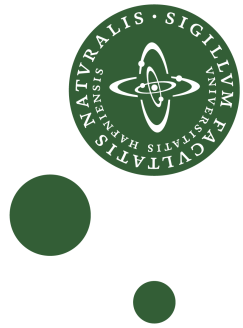
# Statistical Classification and Level Set Methods in Medical Image Analysis

# Quantitative Evaluation of Articular Cartilage in Knee MRI

Jenny Folkesson




University of Copenhagen
Faculty of Science
Department of Computer Science
2007

University of Copenhagen
Faculty of Science
Department of Computer Science
Universitetsparken 1, DK-2100, Copenhagen, Denmark
Phone: (+45) 35 32 14 00
Email: diku@diku.dk
http://www.diku.dk/

*I love deadlines. I like the whooshing sound they make as they fly by.*

Douglas Adams

# Preface

This thesis was prepared at the department of Computer Science at the University of Copenhagen (DIKU), in partial fulfillment of the requirements for acquiring a Ph.D. degree in Computer Science.

The thesis deals with the research and development of methods for segmentation and quantitative evaluation of medical images, with focus on the articular cartilage in knee magnetic resonance imaging (MRI) data in relation to osteoarthritis (OA). It consists of a discussion and motivation for the cartilage evaluation framework, followed by selected research papers that have either been published, accepted for publication during the course of the Ph.D. studies, or are undergoing a review process at the time this thesis was handed in.

The work has been carried out in collaboration with Center for Clinical and Basic Research (CCBR), Ballerup, Denmark, and Nordic Bioscience, Herlev, Denmark. Approximately the first two years of the study period was spent at the IT University of Copenhagen. A part of the research was conducted at Brigham and Women's Hospital (BWH), Harvard Medical School, Boston, USA.

The project was supervised by Ole Fogh Olsen currently at Oticon (formerly Associate Professor at the IT University of Copenhagen), and co-supervised by Dr. Erik B. Dam (Nordic Bioscience), and Professor Mads Nielsen (DIKU, Nordic Bioscience). Funding for the project was provided by CCBR and Nordic Bioscience.

Copenhagen, September 2007

Jenny Folkesson

# Acknowledgements

I would first and foremost like to thank my supervisor, Ole Fogh Olsen, for much guidance and support, and for always being good-humored and encouraging. I would also like to thank my co-supervisor Erik Dam for all his help and advice. During my visit at Brigham and Women's Hospital in Boston, Carl-Fredrik Westin supervised me, which I thank him for. It was nice to have one supervisor who could speak comprehensible Swedish at least...

CCBR and Nordic Bioscience funded this project and provided most of the data used in this thesis, along with expertise in medicine and clinical research. Their contributions made this project possible and for that I am most grateful.

Thanks also goes to the people I have worked with and co-authored research papers with. Their names are listed in the Publications section and underneath the title in the chapters they concern.

I thank my colleagues, mostly people at the Image Group at DIKU, and at the Surgical Planning Laboratory at Brigham and Women's Hospital for inspirational, and often entertaining, work environments. In particular I would like to thank Karl Sjöstrand for kindly volunteering to proof-read this thesis.

Last but not least, I would like to thank my family and friends for all their love and support.

# Abstract

This thesis presents research and development of methods for automatic segmentation and quantitative evaluation of anatomical structures from medical images. An automatic segmentation algorithm is the foundation on which this framework stands, therefore a substantial part of the thesis concerns this subject. The segmentation methods focus on statistical voxel classification, using feature selection to find representative features for the classes, and incorporation of level set methods into the classification is also investigated. The segmentation methods are evaluated on magnetic resonance imaging (MRI) data of the knee and heart.

The thesis also contains a quantitative evaluation of articular cartilage in relation to osteoarthritis (OA), with the aim of developing and evaluating imaging-based OA biomarkers suited for clinical trials. Factors such as joint incongruity, cartilage lesions, and fibrillation can all be linked to OA, and they can in turn be related to cartilage surface curvature at different scales. Cartilage curvature is analyzed using mean curvature flow and by a comparison of surface normals. Besides curvature, other biomarkers founded on the automatic segmentation algorithm such as volume, thickness, and homogeneity are also evaluated as potential OA biomarkers.

The automatic segmentation methods described in this work perform well compared to gold standard segmentations, and compare favorably to state-of-the-art segmentation methods. The OA biomarkers obtained founded on the automatic cartilage segmentation show potential as disease markers as they are clearly related to radiographic signs of OA.

# List of Publications

Below, publications that were produced throughout the course of the Ph.D. studies are listed. They consist of 6 journal papers, 10 conference papers, 13 conference abstracts, and 2 patents. The first author is the main investigator for these publications.

## Peer-Reviewed Journal papers:

1. Jenny Folkesson, Erik B. Dam, Ole F. Olsen, Paola Pettersen and Claus Christiansen, *Segmenting Articular Cartilage Automatically Using a Voxel Classification Approach*, IEEE Transactions on Medical Imaging, 26: 106-115, 2007.

2. Jenny Folkesson, Erik B. Dam, Ole F. Olsen, and Claus Christiansen, *Accuracy Evaluation of Automatic Quantification of the Articular Cartilage Surface Curvature from MRI*, Academic Radiology, 14 (10): 1221-1228, 2007.

3. Jenny Folkesson, Erik B. Dam, Ole F. Olsen, Morten A. Karsdal, Paola Pettersen and Claus Christiansen, *Automatic Quantification of Global and Local Articular Cartilage Surface Curvature: Markers of Smoothness and Congruity?*, Submitted to Magnetic Resonance in Medicine, 2007.

4. Jenny Folkesson, Eigil Samset, Raymond Y. Kwong, and Carl-Fredrik Westin, *Unifying Statistical Classification and Geodesic Active Regions for Segmentation of Cardiac MRI*, IEEE Transactions on Information Technology in Biomedicine, 2007. *In press.*

5. Erik B. Dam, Jenny Folkesson, Paola Pettersen and Claus Christiansen, *Automatic Morphometric Cartilage Quantification in the Medial Tibial*

*Plateau from MRI for Osteoarthritis Grading*, Osteoarthritis & Cartilage, vol. 15 (7): 808-818, 2007.

6. Arish A. Qasi, Jenny Folkesson, Paola Pettersen, Morten A. Karsdal, Claus Christiansen and Erik B. Dam *Separation of healthy and early osteoarthritis by automatic quantification of cartilage homogeneity*, Osteoarthritis & Cartilage, 15 (10): 1199-1206, 2007.

## Peer-Reviewed Conference Papers:

7. Jenny Folkesson and Carl-Fredrik Westin, *Spatially Varying Classification with Localization Certainty for Level Set Segmentation*, MMBIA Workshop in conjunction with ICCV, 2007.

8. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Automatic Curvature Analysis of the Articular Cartilage Surface*, MICCAI Joint Disease Workshop, 2006.

9. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Position Normalization in Automatic Cartilage Segmentation*, MICCAI Joint Disease Workshop, 2006.

10. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Automatic Segmentation of the Articular Cartilage in Knee MRI Using a Hierarchical Multi-class Classification Scheme*, MICCAI, pp 327-334, 2005.

11. Jenny Folkesson, Ole Fogh Olsen, Erik B. Dam, Paola Pettersen and Claus Christiansen, *Combining Binary Classifiers for Automatic Cartilage Segmentation in Knee MRI*, ICCV First International Workshop CVBIA, pp 230-239, 2005.

12. Jenny Folkesson, Erik B. Dam, Paola Pettersen, Mads Nielsen, Ole Fogh Olsen and Claus Christiansen, *Locating Articular Cartilage in MR Images*, Proceedings of SPIE Medical Imaging: 1484-1490, 2005.

13. Erik B. Dam, Jenny Folkesson, Paola Pettersen and Claus Christiansen, *Semi-Automatic Knee Cartilage Segmentation*, Proceedings of SPIE Medical Imaging, 2006.

14. Erik B. Dam, Jenny Folkesson, Paola Pettersen and Claus Christiansen, *Automatic Cartilage Thickness Quantification using a Statistical Shape Model*, MICCAI Joint Disease Workshop, 2006.

15. Erik B. Dam, Jenny Folkesson, Marco Loog, Paola Pettersen and Claus Christiansen, *Efficient Automatic Cartilage Segmentation*, MICCAI Joint Disease Workshop, 2006.

16. Arish A. Qazi, Ole Fogh Olsen, Erik B. Dam, Jenny Folkesson, Paola Pettersen and Claus Christiansen, *Automatic Quantification of Knee Cartilage Homogeneity*, MICCAI Joint Disease Workshop, 2006.

## Peer-Reviewed Conference Abstracts:

17. Jenny Folkesson, Ole Fogh Olsen, Erik B. Dam, Morten A. Karsdal, Paola Pettersen and Claus Christiansen, *Cartilage surface incongruity: a biomarker for OA?*, Workshop on Imaging-based Measures of Osteoarthritis, Salzburg 2007.

18. Jenny Folkesson, Ole Fogh Olsen, Erik B. Dam, Morten A. Karsdal, Paola Pettersen and Claus Christiansen, *Longitudinal changes in cartilage surface smoothness: a marker of progression during moderate OA?*, Workshop on Imaging-based Measures of Osteoarthritis, Salzburg 2007.

19. Jenny Folkesson, Ole Fogh Olsen, Erik B. Dam, Paola Pettersen and Claus Christiansen, *Improving Automatic Tibial Medial Cartilage Segmentation by Incorporating Femoral Cartilage Compartment*, Proc. Intl. Soc. Mag. Res. Med., 2006.

20. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Robust automatic estimation of articular cartilage surface area from knee MRI*, Osteoarthritis & Cartilage, Supplement, 2006.

21. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Automatic volume estimation of articular cartilage from low-field knee MRI*, Osteoarthritis & Cartilage, Supplement, 2006.

22. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Automatic cartilage volume estimation from low-field knee MRI: a longitudinal study*, Osteoarthritis & Cartilage, Supplement, 2006.

23. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Robust Volume Estimation of Articular Cartilage from Knee MR Scans*, Osteoarthritis & Cartilage, supplement, 2005.

24. Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen and Claus Christiansen, *Automatic Quantification of Articular Cartilage from Knee MR Scans Using Surface Smoothness as Osteoarthritis Indicator*, Osteoarthritis & Cartilage, supplement, 2005.

25. Erik B. Dam, Jenny Folkesson, Paola Pettersen, Morten A. Karsdal and Claus Christiansen, *Automatic Knee Cartilage Volume Quantification Compared to Joint Space Width: Biomarkers of Longitudinal Progression?*, Accepted for Osteoarthritis & Cartilage, 2007.

26. Erik B. Dam, Paola Pettersen, Jenny Folkesson and Claus Christiansen, *Automatic Quantification of Cartilage Thickness from MRI for Monitoring Progression of Osteoarthritis - a Longitudinal Study*, Osteoarthritis & Cartilage, supplement, 2006.

27. Erik B. Dam, Paola Pettersen, Jenny Folkesson and Claus Christiansen, *Is Cartilage Thinning during Osteoarthritis Focal or Global?*, Osteoarthritis & Cartilage, supplement, 2006.

28. Arish A. Qazi, Erik B. Dam, Paola Pettersen, Ole Fogh Olsen, Jenny Folkesson, Claus Christiansen and Morten A. Karsdal, *Is Diseased Cartilage More Homogenous Than Healthy Cartilage?*, Osteoarthritis & Cartilage, supplement, 2006.

29. Erik B. Dam, Jenny Folkesson, Paola Pettersen and Claus Christiansen, *Automatic Cartilage Thickness Quantification from Knee MRI*, Osteoarthritis & Cartilage, Supplement, 2005.

## Patents:

30. Erik B. Dam, Jenny Folkesson, Paola Pettersen, Ole Fogh Olsen, Mads Nielsen and Claus Christiansen, *Automatic Quantification of a Pathology Indicating Measure from Cartilage Scan Data*, patent number PCT/EP2006/008606. World Patent submitted 2006.

31. Erik B. Dam, Jenny Folkesson, Paola Pettersen, Ole Fogh Olsen, Mads Nielsen and Claus Christiansen, *Automatic Quantification of a Pathology Indicating Measure from Cartilage Scan Data*, patent number GB 0521640.3. Submitted at British Patent Agency 2005.

# Contents

Chapter 1

# Introduction

This thesis presents the research and development of methods for segmentation and quantitative analysis of medical images, with focus on the articular cartilage in knee magnetic resonance imaging (MRI) data. The contents of this thesis are divided into two parts; the first part treats automatic segmentation methods for knee and cardiac MRI data, and the second part concerns biomarkers from articular cartilage segmented with a method described in the first part, and the relation between the biomarkers and osteoarthritis (OA).

The two parts are based on research papers of which most are published or accepted for publication over the course over the Ph.D. project. This thesis only include a selection of publications of which the author of this thesis is the first author. Each part starts with a chapter which provides a general overview of the subject, a presentation of the main results in the different papers, and a discussion of how the different papers relate to each other.

As the work progressed an increasing number of knee MRI scans were available and added to the data collection, therefore the number of scans mentioned in the different chapters varies. The chapters containing research papers are to some extent self-contained, and there is some overlap between the different chapters.

## 1.1 Purpose

The purpose of this Ph.D. project is to find methods for automatic quantitative evaluation of the articular cartilage from knee MRI data, and evaluate potential biomarkers for OA. In particular, the potential of cartilage surface curvature as a biomarker for OA is investigated. Because a segmentation algorithm is the foundation in the cartilage assessment framework, this thesis first describes automated segmentation methods in Part I. The methods are evaluated by how well they perform the task of segmenting the articular cartilage in knee MRI and the myocardium in late contrast-enhanced cardiac MRI compared to gold standard segmentations. In Part II, biomarkers for OA which are all founded on the automatic segmentation algorithm for cartilage described in Part I are evaluated, with emphasis on cartilage surface curvature. In Part II, the cartilage imaging sequence and clinical aspects are also discussed.

## 1.2 Scope

The first and most fundamental step in our quantitative cartilage evaluation framework is an automatic segmentation of the cartilage. The cartilage shape can vary significantly with the severity of signs of OA, in late stages of the disease the cartilage loss may even cause a change in topology of the cartilage sheets. We employ a voxel-based statistical classification approach to cartilage segmentation since cartilage may be adequately described by local image structure without enforcing assumptions on global shape, which is non-trivial to model due to possible changes in object topology. Feature selection allows us to find suitable features based on prior information. This automatic segmentation framework is described in Chapters 3 - 6 with an overview in Chapter 2. Level set methods can naturally deal with topology changes, therefore we explore ways to incorporate shape and regional information in a geodesic active region framework described in Chapters 7 and 8. These methods are evaluated on late gadolinium enhanced cardiac MRI data, where ambiguous boundary information and irregular appearance occur frequently.

When affected with OA the articular cartilage degenerates by loss of integrity and structure, and by thinning. Standard cartilage imaging biomarkers are volume, thickness and surface area, which quantify overall cartilage loss. From the cartilage segmentation, the cartilage volume and surface area can be measured directly, but for other biomarkers additional steps are required. This thesis focus on an analysis of the cartilage surface curvature, which is not among the more traditional OA related cartilage biomarkers but in Chapters 9 - 11 we describe the relation to OA and show results using curvature estimation based on mean

curvature flow and by comparison of surface normals from a shape model. Other biomarkers that are founded on the same automatic segmentation framework is a thickness estimation using a shape model and a cartilage homogeneity estimation quantifying cartilage structure. Research papers for these two estimates are not included in this thesis, they are referenced in Part II and listed in the Publications Section.

## 1.3 Contributions

One of the main contributions of this thesis is the development of a fully automatic segmentation of articular cartilage from MRI, which is described in Chapters 2 - 6 (references [48], [50], [52], [55]). Our cartilage segmentation algorithm shows good agreement with gold standard segmentations and compares favorably to state-of-the-art methods. Feature selection is performed on manually segmented training data for an evaluation of which features that can best describe the articular cartilage. The feature selection is also discussed in Chapters 2 - 6.

Automatic, quantitative evaluation of cartilage is here performed on low-field MRI data, unlike most other studies that use high-field scanners. A discussion on how it compares to more well-established cartilage imaging sequences and how low-field MRI-based measures can be used in clinical studies of OA can be found in Chapter 9.

The potential of articular cartilage surface curvature as a biomarker for OA is investigated by performing an analysis of both high- and low-scale surface curvature. We demonstrate that these curvature estimates are related to signs of OA in Chapters 10 - 12 (references [51], [54], [53]).

A framework for combining statistical classification with geodesic active regions is developed, which are components that benefit from each other and fit well together. These methods are described in Chapters 7 - 8 (references [56], [47]). The potential of these automatic methods are evaluated on segmentation of the myocardium in late gadolinium enhanced cardiac MRI.

# Part I

# Medical Image Segmentation

*If we knew what it was we were doing, it would not be called research, would it?*

Albert Einstein

Chapter 2

# Automatic Segmentation of Medical Images

The chapters of Part I consist of research papers describing statistical classification based methods for cartilage segmentation (Chapters 3-6) and combining statistical classification and global shape in Chapters 7-8 with examples on cardiac MRI. After a brief discussion of medical image segmentation and validation in Sections 2.2-2.3, this chapter is dedicated to a discussion on statistical classification (Chapters 2.4-2.4.4) and how the research papers in this subject (Chapters 3-8) relate to each other.

## 2.1 Introduction

Medical imaging allows for visualization of anatomy with no or little invasiveness, and is an invaluable tool in diagnostics, treatment planning, and clinical studies. The vast amount of data thus generated has been accompanied with time-consuming analysis, and the desire to reduce analysis time along with inter- and intra-observer variability has lead to a search for ways of automating the analysis. In particular, computer algorithms for automated delineation of anatomical structures, or segmentation algorithms, have become increasingly important in tissue quantification [106].

When developing automatic medical image segmentation algorithms one faces a number of challenges such as spatial resolution, noise, ambiguous boundary information, poor contrast between tissues, and variations in shape and intensity. This work deals with magnetic resonance imaging (MRI) data where typical artifacts are partial volume effects, for which different tissue types contribute to the intensity of one voxel, and intensity non-uniformity across the image, which depends on a number of factors such as electromagnetic properties of the test subject and spatial inhomogeneity of the excitation field [120]. Due to these challenges, the task of automatic segmentation is usually non-trivial and incorporation of prior knowledge needs to be considered.

The segmentation objectives in this thesis are to develop fully automatic segmentation methods for the articular cartilage in knee MRI and the myocardium in cardiac MRI. A priori knowledge is incorporated in terms of local features determined from feature selection for a segmentation framework based on supervised voxel-based statistical classification. Incorporation of global information using level set methods is also investigated.

## 2.2 Comments on Medical Image Segmentation in the Literature

The subject of medical image segmentation have been discussed extensively in numerous articles and books. Here are some brief comments on some of the existing literature, for more thorough overviews on the subject of medical image segmentation methods readers are referred to [120],[106].

An image can be seen as an ordered set of measurements, where the grid point for each measurement is called a pixel in 2D and voxel in 3D, and medical image segmentation is defined as the partitioning of the image into regions that correspond to distinct anatomical structures or regions of interest.

Image segmentation methods may be *edge-based*, *region-based*, or *classification-based*. Typically, edge-based methods exploit differences between regions and region-based methods exploit similarities within regions, whereas classification assigns class labels to individual voxels based on a combination of local image features. Deformable models, *i.e.* curves or surfaces deforming according to internal and external forces, can be robust to noise and ambiguous boundary information. *Parametric deformable models* (e.g. active contour models or snakes, Kass *et al.* [72]) represent curves or surfaces explicitly in their parametric forms during deformation and can have compact representations and fast implementations. *Geometric deformable models* on the other hand (e.g. geodesic active contours by Caselles *et al.* [15], and Kichenassamy *et al.* [77]) represent curves

or surfaces implicitly as a level set of a higher dimensional function, thereby allowing topological adaptivity.

Prior knowledge is often available in medical image segmentation tasks, and can greatly improve the algorithm performance by resolving ambiguities in image information. Incorporation of prior knowledge requires a training phase which typically involves interaction from a human expert in terms of manual delineation of structures of interest. In deformable models prior information of shape and/or regional statistics can be incorporated, examples are active shape models (ASM) [23], active appearance models by Cootes *et al.* [24], and deformable Fourier models (Staib and Duncan [122]) in the parametric case, and hybrid geometric deformable models (Guo and Vemuri, [63]) and geodesic active regions (Paragios and Deriche [102]) in the geometric case. Statistical classification seek to partition a feature space derived from the image using data with known labels. Atlas approaches treats segmentation as a registration problem, where a pre-segmented atlas image is mapped to the image, and are conceptually similar to classification but implemented in a spatial domain instead of a feature space.

## 2.3 Evaluating the Performance of the Segmentation Method

One approach to evaluation of a segmentation algorithm performance is how close it is to a truth model. Ideally, one would like to validate algorithms against a perfect model, *ground truth*. Phantoms with known ground truth can be generated, but they do typically not offer realistic representations of subject anatomy nor simulations of the image acquisition process. One could perform imaging *in vitro* and surgically remove the anatomical structure of interest and measure it after imaging. For example, knees could be scanned and the articular cartilage volume could be measured, *e.g.* using water replacement, and compared to volumes from automatic segmentations, but there are variations in measurements such as water replacement as well which means it may not lead to a perfect truth model.

A very common approach is to validate segmentation algorithms against *gold standard* segmentations, typically determined by manual expert delineations. This approach does not guarantee a perfect truth model due to inter- and intra-rater variability. There are methods designed to reduce variations in the truth model if one has access to a collection of human expert segmentations (*e.g.* simultaneous truth and performance level estimation (STAPLE) by Warfield *et al.* [133]). The data sets described in this work consists of *in vivo* data, and in

the segmentation part of this thesis we evaluate how the algorithms compare to gold standard segmentations by a radiologist. This allows for a comparison of the performance to other algorithms to some extent, however a full comparison can only be achieved if the algorithms are evaluated on the same data using the same truth model. In this part of the thesis gold standard segmentations by a radiologist are used as a truth model.

Other approaches to performance evaluation are how often the segmentation outcome results in an accurate diagnostic decision or examine how well the algorithms perform in a clinical setting. In this work the purpose is to evaluate potential biomarkers for OA, and in Part II we evaluate the biomarkers obtained from our quantitative cartilage evaluation framework by a comparison to a well-established disease score for OA. We also evaluate the inter-scan reproducibility of the automatically obtained biomarkers.

### 2.3.1 Performance Evaluation Compared to Segmentation Truth Model

Figures of merit need to be established for quantifying the performance of the segmentation algorithm compared to the truth model. The choice depends on the application and can be based on region information, such as number of accurately classified voxels, or boundary information, such as distance to the boundary of the truth model.

Even though data may contain multiple labels, they can be seen as multiple images for which each tissue class is represented as *foreground* (labeled "1") and the locations outside the tissue is *background* (labeled "0"). If we represent two different segmentation by two images and let $A$ and $B$ be the voxels in the respective images with label "1" with $|\cdot|$ denoting the number of voxels, then the number of counts of all possibilities of agreements between the segmentations can be summarized in a *confusion matrix*:

| | 1 | 0 |
|---|---|---|
| 1 | $a_{11}$ | $a_{12}$ |
| 0 | $a_{21}$ | $a_{22}$ |

Here, $a_{11} = |A \cap B|$, $a_{12} = |A - A \cap B|$, $a_{21} = |B - A \cap B|$ and $a_{22} = |\bar{A} \cap \bar{B}|$. If $A$ represents the reference segmentation, $a_{11}$ are *true positives*, $a_{21}$ *false positives*, $a_{11}$ *false negatives* and $a_{22}$ *true negatives*. Two of the simplest measures of agreement are the *sensitivity* and *specificity*

$$\mathrm{Se} = \frac{a_{11}}{a_{11} + a_{12}},$$

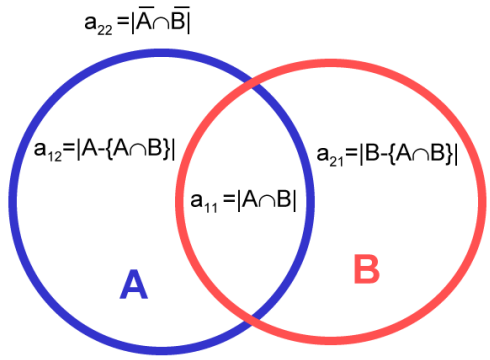



**Figure 2.1:** Segmentations $A$ and $B$ and scalar values $a_{11}$, $a_{12}$, $a_{21}$, $a_{22}$.

$$\mathrm{Sp} = \frac{a_{22}}{a_{21} + a_{22}},$$

which are the fraction correctly labeled foreground and background voxels respectively. Similarly, the *accuracy* can be defined as the fraction correctly labeled voxels in the image,

$$\mathrm{Ac} = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}},$$

where $T$ are the true class labels and $D$ are the labels from classification. However, such a sum can be misleading if one class is dominant in the image since it treats the correct classification rate the same for each class. Cartilage, for instance, covers only a few percent of the image hence even if the entire image is classified as background the accuracy can be quite high. An alternative definition of accuracy that accounts for the relative proportion of each class is to use the *positive and negative predictive values*, which extends the above formulation for accuracy by taking prior probabilities for each class into account, thus accounting for the relative proportion of each class.

For this type of problems, measures of overlap of the foreground components can be more interesting. Two well-established overlap measures are the *Jaccard coefficient* [67] (JC) and the *Dice similarity coefficient* [36] (DSC), defined as:

$$\mathrm{JC}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{a_{11}}{a_{11} + a_{12} + a_{21}},$$

$$\mathrm{DSC}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2a_{11}}{2a_{11} + a_{12} + a_{21}}.$$

Both DSC and JC range between 0 and 1, where 1 signifies perfect overlap, and 0 is no overlap. JC is almost linearly proportional to DSC and essentially captures the same information [9], therefore we use DSC exclusively in this work. For our knee MRI data we evaluate the segmentation algorithm also using sensitivity and specificity, since they have been used in related methods in literature.

The measures above assumes binary segmentations, however classifiers like the $k$NNcan produce soft labels. Soft segmentations may reduce partial volume effects in segmentations, and they can be converted to hard labels by thresholding or assigning the voxels to the class with highest probability. The *Receiver operating characteristics* [70] (ROC) can be worth examining since it evaluates the performance of a classifier for each possible threshold for the two-class case. This is achieved by calculating the fraction of true positives against the fraction of false positives for each threshold, and the area under the curve (AUC) gives a measure of the performance. The AUC can be misrepresentative for images with much background compared to foreground, as in the case of accuracy and specificity. We use the AUC as a criterion function in feature selection, since we assume no prior knowledge regarding thresholds.

Among the most common methods for evaluation of differences in boundary locations between different segmentations is the mean point to curve distance, which for each contour point measures the shortest distance to the contour of the truth model (in this case $A$):

$$d_M(A,B) = 1/|B| \sum_{b \in B} \{ \inf_{a \in A} d\{a,b\}\},$$

where $d\{\cdot,\cdot\}$ is a distance between two points. It can be noted that $d_M(A,B)$ is not necessarily equal to $d_M(B,A)$. Another is the Hausdorff distance, which measures the maximum distance of a set of boundary points to the nearest point in the other set:

$$d_H(A,B) = \max\{ \sup_{b \in B} \inf_{a \in A} d\{a,b\}, \sup_{a \in A} \inf_{b \in B} d\{a,b\}\}.$$

In cardiac segmentation tasks it can be important to keep track of boundary locations since clinical procedures such as catheter ablation involves locating heart walls. In our cardiac data we evaluate using mean point to curve distance, since it has been used in related work in the literature.

## 2.4 Segmentation Using Statistical Classification

One of the main objectives of this work is to design an automatic cartilage segmentation algorithm. The main challenges in developing an algorithm for

this purpose are that the cartilage compartments are thin, the low contrast between the cartilage and surrounding soft tissues, and the large variations in cartilage shape depending on the disease stage. During OA progression the cartilage is thinning, and in severe OA it may change topology.

In statistical voxel classification, each voxel is represented by $D$ features in a $D$-dimensional space. Supervised classification consists of two phases: training and testing. In the training phase, feature extraction or selection finds the appropriate features for representing the training data and the classifier is trained to partition the feature space. In the test phase, the trained classifier assigns a class probability or label based on the measured features. Using statistical classification for cartilage segmentation avoids the problem of modeling the global shape of the cartilage, and the local structure of iso-intensity surfaces in the image can capture cartilage characteristics and thus be useful as features for the classifier as described later in this section.

Examples of statistical classification systems are the *Nearest Neighbor* (1NN) method [25], which assigns voxels to the class of the nearest training pattern, and which can be extended to the *k Nearest Neighbors* ($k$NN) where voxels are assigned to the class with most neighbors among $k$ nearest training patterns. The *Parzen-window* method is closely related to $k$NN but instead of examining a fixed number of neighbors it examines the patterns inside a hypercube centered at the test voxel and with fixed side-length. Other examples are *support vector machines* (SVM) [8], that maximizes the margin between the classes in feature space by selecting a minimum number of support vectors, and *artificial neural networks* (ANN), which consists of a large number of simple processors with inter-connections and can change its structure based on information passing through the network during the training phase. For thorough overviews on statistical pattern classification the reader is referred to [37],[69].

The No-Free-Lunch Theorem [37] states that one cannot make claims that the generalization performance of one classifier is better than another, or for a particular application one cannot favor a certain system without prior knowledge. The $k$NN classifier has been found to perform well in MRI segmentation tasks, both in terms of stability for selecting different training data and in classification performance [21], which is a motivation why throughout this thesis we make use of a $k$NN classifier.

### 2.4.1 *k*NN Classification

For the $k$NN classifier, each pixel is described by a feature vector $\mathbf{u}$, and the probability of belonging to class $\omega_c$ is determined by calculating the distance (often Euclidean) to the $k$ nearest labeled neighbors in feature space. The $k$NN

classifier is considered non-parametric because it makes no underlying assumptions on class density functions, hence can be used with arbitrary distributions. The $k$NN algorithm estimates the *a posteriori* probabilities $P(\omega_c|\mathbf{u})$ from a set of $n$ labeled samples: suppose a cell of volume $V$ is placed around $\mathbf{u}$ and captures $k$ samples of which $k_c$ are labeled $\omega_c$ ($\omega = 1, \dots C$). Then, an estimate for the joint probability $p(\mathbf{u}, \omega_c)$ is

$$p_n(\mathbf{u}, \omega_c) = \frac{k_c/n}{V},$$

and thus an estimate the posterior probability is

$$P(\omega_c|\mathbf{u}) = \frac{p_n(\mathbf{u}, \omega_c)}{\sum_{i=1}^{N} p_n(\mathbf{u}, \omega_i)} = \frac{k_c}{k}.$$

### 2.4.2 Features

The performance of the classifier depends on the feature space, and the features should be representative for the classes. A common approach approach to examine the local structure of an image $I(\mathbf{x})$ is to consider a Taylor expansion around the point $\mathbf{x_0}$,

$$I(\mathbf{x}) \approx I(\mathbf{x_0}) + (\mathbf{x} - \mathbf{x_0})^T \nabla I(\mathbf{x_0}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_0})^T \nabla^2 I(\mathbf{x_0})(\mathbf{x} - \mathbf{x_0}),$$

where $\nabla I(\mathbf{x_0})$ and $\nabla^2 I(\mathbf{x_0})$ denote the gradient vector and the Hessian respectively. Hence the local structure of an image up to second order can be described by the intensity, the gradient, and the Hessian.

Both the image intensity and the intensity convolved with a Gaussian according to the scale-space framework [82] on different scales are considered as features in this work. The 3-jet, which consists of all first, second and third order derivatives forms a basis which describes all geometric features up to third order [46] and are considered as features. The spatial averaging in scale-space causes a noise reduction that counteracts the noise amplification caused by differentiation for Gaussian derivatives [7], defined as $I_{x_1,\dots,x_n} = I * \mathcal{D}_{x_1,\dots,x_n} G(\sigma)$, where $G$ is a Gaussian, $\mathcal{D}$ a differential operator and $\sigma$ is the scale.

The *structure tensor* ($\mathbf{T}$) [136],[81] and the *Hessian* ($\mathbf{H}$) [113] encode important first- and second-order intensity information. For a 3D image they are defined as:

$$\mathbf{H}(\sigma) = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix},$$

$$\mathbf{T}(\sigma_T, \sigma) = G_{\sigma_T} * \begin{pmatrix} I_x I_x & I_x I_y & I_x I_z \\ I_y I_x & I_y I_y & I_y I_z \\ I_z I_x & I_z I_y & I_z I_z \end{pmatrix}.$$

The scale $\sigma$ determines the size of the neighborhood in which the intensity iso-surface structure is examined, and $\sigma_T$ smooths the tensor without hampering image signal. Eigenvalue decomposition of a $3 \times 3$ symmetric matrix as the structure tensor or the Hessian gives three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ ($\lambda_1 \geq \lambda_2 \geq \lambda_3$) with corresponding eigenvectors $\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}$ respectively. The eigenvector $\mathbf{e_1}$ corresponding to the largest eigenvalue $\lambda_1$ represents the direction of maximum value of the first- respectively second-order derivative for the structure tensor respectively the Hessian, and $\lambda_1$ gives this maximum value. Similarly, $\lambda_3$ and $\mathbf{e_3}$ gives the minimum directional derivative value and its corresponding direction, and $\lambda_2$ and $\mathbf{e_2}$ gives the minimum value in a direction orthogonal to $\mathbf{e_3}$ and maximum value in a direction orthogonal to $\mathbf{e_1}$.

Since the structure tensor is based on first-order derivatives, it has a strong response at image boundaries at a given scale. Combinations of the eigenvalues can be used to determine the degree to which the local structure resembles a line ($\lambda_1 \simeq \lambda_2 \gg \lambda_3$), a plane ($\lambda_1 \gg \lambda_2 \simeq \lambda_3$) or a sphere ($\lambda_1 \simeq \lambda_2 \simeq \lambda_3$) [137]. The Hessian, which is based on second-order derivatives, have a strong response at the center of the iso-intensity structure over the scale $\sigma$. The eigenvalues of the Hessian can characterize the resemblance of the local structure to a tube ($\lambda_1 \simeq \lambda_2 \gg \lambda_3$), a sheet ($\lambda_1 \gg \lambda_2 \simeq \lambda_3$) and a blob ($\lambda_1 \simeq \lambda_2 \simeq \lambda_3$) [113]. This is summarized in Table 2.1. The eigenvalues of the Hessian have for instance been used to characterize tube-like objects for vessel segmentation [57], and sheet-like objects in sinus bone segmentation [34]. Prior

**Table 2.1:** Local structure resemblance using eigenvalues of the structure tensor and Hessian, assuming $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3|$.

| Eigenvalue condition | **T** | **H** |
|---|---|---|
| $\lambda_1 \simeq \lambda_2 \gg \lambda_3$ | Line | Tube |
| $\lambda_1 \gg \lambda_2 \simeq \lambda_3$ | Plane | Sheet |
| $\lambda_1 \simeq \lambda_2 \simeq \lambda_3$ | Sphere | Blob |

knowledge could be incorporated *e.g.* by assuming that cartilage is a sheet-like structure and incorporating measures of sheetness, however the feature sets described in this work are kept more generic by not making combinations of the eigenvalues but to use the eigenvalues with their relative magnitudes preserved as three-dimensional feature components. The eigenvectors are also used as three-dimensional feature components. According to scale-space theory [46], the scales should be selected using an exponential sampling, $\sigma_i = \sigma_0 \cdot e^{i\rho}$ where $\rho$ is a constant, however in this work the scales $\sigma$ and $\sigma_T$ were chosen empirically to cover a range of thicknesses of objects of interest in the data sets.

### 2.4.3 Feature Selection

The outcome of the classifier is dependent on the features the classifier uses to partition the feature space and thus the classifier training. There are several reasons to keep the dimensionality of the feature space as low as possible. One is related to the *curse of dimensionality*. If the class-conditional densities are known the probability of misclassification does not increase with the number of features, yet it is often observed that added features may degrade the classifier performance. The reason for this phenomenon is that high-dimensional functions have the potential to be much more complicated than low-dimensional ones, and that those complications are harder to discern. A rule of thumb is to use at least ten times as many training samples ($n$) per class as the number of features ($d$), $n/d > 10$, in order to avoid the curse of dimensionality, and for complex classifiers the demand for number of samples grows exponentially with the dimensionality of the feature space [69]. Other reasons for keeping the number of features as low as possible are removing redundant or even misleading features, and reducing computational time and memory usage. Methods for dimensionality reduction are *feature extraction* and *feature selection*. In feature extraction, a subspace of dimensionality $m \leq d$ is determined based on transformations and/or weighted combinations of the original feature set, which may lead to good discriminative abilities but the generated features may not have a clear physical meaning. In feature selection, the $m$ selected features retain their original values and interpretations, which is a motivation of the usage of feature selection in this work.

In feature selection, given a set of $d$ features the task is to find a feature subset $m \leq d$ that leads to the smallest classification error according to a criterion function. There are no non-exhaustive sequential feature selection method that guarantees an optimal feature subset, but exhaustive search is infeasible for everything but small feature sets. For systems with large amounts of data, number of features and a time demanding classifier, there are suboptimal selection methods that offer a tradeoff between the optimality of the selected subset and computational efficiency. One method is to evaluate each feature individually and select the $m$ features with best score according to the criterion function, however that does not consider combination of features. Two methods that evaluate feature dependencies are *sequential forward selection* (SFS) and *sequential backward selection* (SBS), where one feature at the time is added respectively deleted from the feature set. A drawback is that once a feature has been added (deleted) it cannot be discarded (brought back). By applying first SFS followed by SBS this problem is partially avoided, and this is the method used throughout this thesis. Additionally, in each iteration of the SFS we allow all $d$ features to be selected, not only those that have not already been selected, which allows for an indirect importance weighting of the features by the number of times

they were selected. SFS and SBS have been extended to sequential floating forward (SFFS) and backward (SFBS) search methods, where SFFS starts with an empty feature set and for each step, one step of SFS is performed. The algorithm also verifies the possibility of improvement of the criterion if some feature is excluded by performing one step of SBS. The SFFS proceeds by dynamically increasing and decreasing the number of features until the desired $m$ is reached.

The choice of criterion function is important in feature selection. For two-class classifiers we use the AUC since it evaluates the classifier performance for all threshold simultaneously. However there is no direct extension of the AUC for classification tasks with more than two classes. For three-class classification where all thresholds of the classifier are evaluated, we use Mossman's three-way ROC [94], however it does not evaluate all misclassification possibilities.

### 2.4.4 Multi-Class Classification

The $k$NN system is inherently a multi-class classifier, and the segmentation of different cartilage compartments in the knee is a multi-class (number of classes $C > 2$) classification task. In this thesis we focus on the medial cartilage compartments, which is a three-class task; medial tibial ($\omega_t$) and femoral cartilage ($\omega_f$), and background ($\omega_b$), however, there are several other cartilage compartments in the knee. A multi-class classification system can be decomposed into a combination of two-class classifiers, and a motivation for examining this option is to examine if the different classes exhibit different features. *Class-dependent features* have been defined by Oh *et al.* [98] and Bailey [6] as features whose discrimination ability varies significantly depending on the classes which are to be discriminated. In this context, each class is considered to have its own optimal feature vector which suggests that features should be selected locally for each class instead of globally for all classes. This approach may reduce features that are redundant for a specific class which otherwise may have impaired the classification performance.

Three categories for conversion between two-class and $C$-class ($C > 2$) frameworks are: *one vs. all* where $C$ two-class systems compete against each other, *hierarchical* where the system is separated into $C-1$ recursive two-class systems, and *pairwise* where the classifications are separated into $C(C-1)/2$ between all pairs of classes. For the medial cartilage classification, there are two cartilage classes ($\omega_t$ and $\omega_f$) and the background class ($\omega_b$). This makes an hierarchical approach interesting, where the voxels are first divided into cartilage or background, then into $\omega_t$ or $\omega_f$. Though two cartilage compartments are studied in this work, the framework will be extended to incorporate the other compartments in the knee, therefore one can consider extensions to more classes. If one more class is incorporated, the pairwise framework needs to train three new clas-

sifiers. The hierarchical framework needs to train a new classifier and to retrain the foreground *vs.* background classifier, whereas the one *vs.* all framework only needs to train one new classifier and the rest of the classifiers remain unaltered. Hence the one *vs.* all approach may be advantageous in case of class-dependent features and if the framework needs to be extended to more classes.

### 2.4.5 Statistical Classification for Cartilage Segmentation

The remaining chapters in Part I are a selection of research papers on the subject of automatic segmentation of MRI data, modified to fit into this thesis. This section provides an overview of the statistical classification methods in the different chapters and a discussion on how they relate to each other.

In Chapter 3, which corresponds to the publication with reference number [48], we describe a hierarchical approach to segmentation using the classification framework. First, a two-class classifier roughly separates cartilage from background, then a three-class classifier separates the voxels classified as cartilage in the first instance into tibial or femoral cartilage or background. Hence the hierarchical approach does not reduce the three-class system into a combination of two-class systems, but is rather a way of reducing computational time by using a low-dimensional feature in the first step. The first two-class system roughly separates the image into foreground and background using relatively few features reduces the number of voxels investigated in the second classification round to a fraction of the entire image, and in the second round approximately three times as many features are used. Using this methodology, the sensitivity, specificity and DSC are 90.0%, 99.8%, and 0.80 respectively for the medial cartilage, trained on 25 knee scans and evaluated on 46 scans. The data set contains two scans with clear signs of OA.

The concept of class-dependent features suggests that it might be interesting to search for features that are specific for certain classes, rather than to search for features that represent all classes simultaneously. In Chapter 4 (reference [50]) we investigate combinations two-class classifiers for separating the tibial and femoral cartilage from the background. Here we use a one vs. all approach, with only foreground classifiers ($\omega_t$ *vs.* rest and $\omega_f$ *vs.* rest classifiers), replacing the $\omega_b$ classifier with a threshold and therefore only two classifiers are necessary ($C - 1$ two-class systems). This combination of two-class classifiers produces better results than a $k$NN classifier that is trained to partition the voxels directly into three classes. Though one cannot state that the combination of two-class classifiers are better than a direct three-class classification since the three-class classifier has less than twice as many features as a two-class classifier, and the two-class classifiers have many features in common (few class-dependent features), the combination of two-class classification system can more easily be

extended to incorporate more classes, and the two-class classifiers use fewer features each meaning they are less sensitive to curse of dimensionality and memory constraints. The resulting sensitivity, specificity and DSC of the algorithm described in this Chapter are 84%, 99.9% and 0.81 respectively, and the test set contains 86 scans. Comparisons of sensitivity and specificity with the ones reported in Chapter 3 is not straight-forward, because it seems like there is a great decrease in sensitivity with only a marginal increase in specificity which depends on the threshold selected for evaluation. With the threshold selected to optimize the DSC, the DSC offers a more direct comparison between the two methods. Even if the improvement seems only marginal (0.81 in this work compared to 0.80 in the previous) one should keep in mind that the data set is increased to 25+86 instead of 25+46 scans, and it contains 11% knees with clear signs of OA compared to 3% in the previous chapter, hence the segmentation is improved despite increased anatomical and pathological variation.

In Chapter 5 (reference [52]) the aim is to make the segmentation described in Chapter 4 less sensitive the location of the cartilage in the scan, since it tends to vary in clinical studies and the position is determined to be one of the most important features by feature selection in the first place. Localization normalization is achieved by iteratively correcting for position variations using the two following steps: shifting the cartilage towards the expected position from training data, and performing classification with the normalized position as a feature. By applying the position adjustment scheme to our cartilage segmentation algorithm, there is a small improvement in DSC, but the main outcome is that the inter-scan reproducibility of the cartilage volume is much improved and becomes comparable to that of a highly trained radiologist. The test set in this study consists of 114 scans, of which 55% have at least some radiographic signs of OA (31% have clear OA, see Chapter 9 for a definition of the degrees of OA). Also, 31 knees were re-scanned approximately one week after baseline for a reproducibility study. After position adjustment the average medial cartilage paired absolute volume difference decreases from 9.3% to 6.2%. These reproducibility values can be compared to the volumes from the manual segmentations by a radiologist for the same data set, where the radiologist has a paired absolute volume difference of 6.8%.

Chapter 6 (reference [55]) describes the segmentation method based on Chapter 4 with the addition of the position adjustment scheme in Chapter 5. Additionally, a scheme is included that allows for faster classification by sampling a number of voxels randomly and classifying them, and if classified as cartilage extending the classification to neighboring voxels until no more cartilage voxels are found [29]. The data set are the same 114 scans as in Chapter 5, and the resulting sensitivity, specificity, and DSC are 84%, 99.9%, and 0.80. Besides evaluating the method by comparisons to gold standard segmentations, the cartilage volume and surface area are compared to radiographic signs of OA and

significant differences between healthy and OA populations are demonstrated.

Currently the threshold for labeling voxels as foreground from the class probability maps is quite high, and an evaluation of the classifier performance using a denser sampling of background training points, particularly in the vicinity of foreground voxels, would be of interest in order to lower the threshold. The current cartilage segmentation is evaluated on the medial tibial and femoral cartilage compartments, and in the future there are plans to incorporate the other cartilage compartments in the knee as well. Using two-class classifiers trained to distinguish one foreground class each, the incorporation of more classes is straight-forward and the whole classification system does not need to be retrained. Using the strategy of fast classification, only voxels with high likelihood of being foreground are classified for each two-class system and not all the voxels in an image. The scale perspective could be more thoroughly examined, more scales could have been evaluated with exponential increase between scales, and more prior information could have been incorporated by examining features that describes resemblance to a plane or a sheet that is brighter than surrounding tissues. Combining values of the eigenvalues of the structure tensor and Hessian to 'plane-like' and 'sheet-like' measures may reduce the dimensionality of the feature space. It would also be interesting to improve the gold standard segmentation for the training data and evaluate different feature selection and feature normalization techniques.

## 2.5 Statistical Classification and Level Set Methods

Statistical classification has been shown to perform well for cartilage segmentation, but it is possible that the classifier performance can improve by incorporating global information. Level set methods are useful tools in image segmentation because they can naturally deal with topology changes, and can incorporate region, boundary, and shape information. Here, we investigate ways of combining statistical classification and level set methods by incorporating class probability maps in a regional module thus adding global information to the classification system.

### 2.5.1 Level Set Methods in Image Segmentation

The level set method by Osher and Sethian [99] is a numerical technique for tracking implicit curves and surfaces. A surface $\Gamma$ of codimension one is implicitly represented as the zero level set of a function so that $\Gamma = \{\mathbf{x}|\Phi(\mathbf{x}) = 0\}$.

If

$$\Phi = \begin{cases} -1 & \text{inside } \Gamma \\ 0, & \text{on } \Gamma \\ 1, & \text{outside } \Gamma \end{cases}$$

then the outward unit normal $\mathbf{n}$ and the mean curvature $\kappa$ are given by $\mathbf{n} = \frac{\nabla\Phi}{|\nabla\Phi|}$ and $\kappa = \nabla \cdot \frac{\nabla\Phi}{|\nabla\Phi|}$. The function $\Phi$ defines the surface $\Gamma$, but for comparison between different surfaces there has to be a one-to-one correspondence between $\Gamma$ and $\Phi$. This can be achieved by a *signed distance function*, which imposes the extra constraint that $|\nabla\Phi| = 1$ everywhere, except for the skeleton of the region bounded by $\Gamma$. The use of the signed distance function also increases numerical stability, and the shortest distance from any point in the image to $\Gamma$ can easily be determined, however during surface evolution the function needs to be reinitialized in order to keep the signed distance function property.

A surface evolving with velocity $v$ over time $t$ is described by $\Gamma_t = v\mathbf{n}$ can be shown to have, in the level set formulation, the motion equation

$$\frac{\partial\Phi}{\partial t} = v|\nabla\Phi|.$$

Methods for implicit surface motions are typically based on numerical solutions of certain partial differential equations (PDE's). Motion equations for $\Gamma$ can be found by minimization of an energy functional $E$ by variational calculus using a gradient descent procedure, and it can be shown that the evolution is $\frac{\partial\Phi}{\partial t} = \nabla_M E(\Gamma)|\nabla\Phi|$ where $M$ is the Riemannian manifold in which the surface is defined [118]. Motion equations can be internally generated, as in the case of mean curvature motion: $\frac{\partial\Phi}{\partial t} = \kappa|\nabla\Phi|$, or externally generated *e.g.* by some functional aiming to attract to image boundaries. Mean curvature motion will be discussed further in Part II, here we briefly describe level set segmentation techniques which typically include both external and internal motion modules.

The snake or active contour formulation [72] minimizes an energy functional for a parametric curve $\gamma(s) : [0, 1] \to \mathbb{R}^2$:

$$E(\Gamma) = \int_0^1 \alpha\ |\Gamma'(s)|^2 + \beta\ |\Gamma''(s)|^2 + g(\Gamma(s))dp,$$

where $g$ is a function indicating edges in the image $I : \mathbb{R}^2 \to \mathbb{R}$. The first two terms are the internal energy that control the smoothness of the curve, while the third term attracts the contour to image boundaries. In order to find a smooth curve that attracts to image boundaries in the geometric case, Caselles *et al.* [15] proposed the functional:

$$E(\Gamma) = \int_\Gamma g(\Gamma(s))ds,$$

where $s$ is the arclength. This approach is called *geodesic active contours* (GAC) since it involves geodesic computations in a Riemannian space according to a metric induced by $I$, and the solution involves finding the path of minimal length in that metric. The level set formulation of the geodesic active contour model is

$$\frac{\partial \Phi}{\partial t} = \mathrm{div}\Big(g(|\nabla I|)\frac{\nabla \Phi}{|\nabla \Phi|}\Big)|\nabla \Phi|.$$

The geodesic active contour model depend on well-defined edges and need to be initialized close to the solution in order to avoid getting trapped in local minima. The *Chan-Vese model* [16] aims to overcome these obstacles by incorporating a region-based functional, the *Mumford-Shah functional* [96], in a level set framework. Let $H$ be the Heaviside function

$$H(z) = \begin{cases} 1 & z \leq 0 \\ 0, & z > 0 \end{cases}$$

and $\delta(z) = \frac{d}{dz}H(z)$. The Chan-Vese functional assumes that the image is formed by two regions of approximately piecewise constant intensities of distinct values, and is for the curve $\Gamma$ in the domain $\Omega$ defined by

$$E(c_1, c_2, \Phi) = \mu \int_\Omega \delta(\Phi)|\nabla \Phi| d\mathbf{x} + v \int_\Omega H(\Phi) d\mathbf{x} +$$
$$\lambda_1 \int_\Omega |I(\mathbf{x}) - c_1|^2 H(\Phi) d\mathbf{x} + \lambda_2 \int_\Omega |I(\mathbf{x}) - c_2|^2 (1 - H(\Phi)) d\mathbf{x}. \quad (2.1)$$

Here, the constants $c_1$ and $c_2$ are the averages of $I$ inside and outside $\Gamma$ respectively. The first and second term minimizes the length and area of $\Gamma$ respectively, whereas the two last terms partitions the image into homogeneous regions. Paragios and Deriche extended the GAC method to a geodesic active region (GAR) formalism [102] which enables a partitioning of the image using a boundary-based functional as for the GAC as well as a region-based one. Let $p_1(I(\mathbf{x}))$ and $p_0(I(\mathbf{x}))$ denote the probabilities for a pixel being inside and outside a structure of interest respectively, and by assuming independence between pixels and taking the -log of the *a posteriori* segmentation probability, the region energy functional is

$$E(\Gamma) = -\int_\Omega H(\Phi) \log\big(p_1(I(\mathbf{x}))\big) d\mathbf{x} - \int_\Omega (1 - H(\Phi)) \log\big(p_0(I(\mathbf{x}))\big) d\mathbf{x}.$$

### 2.5.2 Statistical Classification in Level Set Segmentation

Since the geodesic active regions framework employ a probability based approach for incorporation of regional information, it is straight forward to incorporate a statistical classifier in it. Numerically the methods fit well together since all

calculations are done at pixel level, without the need to switch between different representations. A segmentation algorithm based on GAR and statistical classification is evaluated on late gadolinium enhanced cardiac MRI (CE-MRI) [84], which is a challenging task due to ambiguous boundary information and local intensity variations.

In Chapter 7 (reference [56]), the classification probability maps are used as regional information in the geodesic region framework of Paragios [102]. The myocardium in cardiac CE-MRI is segmented using two coupled contours, and the classifier is incorporated both directly in the region term, and indirectly in a shape term where shape inference is based on the class probability maps. The shape model uses a distance function shape representation, and is described by a mean shape, which is a distance function, and local shape variability [112]. Starting from a rigid transform of the mean shape, two coupled contours deform given local region and boundary information. The method is evaluated on 30 slices of late gadolinium enhanced cardiac MRI, it succeeds in segmenting the myocardium in 28 of those images. The mean DSC is 0.79 and the average point distance to contour is 1.44 pixels.

In Chapter 8 (reference [47]), the method in Chapter 7 is extended to a spatially varying classification (SVC) framework. SVC iterates between global shape matching and statistical classification, a combination that resolves ambiguities in feature space with spatial context. The method in Chapter 5 can be seen a more primitive example of the same principle. In the classification step, additional features encoding spatial context are computed from the inferred shape. Then, shape is inferred based on the classification outcome. The algorithm successfully segments the same 28 images with an average DSC and distance to contour of 0.80 and 1.37 pixels respectively. This method is iterative instead of sequential as that in the previous chapter, which means the region term is iteratively updated. The results show some improvement though only significant at a 90% confidence level ($p = 0.06$ for DSC according to a paired t-test). By visual inspection (the right column in Figure 8.2 correspond to Figure 7.3) it appears that the contours correspond better to manual delineations. However, the data set is rather limited and further validation is necessary to determine how this method can be useful in relation to clinical studies of scar tissue.

Level set methods with their topological adaptivity may be of interest in cartilage segmentation, a structure which may not have a fixed topology. It is possible that the voxel classification algorithm may improve by incorporating global information, and measures from the automatic segmentation could become more reproducible. The shape model module in the geodesic active regions framework with its rigid transformation of a mean shape might not be sufficient for cartilage not due to large shape variations in the training data, however it could provide spatial context to the statistical classifier which could resolve am-

biguities in feature space using the spatially varying classification framework. An alternative shape model or a shape template could also be considered.

Chapter 3

# Automatic Segmentation of the Articular Cartilage in Knee MRI Using a Hierarchical Multi-class Classification Scheme

*Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen*

**Abstract**

Osteoarthritis is characterized by the degeneration of the articular cartilage in joints. We have developed a fully automatic method for segmenting the articular cartilage in knee magnetic resonance (MR) scans based on supervised learning. A two-class $k$NN classifier first roughly separates cartilage from background voxels, then a three-class classifier assigns one of three classes to each voxel that is classified as cartilage by the two-class classifier. The resulting sensitivity and specificity are 90.0% and 99.8% respectively for the medial cartilage compartments, which shows that automatic cartilage segmentation is feasible in low-field MR data.

## 3.1 Introduction

Osteoarthritis (OA) is one of the major health concerns among the elderly today [140]. The main effects of OA is the degradation of the articular cartilage together with remodeling and overgrowth of bone, a process causing loss of mobility of the joints. OA can occur in any joint but typically affects large weight bearing joints such as hips and knees. Currently there are no disease-modifying drugs and treatment of OA is restricted to symptom control [27]. Therefore much effort is spent on research concerning disease modifying drugs and analysis of articular cartilage and its relation to disease progression.

MRI allows for non-invasive quantitative evaluation of the cartilage [95],[60], and cartilage deterioration can be detected using this technique [87]. The MR scans can be manually segmented slice-by-slice by experts for cartilage analysis, but for clinical studies manual methods are too time consuming and are also prone to inter- and intra-observer variability. When automating the cartilage segmentation, the main challenges are the thin structure of the cartilage and the low contrast between the cartilage and surrounding soft tissues.

Several groups have developed automated methods for cartilage segmentation. 2D methods has limited continuation between slices and since they have to be converted into a 3D segmentation when finding for example thickness maps, it is advantageous to perform segmentation in 3D directly. Among the 3D techniques that have been developed, Grau *et al.* [61] use a semi-automatic segmentation method that is based on a watershed approach. The method is evaluated on 7 scans from 4 subjects and has an average sensitivity and specificity of 90.0% and 99.9% respectively. Pakin *et al.* [101] have developed an automatic segmentation method based on region growing followed by two-class clustering. It is evaluated on one scan with resulting sensitivity and specificity of 66.2% and 99.6%. The semi-automatic segmentation method of Warfield *et al.* [135], [134] iterates between a classification step and a template registration step, and has a lower variability compared to repeated manual segmentations on the scan it was evaluated on. Naish *et al.* [97] use a data set that consists of a longitudinal study of OA patients and local intensity changes over time is used as a measure of cartilage degradation. However, the cartilage is manually or semi-automatically segmented.

All of the methods mentioned (except for the one of Naish *et al.* but they have not focused on the segmentation part) have only been evaluated on a handful of scans, and the only fully automatic segmentation yields low sensitivity and

specificity values compared to the semi-automatic methods.

In this paper, we present a method for segmenting the tibial and femoral medial cartilage in 3D MR scans of knees. The segmentation is based on a three-class approximate $k$NN classification scheme and is improved by selecting the largest connected component from the result of the classification. The segmentation method works directly in 3D and is fully automatic. This is an improvement of previous work [49] where tibial medial cartilage is roughly located with the purpose of providing an initialization of a shape model, a method based on a two-class classification without any feature selection incorporated.

Our segmentation algorithm is an important step towards an automatic cartilage evaluation framework and is intended for clinical studies on a low-field MR scanner. Though the image quality of the scanner we are using is lower compared to the conventional high-field scanners, we propose to examine if accurate automatic cartilage segmentation is achievable also on a low-field scanner. If such a scanner can be used in clinical studies it would reduce the costs significantly. It has been shown that low-field dedicated extremity MRI can provide similar information on bone erosions and synovitis as expensive high-field MRI units [41] comparing manual segmentations, but there has to our knowledge not been published any work on automatic segmentation of cartilage on low-field MRI. From the automatic segmentation, relevant quantitative measures such as the cartilage volume and thickness can be calculated either globally or locally at a point or a small area. In the latter case comparison between patients or temporal studies of the same patient will require establishing geometric or anatomical correspondence either by expert annotations or by automated shape modeling.

## 3.2 Methods

### 3.2.1 Image Acquisition

An Esaote C-Span low-field 0.18 T scanner dedicated to imaging of extremities acquires Turbo 3D T1 scans (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms). The scans are made through the sagittal plane with a voxel size in the range $0.7031 \times 0.7031 \times (0.7031/0.7813/0.8594)\text{mm}^3$. Among the total of 71 scans, 50 have the resolution $0.7031 \times 0.7031 \times 0.7813\text{mm}^3$, 3 the resolution $0.7031 \times 0.7031 \times 0.7031\text{mm}^3$ and the remaining 18 scans have the resolution $0.7031 \times 0.7031 \times 0.8594\text{mm}^3$. The original scan size is $256 \times 256 \times 104$ voxels, but we only use the central $170 \times 170 \times 104$ voxels because only they contain information. The scans have been manually segmented on a slice-by-slice basis by a radiologist.

A scan slice with the tibial and femoral medial cartilage manually outlined is shown in Figure 4.1.

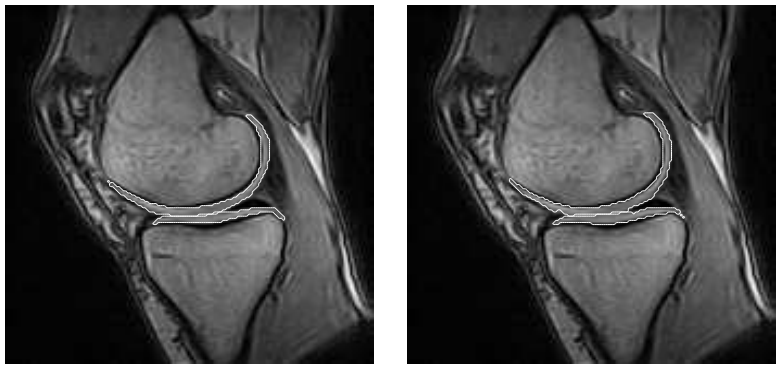


**Figure 3.1:** To the left, a slice from a knee MR scan where the tibial medial and femoral medial cartilage is segmented manually by radiologists. The size of this slice is $170 \times 170$ pixels. To the right is the result from our automatic segmentation for the corresponding slice. The sensitivity and specificity for this scan are 92.52% and 99.82% respectively, with a DSC of 0.83.

The 71 scans in the data set are of both left and right knees. In order to treat all scans analogously, the right knees are reflected about the center in the sagittal plane. The test subjects are both males and females aged between 21 and 72 years and have no or mild OA symptoms.

### 3.2.2 Cartilage Classification

For the segmentation of cartilage we use an approximate $k$NN classifier developed by Arya *et al.* [4]. The approximate $k$NN classifier is in principle a $k$NN classifier, but with the modification that by allowing a small amount of error in the search for nearest neighbors the run time may improve significantly. Instead of returning the $k$ nearest neighbors from a data set, the approximate search algorithm returns $k$ points such that the ratio of the distance between the $i$th reported point ($1 \leq\ i \leq k$) and the true $i$th nearest neighbor is at most $1 + \epsilon$, where $\epsilon$ is the error bound. The approximate $k$NN algorithm preprocesses a set of $n$ data points in $R^d$ in $O(dn\log\ n)$ time and $O(dn)$ space, so that given a query point $q$ and $\epsilon > 0$, the $k$ $(1 + \epsilon)$-approximate nearest neighbors of $q$ can be computed in $O((c_{d,\epsilon} + kd)\log\ n)$ time, where $c_{d,\epsilon} = d[1 + 6d/\epsilon]^d$. Hence, the algorithm is efficient for large values of $n$, but the exponential factor implies that it may not practical for large values of $d$. Therefore, exploiting the benefits of this algorithm would involve using a large training data set and keeping the

dimensionality of the feature space fairly low. Arya *et al.* show empirically that the factor is much smaller than the bound given by $c_{d,\epsilon}$, and that the average distance error is at least an order of magnitude smaller than $1+\epsilon$. The algorithm supports any Minkowski metric, we use Euclidean distance in the $k$NN search. We have found empirically that examining the 100 nearest neighbors yields a good balance between computational complexity and accuracy, and we set $\epsilon = 2$, a value that only marginally lowers the accuracy while reducing computational time significantly.

In this work we examine the medial cartilage since OA is more often observed in this compartment [38] and in particular in the medial tibial part [71], thus these compartments are of major interest when it comes to finding disease markers for OA. The three classes considered are tibial medial cartilage, femoral medial cartilage and background.

The classification is hierarchical, and the first step is a two-class classification where the voxels are roughly separated into cartilage or background. The $k$NN algorithm produces class probabilities for every voxel, and in this step we set the threshold at 0.65 yielding a sensitivity for medial cartilage close to 99% but which also results in a large amount of false positives. Since typically only a few percent of the total volume within the knee belongs to the cartilage, this first step is a way of reducing the voxels to be classified significantly. In the second step, the voxels classified as cartilage in the first step are reconsidered. This time we use a three-class classifier, where the three classes are tibial and femoral medial cartilage and background, and class membership is decided based on a majority vote combined with a threshold. The three-class classifier contains more features and the features are selected for direct three-class separation whereas the classifier in the first step has features selected to separate cartilage from background. It could also be interesting to train a two-class classifier combined with a threshold for the background class in this second step. A sketch of the hierarchical classification scheme is illustrated in Figure 3.2.

The hierarchical approach is faster compared to using the three-class classifier directly, since the first step has fewer features and the number of voxels classified in the second step is only a fraction of the total number in the image. The two-class classifier has a set of 28 features and the three-class classifier that has 52 features.

### 3.2.3 Features and Feature Selection

In order to find a feature set that performs well for our classification scheme, we here introduce our set of candidate features. Later we describe the subsets of the features that were found from our feature selection, which consists of

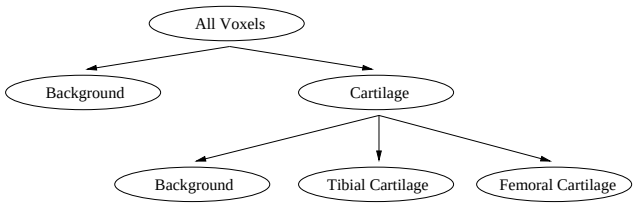



**Figure 3.2:** Tree representation of the hierarchical classification scheme.

sequential forward selection followed by sequential backward selection. Feature selection was performed on 25 scans and within this data set 18 scan were used for training and 7 for evaluation.

When a radiologist examines an MR scan for cartilage, she or he takes the location and the intensity in the image into consideration. We therefore consider these as candidate features. Both the raw intensity and the Gaussian smoothed intensities on three different scales $(0.65\text{mm}, 1.1\text{mm}, 2.5\text{mm})$ are examined.

One can also consider features that are related to the geometry of the object in question. The 3-jet, which is all first, second and third order derivatives with respect to $(x, y, z)$ forms a basis which can describe all geometric features up to third order [46] and are listed as candidate features. All the derivatives mentioned are Gaussian derivatives. All features are examined on the three scales, selected to cover the range of different cartilage thicknesses, mentioned above.

The local first- and second-order structure can be described of eigenvectors and corresponding eigenvalues of the structure tensor ($\mathbf{T}$), and the Hessian($\mathbf{H}$), which are considered as potential features using three different scales for $\sigma$ and three scales of $\sigma_T$.

Third-order structure is also examined by the third-order tensor, $I_{ijk}$. Examining the third-order structure in the local gradient direction $(I_x, I_y, I_z)$ can be described using Einstein summation as

$$L_{www} = I_{ijk} I_i I_j I_k / (I_i I_i)^{3/2}.$$

The third order tensor is examined in the gradient direction on three different scales, which were considered as potential features.

### 3.2.4 Selected Features

The features used in the two-class classifier are the position in the image, the Gaussian smoothed intensities on three different scales (0.65mm, 1.1mm, 2.5mm) and the raw intensities, the first order Gaussian derivatives on scales 0.65mm and 2.5mm, the eigenvalues and the eigenvector corresponding to the largest eigenvalue of the structure tensor with $\sigma = 0.65$mm and $\sigma_T = 2.5$mm, and the eigenvalues of the Hessian on scales 1.1mm and 2.5mm.

The features in the three-class classifier consist of combinations of first, second and third order Gaussian derivatives on the three different scales mentioned, the Gaussian smoothed intensities on three different scales (0.65mm, 1.1mm, 2.5mm) and the raw intensities, the position, the eigenvector corresponding to the largest eigenvalue of the structure tensor with $\sigma_T = 0.65$mm and $\sigma = 1.1$mm, the eigenvalues of the structure tensor with $\sigma_T = 1.1$mm and $\sigma = 2.5$mm, the eigenvalues of the Hessian on scales 1.1mm and 2.5mm. The features selected as most significant are the intensity and the position in the image, followed by eigenvalues of the Hessian and the structure tensor.

All features except the intensity are coupled three by three so that each feature addition results in an equal increase in dimensionality of the feature space. Feature components consisting of three features each comes natural since the coordinates, first order derivatives and the eigenvalues and eigenvectors of the Hessian and the structure tensor are directly grouped. The other features are grouped using the three scales.

The Euclidean distance measure assign more weighting to features with large ranges than those with small ranges, therefore features values should be normalized prior to classification. A simple approach is to perform linear scaling to unit range of the training feature set: given a lower bound $lb$ and an upper bound $ub$ for each feature component $u$,

$$\tilde{u} = \frac{u - lb}{ub - lb},$$

with the resulting $\tilde{u}$ being in the $[0, 1]$ range. Features from test data are then normalized using the same $lb$ and $ub$. Another approach is linear scaling to zero mean and unit variance:

$$\tilde{u} = \frac{u - \mu}{\sigma},$$

where $\mu$ and $\sigma$ are the sample mean and variance. After examining both normalization methods with very similar results, we decided to use linear scaling to unit range. The use of a large training data set reduces the sensitivity to outliers. Spreading the feature values as much as possible in the $[0, 1]$ range may increase the discrimination abilities of the Euclidean metric compared to linear scaling

of the feature components. Taking the distribution of the features into consideration in the feature normalization may improve classification performance [3] which makes it interesting as a potential future research direction.

## 3.3 Results

From our data set of 71 scans we use 25 for training (the same 25 scans that were used in the feature selection) and 46 for the evaluation of our algorithm. For the training data, all cartilage voxels and 0.5% (randomly sampled) background voxels according to the gold standard segmentations were used as prototypes for the classifier.

The results of our automatic segmentation is compared to the manual segmentation made by radiologists, resulting in an average sensitivity and specificity of 90.0% (±2.6% std) and 99.8% (±0.06% std) respectively for the test set for the medial cartilage compartments. The Dice similarity coefficient between our automatic segmentation and the gold standard segmentation is for the test set on average 0.80 (±0.03 std). A comparison between a gold standard segmentation and an automatic 3D segmentation can be seen in Figure 3.3. A slice-by-slice comparison is displayed in Figure 4.1. A typical source of errors for the classifier is where the tibial cartilage curves over the edge of the tibial plateau, which is most likely due to features such as eigenvectors of Hessian and structure tensor which will be different in this small part of the cartilage compared to the rest of the tibial compartment which is relatively flat.

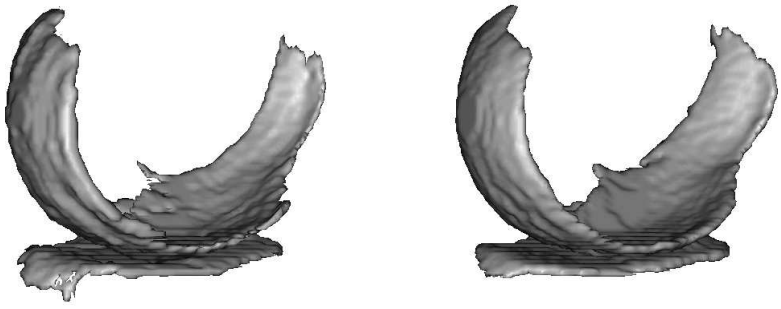


**Figure 3.3:** On the left is the manually segmented medial cartilage from a knee MR scan. To the right is the corresponding automatic segmentation. For this scan, the sensitivity and specificity are 94.82% and 99.79% respectively, with a DSC of 0.81.

## 3.4 Discussion

The average sensitivity and specificity of our method compared to the results of methods with similar evaluation, though on different data, is presented in Table 12.3. Compared to the fully automatic segmentation algorithm (Pakin *et al.* [101]), we get a distinctly higher sensitivity as well as a better specificity. The sensitivity and specificity of Grau *et al.* [61] are higher, and they have a better volume overlap (DSC=0.90), however their method requires manual interaction. Also, we have evaluated our segmentation algorithm on more scans compared to the other two. Some of the semi-automated segmentation techniques described

**Table 3.1:** Evaluation of the automatic segmentation in comparison with state-of-the-art cartilage segmentation algorithms in the literature.

| | Our method | Pakin *et al.* [101] | Grau *et al.* [61] |
|---|---|---|---|
| Sensitivity | 90.01% | 66.22% | 90.03% |
| Specificity | 99.80% | 99.56% | 99.87% |
| Data set | 25+46 | 1 | 7 |
| Interaction time | 0 | 0 | 5-10 min |

in Section 10.1 have been evaluated in terms of inter- and intra-observer variability of the method compared to manual segmentation. The progression of OA is very often slow and it can take many years before the cartilage is reduced from its typical thickness of a few millimeters to possible total loss. It is therefore important to have high accuracy and precision of the cartilage evaluation technique in order to detect statistically significant changes longitudinally. This enables the correlation of the method with the effects of drugs, and the evaluation of their benefit to the joint in reducing signs or symptoms of the disease. Future work will involve an evaluation of inter-scan variability of the segmentation algorithm. We will also examine intra-user variability for the manual segmentations.

Our segmentation algorithm performs well compared to two leading cartilage segmentation schemes, which leads us to conclude that a fully automatic cartilage segmentation in low-field MR data is feasible.

Chapter 4

# Combining Binary Classifiers for Automatic Cartilage Segmentation in Knee MRI

*Jenny Folkesson, Ole F. Olsen, Paola C. Pettersen, Erik B. Dam and Claus Christiansen*

**Abstract**

We present a method for segmenting tibial and femoral medial cartilage in MR knee scans by combining two $k$ Nearest Neighbors ($k$NN classifiers for the cartilage classes with a rejection threshold for the background class. We show that with this threshold, two two-class classifiers are sufficient compared to three classifiers in the traditional one-versus-all approach. The combination of two-class classifiers produces better results than a $k$NN classifier that is trained to partition the voxels directly into three classes. The resulting sensitivity, specificity and Dice similarity coefficient of our method are 84.2%, 99.9% and 0.81 respectively. Compared to state-of-the-art segmentation methods, our method outperforms the only, to our knowledge, fully automatic method.

## 4.1 Introduction

One of the most common health defects among elderly today is osteoarthritis (OA), a disease which most often affect weight bearing joints such as knees and hips and is characterized by the degeneration of the articular cartilage. As of today, the treatment is restricted to symptom control because there are no disease-modifying drugs [43]. Therefore, much effort is put into OA drug development and consequently for finding quantitative measures of disease progression.

Magnetic resonance imaging (MRI) allows for quantitative evaluation of the articular cartilage and OA pathology [60],[104]. Recently it has been shown that low-field dedicated extremity MRI can provide similar information on bone erosions and synovitis as expensive high-field MRI units [41]. If low-field scanners can replace high-field scanners in clinical studies, the costs of making such studies would be reduced significantly.

The cartilage in joints affected by OA degenerates by losing thickness and integrity, and typical relevant quantitative measures of the cartilage status are volume and thickness maps. When finding such measures the segmentation of the cartilage is a crucial step. Many segmentation methods rely heavily on expert user interaction, but having experts to perform manual slice-by-slice delineation of the cartilage is too time consuming for routine clinical use and is inclined to inter- and intra-user variability. In order to find a cost-effective and precise cartilage segmentation, methods that are partly or fully automated are being developed. The main challenges in cartilage segmentation are the thin structure of the cartilage and the low contrast, both between the cartilage and surrounding soft tissues and between different cartilage compartments.

### 4.1.1 Related Work

Several groups have developed automated methods for cartilage segmentation. Lynch *et al.* [88] combine user interaction with active contours, and Solloway *et al.* [119] use active shape models for slice-by-slice cartilage segmentation. 2D methods has limited continuation between slices and since they have to be converted into a 3D segmentation when finding for example thickness maps, it is advantageous to perform segmentation in 3D directly.

Among the 3D techniques that have been developed, Grau *et al.* [61] have a

segmentation method that is based on a watershed approach. Pakin *et al.* [101] have developed a region growing scheme that is followed by a two-class clustering for automatic segmentation. Warfield *et al.* [135] presents a semi-automatic segmentation method that iterates between a classification step and a template registration step, which is shown to a lower variability compared to repeated manual segmentations.

### 4.1.2 Overview of Our Work

The 3D segmentation techniques described in section 4.1.1 all require some amount of manual interaction except for the method of Pakin et al. [101], and the methods are evaluated only on small data sets. Neither Grau *et al.* nor Pakin *et al.* evaluate their methods on scans from OA patients. In this paper, we present a segmentation method that combines a tibial cartilage vs. rest and a femoral cartilage vs. rest approximate $k$NN classifier with a rejection threshold for the background class for the segmentation of tibial and femoral medial cartilage. Our method is evaluated on a large data set containing both healthy and osteoarthritic knees. We show that by introducing the rejection threshold, two binary classifiers are sufficient for our three-class classification task. OA is more often observed in the medial compartments [38], therefore we focus on medial cartilage in this work.

Besides comparing our method to state-of-the-art cartilage segmentation methods, we do a comparison between the combination of binary classifiers and previous work where the cartilage is segmented using a hierarchical approach, combining one two-class and one three-class $k$NN classifier [48]. We also compare our results to a direct three-class classifier. Features are selected by forward selection followed by backward selection, and we consider the importance of of a suitable criterion function in the feature selection for the classifier performance.

## 4.2 Methods

### 4.2.1 Image Acquisition

An Esaote C-Span low-field 0.18 T scanner dedicated to imaging of extremities acquires Turbo 3D T1 scans (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms). The scans are made through the sagittal plane and have four different voxel sizes, $(0.7031/0.7813/0.8594/0.9375) \times 0.7031 \times 0.7031\text{mm}^3$. After automatically removing boundaries that contain no information, the scan size is $104 \times 170 \times 170$ voxels.

The scans have been manually segmented on a slice-by-slice basis by a radiologist. A scan slice with the tibial and femoral medial cartilage manually segmented is shown in Figure 4.1. We use 111 scans of both left and right knees,

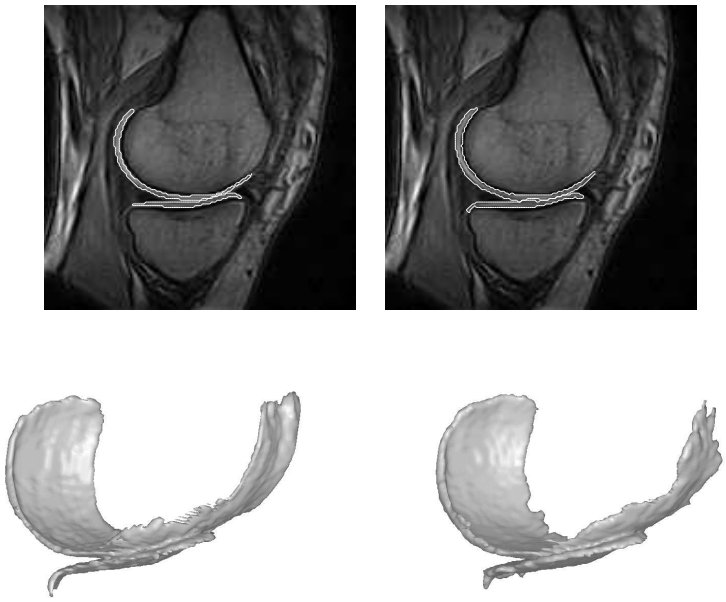


**Figure 4.1:** The segmentation of tibial and femoral medial cartilage. Top row shows one slice (170x170 pixels) and the bottom row are the 3D visualizations. Manual segmentations are to the left and automatic segmentations to the right. The sensitivity, specificity and DSC of our automatic segmentation are 88.68%, 99.88% and 0.812 for this scan.

the right knees are reflected about the center in the sagittal plane in order to treat all scans analogously. The test subjects are both males and females, aged between 22 and 79 years, and diagnosed as having a Kellgren and Lawrence Index (KL) [73] between 0 and 3. There are 98 scans of healthy or borderline OA knees (KL $\leq 1$) and 13 of osteoarthritic knees (KL $> 1$).

### 4.2.2 Cartilage Classification

We implement our classifier in an approximate $k$NN framework developed by Mount and colleagues [4]. The approximate classifier is in principle a $k$NN-classifier, but allows for faster computations if an error is tolerated. The search algorithm returns $k$ points such that the ratio of the distance between the $i$th

reported point ($1 \leq \ i \leq k$) and the true $i$th nearest neighbor is at most $1+\epsilon$. One difficulty in classification tasks is the tradeoff between computational complexity and accuracy. We found empirically that $\epsilon = 2$ and $k = 100$ offers a reasonable such tradeoff. There are three classes we wish to separate, tibial medial cartilage, femoral medial cartilage and background. We combine one two-class classifier trained to separate tibial cartilage from the rest and one trained to separate femoral cartilage from the rest with a rejection threshold, described in section 4.2.6.

### 4.2.3 Features

We here introduce a set of candidate features, and in section 6.3.3 we describe our feature selection scheme. We are interested in features that describe the geometry of the object in question, and therefore have the 3-jet as candidate features since they can describe all local 3D geometric features up to third order [46]. All derivatives are Gaussian derivatives, and all features are examined on three scales, (0.65mm, 1.1mm and 2.5mm), to cover the range of different cartilage thicknesses. The $x$-, $y$ and $z$ axes are defined as the sagittal-, coronal- and axial axes. The eigenvalues and the eigenvectors of the Hessian ($\mathbf{H}$) and the structure tensor ($\mathbf{T}$) are also candidate features.

We would also like to examine third-order properties and therefore evaluate the third-order tensor, in Einstein annotation $L_{www} = I_{ijk}I_iI_jI_k/(I_iI_i)^{3/2}$, in the gradient direction as candidate features.

Besides these features related to the geometry, we also include the location in the image and the intensity as candidate features, because these are features that are highly relevant for a radiologist when visually inspecting the images. The intensities smoothed on the three scales are also candidate features.

### 4.2.4 Training of the Classifier

Feature selection can not only reduce computational time, it can also make the classification system less sensitive to the curse of dimensionality. The features of the classifiers are selected by sequential forward selection (SFS) followed by sequential backward selection (SBS). In the forward selection we start with an empty feature set and expand the search space by adding one feature at the time according to the outcome of a criterion function. The backward selection starts with the features found by the sequential forward selection and iteratively excludes the least significant feature according to the criterion function. All features are examined in every iteration, which means that the same feature

can be selected several times.

For the two-class classifier we perform forward selection until there are 60 features in the feature set, then the backward selection continues until there are 39 features left in the set. As criterion function for the two-class classifiers we use the area under the ROC curve (AUC). The ROC curve is determined by varying the threshold for the classifier and plotting the ratio of false positives against the ratio of true positives. A perfect classifier has an $AUC = 1$ and a random performance yields $AUC = 0.5$. In the two-class classification case, the AUC is a well established criterion function when the costs of misclassifications are unknown. However, this criterion function is not easily extendable to N-class ($N > 2$) classification tasks. Mossman [94] has developed ROC analysis for multi-class classification by varying the class sensitivities as decision thresholds to form 3D surfaces. For three-class classification there are 6 such surfaces, and the volumes under the surfaces (VUS) are the performance metrics, similar to the AUC in 2D. A perfect classifier has $VUS = 1$ and chance performance yields $VUS = 0.17$. However the sensitivities does not provide a complete description of the operator performance, and all $N^2 - N$ misclassification options are not taken into account. Edwards *et al.* [40] has generalized ROC analysis for multi-class classification tasks into what they call the hypervolume under the ROC hypersurface, a performance metric that take all misclassification probabilities into account. However they conclude that the hypervolume may not be a useful performance metric when $N > 2$. For the three-class classifier, we take the average of the six Mossman's VUS's as the criterion function. In the SFS, we expand the feature space until it is 75 dimensional, and in the SBS the dimensionality is decreased until 54 dimensions remain. We use 25 scans for the training of the classifier, the same 25 scans are used in the feature selection, threshold selection and for the training data set for the final classifier.

### 4.2.5 Selected Features

We train one tibial cartilage $\omega_t$ vs. rest, one femoral cartilage $\omega_f$ vs. rest and one background $\omega_b$ vs. rest classifier. In addition we train a classifier that directly separates the three classes. The selected features for the different classifiers are presented in Table 4.1. The position, the intensity smoothed on the three scales and the first order derivatives on $\sigma = 2.5$mm are highly ranked by all classifiers. The structure tensor and the Hessian are repeatedly selected by all classifiers, but the direct three-class classifier has selected features from the 3-jet more frequently than the others.

**Table 4.1:** The features selected for the different classifiers. The number corresponds to the significance with 1 as the most significant feature. $\mathbf{T}(\sigma_T, \sigma)$ and $\mathbf{H}(\sigma)$ stand for the eigenvalues unless otherwise stated.

| Feature/classifier | $\omega_{tm}$ | $\omega_{fm}$ | $\omega_b$ | Direct 3-class |
|---|---|---|---|---|
| $I_x, I_y, I_z$ $\sigma = 0.65$ | 5 | | 9 | 9 |
| $I_x, I_y, I_z$ $\sigma = 1.1$ | 7 | 3 | 13 | 14 |
| $I_x, I_y, I_z$ $\sigma = 2.5$ | 4 | 4 | 4 | 1 |
| Position | 1 | 1 | 1 | 2 |
| Intensity | | | | 4 |
| Intensity smoothed | 2 | 5 | 2 | 5 |
| $\mathbf{H}$(0.65mm) | | 8 | | |
| $\mathbf{H}$(1.1mm) | 8 | 10,11 | 6 | 13, 14, 15 |
| $\mathbf{H}$(2.5mm) | 10 | 9,13 | 5 | 8 |
| $\mathbf{T}$(1.1mm, 0.65mm) | | 2* | | 16* |
| $\mathbf{T}$(2.5mm, 0.65mm) | 11, 12, 13 | 12 | 10, 11, 12 | |
| $\mathbf{T}$(2.5mm, 1.1mm) | | | | 7 |
| $I_{xxx}$ all scales | | | | 6 |
| $I_{zzz}$ all scales | 6 | 6 | | |
| $I_{xxz}$ all scales | | | | 11 |
| $I_{xx}$ all scales | | | | 3 |
| $I_{yy}$ all scales | 9 | | 7 | 17,18 |
| $I_{zz}$ all scales | 3 | 7 | 3 | |
| $I_{xy}$ all scales | | | 8 | 12 |
| $I_{yz}$ all scales | | | | 10 |

* eigenvector corresponding to largest eigenvalue

### 4.2.6 Setting a Threshold for the Classifier

The outcome of an $\omega_i$ vs. rest classifier can be seen as the posterior probabilities that, for all the voxels in the image, a voxel $j$ with feature vector $\mathbf{u}_{i,j}$ belongs to class $\omega_i$. We denote it $p(\omega_i|\mathbf{u}_{i,j})$ or $p_{i,j}$ in short. In one-versus-all classification, which is commonly used for multi-class classification [143], one builds $\omega_i$ vs. rest classifiers for $i = 1, \ldots N$ and perform a winner-takes-all vote between them, assigning $j$ to the class $\omega_i$ with the highest posterior probability. In the scans, only a few percent of the voxels belong to cartilage, making the background the by far largest class. Our approach is similar to one-versus-all, but due to the dominance of the background class we replace the background vs. rest classifier by a rejection threshold, so that a the posterior probability should be higher than a threshold $T$ before it can be assigned to a cartilage class. The decision

rule is:

$$j \in \begin{cases} \omega_{tm}, & p_{tm,j} > p_{fm,j} \quad \text{and} \quad p_{tm,j} > T; \\ \omega_{fm}, & p_{fm,j} > p_{tm,j} \quad \text{and} \quad p_{fm,j} > T; \\ \omega_b & \text{otherwise,} \end{cases} \tag{4.1}$$

Using the features found from feature selection, the 25 training scans are leave-one-out evaluated. The results of varying the threshold for the tibial cartilage vs. rest and femoral cartilage vs. rest classifiers are seen in Figure 4.2. The Dice similarity coefficient (DSC) is considered a useful statistical measure for studying accuracy in image segmentation [149]. The rejection threshold is set to 0.98, because as demonstrated in Figure 4.2, the maximum DSC on the training data occurs then. The fairly high threshold is a consequence of the prototype selection, where all cartilage voxels and 0.5% randomly sampled background voxels according to gold standard segmentations are used.

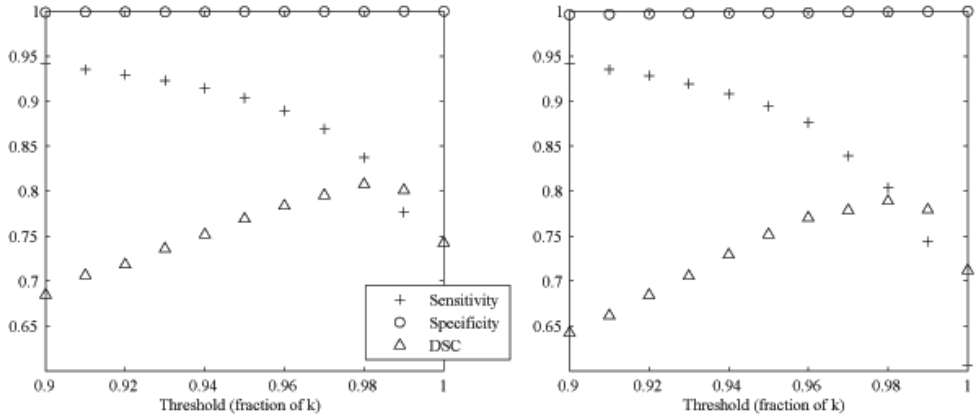



**Figure 4.2:** The results of varying the threshold in leave-one-out evaluation of the training set. The tibial vs. rest classifier performance is demonstrated to the left, and the femoral vs. rest classifier to the right.

## 4.3 Results

We classify the scans according to the decision rule in section 4.2.6, and the largest connected component is selected for the segmentation. The sensitivity, specificity and DSC are calculated using manual segmentations by a radiologist as gold standard. The results are displayed in Table 12.3 and are visualized in Figures 4.1 and 4.3. The results of our method is distinctly higher than those of the fully automatic method of Pakin *et al.* [101]. The sensitivity and

DSC of our method is not as high as that of Grau et al. [61], however we have evaluated our method on far more scans, including osteoarthritic knees, and without manual interaction. In Table 12.3 we also compare the results of our

**Table 4.2:** The results of our combined binary classifier method (denoted 'This work') compared to two state-of-the-art methods, the hierarchical method in [48] and a direct three-class classifier, evaluated with comparisons to manual segmentations.

| Method | This work | Pakin | Grau | Hierarchical | 3-class |
|---|---|---|---|---|---|
| Sensitivity | 84.17% | 66.22% | 90.03% | 90.01 | 81.25% |
| Specificity | 99.89% | 99.56% | 99.87% | 99.80% | 99.85% |
| DSC | 0.81 | N/A | 0.90 | 0.80 | 0.77 |
| Test set | 86 | 1 | 7 | 46 | 86 |
| Interaction time | 0 | 0 | 5-10 min | 0 | 0 |

segmentation method to the results we obtained using a hierarchical approach [48], where the voxels are first roughly partitioned into cartilage/non-cartilage using a two-class classifier, then a three-class classifier partitions the voxels that are classified as cartilage in the first round into tibial cartilage, femoral cartilage and background. The classifiers are trained similar to what is described in this paper. The data set in [48] contains 3% knees with clear signs of OA (KL $> 1$) compared to 11% in this work. The direct three-class classifier performs distinctly worse than the hierarchical classifier and our combined binary classifier method.

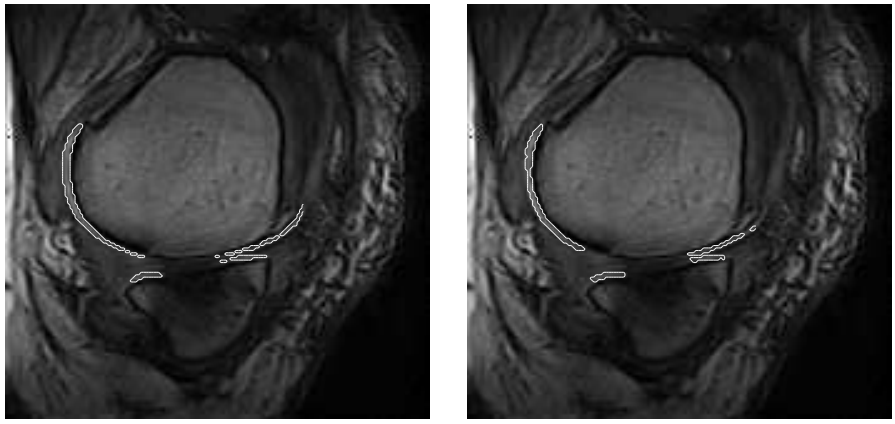


**Figure 4.3:** The segmentation of an osteoarthritic knee, diagnosed as having KL 3, in a slice where the cartilage is degraded. To the left is a gold standard segmentation and to the right our automatic segmentation.

The radiologist re-segmented the tibial medial cartilage in 31 scans in order to determine intra-rater variability for the gold standard segmentations. The average DSC between the two manual segmentations is 0.86, which explains the fairly low values of the DSC in our evaluation since we cannot expect our method to resemble the expert better than the expert itself. The corresponding DSC of the automatic segmentation versus expert for the tibial cartilage of the 31 scans is 0.82. Ideally, several trained experts should segment every scan several times and a more robust gold standard could then be determined *e.g.* using STAPLE [133], but with large amounts of data it is too labor intensive.

We believe that the $\omega_b$ vs. rest classifier could add little more information to our final class decisions. Still, we examine its effect by temporarily including it in the decision rules, making them:

$$j \in \begin{cases} \omega_{tm}, & p_{tm,j} > p_{fm,j} \quad \text{and} \quad p_{tm,j} > p_{b,j} \quad \text{and} \quad p_{tm,j} > T; \\ \omega_{fm}, & p_{fm,j} > p_{tm,j} \quad \text{and} \quad p_{fm,j} > p_{b,j} \quad \text{and} \quad p_{fm,j} > T; \\ \omega_b & \text{otherwise.} \end{cases}$$

The average sensitivity, specificity and DSC using this rule remained exactly the same as for the two binary classifiers with rejection threshold. Using a pure one-versus-all approach with a winner-takes-all vote (assigning the $j$ to the class with the highest probability of the classifiers) instead of the threshold $T$ resulted in a large number of false positives, and yielded an average sensitivity, specificity and DSC of 98.5%, 98.9% and 0.48 for the medial cartilage compartments.

## 4.4 Discussion

Our segmentation method combines two two-class $k$NN classifiers trained to distinguish tibial cartilage respectively femoral cartilage from the rest of the image, and a majority vote between the two-class classifiers are combined with a threshold for the background class. The threshold is optimized for maximizing the DSC. We have shown that this classification system can replace a one-versus-all system without degrading the performance. This means that the background vs. rest classification is unnecessary, which saves approximately one third of the computational time compared to a one-versus-all approach.

In the feature selection, the criterion function determines what features that will be included in the classifier, and thus has a major impact on the classifier performance. The $k$NN classifier is inherently a multi-class classifier, but the combination of binary classifiers yields better results than the direct three-class classification approaches. The direct three-class systems have a lower total amount of features due to memory limitations, given a certain amount of allowed

memory the two-class systems with lower dimensional feature spaces can include more training data points.

We have demonstrated a cartilage segmentation algorithm that compared to state-of-the-art methods outperforms the to our knowledge only fully automatic method, and performance close to that a semi-automatic method though our method is fully automatic, is evaluated on a larger data set, and includes knees with signs of OA.

Chapter 5

# Position Normalization in Automatic Cartilage Segmentation

*Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen*

**Abstract**

In clinical studies of osteoarthritis using magnetic resonance imaging, the placement of the test subject in the scanner tends to vary and this can affect the outcome of automatic image analysis methods for articular cartilage assessment, particularly in multi-center studies. We have developed an automatic iterative method that corrects for position variations by combining the two steps: shifting the cartilage towards the expected position and performing a voxel classification with the normalized position as a feature. By applying this placement adjustment scheme to an automatic knee cartilage segmentation method we show that the inter-scan reproducibility is much improved and is now as good as that of a highly trained radiologist.

## 5.1 Introduction

Osteoarthritis (OA) is one of the major health issues among the elderly population [43]. One of the main effects of OA is the degradation of the articular cartilage, causing pain and loss of mobility of the joints.

Magnetic resonance imaging (MRI) is the only imaging modality for direct, non-invasive visualization of the articular cartilage [60] where cartilage deterioration can be detected [104]. Among MRI sequences, the most established are fat-suppressed gradient-echo T1 sequence using a 1.5T or a 3T magnet which yields high image quality. Low-field dedicated extremity MRI produces images with lower quality but at a very low cost. They can provide similar information on bone erosions and synovitis as expensive high-field MRI units [41], and if a low-field scanner can be used for articular cartilage assessment as well, costs for making clinical studies would be reduced significantly.

In quantitative assessment of articular cartilage using MRI data, the most crucial step is the segmentation. The cartilage can be manually segmented slice-by-slice by experts, but for routine clinical use manual methods are too time consuming and they are prone to inter- and intra-observer variability. In order to overcome these problems much effort has been put into development of semi- or fully automatic segmentation methods, both in 2D [88],[124] and directly in 3D [61],[135],[48]. When assessing the cartilage directly in 3D the problem of limited continuation between slices that is present in 2D techniques is eliminated.

An uncommitted segmentation scheme is often desired to achieve invariant performance under irrelevant transformations of the images such as translation, rotation, resolution of and orientation and position in the measuring device. A mathematical well-founded and operational approach to this is the scale space methodologies as is also applied in the automatic segmentation framework described in [50]. In segmentation by classification it is also well recognized that the geometrical covariance between feature points on a space of features derived mainly from Gaussian scale space derivatives can support the task significantly [74]. Similar we would like to introduce position variance relative to the training data in a normalized way. Images and in particular medical scans are obviously not acquired with random relation between objects of interest and the field of view. This manifests itself in the segmentation algorithm [50] by having non-normalized position as a very strong feature for classification between cartilage and background when compared to invariant geometric scale space features. On

the other hand patients will be placed slightly different depending on clinical staff. This can be corrected using an iterative scheme for position normalization and at the same keep the relative position as strong feature. In this paper, we introduce an iterative position normalization scheme for automatic segmentation tasks and our evaluation shows that it increases the robustness of an automatic segmentation method significantly.

## 5.2 Methods

### 5.2.1 Acquisition and Population

The test subjects are between 22-79 years old with an average age of 56 years, 59% females, and there are both healthy and osteoarthritic knees according to the Kellgren-Lawrence index (KL) [73], a radiographic score from 0-4 where KL 0 is healthy, KL 1 is considered borderline or mild OA, and KL $\geq$ 2 is severe OA. We examine 25+114 knees, 25 for training of the segmentation method and 114 for evaluation. Of the 114, 31 knees have been re-scanned approximately a week after the first scan occasion for reproducibility evaluation. In the test set there are 51, 28, 13 and 22 scans that have KL 0, 1, 2 and 3 respectively.

MRI is performed with an Esaote C-Span low-field 0.18T scanner dedicated to imaging of extremities yielding a Turbo 3D T1 sequence (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms). Approximate acquisition time is 10 minutes and the scan size, after automatically removing background that contain no information, is $104 \times 170 \times 170$ voxels. The spatial resolution of the scans is approximately $0.8 \times 0.7 \times 0.7\text{mm}^3$.

31 knees were re-scanned after approximately one week in order to examine segmentation precision. All the scans have been manually delineated by a radiologist in order to establish the accuracy of the automatic method and the same 31 scans were delineated twice with the purpose of examining the intra-rater variability of the manual delineations. An example of how an MRI slice and the manual delineation looks like can be seen in the top rows of Figures 5.2 and 5.3.

### 5.2.2 The Cartilage Segmentation Method

The cartilage segmentation method we examine is fully automatic and consists of two binary approximate $k$NN classifiers [48] implemented in an approximate $k$NN framework developed by Mount et al. [4].

Three classes are separated, tibial medial cartilage, femoral medial cartilage and background. The method focuses on the medial compartments since OA is more often observed there [38]. One two-class classifier is trained to separate tibial cartilage from the rest and one is trained to separate femoral cartilage from the rest, and cartilage is segmented by a vote between the two classifiers combined with a rejection threshold for the background [50].

The classifiers are trained on 25 scans using feature selection, which is sequential forward selection followed by sequential backward selection with the area under the ROC curve [70] as criterion function. The selected features are: the image intensities smoothed (Gaussian) on three different scales, the position in the image, eigenvalues and the eigenvector corresponding to the largest eigenvalue of the structure tensor and the Hessian, and Gaussian derivatives up to third order. These features are ordered in decreasing significance determined by the feature selection.

### 5.2.3 Position Normalization Applied to the Segmentation

The placement of the knee varies in clinical studies but is still a strong cue to the location of cartilage, which is evident in the described segmentation method where the position in the scan is selected as one of the most significant features. Even though the global location is a strong cue the minor variation in placement is a source of errors. Segmentation methods that rely on manual interaction are typically less sensitive to knee placement, we however wish to eliminate manual interaction in the segmentation algorithm, hence placement variations is an issue that needs attention. Figure 5.1 shows how knee position in the scan affects the automatic segmentation method. One way of correcting for knee placement is to manually determine where in the scan the cartilage is, but this can take time with 3D images since a human expert typically search through the scans on a slice-by-slice basis. And when the segmentation method itself is automatic, an automatic adjustment is advantageous.

In order to adjust the segmentation method to become more robust to variations in knee placement we have developed an iterative scheme, which consists of two steps. First, the coordinates of the scan are shifted so that the cartilage center of mass found from the segmentation is positioned at the location for the center of mass for the cartilage points in the training set. Then the dilated volume of the segmentation is classified with the other features unchanged. The dilation extends the boundary outwards by three voxels and by only classifying the voxels inside this volume, which is typically only a few percent of the total scan volume, the computation time is not significantly increased. In order to determine if the

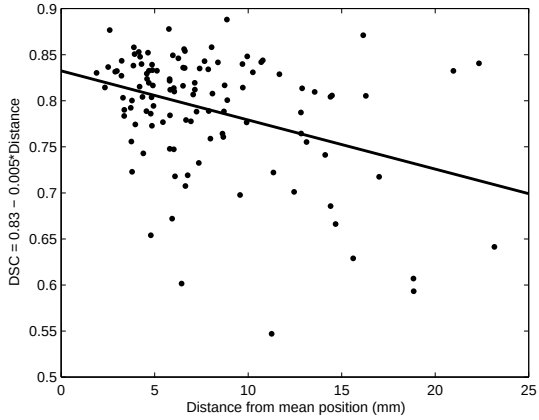



**Figure 5.1:** The DSC between manual and automatic segmentation as a function of the distance to the mean position for the 114 scans with a line least-squares fitted to the points, illustrating how position variance affects the segmentation performance.

selected region is a reasonable choice we repeated the classification with all the voxels in the image, yielding the same results with much longer computation time. The outcome is combined according to (6.1) and the largest connected component is selected as the cartilage segmentation.

## 5.3 Results

The automatic segmentation yields an average sensitivity, specificity and DSC are 81.1% (±11.0% std), 99.9% (±0.04% std) and 0.79 (±0.07 std) respectively in comparison with manual segmentations. The inter-scan reproducibility of the volumes from the automatic segmentations is evaluated on the 31 scan-rescan pairs using linear correlation coefficient (0.86), and pairwise volume difference of 9.3%.

After applying position normalization, the average sensitivity, specificity and DSC are 83.9% (±8.37% std), 99.9% (±0.04% std) and 0.80 (±0.06% std) respectively and it converges in two iterations. Compared to the initial segmentation there is a significant increase in sensitivity ($p < 1.0 * 10^{-7}$) and in DSC ($p < 2.5 * 10^{-3}$) according to a paired t-test. In order to illustrate how the segmentations are affected, the best and the worst results from the position correction scheme are shown in Figures 5.2 and 5.3. In the best case the DSC increases with 0.17 and for the worst scan it decreases with 0.017.

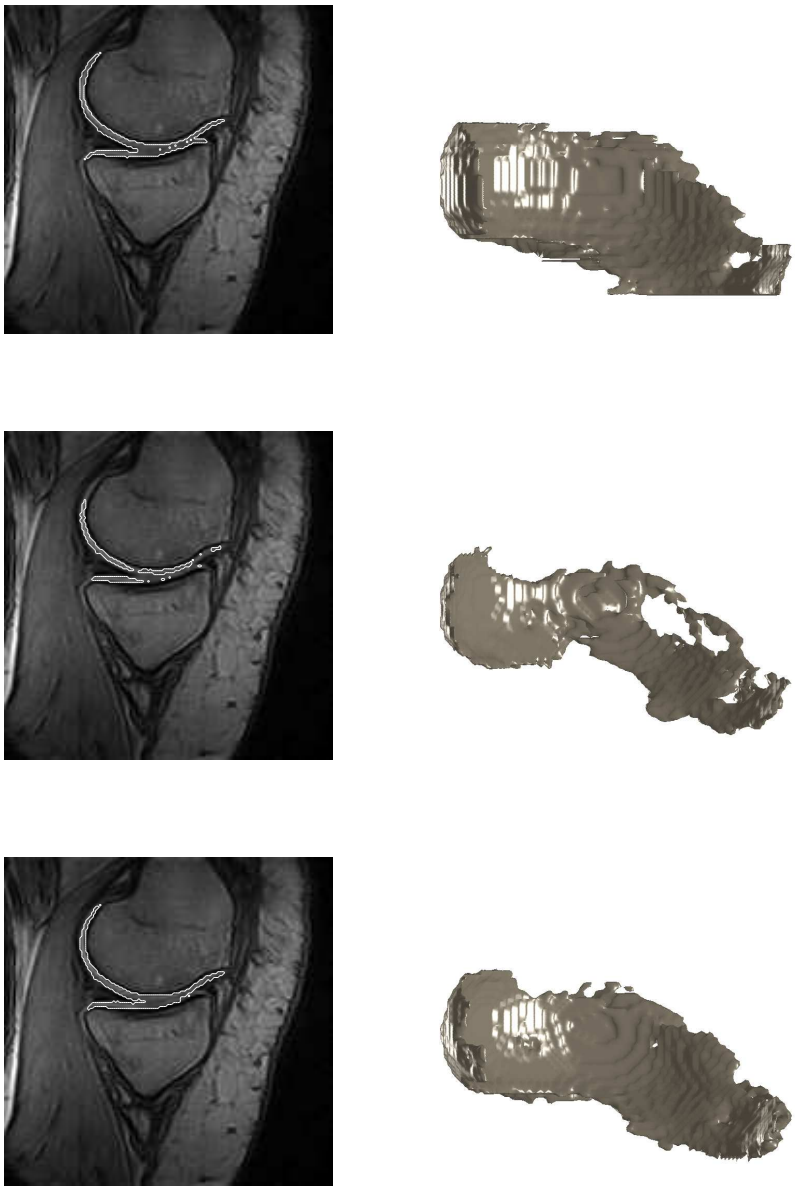


**Figure 5.2:** The scan most improved by the position correction scheme, where the DSC increases from 0.61 to 0.77. Top row shows the manual segmentation, the second row shows the original segmentation and the third row shows the segmentation after position correction. The 3D views are seen from above, and the 2D images are a sagittal slice of the segmentation.

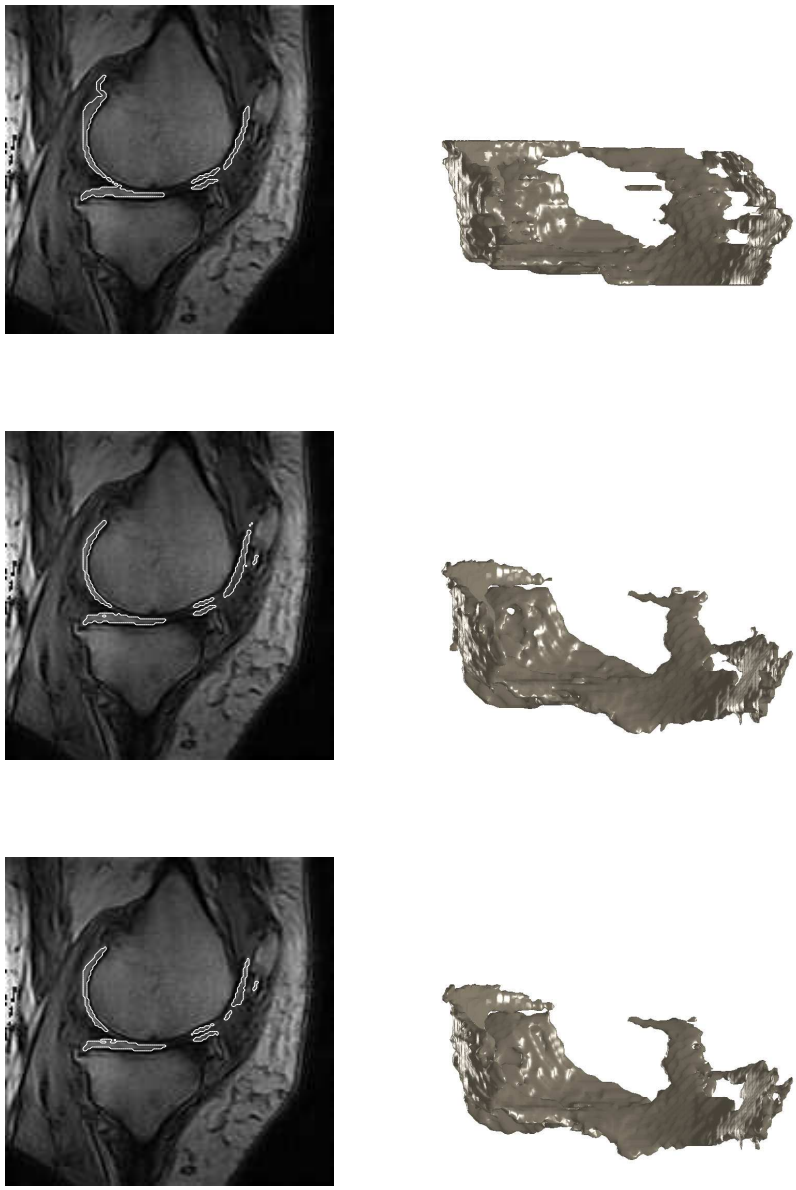


**Figure 5.3:** The worst case scenario of applying position correction. The knee is severely osteoarthritic (KL 3). For this scan there is no improvement in DSC. The manual segmentation is in the top row, the second row shows initial segmentation and the third row shows the segmentation after position correction.

The reproducibility of the segmentation is improved, with an increase of the linear correlation coefficient from 0.86 to 0.93 and the paired volume difference decreases from 9.3% to 6.2%. These reproducibility values can be compared to the volumes from the manual segmentations by a radiologist for the same data set. The linear correlation coefficient is 0.95, and the radiologist has an average volume difference of 6.8%. The radiologist re-delineated the tibial medial cartilage in 31 scans in order to determine intra-rater variability for the manual segmentations. The average DSC between the two manual segmentations is 0.86, and the corresponding DSC of the automatic segmentation versus expert for the tibial cartilage of the 31 scans is 0.82. Due to the variability in the gold standard segmentations the segmentation algorithm may not be much further improved.

## 5.4 Summary and Discussion

We have developed an iterative method for normalization of the position in segmentation of the articular cartilage in knee MRI data. We maintain the strong cue given by position and simultaneously achieve robustness (but not mathematical invariance) to the variance in the operators placement of the knee in the scanner.

Our position normalization scheme converges in only two iterations, after which the inter-scan reproducibility is improved, the linear correlation coefficient for the volumes between the first and second scan occasion increases from 0.86 to 0.93, and the volume difference decreases from 9.3% to 6.2%. The corresponding values for the radiologist are 0.95 for the correlation coefficient and 6.8% volume difference. There is a small but significant increase in both sensitivity and in DSC for the 114 scans evaluated. From Figures 5.2 and 5.3 it can be seen that the best case is an efficient improvement the segmentation and the worst case is a practically unaltered segmentation, thus the segmentation is often positively and never negatively affected by our method. The scan with the worst result is from a severely osteoarthritic knee which can be difficult even for a highly trained expert to segment.

Thus by only slightly increasing computation time our position normalization scheme increases the sensitivity and DSC of the segmentation method, but more importantly, the inter-scan reproducibility of the method is much improved and is now as good as that of a radiologist, which is highly relevant in clinical studies where the knee placement in the scanner will vary. Because the method is completely automatic and improves the segmentation reproducibility, it can become a cost efficient tool in clinical studies where automated segmentation methods are being used.

Chapter 6

# Segmenting Articular Cartilage Automatically Using a Voxel Classification Approach

*Jenny Folkesson, Erik B. Dam, Ole F. Olsen, Paola C. Pettersen, and Claus Christiansen*

**Abstract**

In this paper we describe a fully automatic method for articular cartilage segmentation from MRI which we use as the foundation of a quantitative cartilage evaluation framework. We evaluate our method by comparisons to manual segmentations by a radiologist and by examining the inter-scan reproducibility and relation to signs of osteoarthritis (OA) of the cartilage volume and area estimates. Training and evaluation of the method is performed on a data set consisting of 139 scans of knees with a status ranging from healthy to severe OA. This is to our knowledge the only fully automatic cartilage segmentation method that has good agreement with manual segmentations, an inter-scan reproducibility as good as that of a human expert, and enables the separation between healthy and OA populations.

While high-field scanners offer high quality imaging from which the articular cartilage have been evaluated extensively using manual and automated image analysis techniques, low-field scanners on the other hand produce lower quality images but to a fraction of the cost of their high-field counterparts. For low-field MRI there is no well established accuracy validation for quantitative cartilage estimates, but the significant differences between healthy and OA populations found from cartilage volume and surface area

estimates suggests that low-field MRI analysis can become a useful, affordable tool in clinical studies.

## 6.1 Introduction

Osteoarthritis (OA) is one of the major health issues among the elderly population, it is second to heart disease in causing work disability and is associated with a large socioeconomic impact on health care systems [68]. One of the main effects of OA is the degradation of the articular cartilage, causing pain and loss of mobility of the joints. Currently the treatment of OA is mainly restricted to symptom control [43], and in the search for disease modifying drugs much research is dedicated to analysis of articular cartilage and its relation to disease progression.

Magnetic resonance imaging (MRI) is the leading imaging modality for direct, non-invasive assessment of the articular cartilage [60], and cartilage deterioration can be detected using quantitative MRI analysis [104]. Among MRI sequences, the most established are fat-suppressed gradient-echo T1 sequences using a 1.5T or a 3T magnet. The standard sequences in literature for these scanners have high in-plane resolution but usually have a larger inter-slice distance, and many assessment methods developed for such sequences are on a slice-by-slice basis. For a thorough review of MRI scan protocols for knee OA assessment, see [105].

A recent study shows that low-field dedicated extremity MRI can provide similar information on bone erosions and synovitis as expensive high-field MRI units [41]. There have been several comparisons of diagnostic performance of diagnosing meniscal tears, cruciate ligaments and cartilage lesions between low- and high-field MRI data [75, 80, 139] reporting everything from compatible performance to the high-field unit outperforming the low-field unit. There has also been a comparison between low-field MRI and arthroscopy [111] finding a good correspondence between the two for cruciate ligament and lesion detection in the knee.

The use of a dedicated low-field MRI has its advantages and disadvantages. The drawbacks are related to image quality with lower resolution and more difficulties in incorporating features such as fat suppression, however fat suppression has been successfully implemented lately for low-field MRI [66]. The main advantages are cost-effectiveness with much lower cost per scan, lower installation and maintenance costs, and higher patient comfort without claustrophobic feelings and minimal noise level. So far there has not been any accuracy validation of quantitative cartilage measures from a low-field scanner, but if a low-field scanner can be used for quantitative articular cartilage assessment, costs for making clinical studies would be reduced significantly. If manual labor is con-

nected with the analysis and quantification of MRI data in clinical studies, one more cost factor is introduced. In this work we present a fully automatic segmentation based cartilage assessment framework, and we evaluate it on low-field MRI by comparison to manual delineations by a radiologist, we evaluate the robustness in terms of inter-scan reproducibility, and the ability to detect changes between healthy and osteoarthritic groups using generated cartilage volume and area estimates.

### 6.1.1 Related Work

As in most quantitative studies in medical imaging, the first and most crucial step in our articular cartilage evaluation framework is the segmentation. The cartilage can be manually segmented slice-by-slice by experts, but for routine clinical use manual methods are too time consuming and they are prone to inter- and intra-observer variability. It is thus advantageous to automate the segmentation method and the main challenges in developing an automatic method are the thin structure of the cartilage and the low contrast between the cartilage and surrounding soft tissues.

Several groups have developed semi-automatic/automatic methods for cartilage segmentation. Among 2D techniques, Stammberger and colleagues [124] segments the cartilage by fitting a b-spline snake to each slice. A 2D method combining user interaction with active contours is described by Lynch *et al.* [88]. They combine the segmentation technique with 3D image registration to detect changes in cartilage volume [89]. Solloway *et al.* [119] use active shape models for slice-by-slice cartilage segmentation, and estimate cartilage thickness in the direction perpendicular to the medial axis in each slice.

When working with a 2D technique, continuation between slices is lost and some regularization between the slices is required. Also, since the series of 2D segmentations have to be converted into a 3D segmentation when finding for example thickness maps, it is advantageous to perform segmentation in 3D directly.

Looking at the 3D techniques that have been developed, Grau *et al.* [61] use a watershed based approach, where the watershed is extended to examining difference in class probability of neighboring pixels. The mean sensitivity, specificity and DSC of the segmentation are 90.03%, 99.86%, and 0.90 respectively. The method is evaluated on 7 scans from 4 healthy knees and requires 5-10 minutes of manual labor for selecting markers before initializing the watershed.

Pakin *et al.* [101] has developed a region growing scheme that is followed by a two-class clustering for segmenting the cartilage. However, the method assumes

that the bones are already segmented. The sensitivity and specificity of the method are 66.22% and 99.56% respectively, and it is evaluated on one scan. The method has been further developed to incorporate a trained user for correcting misclassifications [126], and this semi-automatic method is evaluated in terms of intra-user reproducibility.

Another classification approach to segmentation is presented by Warfield *et al.* [135],[134], where a user performs interactive registration of a knee template to a test scan. The method then iterates between a classification step and a template registration step to produce a segmentation. The method has a lower intra-scan variability of the volume compared to repeated manual segmentations on the scan it is evaluated on.

A semi-automatic method based on a graph searching segmentation algorithm by Li *et al.* [86] followed by mean thickness quantification is evaluated on ankle joints in [92]. The method requires only a small amount of manual initialization and shows accurate thickness measurements on 8 cadaveric ankles. Presumably the method could also be adapted to knees.

### 6.1.2 Overview of the Work Presented

The segmentation techniques described in section 4.1.1 all require some amount of manual interaction except for the method of Pakin *et al.* [101], the 3D techniques are evaluated only on relatively small data sets and neither Grau *et al.*nor Pakin *et al.*evaluate their methods on scans from OA test subjects.

In this paper, we propose a method that can fully automatically segment cartilage in MR scans of both healthy and osteoarthritic knees. The segmentation method is the first step in a quantitative, fully automatic cartilage evaluation framework and is primarily intended for clinical studies using low-field MR scanners. The segmentation algorithm is based on a one-versus-all approach of combining binary approximate $k$NN classifiers which is described in sections 6.3.1 and 6.3.2, followed by an iterative position adjustment method that is intended to correct for the variations of the placement of the test subject in the scanner, something that is bound to occur in any clinical study and is described in section 6.3.7. Since $k$NN classification is a slow process we propose to use an efficient voxel classification algorithm which is described in section 6.3.6.

Since we cannot obtain ground truth for an *in vivo* study with both healthy and OA test subjects and ground truth accuracy of low-field MRI analysis is yet to be established, we evaluate our method not only compared to manual tracings of a radiologist, but also in terms of precision and relation to radiographic signs of OA. We evaluate the inter-scan reproducibility using the volume and

surface area estimate, and the ability to detect changes between healthy and osteoarthritic populations by performing unpaired Student's t-tests between populations with different OA severity using the volume and area estimates. OA is more frequently observed in the medial compartment [38], therefore we focus on the medial cartilage compartment in this study. The evaluation of the segmentation framework is described in section 6.4 followed by discussion in section 6.5.

## 6.2 Image Acquisition

### 6.2.1 Magnetic Resonance Image Acquisition

MRI was performed with an Esaote C-Span lowfield 0.18T scanner dedicated to imaging of extremities yielding a sagittal Turbo 3D T1 sequence (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms). Approximate acquisition time is 10 minutes and the scan size, after automatically removing boundaries that contain no information, is $104 \times 170 \times 170$ voxels. The spatial in-plane resolution of the scans are $0.70 \times 0.70\text{mm}^2$, with a distance between slices ranging between 0.70mm - 0.94mm, where the most common distance is 0.78mm.

Assessing the cartilage directly in 3D eliminates the problem of limited continuation between slices that is present in 2D techniques. We use a 3D sequence consisting of near isotropic voxels since this is well suited for cartilage quantification [141] and for 3D analysis in general.

### 6.2.2 Test Subject Population

We examine 139 knee joints in vivo, of which 59% are from female test subjects. The ages of the test subjects varies between 22-79 years with an average age of 56 years. The status of the knees range from healthy to osteoarthritic according to the Kellgren-Lawrence index (KL) [73], a radiographic score established by x-rays between 0-4 where KL 0 is healthy, KL 1 is considered borderline or mild OA, and KL $\geq 2$ is severe OA. In our data set, 51 knees have KL 0, 28 have KL 1, 13 have KL 2 and the remaining 22 knees have KL 3. In the x-rays the width of the tibial plateau has also been measured, which we use for normalization of the cartilage volume and surface area so that measures of subjects of different sizes can be compared. The scans are from both left and right knees, and in order to treat all scans analogously with the same methods all the right knees are reflected about the center of the sagittal axis.

The images are transmitted from the MRI unit to a workstation, where they are processed using a medical imaging display and analysis system designed for the task. The software allows for manual segmentation on a slice-by-slice basis. A user marks points on the object boundary, and linear interpolation between the points delineates the boundary. The MR scans have all been manually segmented by a radiologist using this software, and 31 scans are segmented twice with the purpose of examining the intra-rater variability of the manual delineations.

Of the 139 knees, the same 31 knees that were segmented twice were re-scanned after approximately one week in order to examine the segmentation precision, giving a total of 164 MR scans. An example of how a MRI slice and the manual delineation looks like can be seen in the first column of Figures 5.2 and 5.3.

## 6.3 Cartilage Segmentation

### 6.3.1 Voxel Classification

We use a $k$NN classifier, which finds the posterior probability for each voxel $j$ with feature vector $\mathbf{u}_j$ of belonging to class $\omega_c$, where $i = 1, \ldots N$ is the number of classes. The probability is given by

$$p(\omega_c|\mathbf{u}_j) = \frac{k_c}{k},$$

where $k_i$ is the number of neighbors with class label $\omega_c$. We implement our classifier in an Approximate Nearest Neighbor framework developed by Mount and colleagues [4]. The classifier is in principle a $k$NN-classifier, but allows for faster computations if an error in the distance calculations is tolerated. The approximate $k$NN search algorithm returns $k$ points such that the ratio of the distance between the $i$th reported point ($1 \leq \ i \leq k$) and the true $i$th nearest neighbor is at most $1 + \epsilon$. One difficulty in classification tasks is the tradeoff between computational complexity and accuracy. We found empirically that $\epsilon = 2$ and $k = 100$ give a reasonable such tradeoff.

### 6.3.2 Multi-Class Classification by Combining Binary Classifiers

There are three classes we wish to separate, tibial medial cartilage, femoral medial cartilage and background. We combine one binary classifier trained to separate tibial cartilage from the rest and one trained to separate femoral

cartilage from the rest with a rejection threshold ($T$) (Section 4). In one-versus-all classification, which is commonly used for multi-class classification [143], one builds $N$ one vs. rest classifiers and perform a winner-takes-all vote between them, assigning $j$ to the class $\omega_c$ with the highest posterior probability of the classifiers. In the scans, roughly 0.2% of the voxels belong to tibial cartilage and 0.5% to the femoral cartilage, making the background the by far largest class. Our approach is similar to one-versus-all, but due to the dominance of the background class we replace the background vs. rest classifier by a rejection threshold, which states that the posterior probability should be higher than the threshold $T$ before it can be assigned to a cartilage class. The decision rule is:

$$j \in \begin{cases} \omega_{tm}, & p_{tm,j} > p_{fm,j} \quad \text{and} \quad p_{tm,j} > T; \\ \omega_{fm}, & p_{fm,j} > p_{tm,j} \quad \text{and} \quad p_{fm,j} > T; \\ \omega_b & \text{otherwise.} \end{cases} \tag{6.1}$$

where $p_{c,j}$ is short for $p(\omega_c|\mathbf{u}_{c,j})$ ($\mathbf{u}_{c,j}$ is the feature vector for voxel $j$ and classifier $\omega_c$) and the subscripts *tm*, *fm* and *b* stands for *tibial medial*, *femoral medial* cartilage and *background* respectively. The threshold is optimized on the training set to maximize the Dice Similarity Coefficient (DSC), which is considered a useful statistical measure for studying agreement between different segmentations [149].

### 6.3.3 Feature Selection

Feature selection can provide a suitable feature set for the classification task at hand. The features of the classifiers are selected by sequential forward selection followed by sequential backward selection from a large bank of features described below in section 6.3.4 [50]. In the forward selection we start with an empty feature set and expand the search space by adding one feature at the time according to the outcome of a criterion function, the area under the Receiver Operator Characteristics (ROC) curve [70]. The backward selection starts with the features found by the sequential forward selection and iteratively excludes the least significant feature according to the criterion function.

All features are examined in every iteration which means that the same feature can be selected several times, allowing us to establish an indirect weighting of important features. We use 25 scans for the training of the classifier, the same 25 scans are used in the feature selection, threshold selection and for the training data set for the final classifier. We observe a monotonically increase of the outcome of the criterion function and we stop iterating when the feature space is 60 dimensional, which is at a point when the improvement is not significant anymore and the search becomes ineffective due to the increase in computational time, and memory becomes an issue. We perform backward selection

until there are 39 features remaining in the set, and we observe that for these iterations there is no significant decrease in the classifier performance. This feature selection scheme does not guarantee a global optimum, but by doing forward selection followed by backward selection we search a larger number of combinations of features than by only using forward selection.

We combine binary classifiers even though $k$NN is inherently a multi-class classifier. The reason for training a classifier for each foreground class is to find class-dependent features [6], and for allowing for a straight-forward extension to incorporate more cartilage compartments in the classification system. Also, the area under the ROC curve in the feature selection evaluates the classifier performance for all operating points for a two-class task. But there is no obvious extension of ROC analysis for multi-class classification tasks and we have found better results by training and combining binary classifier than we have with direct multi-class classifiers [50].

### 6.3.4 Features

We here introduce the set of candidate features from which the feature selection scheme selects a subset.

The intensity and the position in the image are both features that can be indicators cartilage voxels in a scan, since cartilage voxels are typically brighter than the background and the knee is placed in a similar position for all acquisitions. Both the raw image intensities and intensities from the image convolved with a Gaussian according to the scale space framework [82] on different scales are considered. Three scales are chosen (0.65mm, 1.1mm and 2.5mm) to cover the range of different cartilage thicknesses. Though the location and the shape of the cartilage varies from scan to scan, the coordinates are still an indicator of where cartilage is more likely to be situated.

Other features of interest are those related to the geometry of the object in question. The 3-jet, which consists of all first, second and third order derivatives with respect to $(x, y, z)$, forms a basis which describes all geometric features up to third order [46] and are thus considered as candidate features. The $x$-, $y$- and $z$-axes are here defined as the sagittal-, coronal- and axial axes.

It is well known that numerical differentiation enhances higher spatial frequencies and that the effect increases with the order of the differentiation, meaning that noise may limit the practical use of higher order derivatives. Blom [7] shows that the spatial averaging in scale-space causes a noise reduction that counteracts the noise amplification caused by differentiation. Hence all the derivatives mentioned in this section are achieved by convolution with Gaussian derivatives,

defined as $I_{i_1,\ldots,i_n} = I * \mathcal{D}_{i_1,\ldots,i_n} G(\sigma)$, where $G$ is a Gaussian, $\mathcal{D}$ a differential operator and $\sigma$ is the scale. All features where derivatives or smoothing are involved are examined at the three different scales mentioned above. Though the lowest scale we use (0.65mm) is lower than the resolutions of the scans, there is still some spatial averaging and Gaussian derivatives allow for robust differentiation at that scale.

Eigendecompositions of the Hessian ($\mathbf{H}$) [113],[35] and the structure tensor ($\mathbf{T}$) [136],[81] can provide useful information of the local structure. The eigenvalues and their corresponding eigenvectors of $\mathbf{H}(\sigma)$ and $\mathbf{T}(\sigma_T, \sigma)$ using three different scales on $\sigma_T$ and $\sigma$ are candidate features.

We have features that examine the local first and second order structure in relevant directions. We wish to include a similar feature for the local third order structure as well. The third order derivatives with respect to $(x, y, z)$ can be conveniently represented in the third order tensor $I_{ijk}$. Examining the third order structure in the local gradient direction $(I_x, I_y, I_z)$ can be described using Einstein summation as

$$L_{www} = I_{ijk} I_i I_j I_k / (I_i I_i)^{3/2}.$$

The third order tensor examined in the gradient direction on three different scales are candidate features.

In summary, our candidate features are the intensity, the position, the 3-jet, eigenvalues and eigenvectors of both the Hessian and the structure tensor and the third order tensor in the gradient direction. All features except the position are calculated at three different scales (0.65mm, 1.1mm and 2.5mm), and the scales are in mm instead of number of voxels for handling scans with different resolutions.

All features except the intensity are coupled three by three so that each feature addition results in an equal expansion of the feature space. The three by three grouping comes natural because we have three dimensional images, so the coordinates, first order derivatives and the eigenvalues and eigenvectors of the Hessian and the structure tensor are directly grouped. The other features are grouped using the three scales.

### 6.3.5 Selected Features

After feature selection, the resulting features for the $\omega_t$ vs. rest classifier are (in order of decreasing significance): the position in the image, the intensities smoothed on the three scales, $I_{zz}$ on the three scales, the first order derivatives on all three scales, $I_{zzz}$ on the three scales, the eigenvalues of $\mathbf{H}$(1.1mm),

$I_{yy}$ on all three scales, the eigenvalues of $\mathbf{H}(2.5\text{mm})$ and the eigenvalues of $\mathbf{T}(2.5\text{mm}, 0.65\text{mm})$.

The $\omega_f$ vs. rest classifier contains the following features after feature selection: the position, the eigenvector corresponding to the largest eigenvalue of $\mathbf{T}(1.1\text{mm}, 0.65\text{mm})$, the first order derivatives on scales 1.1mm and 2.5mm, the intensity smoothed on three scales, $I_{zzz}$ on the three scales, $I_{zz}$ on all three scales, the eigenvalues of the Hessian on all three scales, and the eigenvalues of $\mathbf{T}(2.5\text{mm}, 0.65\text{mm})$. For an overview of these features, see Table 4.1 in Chapter 4.

It can be noted that the position is selected as the most significant feature by both classifiers. The intensity smoothed on three scales is also highly ranked by both classifiers. In the training data, the intensity values for cartilage voxels are on average brighter than background voxels. The magnitudes of $I_{zz}$ are a factor 10 times higher for cartilage voxels than background voxels on average, and the values typically negative for cartilage voxels, which signifies an intensity maximum for the axial direction for cartilage. The eigenvalues of the Hessian are selected for both classifiers, on scales 1.1mm and 2.5mm for the $\omega_t$ classifier and on all three scales for the $\omega_f$ classifier. Cartilage voxels typically have a large negative eigenvalue (Figure 6.2), which signifies a maximum for at least one direction. The first order derivatives are selected by both classifiers, on all three scales for the $\omega_t$ classifier and on scales 1.1mm and 2.5mm for the $\omega_f$ classifier. In Figure 6.3, the gradient magnitudes are plotted for the two scales that are common for both classifiers. The eigenvalues of $\mathbf{T}(2.5\text{mm}, 0.65\text{mm})$ are selected by both classifiers, hence the low-scale first order structure is a more useful feature for cartilage when the structure tensor is smoothed on a large scale. In Figures 6.1-6.3, some relevant features for cartilage are illustrated using a sagittal slice of a scan. The original slice with the cartilage delineated, along with smoothed intensities and third order structure are plotted in Figure 6.1. Second order structure is plotted in Figure 6.2, and here the cartilage contours are drawn since the maximum response is inside the cartilage. In Figure 6.3, first order structure is plotted and here the cartilage is not outlined since the maximum response is at the boundaries.

### 6.3.6 Efficient Voxel Classification

Our segmentation method is fully automatic, but due to the high computational complexity of the $k$NN classification it takes approximately 60 minutes to classify all voxels in a scan consisting of around two million voxels by the two binary classifiers. Even though computation power is relatively in-expensive, such long computation times are inconvenient in clinical studies using large numbers of scans.

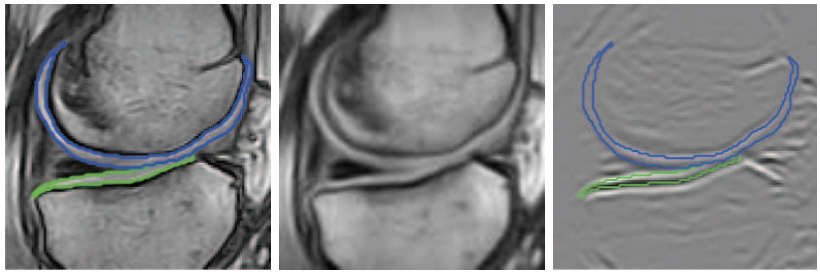


**Figure 6.1:** From left to right: central part of a sagittal slice with manual outline of tibial (green) and femoral (blue) cartilage, intensity smoothed on scale 0.65mm, and $I_{zzz}$ on scale 1.1mm.

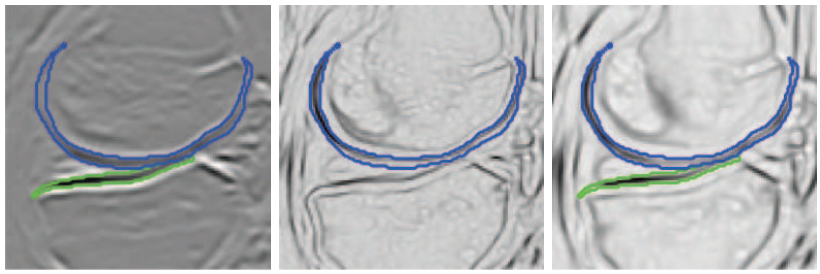


**Figure 6.2:** Second order structure, from left to right: $I_{zz}$ on scale 1.1mm (specific for the $\omega_t$ classifier), smallest eigenvalue of $\mathbf{H}$ on scale 0.65mm (specific for the $\omega_f$ classifier), smallest eigenvalue of $\mathbf{H}$ on scale 1.1mm (typically negative for cartilage).

**Figure 6.3:** First order structure, from left to right: gradient magnitude on scale 1.1mm, gradient magnitude on scale 2.5mm, largest eigenvalue of $\mathbf{T}(2.5\text{mm}, 0.65\text{mm})$.

We have therefore developed an efficient voxel classification algorithm [29], and the basic idea behind it is to not classify all voxels but to focus mainly on the cartilage voxels. The algorithm is conceptually very simple: starting from a set of randomly sampled voxels, we classify them as either cartilage or background. If a voxel is classified as cartilage, we continue with classification of the neighboring voxels and this expansion process continues until no more cartilage voxels are found.

This results in a number of connected regions of cartilage. Provided that our initial sampling of starting voxels hits each cartilage sheet in at least a single voxel, the resulting segmentation will be exactly like the one resulting from a full voxel classification after extraction of the largest connected component. This is ensured by not making the initial random sampling too sparse. Since some parts of the cartilage compartments will be fairly centered in scan we sample fairly densely at the center, with a sampling probability of 5% for each voxel, and gradually more sparsely towards the periphery.

### 6.3.7 Position Adjustment

Besides a large anatomical variation, the placement of the knee in the scanner in clinical studies is a source of variation. Still the position in the scan is a strong cue to the location of cartilage, which is evident in our segmentation method where the position is selected as one of the most significant features. Even though the global location is a strong cue the minor variation in placement is a source of errors. Segmentation methods that rely on manual interaction are usually less sensitive to knee placement since a user can define where in the scan the cartilage is. We however have a segmentation technique that is completely independent of user interaction thus the placement variations that occur in scans in clinical studies is an issue that needs attention.

One way of correcting for knee placement is to manually determine where in the scan the cartilage is, but this can take time with 3D images since a human expert typically search through the scans on a slice-by-slice basis. And when the segmentation method itself is automatic, an automatic adjustment is advantageous.

In order to adjust the segmentation method to become more robust to variations in knee placement we have developed an iterative scheme which consists of two steps that are repeated until convergence (Section 5). The first step consists of shifting the coordinates of the scan so that the cartilage center of mass found from the segmentation is positioned at the location for the center of mass for the cartilage points in the training set. Then in the second step the scan is classified using the sample expand algorithm with the other features unchanged. The

outcome is combined according to (6.1) and the largest connected component is selected as the cartilage segmentation.

The position of the tibial and femoral compartments are shifted individually for the two binary classifiers because the classification depends on the training set, and for the training data the different cartilage compartments have different relative position with respect to each other due to different positioning of the joint for the test subjects.

## 6.4 Results

The average computation time for automatic segmentation of a scan is approximately 10 minutes on a standard desktop 2.8 Ghz PC. For a trained radiologist it takes around two hours to segment the tibial and femoral medial cartilage in a scan with slice-by-slice delineation of the contour by manual selection of boundary points and automatic linear interpolation.

### 6.4.1 Comparison Between Automatic and Manual Segmentations

The methods are trained on 25 scans and evaluated on 114 scans. Of the 114, 31 knees have been re-scanned and the reproducibility is evaluated by comparing the volume and area estimates from the first and second scanning.

Before applying the position adjustment scheme described in section 6.3.7, the automatic segmentation method yields an average sensitivity, specificity and DSC of 81.1%, 99.9%, and 0.79 respectively for the total medial cartilage segmentation, in comparison with manual segmentations.

After applying the automatic position normalization, the average sensitivity, specificity and DSC are 83.9%, 99.9% and 0.80 respectively. The scheme converges after the second iteration. Compared to the initial segmentation there is a significant increase in sensitivity ($p < 1 * 10^{-7}$) and in DSC ($p < 3 * 10^{-3}$) according to a paired t-test. In order to illustrate how the segmentations are affected, the best and worst cases from the position correction scheme are shown in Figures 5.2 and 5.3. In the best case the DSC increases with 0.2 and for the worst scan it decreases with 0.02. The scan with the worst result is from a severely osteoarthritic knee which can be difficult even for a highly trained expert to segment. The results for each compartment is listed in Table 6.1. OA knees are more difficult to segment compared to healthy ones, which can be seen by dividing the test set into healthy and OA groups. The healthy group (KL 0)

has a mean DSC of 0.82, and the OA group (KL $\geq$ 1) has a mean DSC of 0.78. This could improve by including more OA knees in the training data set.

**Table 6.1:** Results from our automatic segmentation method before and after position adjustment (PA) for medial tibial, medial femoral, and the medial compartments together. Sensitivity (Se), specificity (Sp) and DSC are found from comparison with manual segmentations on the 114 scans in the test set. Standard deviations are denoted std.

| Compartment(s) | Se | std | Sp | std | DSC | std |
|---|---|---|---|---|---|---|
| Tibial | 81.1% | ±10.6% | 99.96% | ±0.01% | 0.80 | ±6.7% |
| Tibial PA | 86.8% | ± 7.7% | 99.96% | ±0.01% | 0.81 | ±6.0% |
| Femoral | 77.9% | ±12.8% | 99.92% | ±0.03% | 0.77 | ±8.0% |
| Femoral PA | 80.3% | ±11.6% | 99.91% | ±0.03% | 0.77 | ±8.0% |
| Tibial + Femoral | 81.1% | ±10.9% | 99.88% | ±0.04% | 0.79 | ±6.5% |
| Tibial + Femoral PA | 83.9% | ± 8.4% | 99.87% | ±0.04% | 0.80 | ±5.6% |

When comparing between manual and automatic estimates for the 114 scans, the average pairwise differences for medial volume and area are 8.7% and 0.05% respectively. The volume from the automatic method overestimates the manual with 10% with significant difference between group means ($p = 0.02$) and the area is underestimated by 0.7% with no significant difference. Some of the overestimation of the volume most likely originates from false positives from lateral and patellar cartilage that is adjacent to the medial compartments. Visual inspection supports this, for instance in Figure 5.2 (Chapter 5) it can be seen that the manual segmentation ends more abruptly at the medial/lateral border than the automatic segmentation. Still this remains to be verified statistically in a future study including all compartments. Also, segmentation errors occur along the edges of the cartilage, in particular where the tibial cartilage curves over the tibial plateau. The scans used in this study have low contrast between tissues which may also contribute to segmentation errors.

As for inter-scan reproducibility of the medial cartilage volume from the automatic segmentations, we examine the 31 knees scanned twice within a week. Before position adjustment there is an average absolute volume and area difference of 10% and 6.0% for the total medial cartilage, and after position adjustment the reproducibility of the method is improved, with a decrease of the average absolute volume and area differences to 6.5% and 4.5% respectively. These values can be compared to the reproducibility of the manual segmentation which has an average absolute volume and area difference of 6.5% and 5.5% respectively for the same data set. The reproducibility for the automatic method and human expert for both volume and area for all compartments are listed in Table

6.2, where it can be seen that the tibial volume and area estimates are the most reproducible for the automatic method, possibly because the tibial cartilage has a less complex shape compared to the femoral cartilage. In Figures 6.4 and 6.5 the Bland-Altman plots of inter-scan reproducibility for the automatically obtained tibial volume and area estimates are displayed. The most significant outliers are due to false positives for the femoral class.

The radiologist has a fairly poor precision on volume both tibial and femoral separately, but it improves when the two compartments are combined. This shows that the radiologist is mainly in doubt on the part of the cartilage sheets where tibial and femoral are contacting. These volume precision numbers are lower than what is reported in other studies, something which could be a consequence of the low-field low resolution scans used in this study.

**Table 6.2:** Inter-scan reproducibility of our automatic segmentation method before and after position adjustment (PA) and of the manual segmentations (M), for medial tibial, medial femoral, and the medial compartments together. Linear correlation coefficient (Corr.) and mean absolute pairwise differences (Diff.) for the 31 knees scanned twice.

| Compartment(s) | Corr. Vol | Diff. Vol | Corr. Area | Diff. Area |
|---|---|---|---|---|
| Tibial | 0.82 | 8.8% | 0.88 | 5.9% |
| Tibial PA | 0.91 | 5.8% | 0.95 | 4.3% |
| Tibial M | 0.90 | 10.3% | 0.84 | 9.4% |
| Femoral | 0.90 | 9.9% | 0.91 | 7.0% |
| Femoral PA | 0.94 | 7.4% | 0.93 | 5.3% |
| Femoral M | 0.94 | 8.6% | 0.93 | 6.6% |
| Tibial + Femoral | 0.86 | 10.0% | 0.92 | 6.0% |
| Tibial + Femoral PA | 0.93 | 6.5% | 0.95 | 4.5% |
| Tibial + Femoral M | 0.95 | 6.5% | 0.94 | 5.5% |

The radiologist re-delineated the tibial medial and femoral cartilage in 31 scans in order to determine intra-rater variability for the manual segmentations. The average DSC between the two manual segmentations are 0.87 for the medial cartilage, and the corresponding DSC of the automatic segmentation versus expert for the medial cartilage of the 31 scans is 0.80. which explains the fairly low values of the DSC in our evaluation since one cannot expect to obtain better results given the variability of the gold standard segmentations.

For all the scans the in plane resolution is $0.70 \times 0.70\text{mm}^2$, but the slice distance is either 0.78mm, 0.70mm, 0.94mm or 0.86mm with the first being the most

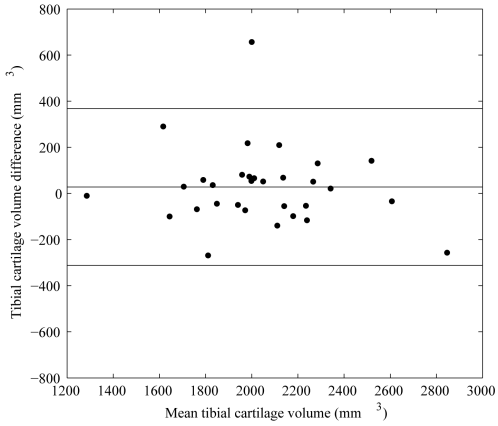



**Figure 6.4:** Bland-Altman plot of the inter-scan reproducibility of the tibial volume from automatic (position adjusted) segmentations. The lines are the mean ±2 std of the difference between measurements.

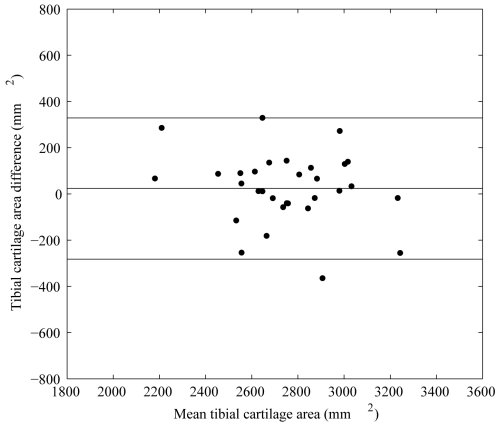



**Figure 6.5:** Bland-Altman plot of the inter-scan reproducibility of the tibial area from automatic (position adjusted) segmentations. The lines are the mean ±2 std of the difference between measurements.

predominant. Of the 25 scans in the training set, 13 scans have slice distance 0.78mm and of the 114 scans in the test set, 72 have that same slice distance. For these 72 scans, the DSC of the medial cartilage compartments is 0.80 ($\pm$0.04) std. For the other resolutions in the test set the average DSC is 0.79 ($\pm$0.07) std. Of these remaining scans, 32 have 0.86mm slice distance, 7 have 0.94mm and 3 have 0.70mm.

### 6.4.2 Relation Between the Volume and Area Estimate and Disease

Typical quantitative disease markers for OA is the articular cartilage volume, thickness and surface area, and several studies have been dedicated to evaluation of them [14],[123],[65]. In this study we evaluate the volume and surface area estimates obtained directly from the automatic segmentation. The volume estimate is directly obtainable by summing all voxels classified as cartilage, and an estimate for the surface area is obtained by creating an isosurface using a smoothed version of the binary segmentation. But a voxel based method alone does not allow for morphometric quantification, and for measuring the thickness we fit a deformable shape model to the cartilage so that thickness can be measured through the normal direction of the cartilage surface at anatomical well defined locations. This is however not within the scope of this paper, for thickness measurements of the data set, see [30].

We examine the ability to separate healthy from osteoarthritic populations of the volume and area estimates using unpaired Student's t-tests. The results are displayed in Tables 6.3 and 6.4, and since knees with KL 1 are borderline cases we evaluate populations both by including these cases to the healthy population and to the OA population. It can be seen that for the volume estimate the most confident separations occurs for tibial cartilage, and for the area estimate statistical significant separation is obtainable only from tibial cartilage.

Since our test subjects come in all shapes and sizes, we normalize the volume by the width of the tibial plateau cubed and the surface area by the tibial plateau width squared. In Figures 6.6 and 6.7 the mean normalized volume and surface area estimates for medial tibial and femoral cartilage together are plotted for the different KL populations.

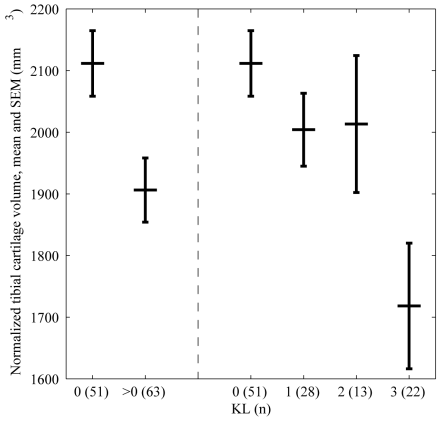



**Figure 6.6:** The separation between different OA populations using the KL and the normalized tibial medial cartilage volume from automatic (position adjusted) segmentations.

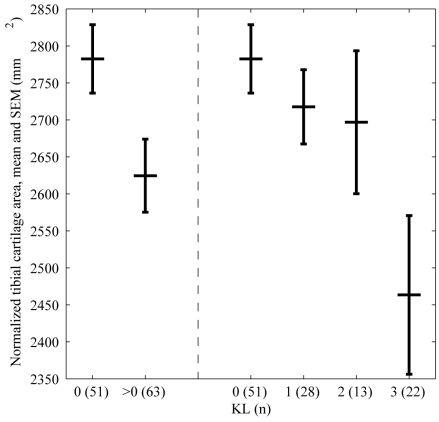



**Figure 6.7:** The separation between different OA populations using the KL and the normalized tibial medial cartilage surface area from automatic (position adjusted) segmentations.

**Table 6.3:** P-values for t-tests of separating groups using the volume estimates. P1 is the p-value for separation of healthy (KL 0) from borderline to OA (KL $> 0$), and P2 is separation of healthy and borderline (KL $\leq 1$) from clear OA cases (KL $> 1$). M stands for manual segmentations and PA are values from automatic segmentation after position adjustment.

| Compartment(s) | P1 | P2 |
|---|---|---|
| Tibial PA | 0.0071 | 0.0027 |
| Tibial M | 0.0016 | 0.00029 |
| Femoral PA | 0.095 | 0.030 |
| Femoral M | 0.35 | 0.80 |
| Tibial + Femoral PA | 0.057 | 0.029 |
| Tibial + Femoral M | 0.094 | 0.20 |

**Table 6.4:** P-values for t-tests of separating groups using the area estimates. P1 is the p-value for separation of healthy (KL 0) from borderline to OA (KL $> 0$), and P2 is separation of healthy and borderline (KL $\leq 1$) from clear OA cases (KL $> 1$). M stands for manual segmentations and PA are values from automatic segmentation after position adjustment.

| Compartment(s) | P1 | P2 |
|---|---|---|
| Tibial PA | 0.024 | 0.005 |
| Tibial M | 0.00011 | 0.0000004 |
| Femoral PA | 0.68 | 0.29 |
| Femoral M | 0.85 | 0.37 |
| Tibial + Femoral PA | 0.32 | 0.099 |
| Tibial + Femoral M | 0.22 | 0.30 |

## 6.5 Discussion

In this paper, we have presented a fully automatic framework for segmentation and quantitative evaluation of the articular cartilage in the knee. This is to our knowledge the only fully automatic cartilage segmentation method that has high precision and agreement with manual segmentations and is evaluated on a fairly large data set (139 scans) consisting of both healthy and osteoarthritic test subjects.

Robustness against the inevitable problem of changes in test subject placement in the scanners is obtained with an iterative scheme, which results in low inter-scan variability of the cartilage estimates.

The medial tibial cartilage gives the the best inter-scan reproducibility with mean absolute difference of 5.8% and 4.3% for the volume and area estimates, and separation between population with p-values of 0.003 and 0.005 for separation between healthy/borderline OA and clear OA populations for volume and area respectively.

Fat suppression and high-field magnets significantly improve image quality with better contrast between tissues and higher resolution. Since our method compares well to manual segmentations using the lower quality images from a low-field scanner, we can hope that the method will perform at least as well on high-field fat suppressed MRI, assuming we would have access to a similar amount of training data. Future work will involve evaluating the method on high-field data. Our segmentation method can handle images with somewhat different resolution, however it is possible and remains to be investigated if features present at higher resolutions can advance the results.

By using two-class classifiers we find class-specific features, and not only is the problem of finding a criterion function for multi-class classification is avoided, but it also establishes a framework for multi-class classification that in the future can be extended to incorporate all cartilage compartments by incorporating two-class classifiers trained separately for the remaining compartments.

Our method is trained and evaluated on low-field MRI, and even though there is no well established accuracy validation for low-field MRI, we show that statistically significant differences between healthy and osteoarthritic populations are detectable using our cartilage volume and area estimates. This suggests that our method combined with low-field MRI data may be useful in clinical studies, particularly multi-center clinical studies since the method is completely automatic, has high reproducibility, and is robust to changes in knee placement in scanner.

Chapter 7

# Unifying Statistical Classification and Geodesic Active Regions for Segmentation of Cardiac MRI

*Jenny Folkesson, Eigil Samset, Raymond Y. Kwong and Carl-Fredrik Westin*

**Abstract**

This paper presents a segmentation method that extends geodesic active region methods by the incorporation of a statistical classifier trained using feature selection. The classifier provides class probability maps based on class representative local features, and the geodesic active region formulation enables partitioning of the image according to the region information. We demonstrate automatic segmentation results of the myocardium in cardiac late gadolinium enhanced magnetic resonance imaging (CE-MRI) data using coupled level set curve evolutions, in which the classifier is incorporated both from a region term, and from a shape term from particle filtering. The results show potential for clinical studies of scar tissue in late CE-MRI data.

## 7.1 Introduction

Atrial fibrillation is a disorder found in about 2.2 million Americans, and about 15 percent of strokes occur in people with the disorder. Ventricular arrhythmias are the major cause of the more than 350,000 sudden cardiac deaths annually in the U.S. [148]. Catheter ablation of paroxysmal atrial fibrillation has emerged as a first line effective therapy which has the potential to prevent arrhythmia recurrences and improve quality of life. With increasing clinical acceptance and experience, catheter ablation of ventricular arrythmia has been increasingly used and can potentially reduce the need for implanted defibrillators.

Catheter ablation involves the selective destruction of the cardiac tissue responsible for the initiation and maintenance of arrhythmias via percutaneously placed catheters. Currently, cardiac and peripheral vascular interventions are typically guided by fluoroscopic x-ray. This imaging modality yields only limited anatomic information, uses ionizing radiation, and iodine based contrast agents are necessary to visualize the vessels. Use of late gadolinium enhanced magnetic resonance imaging (CE-MRI) in such procedures eliminates the concerns of both exposure and contrast, and enables visualization of areas of myocardial scar that contain the arrhythmia substrate [117]. Delineations of scar tissue can be useful both in pre-operative MRI registered to intra-operative imaging during ablation, and in clinical studies of myocardial scar and its impact on cardiac disorders [84].

With sufficient tissue contrast and signal to noise ratio of the current CE-MRI technique, accurate intensity-based delineation of the scar extent is possible if myocardial segmentation can be accurately performed. Currently, clinical analysis of this type of data relies on manual labor and segmentation of cardiac CE-MRI data is considered non-trivial [116], hence significant inter- and intra observer variability can be expected. Automating the myocardium segmentation process not only eliminates observer variability, it also relieves the cardiologists from time-consuming labor.

For an automated segmentation of the myocardium in CE-MRI there are a number of challenges to consider besides low image contrast and noise. Irregular appearance can be found both within the myocardium where scar tissue can occur at random locations, and inside the left ventricle due to the papillary muscles. Scar tissue and protruding papillary muscles can also be a cause for missing or ambiguous boundary information between the myocardium and surrounding tissues.

Voxel based statistical classification can deal with non-Gaussian intensity distributions, but lacks global boundary or shape information. It has been shown that adding global information can improve statistical classification results [135],[32].

Level set methods are useful tools in image segmentation because they can change topology, and geodesic active contours enables evolution towards image boundary information while preserving regularity [99]. They have been extended to an elegant geodesic active region formalism [102] which enables a partitioning of the image according to given region information.

Methods providing prior global shape or appearance information can be helpful when dealing with ambiguous boundary information [129]. However, global appearance models cannot capture intensity variations at random locations inside objects. Active appearance models [24] utilize a linear model of appearance by applying principal component analysis (PCA) on the intensities inside an object, something that can produce unreliable results if the intensities are not Gaussian distributed. For application to data with irregular appearance, regional statistics need to account for local and non-linear intensity variations [33].

### 7.1.1 Related Work

In the literature, there are a number of shape models reported to perform well in image segmentation tasks, however we limit the discussion to level set shape representations due to our level set framework. Leventon *et al.* [85] capture shape variations by applying PCA directly on distance function representations of objects in training data. Linear combinations of distance maps are typically not distance maps, and scale and rotation are handled implicitly in the shape model. Charpiat *et al.* [18] have generalized the distance function shape representation and modeling to a non-linear setting. Tsai *et al.* [129] extended Leventon's model to estimate shapes with the pose parameters explicitly, where the model parameters are optimized via gradient descent using regional intensity statistics. The method does not require point correspondences during training, and the region-based optimization framework is quite general and could be be extended to incorporate other features than intensity. The method has been extended to a mutual information driven shape estimation, which makes no explicit assumptions on the class intensity distributions by estimating them using Parzen windowing [130]. However, PCA-based shape inference allows only combinations of shape variations found in the training data and cannot capture previously unseen fine details in a test image. Huang *et al.* use a distance map shape representation, and employ PCA to capture variations in shape and appearance in their registration-based segmentation which efficiently segments tagged cardiac MRI. However, a global linear model of appearance may be unreliable if the intensities are not normally distributed, and the model fit may be

impaired by local intensity variations.

Rousson and Paragios [112] have developed a level set based shape model by finding a mean distance function from a set of aligned training distance maps in a maximum-likelihood sense. This mean shape is constrained to be a level set representation, and the variability at each grid location is taken into account. A level set function is then evolved that becomes optimal when it is a rigid transformation of the prior shape model. Paragios *et al.* [103] have also developed a geodesic active region method for myocardium segmentation with a region term derived by assuming separable Gaussian class density distributions. Hernandez *et al.* [64] extended the method to non-parametric density function estimation. Cremers *et al.* [28] adopted the distance function shape representation and introduced a dynamic labeling scheme in a Chan-Vese segmentation framework which handles occlusions but assumes intensity homogeneity.

Pichon *et al.* [107] developed a level set segmentation method in which the contour grows towards a maximum a posteriori region. Parzen windows are employed for unsupervised estimation of intensity probability density functions. Parzen windowing is not restricted to normal intensity distributions, however the method does not include shape information, and is semi-automatic. It is restricted to region growing, which might not be ideal when coupled contours are interacting.

Zeng *et al.* segment brain MRI data using level sets evolving towards locations with high probabilities of being tissue boundaries assuming Gaussian within tissue intensity distributions combined with a strong coupling force. [144].

### 7.1.2 Overview of the Presented Work

We propose to incorporate supervised learning in terms of a statistical classifier, which makes no explicit assumptions on underlying class density functions, into a geodesic active region framework in order to deal with irregular appearance. The numerics for the two methods fit well together since all calculations are done directly at pixel level. We use a $k$NN classification framework trained by multi-class feature selection [55] in order to find suitable features for discrimination between classes. Candidate features include position, intensity smoothed on different scales, and local geometric features, as described in sections 7.2.1 and 7.2.2.

The myocardium is segmented using two coupled level sets which constrains the endo- and epicardium to remain within a certain distance relative to each other (section 7.2.5). A geodesic active contour term is included in order to account for boundary information (section 7.2.3).

Due to boundary information ambiguities in the data we incorporate a shape term. We use a distance function shape representation which fits naturally into our level set framework (section 7.2.4). Using a signed distance function as a shape representation could provide some tolerance to slight misalignments of the training shapes since slightly misaligned pixels in a distance map are generally correlated. This can be seen as an attempt to avoid having to solve the general correspondence problem [85], and can be advantageous for objects without well defined anatomical landmarks, as is the case for many objects including the myocardium. Shape is inferred by shape particle filtering [33], which like the region term it is based on the class probability maps from classification.

For automatic myocardium segmentation we initialize two level set representations from the shape estimates. Since these are rigid transformations of a mean shape we allow the contours to deform locally by evolving the coupled level sets using regional information from the classification while still respecting boundary information and global shape.

## 7.2 The Segmentation Method

### 7.2.1 Statistical Pattern Classification

In voxel based statistical classification, each voxel is represented by $d$ features in a $d$-dimensional space and the objective is to establish decision boundaries between different classes in this feature space. Supervised methods typically establish boundaries using manually labeled training data. The $k$ Nearest Neighbor ($k$NN) classifier determines class boundaries from the distance to the $k$ nearest neighboring training data points in feature space, hence makes no assumptions on class distribution functions.

Each pixel $j$ is described by a feature vector $\mathbf{u}_j$, and the posterior probability of being class $\omega_c$ is

$$P(\omega_c|\mathbf{u}_j) = \frac{k_c}{k},$$

where $k_c$ is the number of training points of the $k$ nearest neighbors with class label $\omega_c$. The myocardium segmentation is a three class classification problem; with the myocardium ($myo$), left ventricle ($lv$), and background ($b$). We use an approximate $k$NN classifier [4], which allows for faster computations if an error in the distance calculations is tolerated. The classifier is trained on images throughout the heart from different persons with different positioning in the scanner and different degree of scar tissue in order to cope with these variations that can occur in a test image.

### 7.2.2 Feature Selection

Feature selection can not only increase computational efficiency and reduce memory usage, but also make the classification less vulnerable to the curse of dimensionality. In sequential forward selection (SFS), an initially empty feature set is expanded by iteratively adding a feature from a bank of candidate features according to the outcome of a criterion function. We apply sequential backward selection (SBS) on the features found by the SFS by iteratively excluding the least significant feature. The SFS is repeated until the classifier performance clearly has peaked, the SBS is then iterated until there is a clear performance degradation. All features are examined in every iteration, so the same feature can be selected several times which allows for an indirect weighting of important features [55]. This scheme does not guarantee a global optimum, but SFS followed by SBS examines a larger amount of possible combinations of features than *e.g.* by SFS alone.

The feature selection is performed on the training data, and we use the Dice Similarity coefficient (DSC) [36] as criterion function between the classification outcome (labeled by assigning pixels to the most probable class) and manually delineated training data. For a two-class classification system, the classifier can be evaluated using the area under the Receiver operating characteristics (ROC) curve [70], which evaluates the performance for all possible thresholds of the classifier. However, the myocardium is a three-class classification problem and there are no direct extensions of the ROC for more than two classes [55]. Also, the classifier does not need to be thresholded since both the geodesic active region term and the particle filtering use the probability map, hence soft labels, and not hard labels by thresholding. DSC measures the spatial volume overlap between two different segmentations and is defined as $\mathrm{DSC}(A, B) = \frac{2\times|A\cap B|}{|A|+|B|}$ for segmentations $A$ and $B$.

The feature bank consists of the 3-jet, which consists of all first, second and third order derivatives with respect to $(x, y, z)$, and forms a basis which describes all geometric features up to third order [46]. Numerical differentiation can enhance higher spatial frequencies, an effect which increases with the order of the differentiation. Blom [7] shows that the spatial averaging in Gaussian scale-space derivatives causes a noise reduction that counteracts the noise amplification caused by differentiation. Hence all the derivatives mentioned in this work are Gaussian derivatives. Other candidate features are the intensity (raw and Gaussian smoothed on different scales), the position, and eigenvectors and eigenvalues of the Hessian and the structure tensor [136]. All features that involve smoothing or differentiation are evaluated on all possible combinations of six scales: 0.9, 1.5, 2.5, 3.5, 5 and 8 pixels, chosen to cover characteristics of the left ventricle and myocardium. The Hessian describes the local second-order

structure:

$$\mathcal{H}(\sigma) = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix},$$

The structure tensor [136] examines the local first-order structure:

$$\mathcal{T}(\sigma, \sigma_T) = G_{\sigma_T} * \begin{pmatrix} I_x I_x & I_x I_y \\ I_y I_x & I_y I_y \end{pmatrix}.$$

### 7.2.3 Statistical Classification in Geodesic Active Regions

Caselles *et al.* [15] and Kichenassamy *et al.* [77] derived the geodesic active contour formulation for image segmentation, which is a geometric alternative of snakes [72]. A common approach to implementing the motion equation is level set methods [99], where a curve is represented implicitly by a higher-dimensional function $\Phi$.

Contour- and region-based energetic modules for evolving interfaces can be implemented using the following definitions of approximations of the Dirac and Heaviside distributions [145]:

$$\delta_\alpha(\Phi) = \begin{cases} 0, & |\Phi| < \alpha; \\ \frac{1}{2\alpha}(1 + \cos(\frac{\pi\Phi}{\alpha})), & |\Phi| > \alpha \end{cases}$$

$$H_\alpha(\Phi) = \begin{cases} 0, & \Phi > \alpha; \\ 1, & \Phi < -\alpha \\ \frac{1}{2}(1 + \frac{\Phi}{\alpha} + \frac{1}{\pi}\sin(\frac{\pi\Phi}{\alpha})), & |\Phi| < \alpha \end{cases}$$

In these equations, $\alpha$ is the region in which the distributions are approximated, and $\Phi$ is assumed to be negative inside the contour it represents. Using these definitions, Rousson and Paragios [112] define the following active contour term:

$$\frac{d}{dt}\Phi = \delta_\alpha \mathrm{div}\big(g(|\nabla I|)\frac{\nabla\Phi}{|\nabla\Phi|}\big). \tag{7.1}$$

Assuming the class probabilities from classification are conditionally independent, we seek curve evolutions which maximize the likelihood for each region. This is achieved using the geodesic active regions term described by Paragios *et al.* [103], by minimizing

$$E_(\Phi_I, \Phi_O) = -\iint_\Omega H_\alpha(\Phi_O)(1 - H_\alpha(\Phi_I))\log(P_{myo})$$
$$-\iint_\Omega H_\alpha(\Phi_I)\log(P_{lv}) - \iint_\Omega (1 - H_\alpha(\Phi_O))\log(P_b), j \in \Omega,$$

where $\Phi_I$ and $\Phi_O$ represent the endo- and epicardium, and $P_i$ are the class probability density functions.

Using calculus of variations and Greens theorem, the authors in [147] derived functional derivatives of integrals along contours and over regions. Modifying the geodesic active regions term in [103] with probability density functions estimated by the $k$NN classifier results in the following level set evolution equations for the inner and outer contours:

$$\frac{d}{dt}\Phi_I = -\delta_\alpha(\Phi_I)\log\frac{p(\omega_{lv}|\mathbf{u}_j)}{p(\omega_{myo}|\mathbf{u}_j)}|\nabla\Phi_I|, \tag{7.2}$$

$$\frac{d}{dt}\Phi_O = -\delta_\alpha(\Phi_O)\log\frac{p(\omega_{myo}|\mathbf{u}_j)}{p(\omega_b|\mathbf{u}_j)}|\nabla\Phi_O|. \tag{7.3}$$

These are adaptive inflationary forces that aim to shrink or expand the contours according to the classification outcome towards maximum a posteriori regions. These functions assumes that the contours do not overlap, something which in our framework is made highly unlikely because of the coupling and shape terms.

### 7.2.4 Shape Estimation

We use a distance function shape representation and a shape model which accounts for local variations by assuming a mean shape, $\Phi_M$ with local degrees of shape variability $\sigma_M$. The distribution in each pixel is assumed to be Gaussian, and the mean shape is estimated from a set of aligned training shapes in a maximum likelihood sense under the constraint that the mean shape remains a signed distance function [112].

The training shapes are aligned by finding a rigid transform ($\mathbf{A} = T, \theta, s$) with respect to translation, rotation, and scale. The optimization criterion is the sum of squared differences between the source shape $\mathcal{S}$ and a target shape $\mathcal{T}$:

$$E(\mathbf{A}) = \iint_\Omega \big(\Phi_\mathcal{T}(x,y) - \Phi_\mathcal{S}(\mathbf{A}(x,y))\big)^2 dxdy,$$

and shape alignment can be achieved using gradient descent by keeping the initial pose of one shape and align the remaining shapes to it. For more details, see [112] and references therein. We constrain the shapes for the epi- and endocardium in an image to have the same pose parameters since they are closely connected.

#### 7.2.4.1 Shape Particle Filtering

In a test image, we wish to make region-based shape inference based on the class probability maps. This can be realized in a variety of ways, we use shape particle filtering as in [33] since the method has demonstrated accurate results for shape inference based on $k$NN classification. The particle filtering does not easily get trapped in local maxima nor does it need initialization close to optimal solution. We use shape particle filtering to determine the best transformation of the mean shape model given the appearance model, which in our case is the $k$NN classification. A set of $f = 1, \ldots, N$ myocardium shape hypotheses (particles) are sampled randomly relative to translation, rotation, and scale variations in the training data. Each hypothesis is associated with an image labeling which is compared to the label probability maps from classification. Particles are weighted by these likelihood terms for each region,

$$W = \exp\big[\frac{c}{m}\sum_{j=1}^{m}\log p(\mathbf{u}_j|\omega_c)\big],$$

where $c$ is a constant controlling the randomness of the process and $m$ is the number of pixels inside the template [33]. A new set of $N$ hypotheses are generated from the current set by random sampling proportional to the weights $W_f$ and random perturbation of duplicate particles, so that successful particles will multiply while unlikely shapes will vanish. This resampling is repeated until convergence to the maximum likelihood solution, which is when the change of strongest local mode of the particle distribution becomes negligible.

Particle filtering is employed to find a rigid transformation of the mean shape model (annotated $\Phi_{M,i}(A_i)$) given the classification of a test image. In the evolution we include a shape term which evolves towards the inferred shape weighted by the shape variability, hence the shape term is less influential in locations with large prior shape variability [112]:

$$\frac{d}{dt}\Phi_i = -H_\alpha(\Phi_i)\frac{\Phi_i - \Phi_{M,i}(A_i)}{\sigma^2_{M,i}(A_i)}. \tag{7.4}$$

### 7.2.5 Coupling Force

The distance between the epi- and endocardial contours is approximately constant, therefore we introduce a coupling force that respects this anatomical constraint. The distance between the contours can be found with little computational expense from the distance functions. In [144], a coupling force is designed to slow down and eventually stop the evolution if the contours move too far away from each other given a pre-defined allowed distance. In [103], the

coupling forces are plus/minus one when distances between contours are outside the allowed range and zero otherwise.

We design a coupling force that takes advantage of prior information by calculating the distance between the epi- and endocardial contours for all contour points in the training data. We let the coupling force, $h$, be inactive in the interval of the average distance plus/minus half a standard deviation. Outside that range the magnitude of the force increases as a Gaussian function with variance proportional to that of the training data, as can be seen in Figure 7.1. The coupling terms for the inner and outer contour are described by

$$\frac{d}{dt}\Phi_I = \delta_\alpha(\Phi_I)(-h(-\Phi_O))|\nabla\Phi_I| \tag{7.5}$$

$$\frac{d}{dt}\Phi_O = \delta_\alpha(\Phi_O)(h(\Phi_I))|\nabla\Phi_O|. \tag{7.6}$$

When the distance between the endo- and epicardial contours are within the normal range the coupling terms are passive, and outside that range these forces will act as attraction/repellation between the contours when they are too far away/close to each other. In total, the curve evolution criterion for the inner

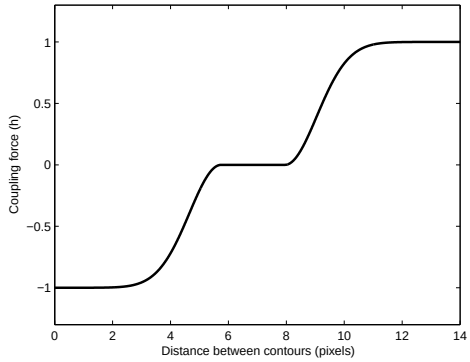



**Figure 7.1:** The coupling force $h$ for the endocardiac and epicardiac contours.

contour is

$$\frac{d}{dt}\Phi_I = \lambda_1\delta_\alpha \mathrm{div}\big(g(|\nabla I|)\frac{\nabla\Phi}{|\nabla\Phi|}\big) - \lambda_2\delta_\alpha(\Phi_I)\log\frac{p(\omega_{lv}|\mathbf{u})}{p(\omega_{myo}|\mathbf{u})}|\nabla\Phi_I| \\ -\lambda_3 H_\alpha(\Phi_I)\frac{\Phi_I - \Phi_{M,I}(A)}{\sigma^2_{M,I}(A)} + \lambda_4\delta_\alpha(\Phi_I)(-h(-\Phi_O))|\nabla\Phi_I|,$$

where the terms on the right hand side are the boundary term, region term from classification, shape from classification, and coupling force. The evolution for

the outer contour can be found in the same manner from equations (8.3), (7.3), (8.5) and (7.6).

The contours are initialized from the shape term, which is restricted to a rigid transformation of the mean shape from the training phase. These initial contours are typically fairly close to convergence, and the evolution will adjust the contours for local boundary and region information while preserving contour regularity and respecting inferred global shape.

## 7.3 Experimental Results

### 7.3.1 Data Set

CE-MRI short axis data was acquired as in [84] using a 1.5T CMR system (Signa CV/*i*, GE Healthcare, USA), with LGE imaging ($T_R/T_E$ 4.8/1.3ms, $T_I$ $200-300$ms) for myocardial scar. An inversion-recovery sequence for LGE was used starting 15 minutes after cumulative 0.15mmol/Kg dose of Gadolinium-DPTA. The in-plane resolution $1.5\times$ 2mm$^2$ with thickness of 8mm. Epicardial and endocardial boundaries were manually outlined. Example slices and manual segmentations can be seen in Figures 7.2 and 7.3. The data set includes 11 patients and after removing three slides on each side it consists of 57 slices from 7 patients for training the method and 30 slices from 4 patients for evaluation.

### 7.3.2 Selected Features

After feature selection on the training data the feature set consists of the following features in decreasing significance: the position, the intensity smoothed on scales 5 and 8, $I_y$ on scales 5 and 8 (selected twice), $I_{yy}$ on scales 5 and 8, $I_x$ on scales 5 and 8, eigenvalues of the structure tensor, $\mathbf{T}(8,8)$, eigenvalues of the Hessian ($\mathbf{H}(3.5)$ and $\mathbf{H}(8)$), $I_{yyy}$ on scales 5 and 8, $I_{xy}$ on scales 2.5 and 3.5, eigenvalues of $\mathbf{T}(0.9,1.5)$, $\mathbf{T}(2.5,0.9)$, $\mathbf{T}(1.5,0.9)$ (selected trice), $\mathbf{T}(5,1.5)$, and $\mathbf{T}(8,5)$, $I_{xyy}$ on scales 5 and 8, eigenvalues of $\mathbf{T}(5,0.9)$, and $\mathbf{T}(3.5,0.9)$. It can be noted that the original image intensity information is never selected as a feature, and that among the most significant features, both first- and second-order structure on high scales are predominant. Eigenvalues of the structure tensor are selected several times hence smoothed low-scale first-order structure is useful for discrimination between classes. The eigenvectors are not selected, which could be related to the changes in gradient direction around the myocardium contours. The features are normalized by linear scaling to unit range prior to classification.

### 7.3.3 Segmentation Evaluation

The automatic segmentation method is evaluated by comparison with manual segmentation by cardiologists. Catheter ablation involves detecting the heart wall inside the left ventricle, therefore we evaluate a distance measure; the mean point to curve distance. In clinical studies, the percent area of the myocardium covered by scar tissue is a biomarker for coronary artery disease, and therefore we also evaluate volume overlap using DSC. Example images along with manual and automatic segmentations and classification results can be seen in Figures 7.2 and 7.3.

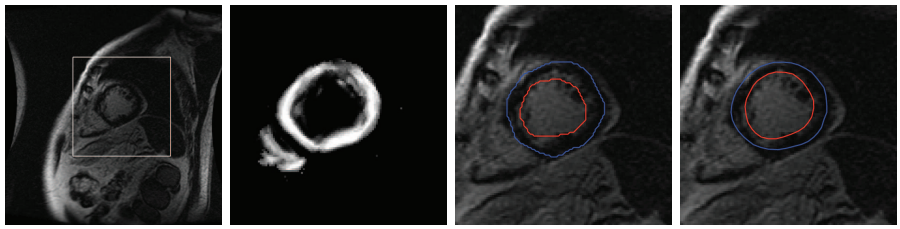


**Figure 7.2:** Example segmentation of the myocardium by the endo- (red) and epi- (blue) cardial contours. From left to right; input image where the box indicates the close-up in the remaining images, probability map for myocardium from classification, manual delineation by cardiologists, and automatic segmentation result.

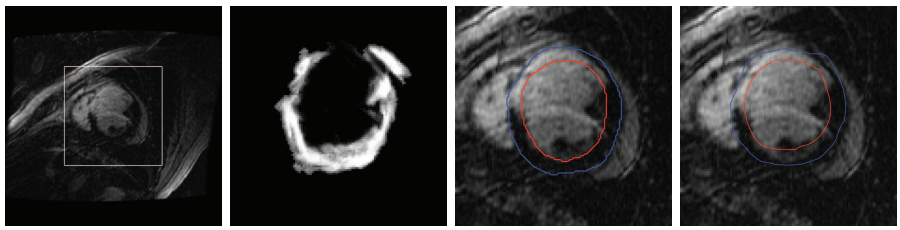


**Figure 7.3:** Example segmentation. From left to right; input image, probability map for myocardium from classification, manual delineation by cardiologists, and automatic segmentation result.

The segmentation failed to converge for 2 of the 30 test images due to inferior image quality and the extent to which scar tissue covers the myocardium. This could be the result of the limited amount of training data and the large appearance variations in the data. For the remaining 28 images the results are presented in Table 7.1 for the the classification, shape estimation, and the automatic segmentation. As can be seen in the figures and the table, the classification alone is not sufficient for a reliable segmentation. The incorporation of

the classification based shape estimate clearly improves the performance, and is used as the initialization for the curve evolution as in Equation (7.7). The shape is a rigid transformation of the mean shape, so local adjustments of the shape to the image information needs to be incorporated. The evolution allows for such local adaptation of the mean shapes given region and boundary image information. Once the evolution has converged there is a significant improvement compared to the shape estimate both in terms of DSC ($p = 0.001$) and distance to contour $d$ ($p = 0.02$) according to a paired t-test.

**Table 7.1:** The results of the segmentation compared to manual segmentations using sensitivity, specificity, DSC for the myocardium, and point distance to curve (pixels) for the left ventricle. Numbers in parenthesis are the standard deviation (std).

| | Classification | Shape estimate | Segmentation |
|---|---|---|---|
| Sensitivity | 70.2% (0.10%) | 70.9% (0.08%) | 75.1% (0.07%) |
| Specificity | 99.1% (0.003%) | 99.7% (0.002%) | 99.7% (0.002%) |
| DSC | 0.64 (0.11) | 0.76 (0.08) | 0.79 (0.07) |
| Distance | 9.23 (6.98) | 1.52 (0.51) | 1.44 (0.54) |

In order to evaluate the influence of the region term on the final segmentation, we ran the experiments the exact same way only without this term. This resulted in a significant decrease of DSC to 0.77 and increase in distance to contour, $d = 1.48$. Leaving out the boundary term lead to a small performance drop for the DSC (DSC = 0.78, $d = 1.44$). Leaving out the coupling force also lead to a small performance deterioration (DSC = 0.78, $d = 1.49$). Leaving out the shape term lead to the most striking performance degradation with DSC = 0.71, and $d = 2.84$, however this is still an improvement compared to the classification on its own. Hence the shape term is the most important term, which is most likely a product of the missing or ambiguous boundary information in the data, followed by the region term. The statistical classification contributes greatly to the results since it is incorporated both in the region term, and in the region-based shape inference.

Selecting the parameters $\lambda_1, \ldots, \lambda_4$ is an issue. Based on experiments, the shape is the most important term due to the ambiguous boundary information in the data, we set $\lambda_3 = 0.2$. The region term is also important, $\lambda_2 = 0.15$. The coupling force is a soft to hard constraint, $\lambda_4 = 0.25$, and the boundary term contributes to some extent to the performance when there is strong gradients close to the contours, $\lambda_1 = 0.25$. For the classifier $k = 50$, and for the particle filtering $N = 2000$.

## 7.4 Summary and Discussion

Spreeuwers and Breeuwer [121] segmented the myocardium semi-automatically using coupled active contours guided by intensity profiles in the normal direction of each node. They report a point to contour distance of 1.5 pixels and an intra-observer variability of 1 pixel, evaluated on 8 slices from 2 test subjects in cine short-axis MRI data. De Bruijne and Nielsen [33] have presented an automatic segmentation method using shape particle filtering based on point distribution models (PDM's) which require point correspondences during the training phase. The method is evaluated on a data set consisting of 14 short-axis end-diastolic cardiac MRI slices with manually placed landmarks on the epi- and endocardial contour, with similar results (mean distance to contour $1.1 \pm 0.3$ pixels). Stegmann *et al.* [125] use an active appearance model approach to segment perfusion cardiac MRI in 4D, and demonstrate results of $1.3 \pm 0.4$ pixels mean point to curve distance on the left and right ventricle combined. Our results in terms of distance to contour ($1.4 \pm 0.5$ pixels) are comparable to the results in [121] though our method is fully automatic. Stegmann *et al.* achieve somewhat better results, however our method does not require point correspondence during training. The relatively higher mean distance of our method compared to [33] can be related to the ambiguous image information, in particular the enhanced scar tissue, in CE-MRI data, making the images difficult to delineate even for a trained human expert. Though we have no inter- or intra-rater data, it is likely that there are variations in these delineations. The evaluated data set shows good agreement with gold standard segmentations, and future work involves evaluating the potential of the fraction of the myocardium covered by scar tissue obtained by the automatic algorithm as a potential predictive marker for coronary artery disease on a larger data set, as is done by manual analysis in [84].

Due to the large distance between slices we focus on a 2D description of the segmentation method in this paper, but the method could be extended to 3D which would be of interest in this application in order to use information from neighboring slices. The method may be adapted to other image segmentation problems provided access to a sufficient amount of labeled training data.

In this work we use particle filtering for shape inference based on probability maps from classification, which demonstrates good performance and is independent of initialization. But any region-based shape inference method could be incorporated, for instance it could be interesting to compare the performance of a gradient descent approach such as in [130], which could be adjusted to incorporate the classification into the regional statistics.

If the myocardium is in a different position than described in the training set, the location feature might not contribute to the classifier performance. In or-

der to adjust the segmentation method to become more robust to variations in myocardium location, the classification could be repeated after shape inference using the location of the inferred shape for an adjustment of the position feature which might improve the classification. Still, the position in the image is determined to be an important feature based on the feature selection which is performed on unaligned images, so the unaligned position in the image is a strong cue to myocardium location.

In summary, we have presented a fully automatic segmentation method that unifies statistical pattern classification and geodesic active regions, by a region term and a shape estimate that are both based on $k$NN classification. The level set shape representation does not require point correspondences during the training phase and makes the registration of the training shapes less sensitive to misalignments. The feature selection ensures that suitable features given the training data are employed in the classification, and makes the method able to cope with non-linear and local appearance variations. The method also handles missing or ambiguous boundary information due to the shape estimation. This suggests that our automatic segmentation method may become a useful tool in image-guided interventions and clinical studies of myocardial scar using late CE-MRI.

## Acknowledgements

The authors would like to thank Marleen de Bruijne at University of Copenhagen for sharing code and insights on shape particle filtering. Thanks also goes to Ole Fogh Olsen at the IT University of Copenhagen for helpful discussions.

Chapter 8

# Spatially Varying Classification with Localization Certainty in Level Set Segmentation

*Jenny Folkesson and Carl-Fredrik Westin*

**Abstract**

We introduce a segmentation framework which extends spatially varying classification to not only incorporate spatial localization from shape estimation, but to also encode certainty of the localization by local shape variability. The method iterates between a classification step where a statistical classifier trained by feature selection is extended with localization features, and a shape estimation step where, given the class probability maps, shape is inferred by particle filtering using a level set shape model that accounts for local degrees of anatomical variability. The spatially varying classification is embedded in a geodesic active region framework which allows for local deviations from the inferred shape using an iteratively updated classification-based region term. The method is evaluated on myocardium segmentation in late gadolinium enhanced cardiac MRI.

## 8.1 Introduction

Automatic medical image segmentation methods face a number of challenges such as low contrast, noise, missing or ambiguous boundary information, and intensity variations occuring e.g. as a result of disease or of the image acquisition technique. Edge-based segmentation methods have been used in numerous applications (e.g. [15], [23]) but typically need to be initialized close to the desired solution. Region-based methods use statistics of entire regions hence are more global than edge-based [129]. In case of ambiguous boundary information the incorporation of prior shape information can be helpful [129] as well as global models for appearance [24], but for irregular appearance regional statistics need to account for local and non-linear intensity variations [33]. The image intensity might not be a sufficient or even the most useful feature for discrimination between regions.

Statistical classification can deal with local intensity variations, but lacks global boundary and shape information. Spatially varying statistical classification (SVC) was presented in [135], and improved classification by adding global information. Level set methods represent contours implicitly and are widely used in image segmentation [99]. They have been extended to a geodesic active region formalism [102] which can create optimal partitions according to expected region properties. In [103], geodesic active regions are used for myocardial segmentation, with a region term that assumes separable Gaussian within-class intensity distributions, and a shape model where the mean shape is a distance function which accounts for local variability [112]. Cremers *et al.* [28] adopted the shape representation in [112] and introduced a dynamic labeling in a Chan-Vese segmentation framework which can handle occlusions but assumes intensity homogeneity. The authors in [85] use a level set shape model where shape variations are found by principal component analysis on aligned distance functions, however they assume the set of distance functions are a linear manifold. This model is used in [109] with a shape representation that enables encoding of boundary certainty by adjusting the slope of the representation, though this property is not directly used in their segmentation method.

We propose a method which exploits the benefits and compatibility of SVC and level set methods. From training data, a mean shape represented by a distance function with local shape variability is estimated, and feature selection is performed in order to find relevant features for the classes. The segmentation iterates between shape inference based on the classification, and classification extended with spatial localization features where we define a weighting of the

classification based on the local shape uncertainty. The inferred shape is a rigid transformation of the mean shape, therefore the SVC is embedded in a geodesic active region framework with a classification-based region term, which allows for local deviations from the mean shape and for a segmentation that takes global shape, local shape variability, and local appearance into account. We demonstrate the method on late gadolinium enhanced cardiac MRI, which contains both irregular intensity patterns due to scar tissue at random locations inside the myocardium, and ambiguous boundary information due to low image contrast, scar tissue and the papillary muscles. The analysis of such data currently relies on manual segmentations of the myocardium.

## 8.2 Method

The main idea of this work is to incorporate localization certainty in spatially varying classification, which is described in sections 8.2.6 and 8.2.8. But first we present the statistical classifier in 8.2.1 followed by the shape model used for localization certainty features in 8.2.3. The geodesic active region framework the SVC is embedded in is described in Section 8.2.7.

### 8.2.1 Statistical Classification

We use a $k$ Nearest Neighbor ($k$NN) classification, where class boundaries are determined from the labels of the $k$ nearest neighboring labeled training points in a $D$ dimensional feature space. Each pixel $j$ is described by a feature vector $\mathbf{u}_j$ and the distance $d$ squared to a training point $\mathbf{v}$ is $d_j^2 = \sum_{f=1}^{D}(u_{j,f} - v_f)^2$. The posterior probability in $j$ for class $\omega_c$ is $p(\omega_c|\mathbf{u}_j) = \frac{k_c}{k}$, where $k_c$ among the $k$ nearest neighbors belongs to class $\omega_c$. We use an approximate $k$NN classifier which allows for a tradeoff between computational efficiency and search error [4].

### 8.2.2 Feature Selection

Feature selection can increase computational efficiency and improve classification performance. The feature space is initially empty and is expanded iteratively (sequential forward selection) until there is no performance improvement, then decreased (sequential backward selection) until there is clear performance degradation according to a criterion function. The feature bank consists of all Gaussian derivatives up to $3^{rd}$ order, the intensity (original and Gaussian

smoothed), the eigenvalues and eigenvectors of the structure tensor [81] and the Hessian, which describes the first- and second-order structure locally, on a range of scales.

### 8.2.3 Shape Representation

We use a distance function shape representation and a shape model which accounts for local variations by assuming a mean shape, $\Phi_M$, with local degrees of shape variability, $\sigma_M$, as described in [112]. This method assumes the distribution of distance function values in each pixel to be Gaussian,

$$p_M(\Phi) = \frac{1}{\sqrt{2\pi}\sigma_M} e^{-\frac{(\Phi-\Phi_M)^2}{2\sigma_M^2}}.$$

Given a set of $N$ aligned (training) distance maps, the aim is to recover the distribution which has maximum support by minimizing (using the $-\log$) the energy $E(\Phi_M, \sigma_M) = -\sum_{n=1}^{N} \iint_\Omega \log[p_M(\Phi_n)]$ subject to the constraint that the mean shape remains a signed distance map ($|\nabla \Phi_M(\mathbf{x})|^2 = 1$, $\mathbf{x} \in \Omega$) and a smoothness constraint on the variability by minimizing $E_v(\sigma_M) = \sum_{n=1}^{N} \iint_\Omega ((\frac{d}{dx}\sigma_M)^2 + (\frac{d}{dy}\sigma_M)^2)$. Training shapes are aligned with respect to translation, rotation and scale, with a rigid transform $\mathbf{A} = (\mathbf{T}, \theta, s)$.

### 8.2.4 Shape Inference

Since SVC segmentation of an image relies on statistical classification we wish to infer shape based on the class probability maps. This can be realized in a variety of ways, for example a regional gradient descent approach as in [130] could be adjusted to incorporate classification. We use shape particle filtering as in [33] since that method describes shape inference from $k$NN classification. A set of shape hypotheses are sampled from transformations of the prior shape model $\Phi_{M,i}$. Each hypothesis is associated with an image labeling which is compared to the probability maps, and weighted by the likelihood for each region: $W = \exp\left[\frac{r}{m}\sum_{j=1}^{m} \log p(\mathbf{u}_j|\omega_c)\right]$, where $r$ is a constant controlling the randomness and $m$ is the number of pixels inside the template. New hypotheses sets are iteratively resampled proportional to the weights generated from the previous set, hence successful shapes will multiply while the unlikely will vanish. The distribution converges to a $\delta$-peak at the maximum likelihood solution, which before convergence can be approximated by the strongest local mode.

### 8.2.5 Spatially Varying Classification

SVC [135] iterates between global shape matching and statistical classification, a combination that resolves ambiguities in feature space with anatomical context. The $k$NN framework allows for a balance between spatial localization and other features using Euclidean distance in a modified feature space. After initial classification and shape inference, the feature space is expanded with $F$ extra features (corresponding to the number of objects of interest) that provides localization from the global shape,

$$d_j^2 = \sum_{f=1}^{D'} (u_{j,f} - v_f)^2, \tag{8.1}$$

with $D' = D + F$. With larger differences between $u$ and $v$ of the localization features, the distance in feature space increases, and this introduces a penalty to the classification. In [135], the object representation is converted to a distance map, giving a penalty that increases with the distance from the object boundary. This formulation fits well with the distance map representation in level set methods. In section 8.2.6 we describe how the shape model can be converted into features encoding not only anatomical localization but also localization certainty.

### 8.2.6 Spatial Localization Features

We perform SVC with $D'$ features of which $F$ are localization features, one for each object in the image, and define them to be the current level set representations of the objects weighted by the shape model variability outside the object contour;

$$F_i = (1 - H_\alpha(\Phi_i)) \cdot \Phi_i / \sigma_{M,i}^2(\mathbf{A}_i). \tag{8.2}$$

The interpretation of such a feature is that it introduces a classification penalty which is proportional to the distance from the contours weighted by the uncertainty of the boundary locations given the shape variability. Inside the contour the feature value is zero, so that the distance in Equation (8.1) is unchanged if both $u$ and $v$ are located inside the anatomical structure. Outside the structure, there is a classification penalty increasing with distance from the object boundary, weighted by shape variability. Locations with large boundary variations in the aligned training set, thus low boundary certainty, leads to a small value of $F_i$ hence a small classification penalty. Small local shape variations, on the other hand, represent relatively high prior location certainty, and a location with low variability leads to a higher classification penalty. E.g., if $u$ is inside the structure and $v$ is outside, the distance in (8.1) will be larger if $v$ is at a location

with low variability compared to one with high variability. Object localization for the epicardium with local variability is demonstrated in Figure 8.1.

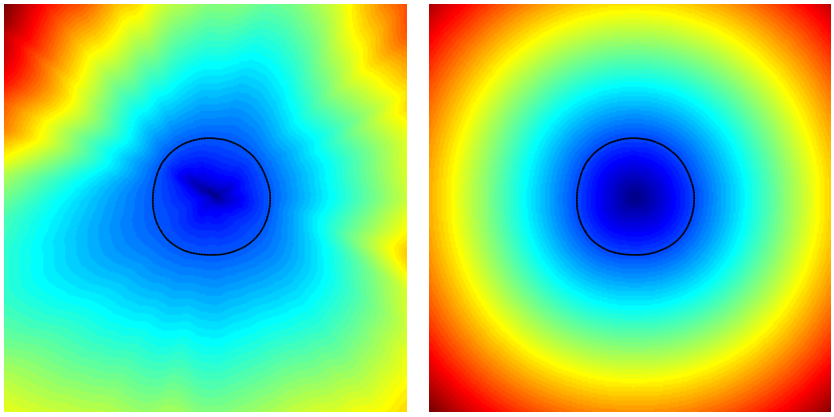


**Figure 8.1:** To the left: color coded mean shape for the epicardium from the data set described in section 10.3 weighted by shape variability ($\Phi_M/\sigma_M^2$). The zero level set is outlined in black. A direction normal to the boundary with more rapidly increasing weighted distance (e.g. towards the upper left corner) signifies high certainty in boundary location compared to directions with slowly increasing weighted distance (e.g. towards the lower left corner). To the right: the mean distance function shape $\Phi_M$.

### 8.2.7 Geodesic Active Regions

The shape estimate itself is a rigid transformation of the mean shape, and we need to allow local deviations from the mean shape given the image information. Therefore the SVC with localization variability is embedded in a geodesic active regions framework described here. Contour- and region-based evolution terms can be implemented using approximations of the Dirac ($\delta_\alpha$) and Heaviside ($H_\alpha$) distributions [145]:

$$\delta_\alpha(\Phi) = \begin{cases} 0, & |\Phi| < \alpha; \\ \frac{1}{2\alpha}(1+\cos(\frac{\pi\Phi}{\alpha})), & |\Phi| > \alpha \end{cases}$$

$$H_\alpha(\Phi) = \begin{cases} 0, & \Phi > \alpha; \\ 1, & \Phi < -\alpha \\ \frac{1}{2}(1+\frac{\Phi}{\alpha}+\frac{1}{\pi}\sin(\frac{\pi\Phi}{\alpha})), & |\Phi| < \alpha \end{cases}$$

In these equations $\alpha$ is the region in which the distributions are approximated, and $\Phi$ is assumed to be negative inside the contour it represents. The classical

active contour energy [15] can be reformulated as in [112],

$$\frac{d}{dt}\Phi_i = \delta_\alpha[g(|\nabla I|)\kappa_i|\nabla\Phi_i| + \nabla\Phi_i \cdot \nabla g], \tag{8.3}$$

where $\Phi_i$ is the level set representation of the $i$th contour, $\kappa_i$ is the curvature and $g$ is a function of the image gradient. Evolution functions for regions are derived in [103], assuming no contour overlap:

$$\frac{d}{dt}\Phi_i = -\delta_\alpha(\Phi_i)\log\frac{p(\omega_I|\mathbf{u})}{p(\omega_O|\mathbf{u})}|\nabla\Phi_i|, \tag{8.4}$$

where $\omega_I$, and $\omega_O$ are the classes inside and outside the object of interest respectively. We use the class probability maps from the supervised classification to estimate $p(\omega_c|\mathbf{u})$. The shape term evolves towards a rigid transformation of the mean shape (annotated $\Phi_{M,i}(A_i)$) but is down-weighted where there is high variability,

$$\frac{d}{dt}\Phi_i = -H_\alpha(\Phi_i)\frac{\Phi_i - \Phi_{M,i}(\mathbf{A}_i)}{\sigma^2_{M,i}(\mathbf{A}_i)}. \tag{8.5}$$

For simultaneous segmentation of several objects, coupling forces or forces that penalize region overlap or unlabeled regions are typically included. For the myocardial segmentation we use a coupling force between the endo- and epicardial contours which keeps them within an allowed range (determined from the training data) from each other and prevents them from overlap.

### 8.2.8 The SVC with Localization Variability Framework

The shape estimate itself is a rigid transformation of the mean shape, and local deviations according to the image information needs to be accounted for. Therefore, we iterate the SVC framework according to Algorithm 8.1 below. In each iteration the contours evolve according to the geodesic active region terms which means the regional information in (8.4) is also iteratively updated from the SVC, hence indirectly benefiting from spatial context. If the image position is one of the $D$ selected features, then the position in a test image can be shifted so that the object center of mass corresponds to the expected center of mass from training data. The SVC is iterated until there is insignificant change of class labels between iterations according to the classification, and when the region term is thus stabilized the geodesic active region terms evolve until low rate of change of the contours.

**Algorithm 8.1** SVC Framework

Classify image with $D$ features found from feature selection.
Infer shapes from class probability maps ($\Phi_{M,i}(\mathbf{A}_i)$).
Initialize the contours $\Phi_i$ as $\Phi_{M,i}(\mathbf{A}_i)$.
**while** Number of changes of class labels between iterations according to $\Phi_i$ are non-negligible **do**
  Evolve curves $\Phi_i$ according to (8.3)-(8.5) and coupling force.
  **if** Number of changes of class labels between iterations according to classification are non-negligible **then**
    Calculate localization features $F$.
    **if** Position is in the $D$ features **then**
      Shift position feature towards center of mass of object in training data.
    **end if**
    Perform SVC with $D' = D + F$.
    Infer shapes from the updated class probability maps ($\Phi_{M,i}(\mathbf{A}_i)$).
  **end if**
**end while**

## 8.3 Experimental Results

### 8.3.1 Data Set

We evaluate the method on late gadolinium contrast enhanced MRI short axis data, with in-plane resolution $1.5 \times 2\text{mm}^2$ and thickness of 8mm, acquired using a 1.5T system for myocardial scar as in [84]. Characterizing the extent of myocardial scar is essential for clinical treatment of patients with a history of infarction or chronic ischemia, and late contrast enhanced MRI is the leading technique for visualization of scar tissue. Segmentation of the myocardium is the initial and most crucial step in scar tissue analysis, which is important both in clinical studies of its impact on cardiac disorders and in computer aided interventions [84]. Example images and segmentations can be seen in Figure 8.2. The data set consists of 87 images from 11 scans, where 57 images (7 scans) are used for training and 30 for evaluation. Due to the large inter-slice distance we perform a 2D analysis on this data set, but the method can be extended to 3D applications.

### 8.3.2 Selected Features

Feature selection is performed with scales of 0.9, 1.5, 2.5, 3.5, 5 and 8 pixels, chosen to cover the myocardium thickness. After feature selection on the train-

ing data the feature set is the following, in decreasing significance: the position, the intensity smoothed on scales 5 and 8, $I_y$ on scales 5 and 8 (selected twice), $I_{yy}$ on scales 5 and 8, $I_x$ on scales 5 and 8, eigenvalues of the structure tensor, $\mathbf{T}(8,8)$, eigenvalues of the Hessian ($\mathbf{H}(3.5)$ and $\mathbf{H}(8)$), $I_{yyy}$ on scales 5 and 8, $I_{xy}$ on scales 2.5 and 3.5, eigenvalues of $\mathbf{T}(1.5,0.9)$, $\mathbf{T}(0.9,2.5)$, $\mathbf{T}(0.9,1.5)$ (selected trice), $\mathbf{T}(1.5,5)$, and $\mathbf{T}(8,5)$, $I_{xyy}$ on scales 5 and 8, eigenvalues of $\mathbf{T}(0.9,5)$, and $\mathbf{T}(0.9,3.5)$. It can be noted that the original image intensity information is never selected as a feature, and that among the most significant features, both first- and second-order structure on high scales are predominant. Eigenvalues of the structure tensor are selected several times hence smoothed low-scale first-order structure is useful for discrimination between classes. The eigenvectors are not selected, which could be related to the changes in gradient direction around the myocardium contours. The features are normalized prior to classification.

### 8.3.3 Segmentation Evaluation

Two coupled contours are used for the myocardium segmentation, hence two localization features were included in the SVC. The SVC converged in 5 iterations or less, determined by low rate of change of classification labels. Based on experiments, we set the weights of the shape and boundary terms to 0.2 and 0.16 respectively. The coupling force has the weight 0.25, and in the presence of strong image gradients, the boundary term contributes to some extent to the performance, with the weight 0.25. For the classifier $k$ is set to 50, and for the particle filtering $N$ is set to 2000.

The segmentation failed to converge to the desired solution for 2 test images due to inferior image quality and the extent to which scar tissue covered the myocardium. This can be related to the limited amount of training data given the large variations in appearance and shape. For the remaining 28 images the mean DSC is 0.80 ($\pm$0.06) and the point distance to contour 1.37 ($\pm$0.50) pixels. Example segmentations are shown in Figure 8.2.

For comparison, the authors in [33] use a point distribution model (PDM) based particle filtering scheme with regional information from a $k$NN classifier for automatic segmentation which they evaluated on 14 short-axis end-diastolic cardiac MRI slices with an average distance to contour of 1.1 ($\pm$0.3) pixels. The somewhat worse results on our data set can be related to the ambiguous image information, in particular the scar tissue, in the late gadolinium enhanced MRI data, making these images difficult to segment even for trained human experts. We implemented the method in [33] on our data set by assuming anatomical correspondence through the aligned shapes, and created a PDM by sampling points along the contours. Using the same parameter settings

and classifier as for our algorithm, the method was successful for the same 28 images with an average distance to contour of 1.6 ($\pm 0.6$) pixels, to which our method compares favorably according to a paired t-test ($p = 0.01$). The PDM requires point correspondences during the training phase, and results may have improved by establishing well-defined landmarks or by the use of a more sophisticated alignment method. We attempted to substitute the region term with the intensity as the only feature (besides localization) in the supervised classification, so that the $k$NN classifier determines class boundaries from the prior intensity information only, but for this data set the intensity as the only feature was insufficient for regional information.

## 8.4 Discussion

We have presented a method for spatially varying classification with spatial localization features from a level set shape model that incorporates shape uncertainty, which gives a classification penalty proportional to local shape uncertainty. The SVC is embedded in a geodesic active region framework which allows for local deviations from the inferred mean shape with an iteratively updated, classification-based, region term. Numerically, level set methods and spatially varying statistical classification fit well together since all calculations are made at pixel level, without need to switch back and forth between different representations. The method is automatic, does not require point correspondence during training, and makes use of prior information both in terms of global shape, local shape variability, and local descriptive features. Though the algorithm needs to be evaluated more extensively on a larger data set, the preliminary results presented here indicates that it may become useful for segmentation of this type of data and relieve cardiologists of time consuming labor in scar tissue analysis.

## Acknowledgements

The authors would like to thank Marleen deBruijne at DIKU for sharing code and insights on shape particle filtering, and Ole Fogh Olsen at IT University of Copenhagen for discussions.

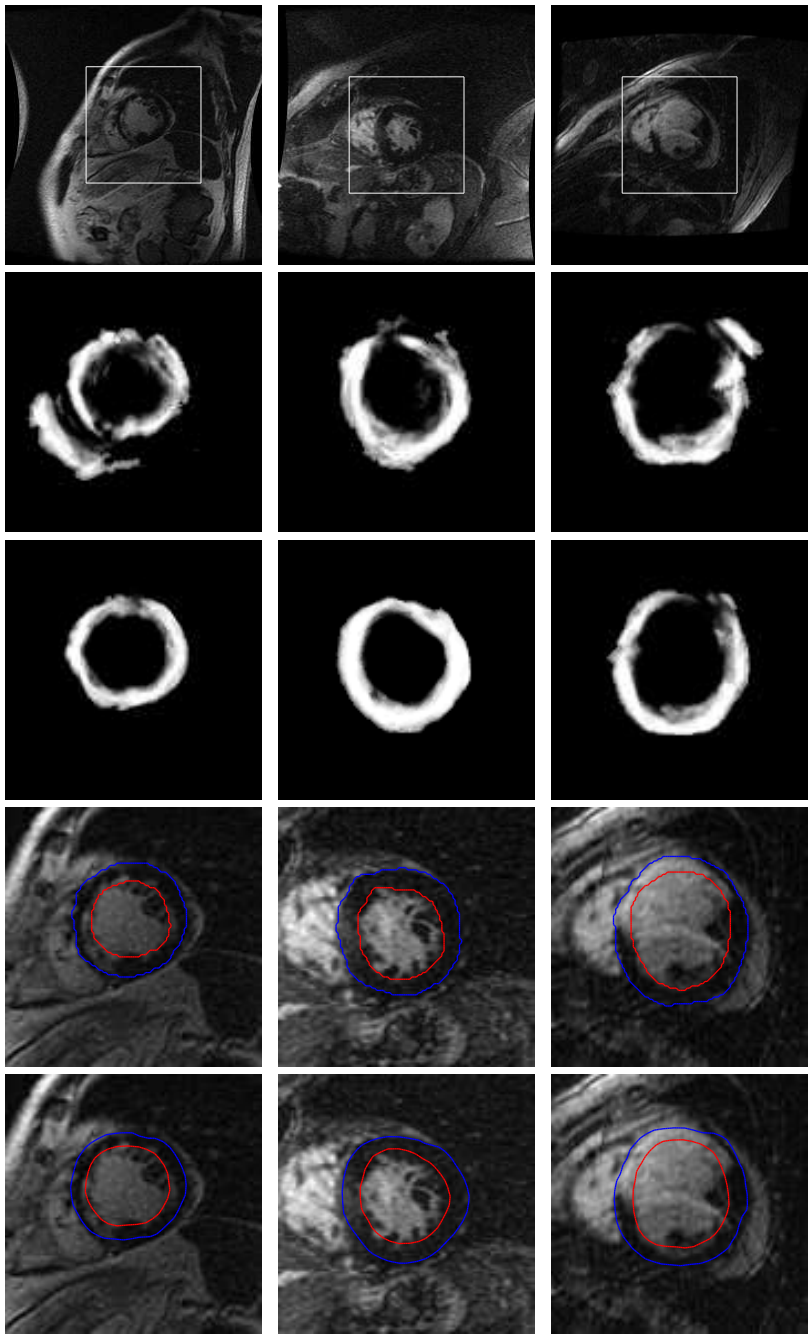


**Figure 8.2:** Three example segmentations from left to right. From top to bottom; input image where the rectangle signifies the region displayed in the following sequence: myocardium probability map from initial classification, same probabilities after convergence of the SVC, manual delineation by cardiologist, and automatic segmentation result. The endo- and epicardium are represented by red and blue contours respectively, and in between, inside the myocardium, scar tissue can be seen as relatively bright areas at varying locations in the images to the left and right.

# Part II

# Segmentation-Based Biomarkers

*Oh, people can come up with statistics to prove anything. 14% of people know that.*

Homer J. Simpson

Chapter 9

# Biomarkers for Osteoarthritis Founded on the Automatic Segmentation

The Chapters of Part II consists of research papers investigating the potential of cartilage surface curvature as a biomarker of osteoarthritis (Chapters 10 - 12). This Chapter serves as an introduction to the disease and its definitions (Sections 9.1, 9.4), imaging-based biomarkers (Sections 9.5), and an overview of curvature estimation can be found in Section 9.6 along with short description of the research papers in the remaining chapters. Chapter 13 briefly provides conclusions and perspectives for this thesis.

## 9.1 Osteoarthritis

*Osteoarthritis* (OA) also known as *osteoarthrosis*, *degenerative arthritis* or *degenerative joint disease* is the most common form of arthritis. It is one of the major health concerns among the elderly today [140], associated with a large socioeconomic impact on health care systems [68]. Current treatment of OA is restricted to symptom control and do not delay progression [93]. These factors, along with the desire to better understand the processes that contribute

to disease initiation and progression, makes OA related research high priority.

Although a single definition of OA remains elusive, it has been agreed that OA is a result of both mechanical and biological events that destabilize the normal coupling of degradation and synthesis of the articular cartilage and subchondral bone. OA diseases are manifested by morphological, biochemical, molecular, and biomechanical changes which lead to softening, fibrillation, ulceration, and loss of articular cartilage; sclerosis and eburnation of subchondral bone (where bone is converted to a dense, smooth substance resembling ivory), osteophytes, and subchondral cysts. When clinically evident, OA diseases can be characterized by joint pain, tenderness, limitation of movement, and variable degrees of inflammation [83].

OA can occur in any joint but is predominant in large weight-bearing joints as hips and knees, and the work in this thesis targets knee OA. Common risk factors are increasing age, obesity, sports injuries, and genetic factors such as sex and prevalence of other inherited joint disorders [26].

## 9.2 Articular Cartilage in The Knee Joint

Since the knees support nearly the entire body weight, they are vulnerable to injuries and development of OA. Cartilage has the purpose of accepting and transferring biomechanical loads between the *femur*, *tibia*, *patella* and *fibula* (see Figure 9.1), with the articular cartilages acting as bearing surfaces. The knee consists of the *femoro-patellar joint*, and the *femoro-tibial joint*, which links the *femur* and the *tibia* bones. Cells inside the cartilage, *chondrocytes*, produce and maintain the cartilaginous matrix, which consist mainly of *collagen* and *proteoglycans*. Collagen is the main protein in connective tissue supporting the matrix, and proteoglycans consists of a core protein with glycosaminoglycan (GAG) chains attached to it and acts as a 'filler' substance. Cartilage is avascular and nutrients are diffused through the matrix, and it has generally limited repair capabilities. In this work the cartilage in the medial side of the femoro-tibial joint is evaluated since OA frequently is observed there [38].

## 9.3 Imaging Methods

Radiography is one of the earliest imaging techniques for quantitative measures for cartilage degradation [11]. However, the cartilage is measured indirectly since it is not visible on standard radiographs, and it is a 2D technique hence cannot fully describe a 3D object. Though radiography is widely used in long-

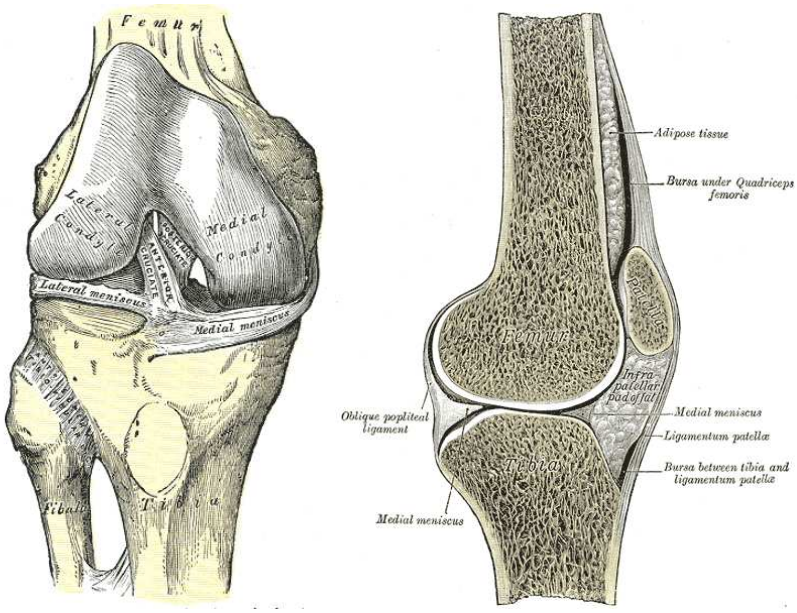



**Figure 9.1:** A right knee joint from the front (left) and from a sagittal view (right). The images are reproductions from the 20th U.S. edition of Gray's Anatomy of the Human Body originally published in 1918 and since lapsed into public domain.

term evaluation of OA, it is of limited use in clinical studies for early detection and assessment of the short-term development of the disease [17]. Arthroscopy can accurately depict swelling and ulcerations of the cartilage [10], but the technique is invasive and only examines the cartilage surfaces, not the entire cartilage. In recent years, magnetic resonance imaging (MRI) has become considered the leading imaging modality for direct, non-invasive assessment of the articular cartilage in 3D. It enables detection of cartilage changes and is less sensitive to joint positioning compared to radiography [104],[124]. An artifact that occurs in muscle-skeleton MRI of tendons and other collagen-containing tissues is the *magic angle effect*. This phenomenon results in an increase in the signal intensity of the fibers in these tissues when they are oriented at 55 degrees relative to the static magnetic field. At this particular angle the interactions among the water protons of collagen fibers that leads to T2 relaxation are nulled. So this will increase the T2 and consequently the signal intensity in the short TE sequences. A knee joint from a radiograph and a slice from a MRI scan of a knee joint can be seen in Figure 9.2. The data set in this thesis

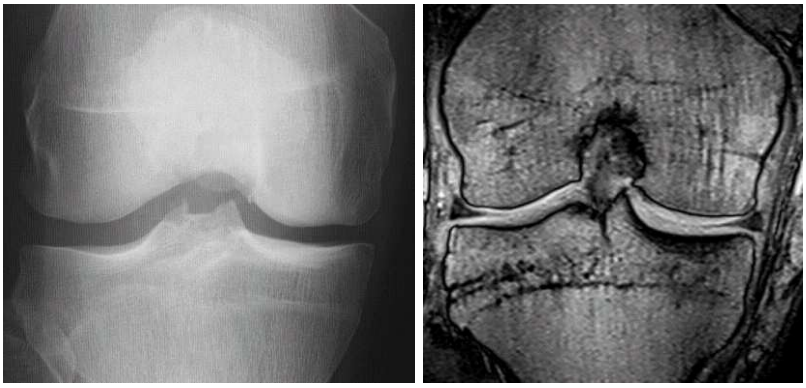


**Figure 9.2:** Coronal view of knee joints. To the left is a radiograph of a joint and to the right is a coronal slice from a MRI scan.

is acquired using a dedicated low-field MRI scanner. This type of scanner has drawbacks related to image quality with lower resolution and more difficulties in incorporating features such as fat suppression compared to high-field (1.5T or 3T) scanners that are commonly used in clinical settings. Though fat suppression has been successfully implemented lately for low-field MRI [66], the scanner used to acquire this data set does not offer that option. The low-field knee data suffer from a chemical shift artifact which appears as a darkening of cartilage/bone interface. The main advantages are cost-effectiveness with much lower cost per scan, lower installation and maintenance costs, and higher patient comfort with minimal noise level and without claustrophobic feelings. So far there has not been any accuracy validation of quantitative cartilage measures

from a low-field scanner as it has for high-field based measures both *in vitro* and *in vivo* [60],[22],[124],[39], but if a low-field scanner can be shown to be useful for quantitative articular cartilage assessment, costs for making clinical studies could be reduced significantly.

Extended MRI-based approaches to identify imaging parameters that target early OA symptoms are delayed gadolinium enhanced MRI of cartilage (dGEMRIC) that measures glycosaminoglycan (GAG) concentrations in cartilage, with GAG loss as an indicator of early OA [128], and diffusion tensor imaging (DTI) where diffusion weighted MRI is acquired in at least six directions. From the acquisitions a diffusion tensor is computed for each voxel, and local diffusion anisotropy can be used for an examination of collagen fiber structure inside the cartilage if the resolution is sufficiently high ($< 0.15$mm) [45],[91].

## 9.4 Definition of OA for Clinical Studies

Much effort has been devoted to development of a definition of OA symptoms for clinical studies. Currently, an MRI-based definition of OA has not been established, and epidemiological studies tend to rely on radiographic definitions of OA.

From radiographs, the *joint space width* is measured as the inter-bone distance at the perceived narrowest point of the joint space, and in longitudinal studies the *joint space narrowing* (JSN), *i.e.* decrease in joint space width, is measured. The gold standard method for grading of OA is the Kellgren-Lawrence index (KL), which is also determined from radiographs [73]. The KL is a score from $0-4$ determined by a radiologist, and the gradings are [93]:

0 Normal joint.

1 Doubtful narrowing of joint space and possible osteophytic lipping.

2 Possible narrowing of joint space and definite osteophytes.

3 Definite narrowing of joint space, moderate multiple osteophytes, some sclerosis, possible deformity of bone contour.

4 Marked narrowing of joint space, large osteophytes, severe sclerosis, and definite deformity of the bone contour.

Currently, JSN from radiography is the gold standard imaging biomarker for cartilage degradation and it is one of the most well-documented method for measuring the impact of OA on cartilage. However, it measures the cartilage

indirectly and in 2D [44], and despite standardization of radiographic techniques the rates of JSN can still vary widely [1]. Most epidemiologic studies use both x-ray and MRI, with a gradual shift toward MR-based outcomes as MRI allows shorter duration studies with fewer participants than if x-ray is used [93].

## 9.5 OA Biomarkers

Biomarkers are used to indicate or measure a biological process. OA progression is a complex process, where increased turnover of cartilage and bone occurs in the early stages, followed by cartilage fibrillation, thickening of the subchondral bone, bone lesion edema, osteophytes, focal cartilage lesions, thinning of the cartilage, and in the final stages, possible denudation of the bone. An endpoint for OA is total joint replacement, which is very long-term. Due to the complexity of the disease process, different biomarkers may be needed for different stages of the disease. A sketch of the disease progression can bee seen in Figure 9.3.

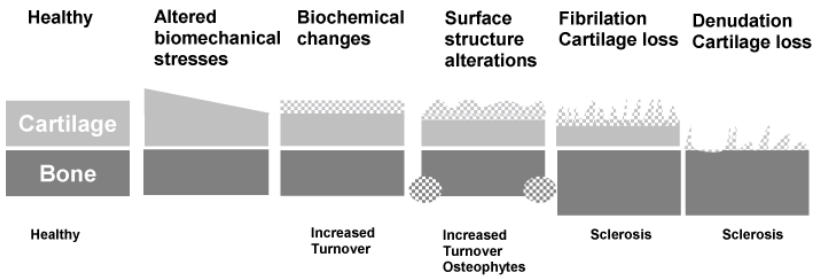



**Figure 9.3:** An illustration of the different disease stages of OA, where the initial stages are associated with altered biomechanical function and biochemical changes, thereafter alternations of structure and integrity, changes in the cartilage surface from lesions and fibrillations, and overall thinning.

Biomarkers for OA in the form of biochemical markers from systemic fluids have been developed targeting bone turnover (*e.g.* CTX-I) or cartilage turnover (*e.g.* CTX-II) [114]. The previously mentioned dGEMRIC is a biomarker that targets cartilage structure by measuring GAG concentrations.

Here, we focus on cartilage biomarkers from MRI without the use of contrast agents or diffusion weighting. The biomarkers described here are all stemming from the automatic cartilage segmentation method described in Part I, Chapter 6.

The cartilage volume and surface area are aimed at a quantification of overall cartilage loss, which occur in moderate to severe OA stages. Volume can be determined directly from the automatic segmentation, and the cartilage surface area can be determined by creating an isosurface from a smoothed version of the segmentation. Cartilage volume and area are evaluated in relation to radiographic signs of OA in Chapter 6 (reference [55]) in this thesis, and in the review [31].

Cartilage thickness also targets cartilage loss. Overall cartilage loss can be measured by the mean thickness across the cartilage, or a thickness map can allow detection of local variations in cartilage loss. Thickness quantification is challenging because healthy cartilage is only $2-3$mm thick, and with a scan resolution of approximately 0.7mm, voxellation effects need to be considered. A statistical shape model is fitted to the automatically segmented cartilage, ensuring a regularized representation of the cartilage segmentation and allows for anatomical correspondence between subjects. We use an m-rep [108], which represents an object by a mesh of medial atoms, each associated with a position, radius and directions to the boundary. With an m-rep the thickness at each atom location is obtained in a straight forward way. For a more thorough description of the shape modeling of cartilage and an evaluation of the cartilage thickness related to signs of OA, see [30]. The shape model can also be used for curvature estimation, which is described later.

The cartilage homogeneity aims at quantifying the cartilage structure, and is determined by estimating the entropy of the intensities of the voxels classified as cartilage. Homogeneity is related to water concentration, and loss of cartilage structure may lead to increase in homogeneity, or a decrease of entropy, something which occur in early OA prior to overall cartilage loss. More details on homogeneity and an evaluation of its relation to signs of OA are described in [110].

Cartilage surface curvature does not target cartilage loss, but changes in the shape of the cartilage. Focal lesions and fibrillation, which occur in early to moderate OA, can increase the cartilage surface curvature locally, whereas bending of the entire cartilage sheet may be related to joint incongruity which in turn is related to biomechanical stress. Cartilage curvature is discussed in Section 9.6 and in Chapters 10-12 (references [51],[54],[53]).

### 9.5.1 Evaluation of Biomarkers

In this thesis, the quantitative measures of cartilage from the knee MRI data are evaluated against the Kellgren-Lawrence index. Though radiographic definitions of OA are not optimal, this way the measures are compared to a well-established

OA score. The MRI-based biomarkers are separated into different populations based on their KL score, and differences between the populations are estimated using a Student's t-test [2], which is a statistical hypothesis test examining if a null hypothesis is true at a certain confidence level. If the p-value is below a threshold determined by the confidence level the null hypothesis is rejected. For the baseline study the samples are assumed to be independent of each other and an unpaired t-test with the null hypothesis $H_0 : \mu_1 = \mu_2$, where $\mu_1$ and $\mu_2$ are the mean values in the different populations, is evaluated for statistical significance at a confidence level of 95%. For the longitudinal data, a paired t-test is employed with the null hypothesis $H_0 : \triangle\mu = 0$, where $\triangle\mu$ is the percent change of values between baseline and follow-up. The t-test assumes normal distributions and equal variances of the population data.

The reproducibility of each biomarker is evaluated by comparing outcomes from knees scanned twice approximately one week apart. The inter-scan reproducibility of a biomarker is estimated using mean relative absolute difference (AD) and mean coefficient of variation CV, where

$$\mathrm{AD} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Q_i^a - Q_i^b|}{(Q_i^a + Q_i^b)/2},$$

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{std}(Q_i^a, Q_i^b)}{(Q_i^a + Q_i^b)/2},$$

where $Q^a$, $Q^b$ are the measured values at the first and second visit, $N$ is the number of scan pairs, and std is the standard deviation.

## 9.6 Cartilage Surface Curvature

### 9.6.1 Curvature Estimation

The evaluation of surface curvature as a potential biomarker for OA is motivated by the development of focal lesions and fibrillation on the cartilage surface after onset of OA, and by the influence of joint incongruity and high biomechanical stress which occur in initial stages of OA and which could be related to more global bending of the cartilage sheet. Since both local and global curvature is relevant, the curvature needs to be analyzed at different scales.

Curvature is a measure of how much a curve in 2D deviates from being a straight line, and is measured in inverse distance ($\mathrm{m}^{-1}$). For a curve in 2D, the curvature for each point on the curve is defined as the inverse of the radius of the osculating

circle, where the osculating circle is uniquely defined by taking three neighboring points along the curve and let the arclength between them go to zero. The curvature can also be described as the rate of change of the tangent direction $\theta$ of the curve with respect to the arclength;

$$\kappa = \frac{d\theta}{ds} = \frac{1}{\rho}$$

where $\rho$ is the radius of the osculating circle, see Figure 9.4.

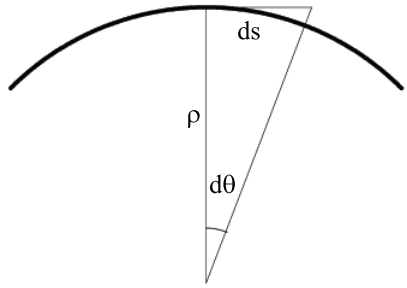



**Figure 9.4:** Definition of curvature as the change of tangent direction $\theta$ with respect to the arclength $s$; $\kappa = \frac{d\theta}{ds} = \frac{1}{\rho}$.

For a surface in 3D, each point $\mathbf{x}$ on the surface a is associated with a surface normal $\mathbf{n}$ and a tangent plane. A normal plane is a plane orthogonal to the tangent plane, and the cut of this normal plane with the surface generates a 2D curve, as illustrated in Figure 9.5. Rotating this normal plane about $\mathbf{n}$ will generate an infinite number of 2D curves and thus an infinite number of curvature values for each point. However, the curvature at each point $\mathbf{x}$ can be fully described by the *principal curvatures*; the *maximum curvature* ($\kappa_1$) and the *minimum curvature* ($\kappa_2$). They are oriented in orthogonal directions called the principal directions. Two classical measures of the curvature in a point is the *Gaussian curvature* which is intrinsic (independent of the embedding), $\kappa_G = \kappa_1\kappa_2$, and the *mean curvature* which is extrinsic (depends on the embedding), $\kappa_M = \frac{\kappa_1+\kappa_2}{2}$.

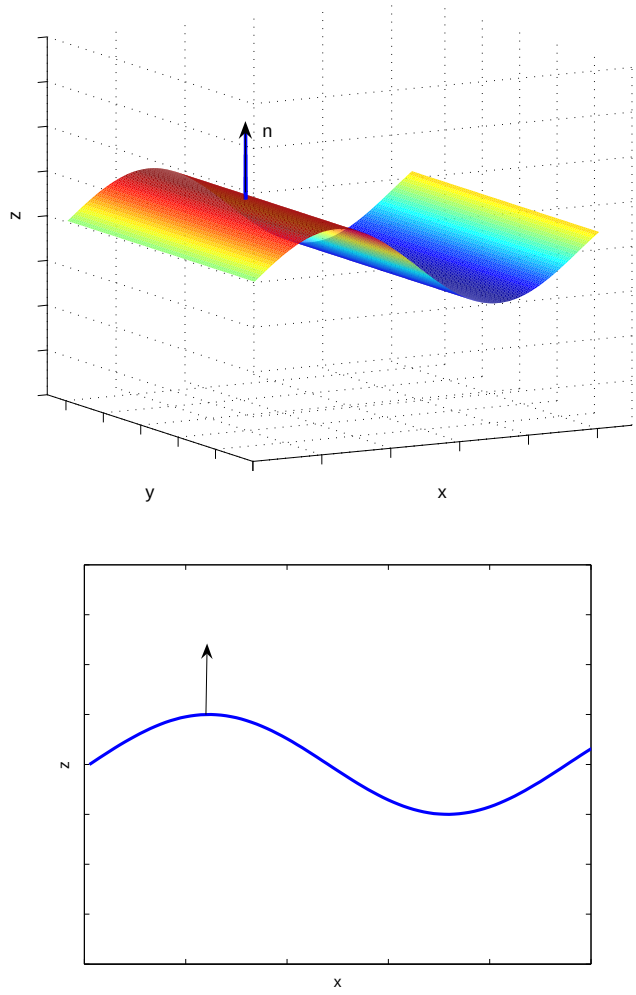



**Figure 9.5:** The top figure is a surface in 3D, and the tangent plane to the point with normal **n** is the $xy$-plane. The bottom figure is a 2D curve generated by cutting the surface with a plane parallel to the $xz$-plane through the point, which is a normal plane to the tangent plane.

### 9.6.2 Cartilage Surface Curvature in Relation to Signs of OA

In this thesis we evaluate the cartilage surface curvature from the binary segmentations produced by the automatic segmentation method. Since the segmentation is voxel-based, some degree of smoothing or regularization needs to be involved. Two techniques are evaluated; one using mean curvature flow, and the other by comparing surface normals on the regularized surface of a deformable shape model. The curvature flow computes the mean curvature of the smoothed surface, whereas the comparison of surface normals approximates the curvature by taking the mean of curvature estimated in 8 discrete directions on the surface. In Chapter 10 (reference [51]), the curvature is estimated using these two methods on 114 scans. Both methods produce measures with statistically significant differences between populations with different KL scores according to unpaired t-tests.

In Chapter 11 (reference [54]), the accuracy of the two curvature estimation techniques are examined by creating digital phantoms with known curvature and comparing them with estimated curvatures. The shape model-based measure is more accurate for high-scale curvature but cannot capture fine-scale curvature due to the regularization of the shape model. For low-scale curvature estimation, the curvature flow-based method is found to be more suitable. The inter-scan reproducibility for the two methods is 2.3% and 6.4% (mean CV), respectively, which is estimated by comparing 31 scan-rescan pairs imaged approximately a week apart. After the estimates are checked for systematic bias and modified compared to the method in Chapter 10, and normalized for joint size, the values in different KL populations for the same 114 scans as in 10 are again compared again with statistically significant differences between populations as a result.

In Chapter 12 (reference [53]), the curvature is evaluated on a larger data set consisting of 288 scans at baseline, and 245 scans during a follow-up study after approximately 21 months. Cross-sectionally, the surface curvature for both estimates are significantly higher in the OA populations compared to the healthy population, with $p < 0.001$ and $p \ll 0.001$ respectively. For the longitudinal study, there is a significant increase in curvature estimated by curvature flow for KL 0-1 populations, which could relate to development of lesions and fibrillation. For the shape model-based curvature estimate there is a significant increase for the KL $\geq 3$ population, which could be related to the collapse of the joint in severe OA.

We hypothesize that high-scale curvature is related to joint incongruity since surfaces with high curvature values globally are more likely to mismatch as the joint bends. However joint congruity is a measure of how well surfaces match each other, and in order to evaluate how congruity relates to high-scale

curvature, a study of contacting cartilage compartments and dynamics of the bending joint is required. A possible measure of joint incongruity could be to compare the principal curvatures and their directions for opposing cartilage surfaces [5]. The cartilage sheets with maximum and minimum coarse-scale curvature are plotted in Chapter 12 and it seems like they are related to a 'flatness' of the surface. It would be interesting to look at the m-rep shape model and see if similar behavior are picked up by certain shape variations. The fine-scale curvature may be related to a 'smoothness' of the cartilage surface and thus to focal lesions and fibrillation. However, clear differences between a 'smooth' and an 'unsmooth' cartilage sheets could not be detected by visual inspection. Though there are significant differences between different populations based on KL, it remains to be verified what anatomical features the fine-scale curvature models. One possibility would be to compare fine-scale curvature to arthroscopy findings or an MRI-based lesion score.

In this work, curvature is estimated based on binary segmentations. It would be interesting to estimate the curvature directly on an isosurface of the probability maps from classification at the level of the threshold for the binary segmentation since the thresholding might discard useful information. The probability map before thresholding provides a soft labeling of the image which could compensate for voxellation effects (see Figure 9.6), however then false positives from neighboring classes needs to be dealt with. Using soft labels could also be interesting in cartilage volume estimations. It would also be interesting to estimate also the principal curvature directions, which could be used for comparing curvature of opposing cartilage sheets in relation to joint congruity.

**Figure 9.6:** A probability map of tibial cartilage in a sagittal slice before (top) and after (bottom) thresholding for a binary segmentation. Estimating curvature on an isosurface of the probability map might reduce voxellation effects.

Chapter 10

# Automatic Curvature Analysis of the Articular Cartilage Surface

*Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen*

**Abstract**

In osteoarthritis (OA), the articular cartilage degenerates, losing its structure and integrity and eventually thinning. Curvature analysis of the cartilage surfaces has been suggested as a potential disease marker for OA but until now there has been few results to evaluate its relation to OA. We present two methods for surface curvature analysis, one that estimates curvature on low scales using mean curvature flow, and one method based on a cartilage shape model, where curvature is estimated by comparison of surface normals for high-scale curvature estimates. We show that both methods can distinguish between healthy and osteoarthritic populations, and the shape model-based method can even distinguish healthy from mild OA populations with high reproducibility, indicating the potential of surface curvature as a disease marker for OA.

## 10.1 Introduction

Magnetic resonance imaging (MRI) is becoming one of the leading imaging modalities in osteoarthritis (OA) research as it allows for non-invasive quantification of the articular cartilage and detection of cartilage degeneration, which is the characteristic symptom of the disease [60]. OA is second to heart disease in causing work disability and is associated with a large socioeconomic impact on health care systems [68].

Typical quantitative disease markers for OA is the articular cartilage volume and thickness, where the volume on its own can be a relatively poor disease marker [58]. Cartilage thickness measures combined with methods for finding correspondences in anatomy can give a more localized measure of cartilage degeneration, and be more sensitive to changes especially in load bearing parts. Williams *et al.* [138] have found sub-millimeter changes in local thickness measurement in a longitudinal study of OA risk subjects, but their method currently relies on manual labor.

Standard clinical treatment of OA today does not reverse the cartilage degeneration which makes it is important to detect the disease at an early stage, and in clinical studies of new treatments it is important to detect changes across populations as soon as possible due to the high costs associated with drug development. This leads to a search for new disease markers that can alone or combined with well established ones detect OA earlier and more reliably.

Joint congruity describes the extent to which contacting joint surfaces match each other. It has been suggested that joint incongruity is related to high peak stress and such mechanical factors can influence the initiation and progression of OA [13]. Curvature analysis of the articular surface could be related to joint incongruity since high curvature on the surface will most likely lead to mismatching of the surfaces as the joint bends.

Hohe *et al.* [65] have analyzed the curvature of knee cartilage surfaces from MRI as an incongruity measure, by first segmenting the cartilage semi-automatically slice-by-slice, then estimating the principal curvatures locally from a b-spline surface interpolation on a $5 \times 5$ neighborhood of sampled surface points 6mm apart. They found an average mean curvature of $29.6\text{m}^{-1}$ ($\pm 9.9\text{m}^{-1}$ std) for the tibial medial cartilage surface and $-0.9\text{m}^{-1}(\pm 3.8\text{m}^{-1}$ std) for the central part of the same surface, with inter-scan reproducibility values up to $4.7\text{m}^{-1}$ stdfor 4 repeated measurements, on 14 healthy subjects.

Terukina *et al.* [127] performed an in-vitro study of the curvature in 2D by slicing the knee joint through the sagittal plane and fitting a circle to three equidistant points within 1cm on the cartilage surface then taking the inverse of the radius as an estimate of the curvature. They found an average curvature of $4.4\text{m}^{-1}$ for the femoral condyle in their study intended for cartilage replacement, where cartilage congruity is of importance.

Even though curvature analysis shows potential as a OA disease marker, there has not as yet been any study showing how or if the curvature differs between healthy and osteoarthritic knees. Joint congruity is related to the global surface shape and mainly concerns high-scale curvature analysis, whereas low-scale curvature could be related to local shape changes and thus to focal cartilage lesions. In this paper, we present two automatic methods for high- and low-scale curvature analysis and demonstrate their abilities to distinguish between healthy and OA populations.

## 10.2 Curvature Estimation

We would like to analyze the curvature of the articular cartilage surface and see if it differs between healthy and OA cartilage, and to examine the curvature at different scales since it could be related to different signs of OA. Looking at existing methods, Terukina *et al.* [127] keep a constant distance between points ($d = 0.5\text{cm}$) for their circle estimates. The curvature is the inverse of the radius of the osculating circle, which is uniquely defined by taking three neighboring points on the curve and let the arclength between them go to zero. By not shrinking the distance at points on the curve with high curvature values, two points can be placed at the same approximate location giving an upper bound for the curvature value of $(d/2)^{-1} = 400\text{m}^{-1}$ for $d = 0.5\text{cm}$ (see Figure 10.1), giving a crude estimate at low scales (high curvature). Another drawback is that the curvature is only estimated in the selected 2D plane of view. For non-invasive 3D curvature analysis, the method of Hohe *et al.* [65] is more interesting. Fitting b-spline surfaces to control points of the segmented surface at specific distances is a way of selecting a scale for the curvature analysis. However, their image voxels are quite anisotropic, which is a disadvantage for surface curvature estimation in 3D where sampling of surface boundary points equally in all directions is essential.

We apply mean curvature flow to super-sampled binary volumes from a fully automatic cartilage segmentation [48] for a relatively smooth deformation of the surface from low towards higher scales, and evaluate the ability to discriminate between a healthy and an OA population during the flow for this analysis which targets low-scale curvature. We also fit a deformable shape model to the carti-

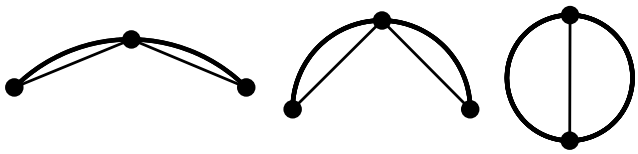


**Figure 10.1:** Fitting a circle to three equidistant points when two of the points are approaching each other (from left to right).

lage and obtain a curvature measure from the normals and locations of boundary points for high-scale curvature analysis, something the shape model is suitable for due to inherent regularization. Using the shape model we also examine the importance of evaluating the curvature in an anatomically well defined region of interest.

### 10.2.1 Curvature Estimation by Mean Curvature Flow

The mean curvature flow for a surface $\mathbf{S}$ is $\mathbf{S}_t = \kappa_M \vec{N}$, where $\kappa_M$ is the mean curvature (the mean of the two principal curvatures), $t$ is time and its subscript denotes differentiation, and $\vec{N}$ is the normal of $\mathbf{S}$. A common approach for implementing surface evolution is level set methods [115]. A function $\phi(x, y, z; t)$ gives an implicit representation of the surface, so that $\mathbf{S}(t) = \big\{(x, y, z) \mid \phi(x, y, z; t) = 0\big\}$. In the level set formulation the mean curvature flow is described by

$$\phi_t = \kappa_M |\nabla\phi| = \Big[ \bigtriangledown \big(\frac{\nabla\phi}{|\nabla\phi|}\big)\Big] |\nabla\phi|,$$

where $\nabla\phi$ is the gradient. Thus the mean curvature can be written

$$\begin{aligned} \kappa_M = \big(\phi_x^2(\phi_{yy} + \phi_{zz}) + \phi_y^2(\phi_{xx} + \phi_{zz}) + \phi_z^2(\phi_{xx} + \phi_{yy}) \\ -2(\phi_x\phi_y\phi_{xy} + \phi_x\phi_z\phi_{xz} + \phi_y\phi_z\phi_{yz})\big) / \big(2\big(\phi_x^2 + \phi_y^2 + \phi_z^2\big)^{3/2}\big) \end{aligned} \tag{10.1}$$

in terms of derivatives of $\phi$, where the derivatives are calculated using finite differences.

For curvature flow of curves in $\mathbb{R}^2$ the Gage-Hamilton and Grayson theorems assures that convex non-intersecting curves will shrink smoothly to a point. This property does not extend to surfaces in $\mathbb{R}^3$ where topology changes can occur.

However existence, stability and uniqueness of viscosity solutions of the mean curvature motion for hypersurfaces in level sets have been found [42] [19].

The time behavior of the curvature for curves undergoing curvature flow is described by a reaction-diffusion equation which is non-trivial to solve [115]. For the cartilage surfaces, we examine the mean curvature throughout the flow, starting from tibial medial cartilage volumes from automatic segmentations [48], by taking the mean of the absolute value of the mean curvature in (10.1) at the tibial medial articular surface. At each iteration, this curvature estimate is evaluated by its ability do discriminate between healthy and OA groups. When the surface is deformed as during curvature flow, the distance map of $\Phi$ needs to be recalculated, and these reinitializations are made every 7 iterations. The time step is 0.15 voxels. We super-sample the scans, dividing each voxel into $5 \times 5 \times 5$, with a new resolution of approximately 0.16mm side length, so that the curvature flow will initially mainly reduce partial volume effects.

### 10.2.2 Curvature Estimation by Shape Model Boundary Points

Using curvature flow the surface curvature can be analyzed at low scales, but as the surface evolves under the curvature flow the surface becomes smoother but the volume loses its cartilage-like shape and therefore might not be suitable for high-scale curvature analysis. Therefore we have developed another scheme for curvature analysis in which a deformable m-rep shape model is fitted to the cartilage. The shape model is an m-rep [108], and its regularized surface representation allows for high-scale curvature analysis. The m-rep represents an object by a mesh of medial atoms, each associated with a position, radius and directions to the boundary. Besides curvature analysis, the shape model is also used for finding anatomical correspondences and local thickness measures in a related study [30]. The tibial medial cartilage model is fitted to the set of automatically segmented volumes, and consists of a mesh of $4 \times 8$ medial atoms from which boundary points are interpolated on the articular surface (see Figure 10.2). We define the load bearing (central) part of the articular surface as a region of interest and estimate the curvature locally in a $5 \times 5$ neighborhood by taking the angle between the normals and dividing it by distance between them. Assuming short distances and small angles due to the regularization, this is an approximation to the curvature, $\kappa = \frac{d\theta}{ds}$, where $\theta$ is the angle between two normals and $s$ is the arclength. The mean the absolute values of local estimations of $\kappa$ of the boundary points in the region of interest is a quantitative estimate of the curvature. The limited number of directions, given by the neighborhood, the curvature is examined in approximates the mean curvature of the surface at that point.

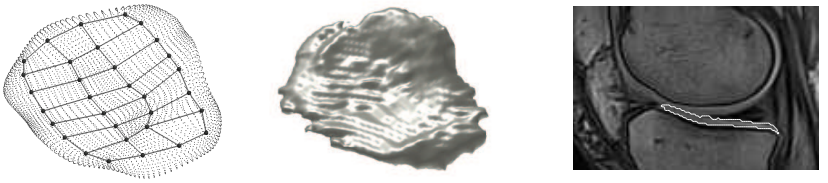


**Figure 10.2:** An m-rep surface of a tibial medial cartilage sheet to the left, large points indicate medial atoms and small indicate boundary points. The segmented cartilage it is fitted to is in the middle, and part of a sagittal slice with the tibial medial cartilage delineated to the right.

## 10.3 Results

Our data set consists of 139 (25 are used for training, 114 for testing) knee Turbo 3D T1 scans acquired from an Esaote C-Span low-field 0.18T scanner dedicated to imaging of extremities (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms), with approximate spatial resolution $0.8 \times 0.7 \times 0.7\text{mm}^3$. The test subjects are between 22-79 years old with an average age of 56 years, 59% females, and there are both healthy and osteoarthritic knees according to the Kellgren-Lawrence index (KL) [73], a radiographic score from 0-4 where KL 0 is healthy, KL 1 is considered borderline or mild OA, and KL $\geq 2$ is severe OA. In the test set there are 51, 28, 13 and 22 scans that have KL 0, 1, 2 and 3 respectively. For reproducibility evaluation, 31 knees were re-scanned after approximately one week, making the total number of scans 170. The spatial resolution is approximately $0.8 \times 0.7 \times 0.7\text{mm}^3$. We analyze the medial tibial cartilage in this study since OA is most often observed in medial compartments [38].

For the mean curvature flow, there is a rapid decrease in mean curvature on the articular surface initially as can be seen in Figure 11.3. During this time the inter-scan reproducibility (defined as absolute percent pairwise measurement differences) is low, and an unpaired t-test cannot separate a healthy population (KL 0) from an OA population (KL $\geq 1$) at significance level of 95%. This could be a result of the cancelation of partial volume effects present in the initial stages. As the surface evolves under the curvature flow, these values stabilize with curvature values of approximately $250\text{m}^{-1}$ and reproducibility of 5% (relative absolute difference). Examining the separability between healthy and OA groups, there is a minimum in p-values after approximately 35 iterations, where the p-value is $= 0.001$ with inter-scan reproducibility of 6.2% pairwise difference with an mean curvature of $330m^{-1}$ ($\pm 38m^{-1}$ std). At no time during the flow is it possible to separate healthy from borderline OA (KL 1) groups. Figure 10.4 illustrates how a cartilage sheet deforms during mean curvature flow.

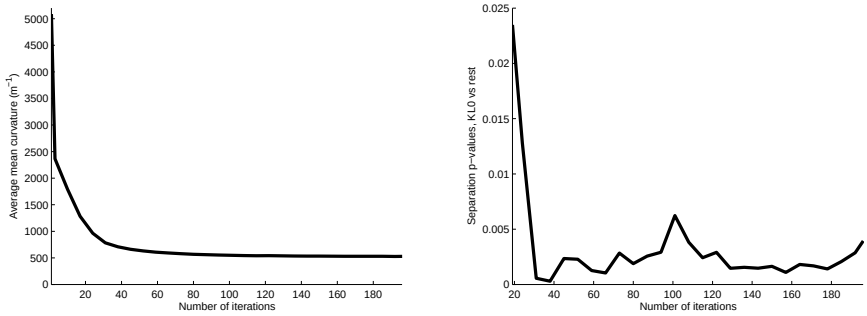



**Figure 10.3:** Left: the mean flow-based curvature estimate of the articular surface for the test data. Right: p-values from t-test between healthy and OA populations.

The curvature estimation of the boundary points of the m-rep gives an average of $45\text{m}^{-1}$ ($\pm 9.7\text{m}^{-1}$ std), which is a factor 10 lower than the values of the curvature flow method. The m-rep based method can separate the healthy from the OA group ($p = 2.6 * 10^{-4}$) and can even separate healthy from borderline OA ($p = 0.009$). The relative absolute difference in measurements is 11% for the reproducibility study. When estimating the curvature on the entire articular surface instead of only the load bearing part, the mean curvature is approximately the same, $46\text{m}^{-1}$ ($\pm 8.1\text{m}^{-1}$ std), but the method is no longer able to separate the healthy from borderline OA group ($p = 0.09$).

These results can be compared to the cartilage volume, which is a well established OA disease marker. Performing an unpaired t-test with the normalized volumes from manual segmentations by a radiologist yields a p-value of 0.095 for the healthy vs. OA groups and 0.22 for the healthy vs. mild OA groups.

## 10.4 Discussion

We have evaluated two methods for curvature analysis on the articular cartilage surface. Using mean curvature flow it is possible to separate a healthy from an OA group ($p = 0.001$) using the mean curvature of the articular cartilage surface. The inter-scan reproducibility at that point is 6.2% (paired absolute difference). The method cannot at any time during the flow separate healthy from borderline OA groups. The model based method can separate the healthy from the OA group ($p = 0.0003$) and from the borderline OA group ($p = 0.009$), with a paired absolute difference of 11%. In comparison, the normalized cartilage volume gives p-values of 0.095 and 0.22 for the healthy vs. OA and mild OA groups

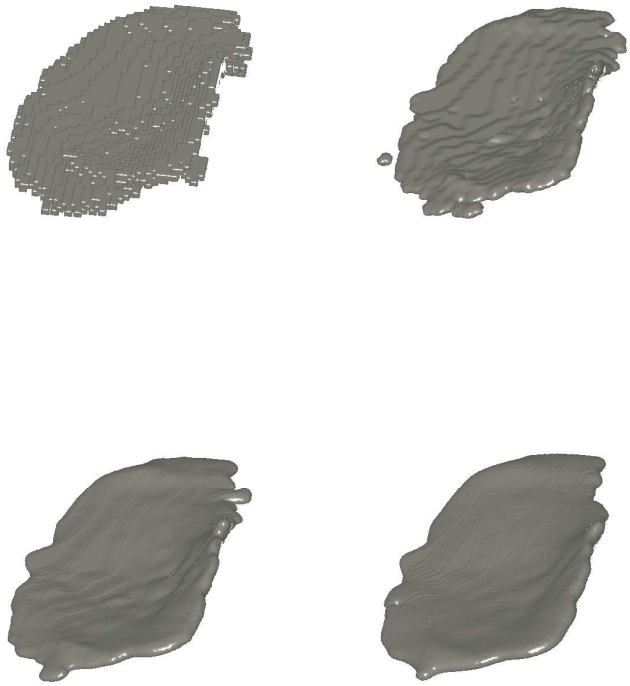


**Figure 10.4:** The appearance of a cartilage sheet during mean curvature flow. From left to right: after 0, 35, 100, and 200 iterations.

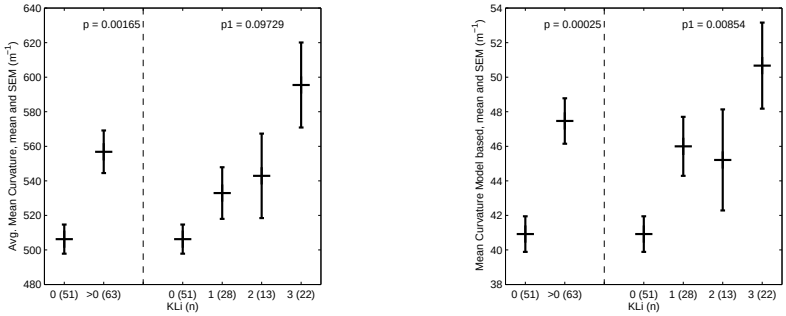



**Figure 10.5:** Separation between healthy and OA groups using curvature estimation. Left: from curvature flow after 35 iterations, right: model-based curvature.

respectively.

The regularization of the model makes it suitable for high-scale curvature analysis, finding curvature values around $50\text{m}^{-1}$ thus structures with a radius in the cm range, which corresponds to evaluating the shape of the surface globally, and could thus be related to congruity. The mean curvature flow evaluates low-scale curvature (curvatures in the range $250\text{m}^{-1}$ thus objects with a radius in the mm range) and could target local changes or lesion on the cartilage surface. Future work will involve a longitudinal study and ground truth evaluation.

Because significant differences between healthy and mild to severe OA populations can be detected using curvature analysis with higher statistical significance than a well established measure such as the cartilage volume normalized for bone size, these fully automatic methods can potentially become useful in clinical studies as markers for OA.

Chapter 11

# Accuracy Evaluation of Automatic Quantification of the Articular Cartilage Surface Curvature from MRI

*Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, and Claus Christiansen*

**Abstract**

**Rationale and Objectives.** To study the articular cartilage surface curvature determined automatically from magnetic resonance (MR) knee scans by evaluating the accuracy of the curvature estimates on digital phantoms, and by their potential as disease markers for different stages of osteoarthritis (OA).

**Materials and Methods.** Knee MR data was acquired using a low-field 0.18T scanner, along with posterior-anterior x-rays for evaluation of radiographic signs of OA according to the Kellgren-Lawrence index (KL). Scans from a total of 114 knees from test subjects with KL 0-3, 59% females, aged 21-79 years were evaluated. The surface curvature for the medial tibial compartment was estimated automatically on a range of scales by two different methods: mean curvature flow, and by comparison of surface normals from a cartilage shape model. The curvature estimates were normalized for joint size for inter-subject comparisons. Digital phantoms were created in order to establish the accuracy of the curvature estimation methods.

**Results.** The phantom study showed that the shape model-based method was more accurate for a coarse scale analysis, while the flow-based method estimated fine scales better. The inter-scan reproducibility for the two methods were 2.3%, and 6.4% (mean CV) respectively. The surface curva-

ture was significantly higher in the OA population (KL $> 0$) compared to the healthy population (KL 0) for both curvature estimates, with p-values of 0.000004 and 0.00006 respectively. The shape model based curvature estimate could also separate healthy from borderline OA (KL 1) populations ($p = 0.005$).

**Conclusion.** Both the fine- and the coarse-scale curvature estimates could distinguish between healthy and OA populations, and the coarse-scale curvature could even distinguish between healthy and borderline OA populations. The highly significant differences between populations demonstrate the potential of cartilage curvature as a disease marker for OA.

## 11.1 Introduction

Osteoarthritis (OA) is a joint disease characterized by degeneration of the articular cartilage and bone deformations. It is second to heart disease in causing work disability and is associated with a large socioeconomic impact on health care systems [68]. Though much effort has focused on development of disease modifying OA drugs, currently there is no standard clinical treatment of OA that can reverse cartilage degeneration. In clinical studies of new treatments the ability to detect even small differences across populations receiving treatment or placebo could significantly reduce the cost of such studies. If OA can be detected at an early stage, patients that would be benefit from non medicinal intervention such as gait adjustment could be identified. Hence reliability and early OA detection are desirable features for OA disease markers.

Radiographs currently presents the gold standard for measuring cartilage degradation in terms of joint space narrowing, and for determining the degree of OA using the Kellgren-Lawrence index (KL) [73]. However standard radiographs are limited to a 2D analysis of the joint and only measures the cartilage indirectly since it is not visible in the images. It is also sensitive to malpositioning, exposes test subjects to radiation, and is insensitive to small changes in the cartilage structure. Magnetic resonance imaging (MRI) has become one of the leading imaging modalities in osteoarthritis research as it allows for non-invasive, three dimensional (3D), quantification of the articular cartilage [60][104][105]. Typical quantitative disease markers from MRI is the articular cartilage volume and thickness [123][30][119][88], where the volume on its own can be a relatively poor disease marker [58].

Joint congruity describes the extent of which contacting joint surfaces match each other. It has been suggested that joint incongruity is related to high peak stress and such mechanical factors can influence the initiation and progression of OA [13][90]. Curvature analysis of the articular surface can be used as an estimation of the joint incongruity since high curvature on the overall surface is

likely to lead to mismatching of the surfaces as the joint bends. Joint congruity is related to the overall surface shape, thus mainly concern coarse-scale curvature analysis. An analysis of fine-scale curvature is related to local shape changes on the articular cartilage surface hence to focal cartilage lesions and fibrillation. Therefore it is interesting to investigate the potential of the curvature as a disease marker for OA

The importance of joint congruity in cartilage replacement has been discussed in an *in vitro* study [127], where the curvature was estimated in 2D by slicing the knee sagittally and fitting a circle to three equidistant points on the cartilage surface within 1cm. Cartilage surface curvature has also been estimated on healthy test subjects in 3D by fitting a b-spline to surface points of the cartilage determined by a semi-automatic slice-wise segmentation [65]. Curvature analysis is also useful in volume rendering [79] and in other medical imaging applications. In [146], polyp detection in virtual colonoscopy is performed using curvature estimation as in [131]. Curvature estimation has also been employed for finding curves of extremal curvature on brain surfaces [76].

In this paper, we present two fully automatic curvature estimation methods. Using digital phantoms we evaluate their accuracy for both fine- and coarse-scale curvature analysis. We describe methods for normalization of the estimates for joint size which enables a comparison across subjects, and demonstrate the abilities of the methods to distinguish between healthy and OA populations using the KL index, since it is the established definition of the degree of OA.

## 11.2 Material and Methods

### 11.2.1 Population and Image Acquisition

The test subjects were between 22-79 years old with an average age of 56 years, with 59% females, and there were both healthy and osteoarthritic subjects according to the KL index. The data set consisted of MR scans and x-rays from 114 knees that were used for the evaluation of the method. In addition, 25 knee scans were used for the training of the automatic method. For reproducibility evaluation, 31 knees were re-scanned after approximately one week, making the total number of knee data sets 170.

Radiographs were acquired in the posterior-anterior position. From these x-rays, the tibial plateau width was measured and the KL index was determined. The KL index is a radiographic score ranging from 0-4 where 0 is healthy, 1 is considered borderline or mild OA, and KL greater than or equal to 2 is severe OA. In the test set consisting of 114 knees there were 51, 28, 13 and 22 scans

that had KL 0, 1, 2 and 3 respectively.

Knee Turbo 3D T1 scans were acquired using an Esaote C-Span low-field 0.18T scanner dedicated to imaging of extremities (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms), with approximate spatial resolution $0.8 \times 0.7 \times 0.7\text{mm}^3$. Slices were acquired in the sagittal direction. The medial tibial cartilage was analyzed in this study since OA is most often observed in medial compartments [38].

### 11.2.2 Cartilage Segmentation

As the first step in the automatic curvature analysis, all MR scans were segmented fully automatically directly in 3D by a supervised learning technique, where each voxel is assigned a probability of being cartilage or background based on prior knowledge from the training data using $k$ Nearest Neighbors ($k$NN) classifiers [55]. The classifiers were trained on 25 scans to find the features that best describe the different classes. The features that were selected were the image intensity, position, and local geometric features based on Gaussian derivatives.

### 11.2.3 Curvature Estimation

For a point on a curve in 2D, the osculating circle is uniquely defined by taking three neighboring points along the curve and let the arclength between them go to zero. The curvature is defined as the inverse of the radius of the osculating circle $\kappa = 1/\rho$ (see Figure 9.4). The curvature is also defined as the rate of change of the tangent direction $\theta$ of the curve with respect to the arclength, $\kappa = \frac{d\theta}{ds}$ [78]. For a surface in 3D, there are two principal curvatures for the surface, the maximal and minimal, and the mean curvature is the mean of the principal curvatures.

#### 11.2.3.1 Curvature Analysis by Mean Curvature Flow

The first curvature estimation method was implemented by first super-sampling the binary segmented images, then evolving them using mean curvature flow in the level set formulation [100][115]. The surface was evolved according to its mean curvature so that regions with high curvature initially will typically move towards lower curvature values as the flow progresses and in the end the object will shrink to one or more spheres and disappear. In the level set formulation the mean curvature flow is described by

$$\phi_t = \kappa_M |\nabla\phi| = \Big[ \bigtriangledown \big(\frac{\nabla\phi}{|\nabla\phi|}\big)\Big] |\nabla\phi|,$$

where $\phi$ is the level set representation of the surface.

The mean curvature on the articular surface was calculated throughout the evolution, and the quantitative curvature measure was obtained at the point during the flow that enabled best separation between groups as described in section 11.3.2 and in [51]. The super-sampling, making each voxel $5 \times 5 \times 5$ voxels, prior to the evolution was done to allow finer changes than the original voxel size. There was no smoothing involved in the super-sampling, so the flow will initially mainly reduce partial volume effects and can be considered a non-linear smoothing process. In late stages of the flow, the object loses its cartilage like form and may even change topology. Figure 10.4 illustrates how a cartilage sheet deforms during mean curvature flow.

#### 11.2.3.2 Curvature Analysis by Shape Model

The second curvature estimation method was based on a deformable shape model, an m-rep [108], which was fitted to each of the binary segmented images [30] as can be seen in Figure 10.2. The m-rep represents an object by a mesh of medial atoms, each associated with a position, radius and directions to the boundary. The shape models that were fitted to the cartilage consisted of a mesh of $4 \times 8$ medial atoms, from which 1425 boundary points were interpolated on the articular surface. The deformable shape model framework ensures a regular, smooth boundary which makes it well suited for coarse scale curvature analysis. We defined the load bearing (central) part of the articular surface as a region of interest and estimated the curvature locally in 5 by 5 neighborhoods of boundary points by comparing the angle between surface normals divided by the distance between them. This yields an approximation to the mean curvature of the surface at that point.

### 11.2.4 Normalization of Curvature Estimates

The data set consists of knees from both male and female subjects with varying weights and ages, therefore it was important to make sure that joint size was not being modeled when doing inter-subject comparisons. The curvature estimates were normalized by the tibial plateau width determined from x-ray. For global surface curvature this aims to factor out joint size from the analysis. However for curvature estimates targeting local curvature, this normalization might not be necessary since the size of lesions might not be related to joint size. We examined the effect of leaving out the normalization in Section 11.3.2, and it did not affect the resulting p-values noticeably.

In this study we have used the tibial plateau width measured from x-rays to normalize the curvature estimates. As long as x-rays are gold standard for cartilage quantification, they will most likely be acquired along with MRI in feasibility studies such as this. Eventually one would want to avoid the x-ray acquisition and rely solely on MRI, however as yet, we have not developed a method for measuring the plateau width automatically from MRI, whereas the plateau width is routinely measured by radiologist for the x-ray analysis.

### 11.2.5 Analysis on Digital Phantoms

In order to determine the accuracy of the curvature estimation methods, a number of phantoms with known curvature were generated. Five phantoms were created by connecting $\pi/2$ parts of cylinders with varying radii as can be seen in Figure 11.1. The mean curvature for the different phantoms were 0.0073, 0.025, 0.05, 0.075, and 0.10 voxel$^{-1}$, which corresponds to 7.35, 25, 50, 75 and 100m$^{-1}$, assuming voxels with a side length of 1mm. The curvatures of the phantoms were estimated using both the flow-based and the shape model-based methods.

For the flow-based method, 5 spheres with varying radii were also analyzed and the automatically estimated curvature values were compared with the analytically calculated ones. The mean curvature of the spheres were $100, 200, 250, 333$, and 500m$^{-1}$ assuming a voxel side length of 1mm.

For the shape model-based method, 5 thin curved plates similar to the first phantom in Figure 11.1 with known curvature were generated. The curvature of these plates (assuming voxel side length 1mm) ranged from $11 - 21$m$^{-1}$ ($11, 14, 15, 17$, and 21m$^{-1}$). The reason spheres were chosen for the flow-based method was that a sphere is an object with known curvature that is constant across the surface. The reason that curved plates are used instead of spheres for the shape-based method was that the shape model was trained on similar objects. For the curved plate phantoms the curvature was evaluated on the central part of the plate.

## 11.3 Results

### 11.3.1 Evaluation on Phantoms

The results of estimating the curvature with both the flow-based and the shape model-based methods on the same five phantoms with increasing known curva-

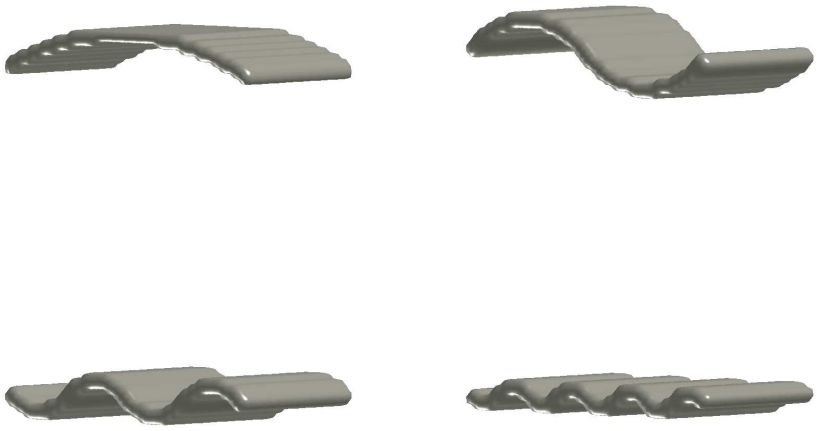


**Figure 11.1:** Phantoms for comparison of the two curvature estimation methods consisting of $\pi/2$ cylinder parts. The mean curvature for the four shown phantoms are 0.0073, 0.025, 0.05, and 0.10 voxel$^{-1}$

ture values can be seen in Figure 11.2. For the flow-based method, the mean curvature was calculated using Gaussian derivatives with scale 0.9, since this generated curvature values closest to ground truth. The results show that the shape model-based curvature is more accurate for estimating low curvature values. This is a result of the regularization properties of the shape model which yields a smooth surface. The flow-based method is more accurate for high curvature values, where the accuracy of the curvature estimation is limited by the voxel size. The spherical phantom with the highest curvature value is made from a sphere with a radius of 5 voxels.

For the curvature of the 5 spheres estimated by the flow-based method, the resulting mean absolute pairwise difference was 1.1% with a mean coefficient of variation (CV) of 0.8% compared to ground truth.

The mean absolute pairwise difference was 4.8% and the mean coefficient of variation was 3.4% for the curvature of the thin plates estimated by the shape model method after a bias in terms of a systematic overestimation of 11% was removed.

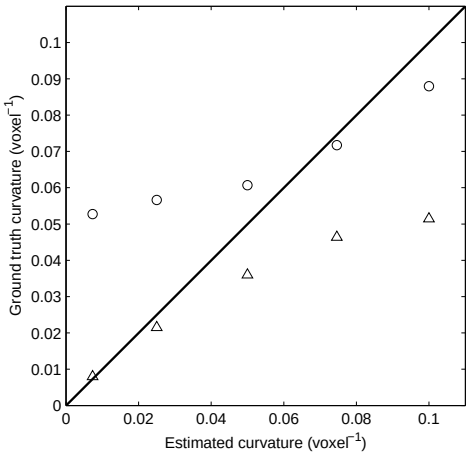



**Figure 11.2:** Phantoms for comparison of the two curvature estimation methods. Triangles represent shape model-based curvature, circles flow-based curvature. Ground truth values are on the y-axis and estimated values on the x-axis.

### 11.3.2 Relation Between Curvature and Signs of OA

In order to determine in what stage during the evolution the curvature should be measured, we estimated the curvature on the articular surface throughout the curve evolution. We examined where in the flow the curvature estimate could differentiate between healthy and OA populations for the 114 evaluation scans, as in Figure 11.3. After 10 iterations we achieved maximum separability with a p-value of 0.00006. The mean curvature at that time was $360\text{m}^{-1}$ ($\pm 32\text{m}^{-1}$ std), which for a sphere corresponds to a radius of 3mm. The flow-based method was not at any stage of the flow able to discriminate between healthy and borderline OA (KL 1) populations. Initially partial volume effects in the super-sampled images were reduced, later in the flow the values stabilize around $150\text{m}^{-1}$. The

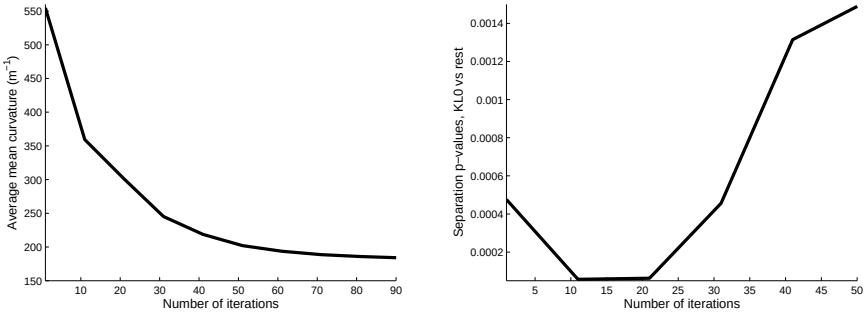



**Figure 11.3:** Left: the average mean curvature of the articular surface as a function of time. Right: p-values from t-test between healthy and OA groups as a function of time.

scale of the Gaussian derivatives for the mean curvature calculation should be increased proportional to the super-sampling (in this case $0.9 \times 5$) for best accuracy, however we also evaluated the effect of less regularization and found that best separation between groups was obtained with the scale 1.5, from which the results are listed above. Using a scale of $0.9 \times 5$ gives values in the same range as when the flow with scale 1.5 has stabilized ($150\text{m}^{-1}$) already in the beginning of the flow, but doesn't achieve as high separation between groups. This suggests that increasing the scale with the super-sampling might give more accurate estimates in an early stage of the flow, but leaving more of smoothing to the curvature flow instead of the Gaussian derivatives preserves the curvature information better. The inter-scan reproducibility of the curvature for the 31 knees scanned twice gave a mean coefficient of variation (CV) of 2.3%, and a relative absolute difference of 3.3%.

The curvature estimation on the boundary points of the m-rep produced a

**Table 11.1:** The mean curvature values for the flow-based (Flow) and shape model-based (Shape) curvature estimation methods in different populations based on KL. Values in parenthesis are standard deviations (std).

| Method | KL 0 | KL 1 | KL 2 | KL 3 |
|---|---|---|---|---|
| Flow ($m^{-1}$) | 341 ($\pm$29) | 339 ($\pm$29) | 390 ($\pm$46) | 410 ($\pm$46) |
| Shape ($m^{-1}$) | 40 ($\pm$5.4) | 45 ($\pm$7.3) | 47 ($\pm$7.1) | 54 ($\pm$13.6) |

mean curvature of $45m^{-1}$ ($\pm 20m^{-1}$ std) on the load bearing part of the articular cartilage surface, which corresponds to a sphere with a radius of 22mm. The shape model based method could separate the healthy from the OA group ($p = 0.000004$) and could even separate healthy from borderline OA ($p = 0.005$). For the reproducibility evaluation on the 31 scan-rescan pairs, the mean coefficient of variation was 6.4% with a relative absolute difference in measurements of 9.0%. The outcome of the two methods are listed in Tables 11.1 and 12.3. For comparison, we have also measured the medial tibial cartilage volume determined directly by the segmentations, and the thickness determined using the method in [30] for the same data set.

**Table 11.2:** Results comparing the performance of the two curvature methods, and cartilage volume and thickness. Mean curvature with standard deviation and maximum and minimum values are listed for the 114 evaluation scans, along with reproducibility estimates; mean CV and relative absolute difference (AD), for the 31 scan-rescan pairs. P-values for separation between the healthy population and OA population (KL $> 0$) and borderline OA (KL 1) are also listed.

| | Flow | Shape | Volume | Thickness |
|---|---|---|---|---|
| Mean | $360m^{-1}$ | $45m^{-1}$ | $1900mm^3$ | 1.30mm |
| std | $320m^{-1}$ | $9.6m^{-1}$ | $440mm^3$ | 0.18mm |
| max | $465m^{-1}$ | $100m^{-1}$ | $3100mm^3$ | 1.70mm |
| min | $302m^{-1}$ | $30m^{-1}$ | $580mm^3$ | 0.79mm |
| CV (%) | 2.3 | 6.4 | 5.3 | 4.6 |
| AD (%) | 3.3 | 9.0 | 7.5 | 6.4 |
| p (KL $> 0$) | 0.00006 | 0.000004 | 0.004 | 0.002 |
| p (KL 1) | 0.75 | 0.005 | 0.27 | 0.32 |

## 11.4 Discussion

In this paper we have presented two methods for automatically estimating the curvature on the articular cartilage surface. One method is based on mean curvature flow, therefore it is essential that the curvature is measured early in the evolution while the object still resembles cartilage. The other method is based on an m-rep shape model, where the curvature is approximated by taking the angle between neighboring boundary point normals divided by their distance.

Ideally, one would like to find the ground truth curvature values of the cartilage sheets for the accuracy study. For an *in vivo* study one option would be to scan patients just before they undergo total joint replacement. But even if one would have access to surgically removed tissue for comparison with the MR-based measures, it would be very difficult to accurately measure the ground truth curvature on the cartilage surface. In this study we have developed methods for curvature estimation and evaluated the accuracy on synthetic images with known curvature. A further step would be to create objects with known curvature and acquire MRI scans of them with the same protocol as for the knees, then evaluate the scans with the curvature estimation techniques, however that is beyond the scope of this paper. Adding different kinds of noise to the current phantoms would be a more direct method for further accuracy analysis.

The phantom study shows that the coarse scale curvature range can most accurately be captured by the shape model method, due to the regularization property of the model which results in a shape representation with a smooth surface. This is also the reason for its inability to capture fine-scale curvature (high curvature values). Fine-scale curvature values on the other hand are better estimated by the flow-based method, with the upper limit of how high curvature values that can be estimated depending on the voxel size. After establishing in what range the methods performed best, we created specific phantoms for them for comparison with ground truth. The flow-based and the shape model-based curvature estimates were accurate, with a relative absolute difference of 1.1% and 4.8% respectively. Future work involves studying the effects of supersampling on the accuracy of the flow-based curvature estimate, and the current curvature flow parameters are selected for best separation between KL population and not for best accuracy.

Both curvature estimation methods can separate healthy from OA populations ($p = 0.00006$ for the flow-based method and $p = 0.000004$ for the shape model), and the shape based curvature can even distinguish healthy from borderline OA groups ($p = 0.005$). This suggests that the coarse scale curvature estimate could potentially become useful as a disease prediction marker. The flow-based

method detects statistically significant differences between groups particularly at high curvature values but cannot detect differences between healthy and borderline OA populations which suggests that the surface undergoes changes in fine scale curvature during the disease progress.

Due to the complex nature of OA there is probably not one marker that can cover all the different stages of the disease. The coarse-scale curvature is aims to detect global structural changes, which could be related to an increase in joint incongruity in early OA, and the fine-scale curvature could be related to fibrillation and lesions occuring in early to moderate OA. Volume decrease and thinning of the cartilage typically occur in moderate to severe OA, and we have shown that the curvature estimates compared favorably to cartilage volume and thickness measurements on the same data in terms of separability between populations.

We suggest that the coarse-scale surface curvature is related to joint incongruity, with the motivation that more curved surfaces are more likely to mismatch as the joint bends. A further development of the method will be to further analyze the relationship to joint incongruity by incorporating the femoral cartilage sheet in the analysis, and determine how the curvatures of the two cartilage surfaces relate to each other.

We compared our results to the Kellgren-Lawrence index, a well-established OA symptom grading from radiographs. X-rays can in many ways can be considered inferior to MRI, however it is important to establish that the markers from MRI can perform at least as well as the gold standard OA measure can. The methods from MRI are fully automatic, non-invasive, and have high reproducibility (no intra-scan and low inter-scan variability), which already are advantages compared to measures based on radiographs. We also compared the results to the cartilage volume from the same segmentations, and showed that curvature estimates enabled better separation between populations.

With our two methods of analyzing the surface curvature on the medial tibial cartilage compartment we have detected highly significant differences between healthy and OA populations, which suggests that surface curvature on different scales can be related to OA symptoms.

Chapter 12

# Automatic Quantification of Global and Local Articular Cartilage Surface Curvature: Markers of Smoothness and Congruity?

*Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Morten A. Karsdal, Paola Pettersen, and Claus Christiansen*

**Abstract**

The objective of this study was to quantitatively assess the surface curvature of the articular cartilage from low-field magnetic resonance imaging (MRI) data, and to investigate its role in populations with varying radiographic signs of osteoarthritis (OA), cross-sectionally and longitudinally. The curvature of the articular surface of the medial tibial compartment was estimated both on fine and coarse scales using two different automatic methods which are both developed from an automatic 3D segmentation algorithm. Cross-sectionally ($n = 288$), the surface curvature for both the fine- and coarse-scale estimates were significantly higher in the OA population compared to the healthy population, with $p < 0.001$ and $p << 0.001$ respectively. For the longitudinal study ($n = 245$), there was a significant increase in fine-scale curvature for healthy and borderline OA populations ($p < 0.001$), and in coarse-scale curvature for severe OA populations ($p < 0.05$). Fine-scale curvature could predict progressors using the estimates of those healthy at baseline ($p < 0.001$). The inter-scan precision was 2.2 % and 6.5 % (mean CV) for the fine- and coarse scale curvature

measures respectively. The results showed that quantitative curvature estimates from low-field MRI at different scales could potentially become biomarkers targeted at different stages of OA.

Key Words: Articular cartilage, automatic curvature analysis, smoothness, congruity, osteoarthritis, low-field MRI.

## 12.1 Introduction

Osteoarthritis (OA) is one of the major health issues among the elderly population and is associated with a large socioeconomic impact on health care systems [68]. Typical characteristics are fibrillation and a general progressive loss of articular cartilage, remodeling and sclerosis of the subchondral bone, the formation of subchondral bone cysts, and marginal osteophytes [59]. Standard clinical treatment of OA today does not reverse the cartilage degeneration [43], hence disease markers for early stages, before the disease progression becomes irreversible, could be useful for detecting patients that would benefit from treatment aiming to prevent further OA progression (such as exercise or gait adjustment by e.g. wedged insoles or similar techniques). Biomarkers targeted for different stages of OA could also be useful for clinical studies of potential disease modifying drugs for OA (DMOAD's). To be able to keep population sizes and time duration in such studies under control, this leads to a search for disease markers that are precise and can detect signs of OA and OA progression.

Macroscopic characteristics, with joint space width from x-rays as gold standard, have long been regarded as measures of outcome in therapeutic trials in OA. But in recent years, magnetic resonance imaging (MRI) has become considered the leading imaging modality for direct, non-invasive assessment of the articular cartilage [60], and MRI analysis can be used to detect cartilage deterioration quantitatively [104],[124]. For a thorough review of MRI scan protocols for knee OA assessment, see Peterfy et al. [105].

Typical quantitative disease markers for OA from MRI data is the articular cartilage volume and thickness [124],[14],[20],[123]. Thickness measures can combined with methods for finding correspondences in anatomy give a more localized measure of cartilage degeneration, and be more sensitive to changes especially in load bearing parts [30],[138],[58].

Elevated biomechanical stress may cause or accelerate the progression of joint degeneration [12],[13],[62], and this may be related to changes in the congruity of the joint, which describes the extent to which contacting joint surfaces match each other. Local irregularities on the cartilage surface, or the smoothness of the surface, could be related to focal lesions and fibrillations.

In this work we investigated the cartilage surface curvature automatically on fine and coarse scales, and evaluated their abilities to detect differences between populations with different severity in signs of OA, both cross-sectionally and longitudinally. Previously we have evaluated coarse- and fine-scale curvature for 114 scans and found significant differences between healthy and OA populations for both estimates [51]. We have also established the accuracy of the two methods using synthetic data [54]. In this work we have extended the evaluation

to a larger baseline data set and, more importantly, presented results from a longitudinal study.

## 12.2 Methods

### 12.2.1 Population

A population of 159 test subjects was prospectively selected. The subjects were randomly selected such that the population had an even distribution between sexes and across ages and such that the number of healthy and subjects with varying degree of OA symptoms were approximately equal. The majority of the subjects were invited from address lists, but the population also contained volunteers reporting pain in at least one of the knees (and therefore likely to have OA). Subjects with previous knee joint replacement, inflammatory arthritis or presenting any contraindication for MRI examination were excluded prior to the study.

At baseline, a total of 302 knees were examined. Of these, 14 scans were excluded prior to evaluation, 10 due to inadequate x-ray data and 4 due to inadequate MRI acquisition results, hence 288 knees were used for the evaluation of the methods. Of these knees, 245 were scanned again in a follow-up study approximately 21 months after the first scan occasion. 31 knees were scanned approximately one week after baseline for a reproducibility study of the methods. Details of the population are listed in Table 12.1. Besides this data set was a set of 25 scans for the purpose of training automatic methods, they were selected to have a fairly even distribution of OA symptoms.

**Table 12.1:** The population at baseline, follow-up, and the scan/rescan set, listing minimum - maximum and (mean) values. N represents number of scans, and Np for number of progressors (knees with higher KL at follow-up).

| | Baseline set | Follow-up set | Scan/rescan set |
|---|---|---|---|
| N | 288 | 245 | 31 |
| Age | 21 – 81, (56) | 21 – 81, (56) | 26 – 75, (61) |
| BMI | 19 – 38, (27) | 19 – 38, (27) | 21 – 34, (27) |
| Gender | 44% female | 42% female | 55% female |
| N KL 0-4 | 143, 88, 29, 23, 1 | 126, 78, 23, 17, 1 | 11, 13, 2, 5, 0 |
| Np KL 0-4 | | 25, 6, 3, 0, 0 | |

All participants signed approved information consent and the study was carried out in accordance with the principles of the Helsinki Declaration II and European Guidelines for Good Clinical Practice [132]. The study protocol was approved by the local Ethical Committee.

### 12.2.2 X-Ray Protocol

Radiographs were acquired in the posterior-anterior position with focus film distance 1.0 m and a tube angulation of 10 degrees. Subjects were standing in a weight bearing position with knees slightly flexed and feet rotated externally using SynaFlex developed by Synarc. Radiographs were acquired in the posterior-anterior (PA) position, while the central beam was displayed directly to the mid point of the line passing through both popliteal regions. Radiographs of both knees were acquired simultaneously.

From these x-rays the Kellgren-Lawrence index (KL) was determined. The KL is a radiographic score ranging from 0-4 based on evidence of osteophytes and joint narrowing [73]. In the text we refer to KL 0 as healthy, KL 1 borderline or mild OA, KL 2 moderate, and KL greater than or equal to 3 as severe OA. The KL distribution is listed in Table 12.1. In addition, the width of the tibial plateau was measured - covering both medial and lateral compartments but excluding possible osteophytes. This is a measure of the size of the knee and it was used for normalization of the quantifications. The distribution of age and BMI in for different KL groups are listed in Table 12.2.

**Table 12.2:** The distribution of age and BMI of the different KL populations at baseline, listing minimum - maximum and (mean) values. N represents number of scans.

| KL | 0 | 1 | 2 | 3 + 4 |
|---|---|---|---|---|
| N | 143 | 88 | 29 | 23 + 1 |
| Age | 21 – 78, (48) | 37 – 81, (62) | 47 – 78, (67) | 58 – 78, (68) |
| BMI | 19 – 38, (25) | 19 – 37,(27) | 22 – 37, (29) | 23 – 34, (29) |

### 12.2.3 MRI Protocol

An Esaote C-Span low-field 0.18T scanner dedicated to imaging of extremities was used for the acquisition of a Turbo 3D T1 sequence (40° flip angle, $T_R$ 50 ms, $T_E$ 16 ms) for each knee of the test subjects. The subjects were scanned

in a lying position with no load-bearing, and there was no load-bearing prior to scanning. Approximate acquisition time was 10 minutes and the scan size, after automatically removing boundaries that contained no information, was $100 \times 170 \times 170$ voxels. The spatial in-plane resolution of the sagittal scans were $0.70 \times 0.70$ mm$^2$, with a distance between slices ranging between 0.70 mm - 0.94 mm depending on joint size, where the most common distance was 0.78 mm. A slice from one of the knees in the data set can be seen in Figure 1. We used a

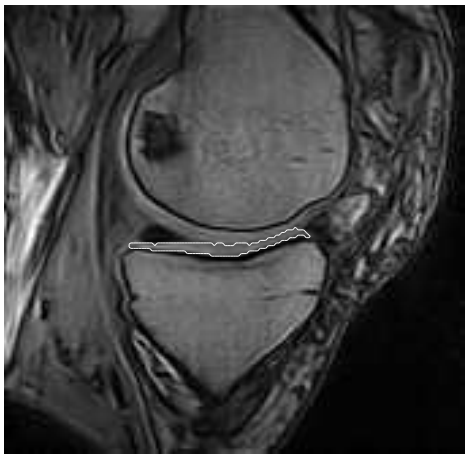


**Figure 12.1:** A slice from a low-field MRI scan, with the medial tibial cartilage automatically segmented.

3D sequence with nearly isotropic voxels since our image analysis methods are 3D, and since morphometric cartilage quantification benefits from near-isotropy [142]. This becomes particularly important for surface curvature estimation in 3D, where sampling of surface boundary points equally in all directions is essential. The medial tibial cartilage was analyzed in this study since OA is most often observed in medial compartments [38].

### 12.2.4 Cartilage Segmentation

As the first step in the automatic curvature analysis, all MR scans were segmented fully automatically directly in 3D by a supervised learning technique, where each voxel is assigned a probability of being cartilage or background based on prior knowledge using $k$ Nearest Neighbors ($k$NN) classifiers. The prior knowledge was a set of features selected to achieve best possible discrimination between the classes (medial femoral and tibial cartilage and background)

from the training data, and were the intensity and position in the image, and local structural features based on Gaussian derivatives. Classification outliers were automatically removed by selecting the largest connected component. Before segmentation, right knees were automatically mirrored around the center sagittal slice in order to manage all knees in the same framework. The automatic segmentation algorithm has shown good agreement with manual segmentations by a radiologists, and as good precision as a radiologist for cartilage volume and area [55]. A slice with automatically segmented medial tibial cartilage outlined can be seen in Figure 1.

### 12.2.5 Fine- and Coarse-Scale Curvature Analysis

An intuitive definition of surface curvature is how much the surface of an object deviates from being flat. We investigated both coarse- and fine-scale curvature since we believed that they could target different joint characteristics (see Figure 2) and could be sensitive to different stages of OA. Fine-scale curvature analysis is related to local changes of curvature values on the surface, which could be related to cartilage lesions and fibrillation. Coarse-scale curvature is associated with global changes across the cartilage surface and could be related to joint incongruity, since a varying curvature across the overall surface will likely lead to mismatching of the surfaces while bending the joint.

We analyzed the articular cartilage surface curvature using two different automatic methods, one based on mean curvature flow, and one based on a shape model [51]. Previously we have established the accuracy of the methods using synthetic data [54]. The shape model-based method was found to be more accurate for estimating coarse-scale curvature and the flow-based method estimated fine-scale curvature values better.

**Figure 12.2:** Illustration of how the curve to the left is perceived on a coarse-scale (top right) and fine-scale (bottom right) curvature curvature estimation.

### 12.2.6 Fine-Scale Curvature Analysis

The fine-scale curvature estimation method was implemented by evolving the segmented volumes according to their mean curvature [51], so that regions on the cartilage surface with high curvature values initially will typically move towards lower curvature values as the flow progresses. Curvature values were obtained after 10 iterations of the flow since we have previously found it to achieve best separability between healthy and OA populations [54]. These iterations performs some smoothing aimed to remove voxellation effects. Since the fine-scale curvature estimate targets surface smoothness, the curvature was evaluated only in the cartilage surface points on the superior surface.

### 12.2.7 Coarse-Scale Curvature Analysis

The second curvature estimation method was based on a deformable shape model, an m-rep [108], which was fitted to each of the binary segmented images [30]. The m-rep represents an object by a mesh of medial atoms, each associated with a position, radius and directions to the boundary. The shape models that were fitted to the cartilage consisted of a mesh of $4 \times 8$ medial atoms, from which 1425 boundary points were interpolated on the articular surface. The deformable shape model framework ensures a regular, smooth boundary which makes it well suited for coarse-scale curvature analysis. We defined the load bearing (central) part of the articular surface as a region of interest, defined in the shape model coordinate system as the boundary points remaining after removing 20% of the boundary points on the medial side and 25% on the lateral side for the medio-lateral direction, and 15% on the anterior side and 20% on the posterior side of in the anterior-posterior direction. The curvature was estimated locally in 5 by 5 neighborhoods of boundary points by comparing the angle between surface normals divided by the distance between them. In Figure 3 the concept of comparing angles between normals for surface curvature estimation is demonstrated.

### 12.2.8 Statistical Methods

We evaluated the diagnostic performance of the curvature estimates by the ability to separate the healthy (KL 0) from OA (KL $> 0$) knees in the baseline data set using an unpaired t-test. The prognostic performance was evaluated by the ability of the baseline values to separate healthy non-progressors (KL 0 at baseline and follow-up) from progressors (KL 0 at baseline and KL $> 0$ at follow-up) using an unpaired t-test.

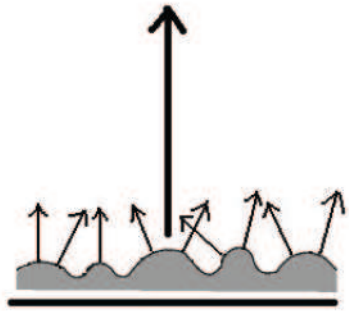


**Figure 12.3:** Curvature estimation using surface boundary point normals. For a flat surface (thick black line), the surface normal direction will be the same across the surface. Hence the angles between the normals will be zero hence the curvature will be zero. For a curved surface (gray area), the angles between the normals will vary and with larger the angles, larger curvature.

Longitudinal changes are evaluated using paired t-tests for the corresponding knees in the baseline and follow-up data sets. The standardized response means (SRM), defined as the mean change divided by the standard deviation of the change, were calculated.

The reproducibility was evaluated using inter-scan variability for the scan/rescan data set by mean and root mean squared coefficient of variation (CV) and relative absolute difference (RD).

In order to investigate a potential correlation between the curvature and age, we examined the linear correlation between age and curvature for the subjects with KL 0 at baseline and corrected linearly for age and gender where appropriate.

## 12.3 Results

### 12.3.1 Evaluation of Baseline Data Set

The curvature estimates from both methods were normalized by the tibial plateau width determined from x-ray in order to enable inter-subject comparisons. Curvature is measured in the unit meter inverse, therefore a measure with unit meter is preferred for normalization. We use the tibial plateau width from x-rays which is routinely measured in our clinical studies

The mean fine-scale curvature was 410 $m^{-1}$ ($\pm 50$ $m^{-1}$ std) for the baseline data

set. The estimate could successfully separate between healthy (KL 0) and OA (KL $> 0$) populations ($p = 2 * 10^{-4}$), but was unable to distinguish healthy from borderline OA (KL 1) knees ($p = 0.6$). The mean coarse-scale curvature estimate was 43 $m^{-1}$ ($\pm 8 m^{-1}$ std). It could distinguish between healthy and OA populations ($p = 7 * 10^{-10}$) and between healthy and borderline OA knees ($p = 1 * 10^{-4}$). Baseline results can be seen in Figure 12.4.

For the coarse-scale curvature, the linear correlation coefficient was 0.05 ($p = 0.6$), therefore we conclude that no age correction was necessary. For the fine-scale curvature estimate, the correlation coefficient was 0.19 ($p = 0.02$). Therefore, we made a linear correction for age and gender. The resulting p-value for separating KL 0 from KL $> 0$ was $1.3 * 10^{-4}$, which is comparable to the value obtained when age was not factored out.

In order to examine what characteristics the curvature estimates capture, the cartilage sheets with maximum respectively minimum values for the coarse- and fine-scale curvature estimates were plotted (Figures 5 and 6). It can be seen that the cartilage sheet with minimum coarse-scale curvature was more flat compared to the sheet with maximum curvature in Figure 5. The knee with minimum coarse-scale curvature value was 30 $m^{-1}$ (the knee had KL 1) and the (KL 3) knee with maximum value was 100 $m^{-1}$. The corresponding values of fine-scale curvature for these knees were 330 $m^{-1}$ and 400 $m^{-1}$ respectively. For the fine-scale curvature it was difficult to visualize differences, in Figure 6 it can be seen that the knees with maximum and minimum fine-scale curvature, 560 $m^{-1}$ (a KL 3 knee) and 310 $m^{-1}$ (a KL 0 knee) respectively, have similar coarse-scale curvature values (43 $m^{-1}$ and 58 $m^{-1}$).

### 12.3.2 Evaluation of Longitudinal Data

In the follow-up study, there was a significant increase of fine-scale curvature per year in the groups at baseline with KL 0 ($p = 3 * 10^{-5}$), KL 1 ($p = 2 * 10^{-5}$), and KL $\geq$ 3 ($p = 0.02$). For these groups the SRM's were 0.4, 0.5, and 0.6 respectively. The coarse-scale curvature showed a significant increase per year only for the severe OA group (KL $\geq 3$), with a p-value of 0.05, and a SRM of 0.4. The fine- and coarse-scale curvature changes in the longitudinal study can be seen in Figure 7. The fine-scale curvature could separate future progressors from non-progressors using the values at baseline ($p = 9 * 10^{-4}$).

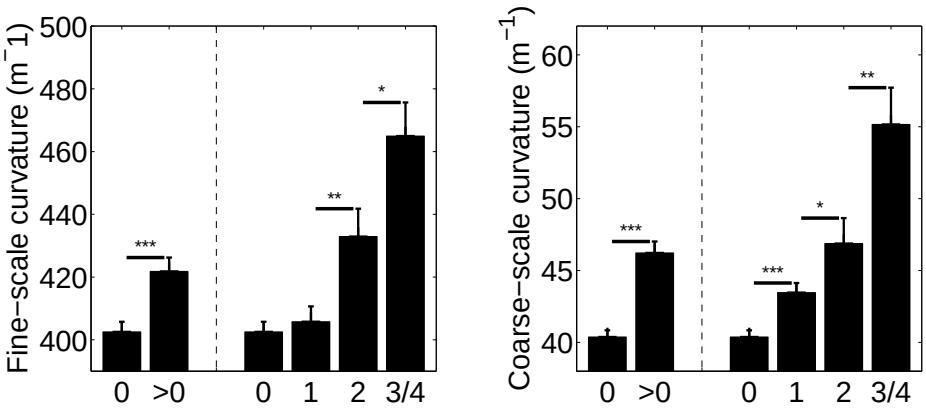



**Figure 12.4:** Mean and standard error of mean (SEM) for the fine- (left) and coarse-scale (right) curvature estimates for different $KL$ populations at baseline. There is a significant separation between $KL0$ and $KL > 0$ populations for both methods ($p = 2{*}10^{-4}$ and $p = 7{*}10^{-10}$ respectively). Smoothness values could not separate between $KL0$ and $KL1$ populations ($p = 0.44$) but congruity values could ($p = 1 * 10^{-4}$).

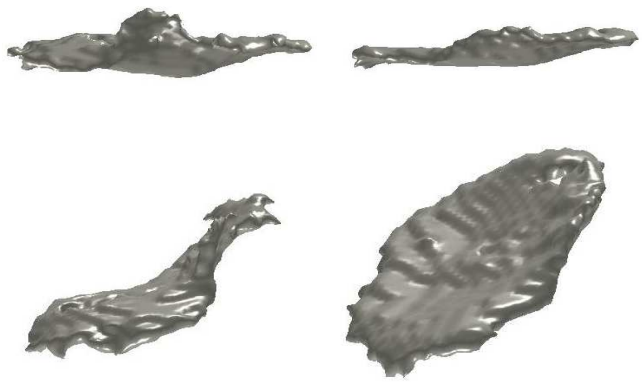


**Figure 12.5:** The upper two figures are the tibial cartilage sheets with maximum (left) and minimum (right) coarse-scale curvature in the baseline data set seen from a sagittal view. The cartilage sheet with minimum curvature (a KL 1 knee) is flat compared to the sheet the maximum curvature (a KL 3 knee). The last two figures are the same cartilage sheets but from a different view.

**Figure 12.6:** Left/right: the tibial cartilage sheets with maximum/minimum fine-scale curvature (from knees with KL 3/ KL 0 respectively) seen from a sagittal view. The sheets have similar coarse-scale curvature values and no observable differences in the overall shape.

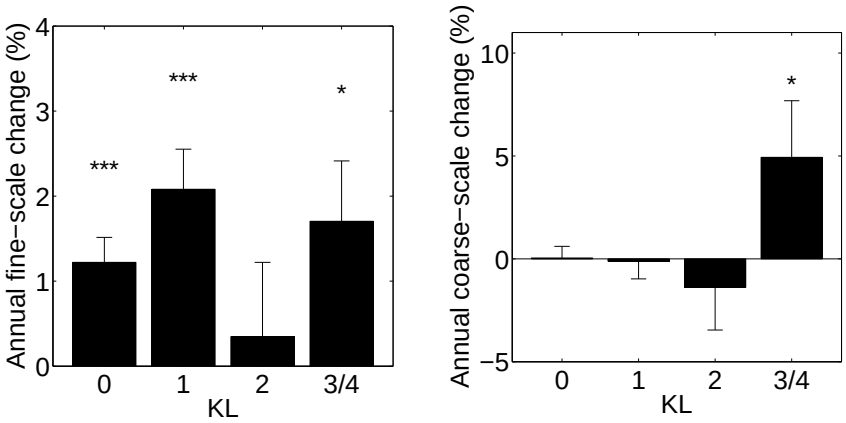



**Figure 12.7:** Mean and SEM annual change in fine- (left) and coarse-scale (right) curvature. For fine-scale curvature there are significant changes for the $KL0-1$ populations, and for the coarse-scale curvature there are significant changes for the $KL \geq 3$ populations.

### 12.3.3 Inter-scan Precision Evaluation

The fine-scale curvature for the 31 knees in the scan/rescan data set had a mean CV of 2.2%, a root mean squared CV of 2.8% and a RD of 3.1%. This can be compared to the annual increase in curvature, which is approximately 2%. For the coarse-scale curvature, the mean CV was 6.5%, the root mean squared CV was 8.7% with a RD of 9.2%. This can be compared to the annual increase in curvature, which is approximately 5% for the KL $> 2$ group. The results of the reproducibility evaluations are summarized in Table 12.3. R1:4

**Table 12.3:** Summary of the evaluation results at baseline related to diagnosis (D), inter-scan reproducibility (R), and prognosis (P) studies. Longitudinal changes are shown in Figure 12.3.2. Baseline results show p-values from separation of healthy/OA populations, precision is shown as mean CV and pairwise difference, and prognosis as p-value between future progressors vs. non-progressors at baseline.

| | | Congruity | Smoothness |
|---|---|---|---|
| KL 0 vs. KL $> 0$ | D | $p = 7 * 10^{-10}$ | $p = 2 * 10^{-4}$ |
| KL 0 vs. KL 1 | D | $p = 1 * 10^{-4}$ | $p = 0.6$ |
| Mean CV | R | 6.5% | 2.2% |
| Diff | R | 9.2% | 3.1% |
| KL 0 | P | 0.2 | $9 * 10^{-4}$ |

## 12.4 Discussion

### 12.4.1 The Relation Between Curvature and Signs of OA

The curvature values can be seen as the inverse of the radii of circles. These radii are on average approximately 25 mm and 2 mm respectively for the coarse- and fine-scale curvature estimates. Hence the two measures analyze the curvature on different scales: the coarse-scale curvature targets more global cartilage shape and the fine-scale curvature targets local characteristics on the cartilage surface.

Fine-scale curvature is related to higher and more localized curvature values on the cartilage surface. A low fine-scale curvature across the surface describes a smooth surface, hence fine-scale curvature is inversely proportional to surface smoothness. Our results showed that the fine-scale curvature could separate healthy from OA knees ($p = 2 * 10^{-4}$) at baseline, and there was a significant

longitudinal increase for the healthy and borderline OA populations. The increase of KL 0 knees can be explained by the number of volunteers with known knee problems but not always detectable signs of OA at baseline. It showed potential as an OA prediction biomarker ($p = 9 * 10^{-4}$). The lower longitudinal increase in the moderate to severe populations could be related to fibrillations and lesions which appear in early stages of OA, and once developed the surface cannot become more 'unsmooth' at later stages of OA.

The coarse-scale surface curvature may be related to joint incongruity since high curvature variation across the surface more likely leads to mismatch as the joint bends. At baseline, healthy knees had on average lower values compared to healthy ($p = 7 * 10^{-10}$) and even the borderline OA knees ($p = 1 * 10^{-4}$). This could be related to a decrease of the 'flatness' of the cartilage sheet with increasing severity in signs of OA. There was a significant longitudinal increase only for the severe OA group ($p = 0.05$), which could correspond to the collapse of the joint in late stages of OA. A further development of the method will be to examine the congruity by extending the curvature analysis to incorporate also the femoral cartilage sheet. This way the relation to incongruity can be further evaluated by developing a method for examining how well the two cartilage surfaces match each other, e.g., by computing a congruence index between two contacting surfaces as in [5].

The MRI sequence employed will provide T1-, T2-, and T2*-weighting. Areas of cartilage damage may therefore exhibit abnormal signal intensity. We have in a previous work established the performance of the segmentation algorithm compared to manual segmentations by a radiologist [55], and the automatic algorithm is trained to recognize cartilage based on manual delineations by a radiologist. Since the radiologist includes cartilage with abnormal signal intensity, so will the automatic algorithm. We are measuring the curvature on the automatic segmentations, and although there is no guarantee that these segmentations include every cartilage voxel with decreased T2, it could be argued that high curvature of the segmentation caused by damaged cartilage will also an indicator of OA. Future work involves evaluating what the curvature estimates models anatomically, possibly by relating them to a radiological reading of cartilage lesions.

### 12.4.2 Related Work

Considering previous studies of curvature analysis of the cartilage, Terukina *et al.* [127] performed an in-vitro study of the curvature in 2D by slicing the knee joint through the sagittal plane and fitting a circle to three equidistant points within 1cm on the cartilage surface then taking the inverse of the radius for the curvature. They found an average curvature of $4.4\text{m}^{-1}$ for the femoral condyle

in their study intended for cartilage replacement, emphasizing the importance of cartilage congruity in such interventions.

Hohe *et al.* [65] have analyzed the curvature of knee cartilage surfaces from MRI as an incongruity measure, by first semi-automatically segmenting the cartilage slice-by-slice then estimating the principal curvatures locally from a b-spline interpolation on a $5 \times 5$ neighborhood of surface points 6 mm apart. They found a mean curvature of 29.6 $m^{-1}$ ($\pm 9.9$ $m^{-1}$ std) for the tibial medial cartilage surface and $-0.9$ $m^{-1}$ ($\pm 3.8$ $m^{-1}$ std) for the central part of the same surface, with inter-scan reproducibility values up to $4.7 m^{-1}$ std for 4 repeated measurements on 14 healthy subjects. They also measured a congruence index based on the work of Ateshian *et al.* [5], where two contacting surfaces are transformed into an equivalent system consisting of a plane and the difference of the profiles of original surfaces.

It is clear that curvature estimation is highly dependent on the modeling, since very different results were obtained with our two curvature estimation techniques. The coarse-scale measure gives an average surface curvature of 42 $m^{-1}$ ($\pm 9$ $m^{-1}$ std), which is comparable to the values of Hohe et al. It corresponds to higher curvature values than Terukina et al. though they examine the femoral cartilage which bends more than the tibial. However their method fitting a circle to three points within 1 cm is quite crude and is restricted to curvuature analysis in one plane. The average fine-scale curvature in our study was 410 $m^{-1}$, which is higher than in related work and targets finer scales.

### 12.4.3 The MRI Sequence

We use a standard gradient echo (GE) sequence for cartilage with a modified flip angle (FA). The advantages of using a GE sequence are that the gradient can restitute the phase faster than the radio frequency (RF) pulses of 180°, and the minimum TE is shorter than in the sequences with spin echo pulses therefore TR and thus examination time can be reduced. The TR can also be reduced without producing saturation since it is possible to use inclination angles lower than 90°, for which the total recovery of the longitudinal magnetization occurs earlier. The lower FA leads to a smaller transverse component of magnetization than in the spin echo sequences where the total longitudinal magnetization is transformed in transverse magnetization, thus a signal is induced containing both T1 and T2 information. Lower FA achieved better contrast by higher T2 weighting. A disadvantage of GE is the lack of compensation for homogeneity disturbances of the magnetic field which in some cases can lead to magnetic susceptibility artifacts.

The use of a dedicated low-field MRI has drawbacks related to image quality

with lower resolution and more difficulties in incorporating features such as fat suppression. However the main advantages are cost-effectiveness with much lower cost per scan, lower installation and maintenance costs, and higher patient comfort without claustrophobic feelings and minimal noise level. So far there has not been any validation of quantitative cartilage measures from a low-field scanner compared to ground truth, but if a low-field scanner can be used for quantitative articular cartilage assessment, costs for making clinical studies could be reduced significantly. If manual labor is connected with the analysis and quantification of MRI data in clinical studies, one more cost factor is introduced. Recently, we published a fully automatic framework for cartilage quantification [31], with disease markers such as volume, thickness, and surface area showing significant differences between healthy and OA populations which suggests that low-field MRI analysis can become a useful tool in clinical studies of OA. In that study as well, we compare findings from low-field MRI data with the Kellgren-Lawrence index which is a well-established score for signs of OA. Future work involves a study of this fully automatic cartilage assessment framework on high-field MRI data.

### 12.4.4 Summary

The progression of OA is comprised of several very different stages. Therefore, it is likely that no single future treatment and no single biomarker is suitable for all stages of OA. Consequently, each potential biomarker should be evaluated for which stages of OA it may be suitable for. In this paper we have presented methods for automatically estimating fine- and coarse-scale curvature of the articular cartilage surface and compared the results to radiographic signs of OA. The results showed that the coarse-scale curvature could separate healthy from OA knees at baseline and even distinguish healthy from borderline OA knees, and may therefore be useful for diagnostic purposes. Coarse-scale curvature may be related to global characteristics of the cartilage surface and to joint congruity. The fine-scale curvature demonstrated significant longitudinal changes in healthy and borderline OA knees, and could thus become useful as a marker of progression at these stages. It also showed potential as an early prognosis marker for use in study population selection. Fine-scale curvature captures local changes on the cartilage surface, possibly related to fibrillation and lesion development and thus to surface smoothness.

In summary, the curvature quantifications could be useful in clinical studies for population selection, as diagnosis and prognosis markers, and for monitoring of disease progression. In particular, they could be of interest when studying early stages of OA, before the disease progression becomes irreversible.

Chapter 13

# Conclusions and Perspectives

This thesis presents a framework for quantitative evaluation of the articular cartilage in knee MRI data. The automatic segmentation method is the foundation of this framework, and challenges such as thin structure of the cartilage and large shape variations in subjects with different OA severity is overcome by the use of a statistical classifier with features capturing the local structure for each class. The automatic segmentation algorithm is shown to perform in agreement with a radiologists gold standard segmentation. Quantitative measures based on the segmentation are shown to have as good inter-scan reproducibility as a human expert. The measures also show (statistically) highly significant differences between populations with and without signs of OA, and they show statistically significant longitudinal changes for some of these populations. To our knowledge, no other fully automatic method for segmentation of articular cartilage can demonstrate such results.

Combining statistical classification and geodesic active regions is examined for automatic segmentation of the myocardium in late contrast enhanced cardiac MRI, a challenging task due to ambiguous boundary information and local intensity variations. The statistical classifier deals with local intensity variations and a global shape model is incorporated in order to cope with ambiguous region information and incorporate spatial context in the classification. Though the data set is rather small and further evaluation is necessary to determine how the method can be used in relation to catheter ablation interventions and clinical

studies of scar tissue, the current results are encouraging. Level set methods can naturally deal with topology changes, something that can occur for cartilage in late stages of OA. Therefore it would be of interest to implement the combined method on cartilage as well. However the room for improvement compared to gold standard segmentation is not large for the cartilage segmentation; when using statistical classification on its own the volume overlap compared to gold standard segmentations is close to the intra-scan volume overlap of the radiologist. But even if there is not much improvement compared to gold standard segmentations to be gained, it is possible that the segmentation-based quantitative measures could become more reproducible. The shape term described in the geodesic active regions framework might not have much influence due to large shape variations in the training data, however an alternative statistical shape model or atlas could contribute with spatial context to the statistical classifier which could resolve ambiguities in feature space using the spatially varying classification framework. Since m-reps are already used in the cartilage evaluation framework, that is an option worth considering.

The cartilage surface curvature is the primary focus among potential biomarkers for OA in this thesis, and it is evaluated on high and low scales. The curvature estimation methods described in this thesis have significantly different means in populations with signs of OA compared to populations without, which demonstrates that surface curvature on both high and low scales have potential as biomarkers for OA. Unlike cartilage volume, it is yet to be established what the curvature models anatomically. The high-scale curvature is related to a 'flatness' of the cartilage sheet, and it could be linked to joint incongruity, assuming that surfaces with more high-scale curvature are more likely to mismatch as the joint bends. However joint congruity is a measure of how well surfaces match each other, and in order to evaluate how it relates to high-scale curvature, contacting cartilage compartments and the dynamics of a bending joint need to be studied. A possible measure of joint incongruity could be to compare the principal curvatures and their corresponding directions for opposing cartilage surfaces. The fine-scale curvature can be related to a 'smoothness' of the cartilage surface and thus to focal lesions and fibrillation. However, clear differences between a 'smooth' and an 'unsmooth' cartilage sheets could not be detected by visual inspection. Low-scale curvature could be evaluated by comparison to arthroscopy findings or an MRI-based lesion score, however that would require establishing some anatomical correspondence between the different modalities.

Though the biomarkers founded on the automatic segmentation algorithm are clearly related to well-established signs of OA, these signs are based on measures from x-rays which is far from an optimal imaging technique for studies of the articular cartilage. Currently, there are no definitions of OA from MRI but such definitions are likely to be developed as clinical studies concerning OA are gradually shifting from x-ray to MRI-based outcomes. How OA pro-

gression is assessed affects not only study duration and population size, which allows for treatment development at lower costs, but could also influence the current opinion of which factors that are linked to the progression. As current research conclusions are based on radiographic techniques, they may undergo some revision as assessment of disease progression evolves.

Most MRI scanners for OA related imaging in clinical settings have field strengths of 1.5T or 3T, and cartilage volume and thickness measures obtained from such data have been extensively validated in terms of accuracy and precision. This is yet to be performed for low-field MRI based measures, but a step towards evaluating the accuracy of the measures from the low-field data is by a comparison to measures obtained by training and running the same algorithms on data from a well-established high-field MRI sequence. This way, the relations between outcomes from a low-field scanner and a high-field scanner could be established, and the precision and cost-effectiveness of low-field MRI based measured could determine if it could become an alternative or complement to high-field MRI in clinical studies.

To conclude, there are several possible future research directions for which the work in this thesis could be improved or further validated. Still, the existing methods builds a foundation for a framework for segmentation and quantitative evaluation of articular cartilage that is automatic, reproducible, in agreement with evaluation of radiologists, and can produce measurements that show great potential as biomarkers for OA.

# List of Tables

# List of Figures

# Bibliography


[1] Eric Abadie, Dominique Ethgen, Bernard Avouac, Gilles Bouvenot, Jaime Branco, Olivier Bruyere, Gonzalo Calvo, Jean-Pierre Devogelaer, Renee L. Dreiser, Gabriel Herrero-Beaumont, Andre Kahan, Godfried Kreutz, Andrea Laslop, Ernst M. Lemme, George Nuki, Leo Van De Putte, Luc Vanhaelst, and Jean-Yves Reginster. Recommendations for the use of new methods to assess the efficacy of disease-modifying drugs in the treatment of osteoarthritis. *Osteoarthritis and Cartilage*, 12:263–268, 2004.

[2] William S. Gosset (a.k.a. Student). The probable error of a mean. *Biometrika*, 6(1): 1–25, 1908.

[3] Selim Aksoy and Robert M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 22(5):563–582, 2001.

[4] S. Arya, D.M. Mount, N. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[5] G.A. Ateshian, M.P. Rosenvasser, and V.C. Mow. Curvature characteristics and congruence of the thumb carpometacarpal joint: differences between female and male joints. *Journal of Biomechanics*, (25):591–607, 1992.

[6] Alex Bailey. *Class-dependent features and multicategory classification.* PhD thesis, University of Southampton, February 2001.

[7] Johan Blom. *Topological and Geometrical Aspects of Image Structure.* PhD thesis, Utrecht University, 1992.

[8] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. pages 144–152. Fifth annual workshop on Computational learning theory, 1992.

[9] S. Bouix, M. Martin-Fernandez, L. Ungar, M. Nakamura, M-S. Koo, R.W. McCarley, and M.E. Shenton. On evaluating brain tissue classifiers without a ground truth. *NeuroImage*, 26:1207–1224, 2007.

[10] T. Boyer and V. Daumen-Legre. Contribution of arthroscopy and magnetic resonance imaging to the evaluation of painful knee osteoarthritis. *Joint Bone Spine*, 67:504–508, 2000.

[11] J.C. Buckland-Wright, D.G. Macfarlane, J.A. Lynch, M.K. Jasani, and C.R. Brandshaw. Joint space width measures cartilage thickness in osteoarthritis of knee. *Annals of the Rheumatic Diseases*, 54(4), 1995.

[12] J.A. Buckwalter, J.A. Martin, and T.D. Brown. Perspectives on chondrocyte mechanobiology and osteoarthritis. *Biorheology*, 43(3-4):603–609, 2006.

[13] P. Bullough. The geometry of diarthrodial joints, its physiologic maintenance, and the possible significance of age related changes in geometry to load distribution and the development of osteoarthritis. *Clinical Orthopaedics*, 156:61–66, 1981.

[14] R. Burgkart, C. Glaser, A. Hyhlik-Durr, K.-H. Englmeier, M. Reiser, and F. Eckstein. Magnetic resonance imaging-based assessment of cartilage loss in severe osteoarthritis. *Arthritis & Rheumatism*, 44(9):2072–2077, September 2001.

[15] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[16] T. Chan and L. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[17] W.P. Chan, P. Lang, M.P. Stevens, K. Sack, S. Majumdar, and D.W. Stoller. Osteoarthritis of the knee, comparison of radiography, ct and mr imaging to assess extent and severity. *American Journal of Roentgenology*, 157:799–806, 1991.

[18] G. Charpiat, O. Faugeras, and R. Keriven. Approximations shape metrics and application to shape warping and empirical shape statistics. *Foundations of Computational Mathematics*, pages 1–58, 2005.

[19] Y.G. Chen, Y. Giga, and S. Goto. Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations. *Journal of Differential Geometry*, 33: 749–786, 1991.

[20] F.M. Cicuttini, A.E. Wluka, and S.L. Stuckey. Tibial and femoral cartilage changes in knee osteoarthritis. *Annals of the Rheumatic Diseases*, (60):977–980, 2001.

[21] L.P. Clarke, R.P. Velthuizen, S. Phuphanich, J.D. Schellenberg, J.A. Arrington, and M. Silbiger. Mri: stability of three supervised segmentation techniques. *Magnetic Resonance Imaging*, 11(1):95–106, 1993.

[22] Z.A. Cohen, D.M. McCarthy, S.D. Kwak, P. Legrand, F. Fogarasi, E.J. Ciaccio, and G.A. Ateshian. Knee cartilage topography, thickness, and contact areas from mri: in-vitro calibration and in vivo measurements. *Osteoarthritis & Cartilage*, 7(1):95–109, 1999.

[23] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. *CVIU*, 61(1):38–59.

[24] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[25] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[26] P. Creamer and M.C. Hochberg. Osteoarthritis. *Lancet*, 350:503–509, 1997.

[27] Paul Creamer and Marc C. Hochberg. Osteoarthritis. *Lancet*, 350:503–509, 1997.

[28] D. Cremers, N. Sochen, and C. Schnorr. Towards recognition-based variational segmentation using shape priors and dynamic labeling. pages 388–400. Scale Space Theories in Computer Vision, 2003.

[29] Erik B Dam, Jenny Folkesson, Marco Loog, Paola C Pettersen, and Claus Christiansen. Efficient automatic cartilage segmentation. In *MICCAI Joint Disease Workshop*, 2006.

[30] Erik B Dam, Jenny Folkesson, Paola C Pettersen, and Claus Christiansen. Automatic cartilage thickness quantification using a statistical shape model. In *MICCAI Joint Disease Workshop*, 2006.

[31] Erik B. Dam, Jenny Folkesson, Paola Pettersen, and Claus Christiansen. Automatic morphometric cartilage quantification in the medial tibial plateau from mri for osteoarthritis grading. *Osteoarthritis & Cartilage*, 15(7):808–818, 2007.

[32] Marleen de Bruijne. Shape particle guided tissue classification. 2006.

[33] Marleen de Bruijne and Mads Nielsen. Shape particle filtering for image segmentation. Seventh International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'04).

[34] M. Descoteaux, M. Audette, K. Chinzei, and K. Siddiqi. Bone enhancement filtering: Application to sinus bone segmentation and simulation of pituitary surgery. pages 9–16. MICCAI, 2005.

[35] Maxime Descoteaux, Louis Collins, and Kaleem Siddiqi. Geometric flows for segmenting vasulature in mri: Theory and validation. volume 3216, pages 500–507. Seventh International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'04), St Malo, France, 2004.

[36] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26: 297–302, 1945.

[37] R.O. Duda, E. Hart, and D.G. Storck. *Pattern Classification*. Wiley-Interscience, second edition, 2001.

[38] T.C. Dunn, Y. Lu, H. Jin, M.D. Ries, and S. Majumdar. T2 relaxation time of cartilage at mr imaging: comparison with severity of knee osteoarthritis. *Radiology*, 232(2): 592–598, 2004.

[39] F. Eckstein, C. Charles, R.J. Buck, V.B. Kraus, A.E. Remmers, M. Hudelmaier, W. Wirth, and J.L. Evelhoch. Accuracy and precision of quantitative assessment of articular cartilage morphology by magnetic resonance imaging at 3.0t. *Arthritis & Rheumatism*, 52(10):3132–3136, October 2005.

[40] Darrin C. Edwards, Charles E. Metz, and Robert M. Nishikawa. The hypervolume under the roc hypersurface of near-guessing and near-perfect observers in n-class classification tasks. *IEEE Transactions on Medical Imaging*, 24(3):293–299, 2005.

[41] B. Ejbjerg, E. Narvestad, S. Jacobsen adn H.S. Thomsen, and M. Ostergaard. Optimised, low cost, low field dedicated extremity mri is highly specific and sensitive for synovitis and bone erosions in rheumatoid arthritis wrist and finger joints: a comparison with conventional high-field mri and radiography. *Annals of the Rheumatic Diseases*, 13, 2005.

[42] L.C. Evans and J. Spruck. Motion of level sets by mean curvature i. *Journal of Differential Geometry*, 33:635–681, 1991.

[43] David T. Felson, Reva C. Lawrence, Marc C. Hochberg, Timothy McAlindon, Paul A. Dieppe, Marian A. Minor, Steven N. Blair, Brian M. Berman, James F. Fries, Morris Weinberger, Kate R. Lorig, Joshua J. Jacobs, and Victor Goldberg. Osteoarthritis: New insights, part 2: Treatment approaches. *Annals of Internal Medicine*, 133(7):726–737, November 2000.

[44] R.S. Fife, K.D. Brandt, E.M. Braunstein, B.P. Catz, K.D. Shelbourn, L.A. Kalasinki, and S. Rayn. Relation between arthroscopic evidence of cartilage damage and radiographic evidence of joint space narrowing in early osteoarthritis of knee. *Arthritis & Rheumatism*, 34:377–382, 1991.

[45] L. Filidoro, O. Dietrich, J. Weber, E. Rauch, T. Oerther, M. Wick, M.F. Reiser, and C. Glaser. High-resolution diffusion tensor imaging of human patellar cartilage: Feasibility and preliminary findings. *Magnetic Resonance in Medicine*, 53:993–998, 2005.

[46] L.M.J. Florack, B.M. ter Haar Romeny, J.J. Koenderink, and M.A. Viergever. Linear scale-space. *Journal of Mathematical Imaging and Vision*, 4(4), 1994.

[47] Jenny Folkesson and Carl-Fredrik Westin. Spatially varying classification with localization certainty for level set segmentation. In *Mathematical Models in Biomedical Image Analysis (MMBIA) Workshop in conjunction with ICCV*, 2007.

[48] Jenny Folkesson, Erik Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen. Automatic segmentation of the articular cartilage in knee mri using a hierarchical multiclass classification scheme. pages 327–334. Eighth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'05), Palm Springs, CA, USA, 2005.

[49] Jenny Folkesson, Erik Dam, Paola Pettersen, Ole Fogh Olsen, Mads Nielsen, and Claus Christiansen. Locating articular cartilage in mr images. volume 5747, pages 1484–1490. Medical Imaging 2005: Image Processing. Proceedings of the SPIE, 2005.

[50] Jenny Folkesson, Ole Fogh Olsen, Paola Pettersen, Erik Dam, and Claus Christiansen. Combining binary classifiers for automatic cartilage segmentation in knee mri. pages 230–239. ICCV First International Workshop: Computer Vision in Biomedical Imaging Applications, 2005.

[51] Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen. Automatic curvature analysis of the articular cartilage surface. In *MICCAI Joint Disease Workshop*, 2006.

[52] Jenny Folkesson, Erik B Dam, Ole Fogh Olsen, Paola C Pettersen, and Claus Christiansen. Position normalization in automatic cartilage segmentation. In *MICCAI Joint Disease Workshop*, 2006.

[53] Jenny Folkesson, Erik B. Dam, Ole F. Olsen, Morten A. Karsdal, Paola C. Pettersen, and Claus Christiansen. Automatic quantification of global and local articular cartilage surface curvature: Markers of smoothness and congruity? *Undergoing review: Magnetic Resonance in Medicine*, 2007.

[54] Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen. Accuracy evaluation of automatic quantification of the articular cartilage surface curvature from mri. *Academic Radiology*, 14(10):1221–1228, 2007.

[55] Jenny Folkesson, Erik B. Dam, Ole Fogh Olsen, Paola Pettersen, and Claus Christiansen. Segmenting articular cartilage automatically using a voxel classification approach. *IEEE Transactions on Medical Imaging*, 26:106–115, 2007.

[56] Jenny Folkesson, Eigil Samset, Raymond Y. Kwond, and Carl-Fredrik Westin. Unifying statistical classification and geodesic active regions for segmentation of cardiac mri. *Undergoing review: IEEE Transactions on Information Technology in Biomedicine*, 2007.

[57] A. Frangi, W. Niessen, K.L. Vincken, and M.A. Viergever. Multiscale vessel enhancement filtering. pages 130–137. MICCAI, 1998.

[58] S.J. Gandy, P.A. Dieppe, M.C. Keen, R.A. Maciewicz, I. Watt, and J.C. Waterton. No loss of cartilae volume over three years in patients with knee osteoarthritis as assessed by magnetic resonance imaging. *Osteoarthritis and Cartilage*, (10):929–937, 2002.

[59] M.B. Goldring. Update on the biology of the chondrocyte and new approaches to treating cartilage diseases. *Clinical Rheumatology*, 20(5):1003–1025, 2006.

[60] H. Graichen, R.V. Eisenhart-Rothe, T. Vogl, K.-H. Englmeier, and F. Eckstein. Quantitative assessment of cartilage status in osteoarthritis by quantitative magnetic resonance imaging: Technical validation for use in analysis of cartilage volume and further morphological parameters. *Arthritis & Rheumatism*, 50(3):811–816, March 2004.

[61] V. Grau, A.U.J. Mewes, M. Alcañiz, R. Kikinis, and S.K. Warfield. Improved watershed transform for medical image segmentation using prior information. 23(4):447–458, 2004.

[62] T.M. Griffin and F. Guilak. The role of mechanical loading in the onset and progression of osteoarthritis. *Exerc. Sport Sci Rev.*, 33(4):195–200, 2005.

[63] Y. Guo and B.C. Vemuri. Hybrid geometric active models for shape recovery in medical images. pages 112–125, 1999.

[64] M. Hernandez, G. Sapiro, and A.F. Frangi. Three-dimensional segmentation of brain aneurysms in cta using non-parametric region-based information and implicit deformable models: Methods and evaluation. volume 2. MICCAI.

[65] Jan Hohe, Gerard Ateshian, Maximillian Reiser, Karl-Hans Englmeier, and Felix Eckstein. Surface size, curvature analysis, and assessment of knee joint incongruity with mri in vivo. *Magnetic Resonance in Medicine*, (47):554–561, 2002.

[66] Rolf W. Huegli, Philip F.J. Tirman, Harald M. Bonel, Harald Staedele, Souhil Zaim, Mikayel Grigorian, and Harry K. Genant. Use of the modified three-point dixon technique in obtaining t1-weighted contrast-enhanced fat-saturated images on an open magnet. *European Journal of Radiology*, 11(7):473–474, July 2003.

[67] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et de jura. *Bull. Soc. Voudouise Sci. Nat.*, 39:547–579, 1901.

[68] D.W. Jackson, T.M. Simon, and H.M. Aberman. Symptomatic articular cartilage degeneration: the impact in the new millenium. *Clinical orthopaedics and related research*, 391:14–25, 2001.

[69] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[70] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: a review. 22(1), 2000.

[71] L. Kamibayashi, U.P. Wyss, T.D. Cooke, and B. Zee. Changes in mean trabecular orientation in the medial condyle of the proximal tibia in osteoarthritis. *Calcif Tissue Int.*, 57(1):69–73, 1995.

[72] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[73] J.H. Kellgren and J.S. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases*, 16(4), 1957.

[74] D.G. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bull. London Math Soc.*, 16:81–121, 1984.

[75] B. Kersting-Sommerhoff, P. Gerhardt, W. Golder, N. Hof, K.A. Riel, H. Helmberger, M. Lentz, and K. Lehner. Mri of the knee joint: first results of a comparison of 0.2-t specialized system and 1.5-t high field strength magnet. *Fortschr. Rontgenstr.*, 162(5): 390–395, 1995.

[76] N. Khaneja, M.I. Miller, and U. Grenander. Dynamic programming generation of curves on brain surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1260–1265, 1998.

[77] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi. Gradient flows and geometric active contour models. pages 810–815. ICCV, 1995.

[78] Ron Kimmel. *Numerical Geometry of Images: Theory, Algorithms and Applications*. Springer, 2 edition, 2000.

[79] G. Kindlmann, R. Whitaker, T. Tasdizen, and T. Moller. Curvature-based transfer functions for direct volume rendering: Methods and applications. pages 513–520. IEEE Visualization, 2003.

[80] B. Kladny, K. Gluckert, B. Swoboda, W. Beyer, and G. Weseloh. Comparison of low-field (0.2 tesla) and high-field (1.5 tesla) magnetic resonance imaging of the knee joint. *Archives of Orthopaedic and Trauma Surgery*, 114(5):281–286, 1995.

[81] H. Knutsson. A tensor representation of 3-D structures. In *5th IEEE-ASSP and EURASIP Workshop on Multidimensional Signal Processing*, 1987.

[82] Jan J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[83] K.E. Kuettner and V.M. Goldberg. Introduction. osteoarthritic disorders. pages xxi–xxv. Rosemont: American Academy of Orthopaedic Surgeons, 1995.

[84] Raymond Y. Kwong, Anna K. Chan, Kenneth A. Brown, Carmen W. Chan, H. Glenn Reynolds, Sui Tsang, and Roger B. Davis. Impact of unrecognized myocardial scar detected by cardiac magnetic resonance imaging on event-free survival in patients presenting with signs or symptoms of coronary artery disease. *Circulation*, 113:2733–2743, 2006.

[85] M. Leventon, E. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contour models. volume I, pages 316–322. IEEE CVPR, 2000.

[86] Kang Li, Steven Millington, Xiaodong Wu, Danny Z. Chen, and Milan Sonka. Simultaneous segmentation of multiple closed surfaces using optimal graph searching. In *Information Processing in Medical Imaging: 19th International Conference*, volume 3565 of *LNCS*. Springer, 2005.

[87] Damien Loeuille, Pierre Olivier, Didier Mainard, Pierre Gillet, Patrick Netter, and Alain Blum. Magnetic resonance imaging of normal and osteoarthritic cartilage. *Arthritis and Rheumatism*, 41(6):963–975, 1998.

[88] John A. Lynch, Souhil Zaim, Jenny Zhao, Alexander Stork, Charles G. Peterfy, and Harry K. Genant. Cartilage segmentation of 3d mri scans of the osteoarthritic knee combining user knowledge and active contours. volume 3979, pages 925–935. Medical Imaging 2000: Image Processing. Proceedings of the SPIE, 2000.

[89] John A. Lynch, Souhil Zaim, Jenny Zhao, Alexander Stork, Charles G. Peterfy, and Harry K. Genant. Automatic measurement of subtle changes in articular cartilage from mri of the knee by combining 3d image registration and segmentation. volume 4322, pages 431–439. Medical Imaging 2001: Image Processing. Proceedings of the SPIE, 2001.

[90] H.J. Mankin, H. Dorfman, L. Lipiella, and A. Zarins. Biochemical and metabolic abnormalities in articular cartilage from osteoarthritic human hips ii. correlation of morphology with biochemical and metabolic data. *J Bone Joint Surg Am*, 53:523–537, 1971.

[91] R. Meder, S.K. de Visser, J.C. Bowden, T. Bostrom, and J.M. Pope. Diffusion tensor imaging of articular cartilage as a measure of tissue microstructure. *Osteoarthritis & Cartilage*, 14(9):875–881, 2006.

[92] S.A. Millington, K. Li, X. Wu, S.R. Hurwitz, and M. Sonka. Automated simultaneous 3d segmentation of multiple cartilage surfaces using optimal graph searching on mri images. *Osteoarthritis and Cartilage*, 13, supplement A, 2005.

[93] R.W. Moskowitz, R.D. Altman, M.C. Hochberg, J.A. Buckwalter, and V.M. Goldberg. *Osteoarthritis: diagnosis and medical/surgical management*. Wolters Kluwer, Lippincott Williams & Wilkins, 4 edition, 2007.

[94] D. Mossman. Three-way rocs. *Med. Decis. Making*, 19:78–79, 1999.

[95] O.J. Muensterer, F. Eckstein, D. Hahn, and R. Putz. Computer-aided three dimensional assessment of knee-joint cartilage with magnetic resonance imaging. *Clinical Biomechanics*, 11(5):260–266, 1996.

[96] D. Mumford and J. Shah. Optimal approximation by piece-wise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[97] Josephine H. Naish, Graham Vincent, Mike Bowes, Manish Kothari, David White, John C. Waterton, and Christopher J.Taylor. A method to monitor local changes in mr signal intensity in articular cartilage: A potential marker for cartilage degeneration in osteoarthritis. volume 3217, pages 959–966. Medical Imaging Computing and Computer-Assisted Intervention, 2004.

[98] I-S. Oh, J-S. Lee, and C.Y. Suen. Analysis of class separation and combination of class-dependent features for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1089–1094, 1999.

[99] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on the hamilton-jacobi formulation. *Journal of Computational Physics*, 79:12–49, 1988.

[100] Stanley Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153. Springer, 2003.

[101] S. Kubilay Pakin, Jose G. Tamez-Pena, Saara Totterman, and Kevin J.Parker. Segmentation, surface extraction and thickness computation of articular cartilage. volume 4684, pages 155–166. Medical Imaging 2002: Image Processing. Proceedings of the SPIE, 2002.

[102] Nikos Paragios and Rachid Deriche. Geodesic active regions: A new framework to deal with frame artition problems in computer vision. *J. Vis. Commun. Image Representation*, 13:249–268, 2002.

[103] Nikos Paragios, Mikael Rousson, and Visvanathan Ramesh. Knowledge-based registration & segmentation of the left ventricle: A level set approach. volume 1, pages 316–322. IEEE Workshop on Applications in Computer Vision, 2002.

[104] E. Pessis, J.-L. Drape, P. Ravaud, A. Chevrot, and M. Dougados X. Ayral. Assessment of progression in knee osteoarthritis: results of a 1 year study comparing arthroscopy and mri. *Osteoarthritis & Cartilage*, 11:361–369, 2003.

[105] C.G Peterfy, G. Gold, F. Eckstein, F. Cicuttini, B. Dardzinski, and R. Stevens. Mri protocols for whole-organ assessment of the knee in osteoarthritis. *Osteoartritis & Cartilage*, 14:95–111, 2006.

[106] Dzung L. Pham, Chenyang Xu, and Jerry L. Prince. Current methods in medical image segmentation. *Annual Rev. Biomed. Eng.*, 2:315–37, 2000.

[107] Eric Pichon, Allen Tannenbaum, and Ron Kikinis. A statistically based surface evolution method for medical image segmentation: Presentation and validiation. pages 277–284. Sixth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'03), 2003.

[108] Stephen M. Pizer, P. Thomas Fletcher, Sarang Joshi, Andrew Thall, James Z. Chen, Yonatan Fridman, Daniel S. Fritsch, Graham Gash, John M. Glotzer, Michael R. Jiroutek, Conglin Lu, Keith E. Muller, Gregg Tracton, Paul yushkevich, and Edward L. Chaney. Deformable m-reps for 3d medical image segmentation. *International Journal on Computer Vision*, 55, 2003.

[109] Killian Pohl, John Fisher, Martha Shenton, Robert W. McCarley, W. Eric L. Grimson, Ron Kikinis, and William M. Wells. Logarithm odds maps for shape representation. pages 995–963. MICCAI, 2006.

[110] Arish A. Qasi, Jenny Folkesson, Paola Pettersen, Morten A. Karsdal, Claus Christiansen, and Erik B. Dam. Separation of healthy and early osteoarthritis by automatic quantification of cartilage homogeneity. *Osteoarthritis & Cartilage (In press)*, 2007.

[111] K.-A. Riel, M. Reinisch, B. Kersting-Sommerhoff, N. Hof, and T. Merl. 0.2-tesla magnetic resonance imaging of internal lesions of the knee joint: a prospective arthroscopically controlled clinical study. *Knee Surg. Sports Traumatol. Arthrosc.*, 7:37–41, 1999.

[112] Mikael Rousson and Nikos Paragios. Shape priors for level set representations. *ECCV*, II:78–92, 2002.

[113] Y. Sato, C-F. Westin, A. Bhalerao, S. Nakajima, N. Shiraga, S. Tamura, and R. Kikinis. Tissue classification based on 3d local intensity structures for volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 6(2):160–180, 2000.

[114] S. Schaller, K. Henriksen, P. Hoegh-Andersen, B.C. Sondergaard, E.U. Sumer, and L.B. Tanko et al. In vitro, ex vivo, and in vivo methodological approaches for studying therapeutic targets of osteoporosis and degenerative joint diseases: how biomarkers can assist? *Assay Drug Dev. Technol.*, 3(5), 2005.

[115] J.A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 2 edition, 1999.

[116] Randolph M. Setser, Daniel G. Bexell, Thomas P. O'Donnell, Arthur E. Stillman, Michael L. Lieber, Paul Schoenhagen, and Richard D. White. Quantitative assessment of myocardial scar in delayed enhanced magnetic resonance imaging. *Journal of Magnetic Resonance Imaging*, 18:434–441, 2003.

[117] K. Soejima, W.G. Stevenson, W.H. Maisel, J.L. Sapp, and L.M. Epstein. Electrically unexcitable scar mapping based on pacing threshold for identification of reentry circuit isthmus. *Circulation*, 106:1678–83, 2002.

[118] Jan Erik Solem. *Variational Problems and Level Set Methods in Computer Vision - Theory and Applications*. PhD thesis, Lund University, 12006.

[119] S. Solloway, C.E. Hutchinson, J.C. Vaterton, and C.J. Taylor. The use of active shape models for making thickness measurements of articular cartilage from mr images. *Magnetic Resonance in Medicine*, 37:943–952, 1997.

[120] Milan Sonka and J. Michael Fitzpatrick. *Handbook of Medical Imaging. Volume 2. Medical Image Processing and Analysis*. SPIE Press, second edition, 2004.

[121] Luuk Spreeuwers and Marcel Breeuwer. Detection of left ventricular and epi- and endocardial borders using coupled active contours. pages 1147–1152. Computer Assisted Radiology and Surgery, 2003.

[122] L.H. Staib and J.S. Duncan. Boundary finding with parametrically deformable models. *IEEE Trans. Patt. Anal. Mach. Intell.*, 14(11):1061–1075, 1992.

[123] Tobias Stammberger, Felix Eckstein, Karl-Hans Englmeier, and Maximillian Reiser. Determination of 3d cartilage thickness data from mr imaging: Compputational method and reproducibility in the living. *Magnetic Resonance in Medicine*, 41:529–536, 1999.

[124] Tobias Stammberger, Felix Eckstein, Markus Michaelis, Karl-Hans Englmeier, and Maximillian Reiser. Interobserver reproducibility of quantitative cartilage measurements: Comparison of b-spline snakes and manual segmentation. *Magnetic Resonance Imaging*, 17(7):1033–1042, 1999.

[125] M.B. Stegmann, H. Olafsdottir, and H.B.W. Larsson. Unsupervised motion compensation of multi-slice cardiac perfusion mri. *Medical Image Analysis*, 9(4):394–410, 2005.

[126] Jose G. Tamez-Pena, Monica Barbu-McInnis, and Saara Totterman. Knee cartilage extraction and bone-cartilage interface analysis from 3d mri data sets. volume 5370, pages 1774–1784. Medical Imaging 2004: Image Processing. Proceedings of the SPIE, 2004.

[127] Mitsunobu Terukina, Hiroyuki Fujioka, Shinichi Yoshiya, Masahiro Kurosaka, Takeshi Makino, Nobuzo Matsui, and Juichi Tanaka. Analysis of the thickness and curvature of articular cartilage of the femoral condyle. *Arthroscopy*, 19(9):969–973, 2003.

[128] C.J. Tiderius, L.E. Olsson, P. Leander, O. Ekberg, and L. Dahlberg. Delayed gadolinium-enhanced mri of cartilage (dgemric) in early osteoarthritis. *Magnetic Resonance in Medicine*, 49:488–492, 2003.

[129] Andy Tsai, Anthony Yezzi Jr., William Wells, Clare Tempany, Dewey Tucker, Ayres Fan, W. Eric Grimson, and Alan Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Transactions on Medical Imaging*, 22(2): 137–154, 2003.

[130] Andy Tsai, William Wells, Clare Tempany, Eric Grimson, and Alan Willsky. Mutual information in coupled multi-shape model for medical image segmentation. *Medical Image Analysis*, 8(4):429–445, 2004.

[131] L.J van Vliet and P.W. Verbeek. Curvature and bending energy in digitized 2d and 3d images. pages 1403–1410. Eight Scandinavian Conference on Image Analysis, 1993.

[132] Gunter Verheugen. Commision directive 2005/28/ec laying down principles and guidelines for good clinical practice as regards investigational medicinal products for human use, as well as the requirements for authorization of the manufacturing or importation of such products. *Official Journal of the European Union*, Legislation 091:13–19, 2005.

[133] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation.

[134] Simon K. Warfield, Carl Winalski, Ferenc A. Jolesz, and Ron Kikinis. Automatic segmentation of mri of the knee. SPL Technical Report 91, ISMRM Sixth Scientific Meeting and Exhibition, Sydney, Australia, July 1998.

[135] Simon K. Warfield, Michael Kaus, Ferenc A. Jolesz, and Ron Kikinis. Adaptive, template moderated, spatially varying statistical classification. *Medical Image Analysis*, (4):43–55, 2000.

[136] Joachim Weickert. *Anisotropic Diffusion in Image Processing*. B. G. Teubner, 1998.

[137] C-F. Westin, A. Bhalerao, H. Knutsson, and R. Kikinis. Using local 3d structure for segmentation of bone from computer tomography images. pages 794–800. CVPR, 1997.

[138] Tomos G. Williams, Christopher J. Taylor, ZaiXiang Gao, and John C. Waterton. Corresponding articular cartilage thickness measurements in the knee joint by modelling the underlying bone. *MICCAI*, pages 480–487, 2003.

[139] Klaus Woertler, Marc Strothmann, Bernd Tombach, and Peter Reimer. Detection of articular cartilage lesions: Experimental evaluation of low- and high-field-strength mr imaging at 0.18 and 1.0 t. *Journal of Magnetic Resonance Imaging*, 11:678–685, 2000.

[140] J. Wyngaarden, Lloyd H.Smith, and J.Claude Bennett. *Cecil Textbook of Medicine*, volume 2. W. B. Saunders, 19 edition, 1992.

[141] Y. Xia. The total volume and the complete thickness of cartilage determined by mri. *Osteoarthritis & Cartilage*, 14:1781–1786, 2004.

[142] Y. Xia. The total volume and the complete thickness of articular cartilage determined by mri. *Osteoarthritis and Cartilage*, 11(7):473–474, 2003.

[143] Chen-Hsiang Yeang, Sridhar Ramaswamy, Pablo Tamayo, Sayan Mukherjee, Ryan M. Rifkin, Michael Angelo, Michael Reich, Eric Lander, Jill Mesirov, and Todd Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 1(1), 2001.

[144] Xiaolan Zeng, Lawrence H. Staib, Robert T. Schultz, and James S. Duncan. Segmentation and measurement of the cortex from 3d mr images using coupled surfaces propagation. *IEEE Transactions on Medical Imaging*, 18(10):100–111, 1999.

[145] Hong-Kai Zhao, T. Chan, B. Merriman, and S. Osher. A variational level set approach to multiphase motion. *Journal of Computational Physics*, 127:179–195, 1996.

[146] L. Zhao, C.P. Botha, J.O. Bescos, R. Truyen, F.M. Vos, and F.H. Prost. Lines of curvature for polyp detection in virtual colonoscopy. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):885–892, 2006.

[147] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

[148] D.P. Zipes and H.J. Wellens. Sudden cardiac death. *Circulation*, 98:2334–2351, 1998.

[149] Kelly H. Zou, Simon K. Warfield, Aditya Bharatha, Clare M.C. Tempany, Michael R. Kaus, Steven J. Haker, William M. Wells III, Ferenc A. Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology*, 11:178–189, 2004.