M] \) of a finite element model is viewed as a single parameter here). Because each physical parameter usually has its own unit, a comparison among different parameters such as the Young’s modulus and the Poisson’s ratio using the dimensional sensitivity is very difficult and potentially could be misleading. In order to remove the dimensional effect, a normalized sensitivity \( (S_n) \) is defined as:

\[
S_n = \frac{p}{H(u,p)} \frac{dH(u,p)}{dp} = \frac{p}{H(u,p)} \left[ \frac{\partial H(u,p)}{\partial p} + \frac{\partial H(u,p)}{\partial u} \frac{du}{dp} \right] \tag{120}
\]

The normalized sensitivity measures the percentage change of the performance function to the percentage change of parameters. Therefore, it allows us to evaluate and rank the importance of different parameters in terms of their impact on the model on a common basis. Note that the normalized sensitivity does not require any changes of the computational
procedure (adjoint state method) developed for the dimensional sensitivity. If the statistical information about the measurement data is available, a normalized sensitivity function weighted by the measurement confidence coefficient can also be considered.

6.2 Experiments with Synthetic Model

6.2.1 Model Configuration and Performance Measure

The model used in this experiment represents a 2D elastic object that has a size of 10x8 (cm$^2$). The object is discretized into a finite element mesh with 357 nodes and 320 elements (Figure 28). The elements are of quadrilateral thin shell type. All elements are assigned with isotropic properties ($E = 50$ kPa, $\nu = 0.460$), except for the 32 elements in the middle of the modeling domain that have higher values ($E = 500$ kPa, $\nu = 0.495$), creating the property heterogeneity. This heterogeneously distributed properties are regarded as the true parameters: $p_t = \{E_t, \nu_t\}$. The Dirichlet boundary conditions are specified as follows: all nodes on the left side boundary of the model are fixed (displacement = 0 cm), while the compression loads are exerted along the right side boundary (displacement = 3 cm). The nodes on the top and bottom boundaries are set free.

With the true parameters and boundary conditions as specified above, the observation data ($\bar{u}$) are generated by running a forward simulation. The data vector is defined on all nodes ($\bar{u} = [\bar{u}_1, \bar{u}_2, ... \bar{u}_N]^T$). It should be noted that, in real applications, the state variable may be measured only on parts of the object.

Let’s first define a performance function in the form of discrete Euclidean norm:

$$ H(u) = \int_{\Omega} f(u) d\Omega $$

$$ = ||f(u)||^2 = \sum_{i=1}^{N} w_i^2 [u_i(p_a) - \bar{u}_i]^2 $$ (121)

where $w_i$ is a weight coefficient assigned to the $i$th node, $p_a$ denotes the assumed parameters, $u(p_a)$ is the solution that is obtained using the assumed parameters and $\bar{u}$ represents the
observation data that is generated with the true parameters. In all of the experiments, the unit weight coefficient: $w_i = 1$ is used.

This performance function consists of the state variable only. So the marginal sensitivity (the second term in (27)) becomes zero. In situations where the parameter measurements are available, they can also be incorporated into the performance function ($H(u,p)$). If stress or strain analysis is of interest, they can also be included to construct a Neumann type performance function.

Since homogeneity is a commonly used assumption, a model is tested that has the homogeneous properties: $p_a = \{E_a = 50 \text{ kPa}, \nu_a = 0.460\}$. The deformed models with the true heterogeneous properties ($p_t = \{E_t, \nu_t\}$) and the assumed homogeneous properties ($p_a = \{E_a, \nu_a\}$) are shown in Figure 29. The discrepancy between the two deformed shapes is clearly visible. Next, it will be demonstrated that (1) the normalized sensitivity can be used to identify the parameter that is most responsible for the discrepancy; (2) the dimensional sensitivity is helpful for determining how good an assumption is.
6.2.2 Normalized Sensitivity Distribution

The normalized sensitivities with respect to the Young’s modulus and the Poisson’s ratio are presented in Figure 30. The sensitivity distribution is plotted element by element as a contour map ($S_n = [S_{n1}, S_{n2}, ..., S_{nM}]^T$). The sensitivity coefficient in an element ($S_{nj}$) represents the percentage change of the performance function due to a 1% change of the parameter value in that element. The difference between the two maps in terms of the magnitude of sensitivity coefficient suggests that the model is more sensitive to the Young’s modulus than to the Poisson’s ratio. In other words, with this particular model configuration, Young’s modulus has a significant impact on the modeling accuracy, while the effect of the Poisson’s ratio is secondary.

Figure 30 (a) reveals that the model is most sensitive at the property discontinuities. The geometry of the high sensitivity area matches the shape of property heterogeneity. So, the sensitivity map is also informative about the location of parameter abnormalities to which effort should be dedicated to correct the original homogeneous assumption.
Figure 30. Normalized Sensitivity Contour Maps.

(a) $S_n$ of the Young’s modulus

(b) $S_n$ of the Poisson’s ratio

Figure 31. Dimensional Sensitivity Contour Maps.

(a) $S$ with $E_{\text{hetero}} = 70$ kPa

(b) $S$ with $E_{\text{hetero}} = 440$ kPa
6.2.3 Dimensional Sensitivity Distribution

It should be pointed out that certain formulations of the performance function may result in a sensitivity value of infinity. For example, with a performance function as defined in (32), the examination of (31) suggests that, as the assumed parameters approach the true parameters \((p_a \to p_t)\), the normalized sensitivity goes to infinity \((S_n \to \infty)\). Although it rarely occurs in real applications that our first assumption is as good as the true parameters, it is still wise to interpret the normalized sensitivity with caution.

In the context of nonrigid motion modeling, a two-step sensitivity analysis procedure is recommended: (1) the normalized sensitivity can be used to quantify the relative importance of parameters in terms of their impact on the model’s behavior; (2) once a parameter is considered as significant, the dimensional sensitivity can be used to assess the assumptions about the parameter. Without much loss of rigor, the closer the assumed parameters \((p_a)\) are to the true parameters \((p_t)\), the smaller the dimensional sensitivity is.

The experiments investigate two assumptions about the Young’s modulus inside the area of heterogeneity (the 32 elements in the center of the model): \(E_1 = 70\) kPa and \(E_2 = 440\) kPa. The Young’s modulus in the background elements and the Poisson’s ratio are the same as the true values. The dimensional sensitivity contours using those two assumptions are shown in Figure 31 (unit = \(m/Pa\)). The smaller magnitude of the second map indicates that the assumption \((E_2 = 440\) kPa) is much closer to the true value \((E_t = 500\) kPa) and is a better assumption than the first one \((E_1 = 70\) kPa). In the study of parameter estimation, the inverse algorithms also utilize the sensitivity information, although the regularization is more important in solving the ill-posed inverse problems [148].

6.2.4 Relative Error Distribution

The distribution of relative modeling error can also characterize the complex relationship between a model’s behavior and its parameters. The relative modeling error is defined as follow:

\[
err = \frac{|u(p_a) - \bar{u}|}{|\bar{u}|}. \tag{122}
\]
To determine the contribution to the relative modeling error by an individual parameter, two experiments were carried out, one with the true heterogeneous Young’s modulus and the assumed homogeneous Poisson’s ratio, and one with the assumed homogeneous Young’s modulus and the true heterogeneous Poisson’s ratio. The results are shown in Figure 32. The errors caused by an incorrect homogeneous assumption of the Young’s modulus is around 10% to 40%, while the errors due to a bad assumption of the Poisson’s ratio is less than 0.08%, which confirms the conclusion drawn from the sensitivity analysis that the Young’s modulus is more significant than the Poisson’s ratio under the configuration of this synthetic model.

Figure 32. Relative Error Distributions.

$E_t$: true Young’s modulus, $E_a$: assumed Young’s modulus, $\nu_t$: true Poisson’s ratio, $\nu_a$: assumed Poisson’s ratio.

Figure 33. Deformation of Burn Scar.
6.3 Experiments with Burn Scar

The experiment with synthetic model has implications to many practical modeling problems, because the property values used in the experiment are based on the published data. For example, in needle insertion simulation, a Young’s modulus in the range of 40-160 kPa was used for prostate tissue [4]. In breast modeling, the variation of $E$ among different tissues can be up to an order of magnitude, 10-88 kPa for the skin tissue and 1-184 kPa for glandular tissue [10, 123]. The Poisson’s ratio used in soft tissue modeling is often in the range of 0.465-0.495. Therefore, the modeling scenario of synthetic model is very realistic and the sensitivity analysis demonstrated with the synthetic model can be readily extended to real objects. In the next experiment, the sensitivity method will be applied to burn scar.

6.3.1 Model Configuration

A patient’s arm with a burn scar was used in the experiment. Using an ink stamp, a rectangle grid was created on the arm’s surface. Two images were then taken before and after the patient’s arm was stretched (see Figure 33). A finite element model with quadrilateral thin shell element was then constructed that matches the rectangle grid (see [135] for detailed descriptions of the modeling procedure). A Dirichlet condition was specified around the boundary using the measured displacement data and the internal nodes were set free.
6.3.2 Normalized Sensitivity

The finite element model covers both the scar (in the middle) and the surrounding normal skins. A homogeneous property distribution throughout the model was assumed for both the Young’s modulus and the Poisson’s ratio ($E = 60$ kPa, $\nu = 0.495$). The normalized sensitivity contours computed with the above setting are shown in Figure 34. Again, the larger values in Figure 34 (a) indicate that the scar model is more sensitive to the change of the Young’s modulus than to that of the Poisson’s ratio, although the difference between the sensitivity values of $E$ and $\nu$ is not as large as those observed in the experiments with synthetic model.

It can be seen that the region of high normalized sensitivity values matched well with the overall shape of scar (in both the Young’s modulus map and the Poisson’s ratio map). This similarity between the scar boundary and the sensitivity contours indicates that the model is very sensitive to the property change between scar and normal skin. It has been observed that the Young’s modulus of the burn scars could be 2-10 times higher than that of the normal skin (5-100 kPa) [135]. Based on the normalized sensitivity map of the Poisson’s ratio in Figure 34 (b), it seems reasonable to expect that the same type of heterogeneity of Poisson’s ratio exists between the scar and the normal skins, although the magnitude of the heterogeneity still remains unclear.

6.3.3 Dimensional Sensitivity

A homogeneous elasticity is certainly not adequate for scar assessment study. The next example is used to demonstrates how to find a better property value based on the dimensional sensitivity information. The Young’s modulus value of scar (8 elements in the middle of the model) was gradually increased from the initial 60 kPa until no further improvement can be observed in terms of the model’s dimensional sensitivity distribution. To quantify the evaluation process, an average dimensional sensitivity value is used:

$$S_{ave} = \frac{1}{M} \sum_{j=1}^{M} S_j.$$  \hspace{1cm} (123)
As discussed in the previous section, the smaller $S_{ave}$ is, the closer the corresponding assumption is to the true distribution. The final Young’s modulus value of scar which resulted in a minimum $S_{ave}$ is 450 kPa. The dimensional sensitivity contour maps of two assumptions ($E_{scar} = 60$ kPa and $E_{scar} = 450$ kPa) are shown in Figure 35. The smaller sensitivity coefficient values in the second map (b) clearly indicates that the heterogeneous assumption ($E_{background} = 60$ kPa, $E_{scar} = 450$ kPa) is a better choice than the original homogeneous assumption ($E_{background} = 60$ kPa, $E_{scar} = 60$ kPa). Note that the landscape of the dimensional sensitivity contour in Figure 35 (b) becomes more “flat” than that in Figure 35 (a), which is a result of using more accurate property values in the model.
6.4 Experiments with Deformed Face Model

Face recognition is usually performed using static intensity images [103, 137]. Recently, the investigation of face biometrics has been extended to the dynamic domain such as facial expression [142, 143], motion field [116], and strain patterns [147]. The new biometrics attempt to explore the uniqueness of soft tissue properties and the associated elastic patterns exhibited during face expressions. However, no quantitative physical link between the hypothesis of new biometrics and the existence of property abnormality has been established. In the next experiment, it is shown that the sensitivity analysis may provide quantitative measures that are useful for the design of new face biometrics.

6.4.1 Face Model Configuration

Two profile images (side view) of a subject’s face were taken using a Minolta 3D range camera, one with mouth closed and the other one with mouth open (see Figure 36). To facilitate correspondence matching, a set of points (8x8) was marked on the subject’s face. A rectangular transparent tape was attached on the face surface close to the lower jaw. The tape is less elastic and is used to simulate property change of facial tissue resulting from either a plastic surgery or an accident. The face model consists of 64 nodes and 49 quadrilateral elements. Face deformation measured from two images will then be used by the finite element model to infer any property change.

6.4.2 Normalized Sensitivity

The normalized sensitivity was used to determine which parameter, the Young’s modulus or the Poisson’s ratio, is more informative about the existence of face property change. In other words, to which parameter variation the face model is more sensitive. As in the previous experiments, a Dirichlet condition was specified along the model boundary and the internal nodes were set free. Homogeneous properties ($E = 50 \text{ kPa}$, $\nu = 0.495$) were also assumed. The resulting normalized sensitivity distributions are plotted in Figure 37.
In both maps, the dense contour lines that surround the area of high $S_n$ values clearly outline the location/shape of the tape. The average sensitivity value inside the tape is 0.148 for the Young’s modulus and 0.013 for the Poisson’s ratio, a difference of about one order of magnitude. So, the face model is more sensitivity to the change of the Young’s modulus than to the change of the Poisson’s ratio. This observation is consistent with that drawn from the experiments using synthetic model and burn scar. The quantitative sensitivity information has following implications: (1) it is easier to detect/estimate the change of the Young’s modulus than change of Poisson’s ratio; (2) The elasticity pattern is a better candidate to design an effective face biometrics.

6.4.3 Dimensional Sensitivity

In the experiment with burn scar, we show that the dimensional sensitivity can be used to improve the initial value by changing the property value inside the abnormal area. The procedure is possible because we know the exact location of the abnormal area. For example, the boundary between the scar and the normal skin is clearly visible on the images. However, in face recognition, the change in visible cues (color, texture and intensity) may not correspond to any underlying property abnormalities, either due to the relatively low
image quality or the make-up put on the face. So, we devise a new procedure that utilizes the dimensional sensitivity itself to guide the search for a better property value.

The experiments start with an initial homogeneous property and compute the dimensional sensitivity. Elements are then selected that have higher sensitivity values as the potential abnormal areas. The property value inside an abnormal area will be changed until its sensitivity value is close to the average of entire model. The above steps will be repeated until the sensitivity values throughout the model are below a user-specified threshold. In other words, the updated value will get closer to the true property values after each step.

A few intermediate results and the final sensitivity contours are shown in Figure 38. Based on the initial sensitivity values, four elements were selected as the abnormal area, all covered by the tape. The Young’s modulus of the four elements were then changed to 230 kPa and 810 kPa, respectively, with the corresponding sensitivity maps shown in Figure 38 (a,b). Note that the sensitivity values inside the taped area are lower than that of the surrounding regions. Using the information in Figure 38 (b), eight more elements that have higher sensitivity values were selected, three below the tape and five above the tape. The Young’s modulus of the eight elements were changed to 150 kPa and 370 kPa, respectively, and the final sensitivity is given in Figure 38 (d). The final sensitivity map has much smaller magnitude and shows relatively uniform landscape, implying that the final Young’s modulus value is more accurate, and therefore more suited for face recognition using a physical model.

Model-based nonrigid and articulated motion analysis requires careful design and calibration of the model being used, whether it is a geometrical model or a physical model. Various assumptions about the model and its parameters must be thoroughly evaluated. *Ad hoc* sensitivity analysis method is impractical for this task, especially when dealing with large scale problems. A systematic sensitivity analysis method is proposed that is capable of making a quantitative and reasoned diagnosis of the model’s performance related to our assumptions about the parameters. The proposed method is formulated using the first-order
(a) $E_{abnormal} = 230$ kPa
in 4 abnormal elements

(b) $E_{abnormal} = 910$ kPa
in 4 abnormal elements

(c) $E_{abnormal} = 150$ kPa
in 8 abnormal elements

(d) $E_{abnormal} = 370$ kPa
in 8 abnormal elements

Figure 38. Dimensional Sensitivity Maps of Face Model.
local gradient information. The adjoint state method is employed to reduce the computational cost of large scale numerical models. A two-step procedure is recommended: (1) the normalized sensitivity is suited for identifying the most significant parameters; (2) the dimensional sensitivity can be used to assess and improve our assumption about a particular parameter.

Based on the experiments with synthetic models, burn scars, and faces, several observations can be made: (1) the models are more sensitive to the change of the Young’s modulus than to the change of the Poisson’s ratio; (2) the models are most sensitive at the property discontinuities (heterogeneity); (3) sensitivity map is informative about the location/geometry of the property abnormalities;

It should be stressed that the experiments reported here were performed using a linear elastic model. The system response of a nonlinear model to the parameter variation is more complex. In future investigations, the proposed method will be applied to address the following issues: (1) How sensitive an elastic model is to an assumed isotropic property? (2) What is the sensitivity distribution of an elastic model under various loading conditions (boundary conditions)? (3) What is the sensitivity response of a nonlinear model? Those cases will be studied that are geometrically nonlinear (Green strain tensor) and materially nonlinear (viscoelastic and plastic materials).
A new physics-based approach for modeling nonrigid motion is proposed. The novelty of the proposed method lies in the fact that the motion parameters are actual physical parameters. The physical parameters have clear physical meanings, and more importantly, they only need to be recovered once. In the tracking experiments, a scheme is demonstrated that uses the recovered elasticity and the boundary correspondence to synthesize the degrees of freedom of the internal nodes, which reduces the tracking complexity significantly. The proposed modeling approach can be applied to a wide range of domains. In particular, its application is demonstrated in burn scar assessment and face recognition.

In burn scar study, an method based on relative elasticity ratio and natural features is developed. The use of natural features and adaptive mesh not only increases the computational efficiency, but also allows us to work with scar images without the need of tagging artificial markers. Burn scar damage is quantified by estimating its Young’s modulus using a simplified regularization method. Experiments show that the proposed method is robust when presented with noisy data. The positive correlation between the physician’s rating and REIs (using both natural features and artificial markers) suggests that the proposed approach can provide a quantitative and objective evaluation of burn scar damage.

In face recognition study, a new biometrics is proposed that is based on the consideration of anatomical and biomechanical characteristics of facial tissues. Elastic strain pattern inferred from facial expression can reveal an individual’s biometrics signature associated with the underlying anatomical structure, and thus has the potential for face recognition. A method based on the continuum mechanics in finite element formulation is employed to compute the strain pattern. Experiments show very promising results. The proposed
method is quite different from other face recognition methods, and both its advantages and limitations, as well as future research for improvement are discussed.

A comparative study is conducted to evaluate the strength and weakness of two recovery algorithms. In all of the experiments, both the CGA and the GNM showed their robustness against data noise by converging to a stable and near-optimal solution, attributed to the constraint imposed through the penalty function. Because of its stochastic nature, the CGA is more demanding than the GNM method in terms of the need of computational resources. Although the CGA is a computationally intensive method, it has a very desirable feature of being less sensitive to the initial conditions. By maintaining a diversified population pool, CGA can explore the solution space on a global scale without being caught in local extrema, which has been the main obstacle to the gradient-based methods.

The proposed motion analysis approach are valid for modeling many deformable objects. But the following areas are identified to which future works need to be dedicated.

**Linear or Nonlinear Model:** Linear model is conceptually and computationally attractive and suffices our need in most cases. However, large amount of biomechanical experiments shown that many live objects, especially soft tissues, do not obey the Hooke’s law strictly. In addition, Cauchy strain tensor derived from infinitesimal deformation assumption is no longer suitable for objects of large deformation. When to abandon linear model is a tricky issue because nonlinear model is far more complex. The validity of using linear model is task-dependent and the choice is often made based on modeler’s knowledge on object’s material characteristics. One possible solution is to derive empirical constitutive formulation (pseudo-elasticity) from tensile test and use it in motion equation [42]. Another possibility is to carry out a thorough numerical experiment to assess the feasibility of using linear model in a particular task. Unfortunately, such kind of numerical exercise is rarely practiced in motion analysis.

**Isotropic or Anisotropic Property:** Similar problem arises in the choice of isotropic or anisotropic property. Many biological materials exhibit anisotropic deformation under stress. Anisotropic property is determined by material composition and substructure (ori-
entation and arrangement of constitutional units). Blood vessels and skeletal muscles are good examples. For a 2D anisotropic object, recovering its Young’s modulus is equivalent to solving an inverse problem with two spatially varying unknowns \((E_x, E_y)\). This should not pose a severe mathematical challenge to the recovery algorithm, if the constitutive and motion equations are properly formulated by taking anisotropic property into account.

**Alternative Regularization Methods:** In addition to the Tikhonov regularization, there exists a large pool of alternative regularization methods that can be used for nonrigid motion analysis, noticeably the iterative regularization and multigrid regularization. As to statistical methods, we refer readers to [124] for classical application-oriented materials and [33] for recent developments.

The fact that the Tikhonov functional of nonlinear problem is difficulty to solve raises the interest of investigating iterative regularization method. The fundamental theory of iterative regularization is based on the "semiconvergence" property observed in the plot of solution error against the iteration count. For example, in a Landweber iteration, solution error gradually decreases with iteration count and then blows out after passing an optimal iteration number. This self-regularization property allows us to stabilize the solution by stopping iteration early. The iteration count plays a role that is similar to the regularization parameter. Choosing an appropriate stopping rule is still a challenging issue and the discrepancy principle seems to be the best choice [15, 30].

Multigrid method is well known for its stability and efficiency in solving large-scale well-posed problems. In finite dimensions, multigrid is expected to yield a stable solution for a discretized inverse problem, but no theory has been rigorously established. Studies have shown that grid refinement itself (dynamic meshing or adaptive meshing) posses the property of regularization [78, 8]. One attractive feature of multigrid method is that parameter space and hence computational complexity can be significantly reduced. Grid refinement can be done in either parameter space (Young’s modulus) or state variable space (displacement or force). If refinement is carried out in parameter space, the finest grid is usually predefined and iteration is implemented through the “zonation” technique where the pa-
rameter in a particular zone is assumed constant. This “zonation” technique effectively reduces the parameter space and has the effect of stabilizing the solution and accelerating the convergence. Applications of multigrid method in estimating tissue elasticity and aquifer hydraulic transmissivity have been reported in [139, 6]. One of the difficulties in implementing multigrid method is to choose an appropriate rule to stop the refinement process. The possibility of combining the Tikhonov regularization with multigrid method is also investigated [8].
REFERENCES


ABOUT THE AUTHOR

Yong Zhang received B.S. degree in Marine Science from Ocean University of Qingdao, China in 1986 and M.S. degree in Marine Science from the same school in 1989. He also received M.S. degree in Computer Science and Engineering from University of South Florida in 2000. He is currently a PhD candidate at University of South Florida. His research interests include computer vision, motion modeling, image processing, pattern recognition and artificial intelligence.