Segmenting Articular Cartilage Automatically Using a Voxel Classification Approach

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

***Abstract*—We present a fully automatic method for articular cartilage segmentation from magnetic resonance imaging (MRI) which we use as the foundation of a quantitative cartilage assessment. We evaluate our method by comparisons to manual segmentations by a radiologist and by examining the interscan reproducibility of the 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 severely osteoarthritic. This is, to our knowledge, the only fully automatic cartilage segmentation method that has good agreement with manual segmentations, an interscan reproducibility as good as that of a human expert, and enables the separation between healthy and osteoarthritic 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 counterpart. For low-field MRI, there is no well-established accuracy validation for quantitative cartilage estimates, but we show that differences between healthy and osteoarthritic populations are statistically significant using our cartilage volume and surface area estimates, which suggests that low-field MRI analysis can become a useful, affordable tool in clinical studies.**

***Index Terms*—Articular cartilage, image segmentation, osteoarthritis, magnetic resonance imaging (MRI), pattern classification.**

I. 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 [1]. 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 [2], 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, noninvasive assessment of the articular cartilage [3], and cartilage deterioration can be detected using quantitative MRI analysis [4]. 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 interslice 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 [5].

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 [6]. There have been several comparisons of diagnostic performance of diagnosing meniscal tears, cruciate ligaments, and cartilage lesions between low-field and high-field MRI data [7]–[9] 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 [10] 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 [11]. 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 would 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. 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 interscan reproducibility, and the ability to detect changes between healthy and osteoarthritic groups using the cartilage volume and area estimates.

A. Related Work

As in most quantitative assessment studies in medical imaging, the first and most crucial step in our articular cartilage assessment is 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 intraobserver variability. It is thus advantageous to automate the segmentation method and the main challenges

Manuscript received June 27, 2006; revised September 15, 2006. *Asterisk indicates corresponding author.*

*J. Folkesson is with the IT University of Copenhagen, DK-2300 Copenhagen S, Denmark (e-mail: jenny@itu.dk).

E. B. Dam and O. F. Olsen are with the IT University of Copenhagen, DK-2300 Copenhagen S, Denmark.

P. C. Pettersen and C. Christiansen are with Center for Clinical and Basic Research, 2750 Ballerup, Denmark.

Digital Object Identifier 10.1109/TMI.2006.886808

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 semiautomated/automated methods for cartilage segmentation. Among two-dimensional (2-D) techniques, Stammberger and colleagues [12] segments the cartilage by fitting a b-spline snake to each slice. A 2-D method combining user interaction with active contours is described by Lynch *et al.* [13]. They combine the segmentation technique with three-dimensional (3-D) image registration to detect changes in cartilage volume [14]. Solloway *et al.* [15] 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 2-D technique, continuation between slices is lost and some regularization between the slices is required. Also, since the series of 2-D segmentations have to be converted into a 3-D segmentation when finding for example thickness maps, it is advantageous to perform segmentation in 3-D directly.

Looking at the 3-D techniques that have been developed, Grau *et al.* [16] use a watershed based approach, where the watershed is extended to examining difference in class probability of neighboring pixels. The sensitivity, specificity and Dice volume overlap of the segmentation are 90.03%, 99.86%, and 0.90, respectively. The method is evaluated on seven scans from four healthy knees and requires 5–10 min of manual labor for selecting markers before initializing the watershed.

Pakin *et al.* [17] 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 [18], and this semiautomatic method is evaluated in terms of intrauser reproducibility.

Another classification approach to segmentation is presented by Warfield *et al.* [19], [20], 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 intrascan variability of the volume compared to repeated manual segmentations on the scan it is evaluated on.

A semiautomatic method based on a graph searching segmentation algorithm [21] followed by mean thickness quantification is evaluated on ankle joints in [22]. The method requires only a small amount of manual initialization and shows accurate thickness measurements on eight cadaveric ankles. Presumably, the method could also be adapted to knees.

B. Overview of the Work Presented

The segmentation techniques described in Section I-A all require some amount of manual interaction except for the method of Pakin *et al.* [17], the 3-D 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 both healthy and osteoarthritic knee scans. The segmentation method is the first step in a quantitative, fully automatic cartilage assessment 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 kNN classifiers which is described in Sections III-A and III-B, 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 III-G. Since kNN classification is a slow process we propose to use an efficient voxel classification algorithm which is described in Section III-F.

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. We evaluate the interscan reproducibility using the volume and surface area estimate, and the ability to detect changes between healthy and osteoarthritic populations by performing unpaired t-tests between the groups using the volume and area estimates and the Kellgren–Lawrence index. OA is more frequently observed in the medial compartment [23], therefore, we focus on the medial cartilage compartment in this study. The evaluation of the segmentation framework is described in Section IV followed by discussion in Section V.

II. Image Acquisition

A. 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 3-D T1 sequence (40° flip angle, T_R 50 ms, T_E 16 ms). Approximate acquisition time is 10 min 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 0.70×0.70 mm^2, with a distance between slices ranging between 0.70 mm–0.94 mm, where the most common distance is 0.78 mm.

Assessing the cartilage directly in 3-D eliminates the problem of limited continuation between slices that is present in 2-D techniques. We use a 3-D sequence consisting of near isotropic voxels since this is well suited for cartilage quantification [24] and for 3-D analysis in general.

B. 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 $(\mathrm{KL}i)$ [25], a radiographic score established by X-rays between 0–4 where $\mathrm{KL}i = 0$ is healthy, $\mathrm{KL}i = 1$ is considered borderline or mild OA, and $\mathrm{KL}i \geq 2$ is severe OA. In our data set, 51 knees have $\mathrm{KL}i = 0$, 28 have $\mathrm{KL}i = 1, 13$ have $\mathrm{KL}i = 2$ and the remaining 22 knees have $\mathrm{KL}i = 3$. In the X-rays the width of the tibial plateau has also

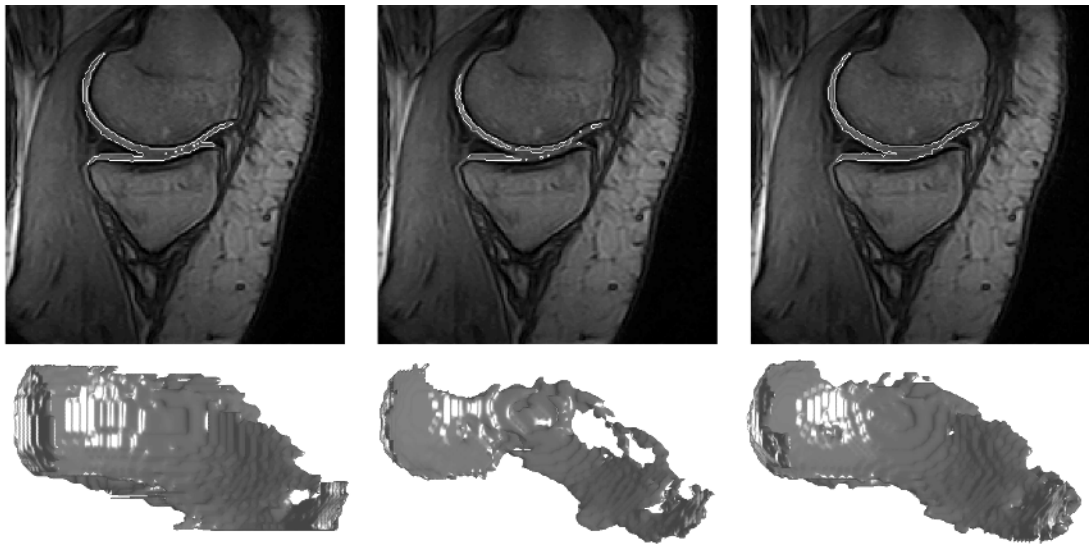

Fig. 1. Scan most improved by the position correction scheme, where the DSC increases from 0.61 to 0.77. First column shows the manual segmentation, the second column shows the original segmentation, and the third column shows the segmentation after position correction. 2-D images in the top row are a sagittal slice of the segmentation and the 3-D views on the second row are of the same segmentation seen from above.

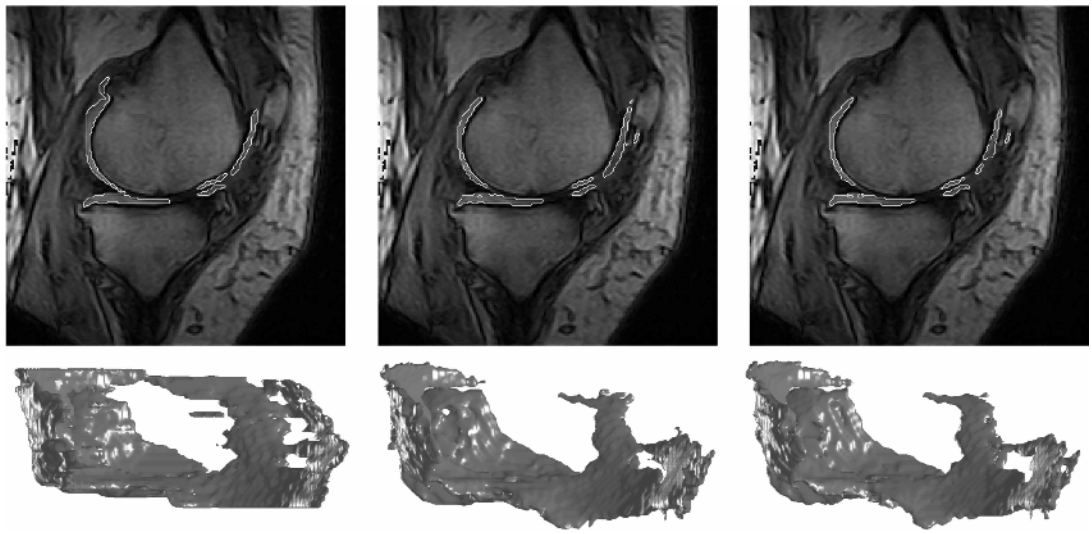

Fig. 2. Worst case scenario of applying position correction. Knee is severely osteoarthritic ($\mathrm{KL}i = 3$). For this scan, there is no improvement in DSC. Manual segmentation is in the first column, the second column shows initial segmentation, and the third column shows the segmentation after position correction. 2-D images in the top row are a sagittal slice of the segmentation and the 3-D views on the second row are of the same segmentation seen from above.

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 intrarater variability of the manual delineations.

Of the 139 knees, the same 31 knees that were segmented twice were rescanned 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 Figs. 1 and 2.

III. Cartilage Segmentation

A. Voxel Classification

We implement our classifier in an approximate nearest neighbor framework developed by Mount and colleagues [26]. The classifier is in principle a kNN-classifier, but allows for faster computations if an error in the distance calculations is tolerated. The approximate kNN search algorithm returns k points

such that the ratio of the distance between the ith reported point $(1 \leq i \leq k)$ and the true ith nearest neighbor is at most $1+\epsilon$. Given a set S of n data points in R^d, the k nearest neighbors of a point in S can be computed in $O((c_{d,\epsilon}+kd)\log n)$ time, where $c_{d,\epsilon} = d[1+6d/\epsilon]^d$, thus the computational complexity increases exponentially with the dimensions. 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.

B. Multiclass 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) [27], [28]. The outcome of a one 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 ω_i, where $i = 1, \ldots, N$ is the number of classes. We denote it $P(\omega_i \,|\, \mathbf{u}_{i,j})$ or $P_{i,j}$ for short. In one-versus-all classification, which is commonly used for multi-class classification [29], one builds N one vs. rest classifiers and perform a winner-takes-all vote between them, assigning j to the class ω_i with the highest posterior probability. 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 versus 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

$$j \in \begin{cases} \omega_{\mathrm{tm}}, & P_{\mathrm{tm},j} > P_{\mathrm{fm},j} \quad \text{and} \quad P_{\mathrm{tm},j} > t; \\ \omega_{\mathrm{fm}}, & P_{\mathrm{fm},j} > P_{\mathrm{tm},j} \quad \text{and} \quad P_{\mathrm{fm},j} > t; \\ \omega_b & \text{otherwise} \end{cases} \tag{1}$$

where $N = 3$ and the subscripts tm, fm, and b stands for *tibial medial, femoral medial* and *background*, respectively. The rejection 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 [30]. It measures the spatial volume overlap between two segmentations A and B and is defined as $\mathrm{DSC}(A,B) = (2 \times |A \cap B|/|A| + |B|)$.

C. 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 III-D, [27]. 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 [31]. 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. Using 25 scans gives us a large enough training set to not be sensitive to the curse of dimensionality—the outcome of the criterion function evaluation is improved after every iteration 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 exponential increase in computational complexity with the number of dimensions. We do backward selection until there are 39 features remaining in the set, and we observed 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 part of the tree consisting of all possible combinations of features (given the number of features one wishes to use, something that is more or less determined by computational complexity) than by only using forward selection.

We combine binary classifiers even though kNN is inherently a multiclass classifier. The reason for so doing is that for feature selection, the area under the ROC curve evaluates the classifier performance for all operating points for a two-class task. But there is no obvious extension of ROC analysis for multiclass classification tasks and we have found better results by training and combining binary classifier than we have with direct multiclass classifiers [27].

D. 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 are highly relevant for a radiologist when visually inspecting a scan, and that is the main motivation for including them as candidate features. Both the raw image intensities and intensities from the image convolved with a Gaussian according to the scale space framework [32] on different scales are considered. Three scales are chosen (0.65, 1.1, and 2.5 mm) 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 three-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 [33] 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 [34] shows that

the spatial averaging in scale-space causes a noise reduction that more than counteracts the noise amplification caused by differentiation. Hence, all the derivatives mentioned in this section are achieved by convolution with Gaussian derivatives, defined as $I_{i_1,\ldots,i_n} = I * D_{i_1,\ldots,i_n} g(\sigma)$, where g is a Gaussian, D a differential operator, and σ 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 (0.65 mm) is lower than the resolutions of the scans, there is still some spatial averaging and Gaussian derivatives allow for robust differentiation at that scale.

In vessel segmentation, the eigenvalues of the Hessian (H) have proven to be useful when looking for central locations inside a tubelike structure [35]. The Hessian is the symmetric matrix containing second-order derivatives with respect to the coordinates (x, y, z)

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

and it describes the second-order structure of local intensity variations. The largest eigenvalue gives the highest second-order derivative value and its corresponding eigenvector is in the direction of the maximum second-order derivative. Cartilage can locally be described as a thin disc, which corresponds to finding positions with one large and two small eigenvalues of the Hessian. The eigenvalues and the three eigenvectors are candidate features.

One feature that has been shown to be significant in the detection of thin structures such as fingerprints is the structure tensor (T) [36]. The structure tensor is a symmetric matrix containing products of the first-order derivatives convolved with a Gaussian

$$T = G_{\sigma_{\text{out}}} * \begin{pmatrix} I_x I_x & I_x I_y & I_x I_z \\ I_y I_x & I_y I_y & I_y I_z \\ I_z I_x & I_z I_y & I_z I_z \end{pmatrix}$$

where the outer scale σ_{out} is not necessarily the same scale as the one used for obtaining the derivatives (σ). The structure tensor examines the local gradient distribution at each location (x, y, z). The directions of the eigenvectors depend on the variation in the neighborhood. The structure tensor eigenvalues and eigenvectors combining three different scales on σ_{out} and σ 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 three-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 (0.65, 1.1, and 2.5 mm), 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 to allow them the same odds of getting picked. The three by three grouping comes natural because we have 3-D images, so the coordinates, first-order derivatives and the eigenvalues and eigenvectors of the Hessian and the structure tensor have a natural grouping. The other features are grouped using the three scales.

E. Selected Features

After feature selection, the resulting features for the ω_{tm} 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 the three scales, I_{zzz} on the three scales, the eigenvalues of H(1.1 mm), I_{yy} on all three scales, the eigenvalues of H(2.5 mm), and the eigenvalues of T(2.5 mm, 0.65 mm).

The ω_{fm} versus. rest classifier contains the following features after feature selection: the position, the eigenvector corresponding to the largest eigenvalue of T(1.1 mm, 0.65 mm), the first-order derivatives on scales 1.1 mm and 2.5 mm, 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 T(2.5 mm, 0.65 mm).

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 ranked high by both classifiers, followed by eigenvalues of both the Hessian and the structure tensor on various scales and second- and third-order derivatives in the direction of the coronal and axial axes.

F. Efficient Voxel Classification

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

We have, therefore, developed an efficient voxel classification algorithm [37], 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

TABLE I
RESULTS FROM OUR AUTOMATIC SEGMENTATION METHOD BEFORE AND AFTER POSITION ADJUSTMENT (PA) FOR MEDIAL TIBIAL, MEDIAL FEMORAL, AND THE MEDIAL COMPARTMENTS TOGETHER. SENSITIVITY, SPECIFICITY, AND DSC ARE FOUND FROM COMPARISON WITH MANUAL SEGMENTATIONS ON THE 114 SCANS IN THE TEST SET. STANDARD DEVIATIONS ARE DENOTED SD AND 95% CONFIDENCE INTERVALS ARE DENOTED CI

Compartment(s)	Sensitivity	SD	CI	Specificity	SD	CI	DSC	SD	CI
Tibial	81.1%	±10.6%	82.0% 85.9%	99.96%	±0.01%	99.96% 99.97%	0.80	±6.7%	0.79 0.81
Tibial PA	86.8%	± 7.7%	85.4% 88.2%	99.96%	±0.01%	99.96% 99.96%	0.81	±6.0%	0.79 0.82
Femoral	77.9%	±12.8%	75.5% 80.2%	99.92%	±0.03%	99.91% 99.92%	0.77	±8.0%	0.75 0.78
Femoral PA	80.3%	±11.6%	78.2% 82.4%	99.91%	±0.03%	99.90% 99.91%	0.77	±8.0%	0.76 0.79
Tibial + Femoral	81.1%	±10.9%	79.1% 83.1%	99.88%	±0.04%	99.87% 99.89%	0.79	±6.5%	0.78 0.80
Tibial + Femoral PA	83.9%	± 8.4%	82.4% 85.5%	99.87%	±0.04%	99.86% 99.88%	0.80	±5.6%	0.79 0.81

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.

G. 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 3-D 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 [38]. 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 (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 there the different cartilage compartments have different relative position with respect to each other due to different positions of the test subjects.

IV. 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.

A. 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 rescanned 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 III-G, 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 in only one iteration. 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 worst cases from the position correction scheme are shown in Figs. 1 and 2. In the best case, the DSC increases with 0.17 and for the worst scan it decreases with 0.017. 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 I.

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 ($p = 0.95$). 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 Fig. 1 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, there is an uncertainty in the segmentation close to the boundary particularly along the crest, and the scans used in this study have low contrast between tissues which may also contribute to false positives compared to manual segmentations.

TABLE II
INTERSCAN 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 AVERAGE 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%

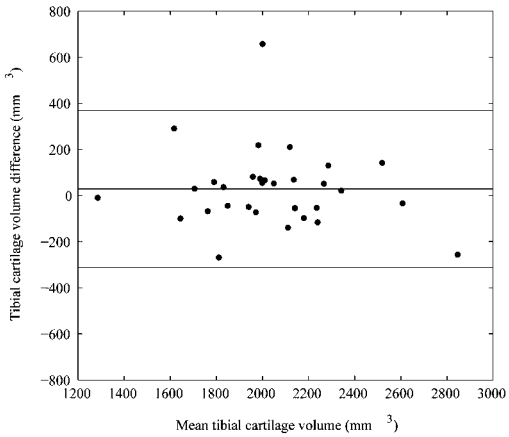

Fig. 3. Bland–Altman plot of the interscan reproducibility of the tibial volume from automatic (position adjusted) segmentations. Lines are the mean ± 2 SD of the difference between measurements.

As to interscan reproducibility of the medial cartilage volume from the automatic segmentations, we examine the 31 knees that were scanned twice. 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 II, 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 Figs. 3 and 4, the Bland–Altman plots of interscan reproducibility for the automatically obtained tibial volume and area estimates are displayed.

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

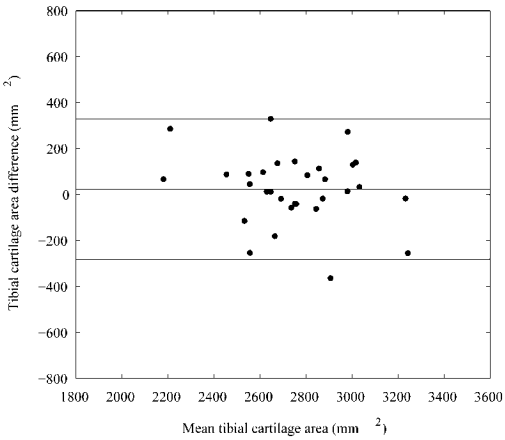

Fig. 4. Bland–Altman plot of the interscan reproducibility of the tibial area from automatic (position adjusted) segmentations. Lines are the mean ± 2 SD of the difference between measurements.

and femoral are touching. 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.

The radiologist redelineated the tibial medial and femoral cartilage in 31 scans in order to determine intrarater variability for the manual segmentations. The average DSC between the two manual segmentations are 0.87 for the medial cartilage, which explains the fairly low values of the DSC in our evaluation because the method is trained on manual segmentations by the expert and therefore attempts to mimic the expert. Also, assuming most misclassifications occur at boundaries, thin structures will typically have relatively low DSC. The corresponding DSC of the automatic segmentation versus expert for the medial cartilage of the 31 scans is 0.80.

For all the scans the in-plane resolution is 0.70×0.70 mm^2, but the slice distance is either 0.78, 0.70, 0.94, or 0.86 mm with the first being the most predominant. Of the 25 scans in the training set, 13 scans have slice distance 0.78 mm 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 $0.80(\pm 0.04)$ SD. For the other resolutions in the test set the average DSC is $0.79(\pm 0.07)$ SD. Of these remaining scans, 32 have 0.86 mm slice distance, seven have 0.94 mm and three have 0.70 mm.

B. Correlation 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 [39]–[41]. 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

TABLE III
P-VALUES FOR T-TESTS OF SEPARATING GROUPS USING THE VOLUME ESTIMATES. P1 IS THE P-VALUE FOR SEPARATION OF HEALTHY ($\mathrm{KL}i = 0$) FROM BORDERLINE TO OA ($\mathrm{KL}i > 0$), AND P2 IS SEPARATION OF HEALTHY AND BORDERLINE ($\mathrm{KL}i \leq 1$) FROM CLEAR OA CASES ($\mathrm{KL}i > 1$). 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 IV
P-VALUES FOR T-TESTS OF SEPARATING GROUPS USING THE AREA ESTIMATES. P1 IS THE P-VALUE FOR SEPARATION OF HEALTHY ($\mathrm{KL} = 0$) FROM BORDERLINE TO OA ($\mathrm{KL} > 0$), AND P2 IS SEPARATION OF HEALTHY AND BORDERLINE ($\mathrm{KL} \leq 1$) FROM CLEAR OA CASES ($\mathrm{KL} > 1$). 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

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 [42].

We examine the ability to separate healthy from osteoarthritic populations of the volume and area estimates using an unpaired students t-test. The results are displayed in Tables III and IV, and since knees with $\mathrm{KL}i = 1$ 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 Figs. 5 and 6, the normalized volume and surface area estimates for medial tibial and femoral cartilage together are plotted against $\mathrm{KL}i$.

V. DISCUSSION

In this paper, we have presented a fully automatic framework for segmentation and quantitative assessment 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.

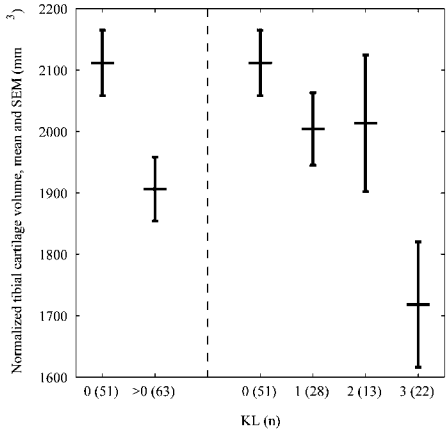

Fig. 5. Separation between different OA populations using the KLi and the normalized tibial medial cartilage volume from automatic (position adjusted) segmentations.

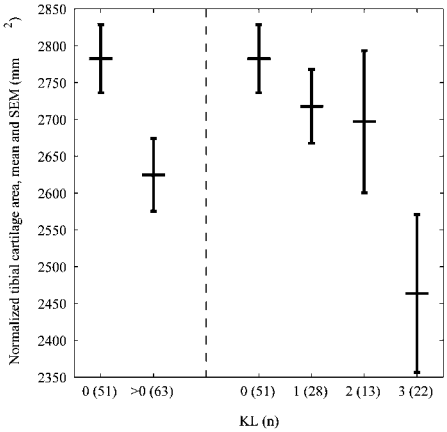

Fig. 6. Separation between different OA populations using the KLi and the normalized tibial medial cartilage surface area from automatic (position adjusted) segmentations.

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

The medial tibial cartilage gives 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 binary classifiers we not only avoid the problem of finding a criterion function for multiclass classification, we have also established a framework for multi-class classification that in the future can be extended to incorporate all cartilage compartments by incorporating binary 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 multicenter clinical studies since the method is completely automatic, has high reproducibility, and is robust to changes in knee placement in scanner.

REFERENCES

[1] D. Jackson, T. Simon, and H. Aberman, "Symptomatic articular cartilage degeneration: The impact in the new millenium," *Clin. Orthopaedics Related Res.*, vol. 391, pp. 14–25, 2001.

[2] D. T. Felson, R. C. Lawrence, M. C. Hochberg, T. McAlindon, P. A. Dieppe, M. A. Minor, S. N. Blair, B. M. Berman, J. F. Fries, M. Weinberger, K. R. Lorig, J. J. Jacobs, and V. Goldberg, "Osteoarthritis: New insights, part 2: Treatment approaches," *Ann. Int. Med.*, vol. 133, no. 7, pp. 726–737, Nov. 2000.

[3] H. Graichen, R. Eisenhart-Rothe, T. Vogl, K.-H. Englmeier, and F. Eckstein, "Quantitative assessment of cartilage status in osteoarthritis by quantitative magnetic resonance imaging," *Arthritis Rheumatism*, vol. 50, no. 3, pp. 811–816, Mar. 2004.

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

[5] C. 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*, vol. 14, supplement A, pp. 95–111, 2006.

[6] B. Ejbjerg, E. Narvestad, S. J. 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," *Ann. Rheumatic Diseases*, vol. 13, 2005.

[7] K. Woertler, M. Strothmann, B. Tombach, and P. Reimer, "Detection of articular cartilage lesions: Experimental evaluation of low- and high-field-strength mr imaging at 0.18 and 1.0 t," *J. Magnetic Resonance Imag.*, vol. 11, pp. 678–685, 2000.

[8] 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," *Arch. Orthopaedic Trauma Surg.*, vol. 114, no. 5, pp. 281–286, 1995.

[9] B. Kersting-Sommerhoff, P. Gerhardt, W. Golder, N. Hof, K. 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.*, vol. 162, no. 5, pp. 390–395, 1995.

[10] 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.*, vol. 7, pp. 37–41, 1999.

[11] R. W. Huegli, P. F. Tirman, H. M. Bonel, H. Staedele, S. Zaim, M. Grigorian, and H. K. Genant, "Use of the modified three-point dixon technique in obtaining t1-weighted contrast-enhanced fat-saturated images on an open magnet," *Eur. J. Radiol.*, vol. 11, no. 7, pp. 473–474, Jul. 2003.

[12] T. Stammberger, F. Eckstein, M. Michaelis, K.-H. Englmeier, and M. Reiser, "Interobserver reproducibility of quantitative cartilage measurements: Comparison of b-spline snakes and manual segmentation," *Magn. Resonance Imag.*, vol. 17, no. 7, pp. 1033–1042, 1999.

[13] J. A. Lynch, S. Zaim, J. Zhao, A. Stork, C. G. Peterfy, and H. K. Genant, "Cartilage segmentation of 3-D mri scans of the osteoarthritic knee combining user knowledge and active contours," *Proc. SPIE Med. Imag. 2000: Image Process.*, vol. 3979, pp. 925–935, 2000.

[14] J. A. Lynch, S. Zaim, J. Zhao, A. Stork, C. G. Peterfy, and H. K. Genant, "Automatic measurement of subtle changes in articular cartilage from mri of the knee by combining 3-D image registration and segmentation," *Proc. SPIE Med. Imag. 2001: Image Process.*, vol. 4322, pp. 431–439, 2001.

[15] S. Solloway, C. Hutchinson, J. Vaterton, and C. Taylor, "The use of active shape models for making thickness measurements of articular cartilage from mr images," *Magnetic Resonance Med.*, vol. 37, pp. 943–952, 1997.

[16] V. Grau, A. Mewes, M. Alcaiz, R. Kikinis, and S. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 447–458, Apr. 2004.

[17] S. K. Pakin, J. G. Tamez-Pena, S. Totterman, and K. J. Parker, "Segmentation, surface extraction and thickness computation of articular cartilage," *Proc. SPIE Med. Imag. 2002: Image Process.*, vol. 4684, pp. 155–166, 2002.

[18] J. G. Tamez-Pena, M. Barbu-McInnis, and S. Totterman, "Knee cartilage extraction and bone-cartilage interface analysis from 3-D mri data sets," *Proc. SPIE Med. Imag. 2004: Image Process.*, vol. 5370, pp. 1774–1784, 2004.

[19] S. K. Warfield, M. Kaus, F. A. Jolesz, and R. Kikinis, "Adaptive, template moderated, spatially varying statistical classification," *Med. Image Anal.*, no. 4, pp. 43–55, 2000.

[20] S. K. Warfield, C. Winalski, F. A. Jolesz, and R. Kikinis, "Automatic segmentation of mri of the knee," presented at the ISMRM 6th Sci. Meeting Exhibition, Sydney, Australia, Jul. 1998.

[21] K. Li, S. Millington, X. Wu, D. Z. Chen, and M. Sonka, "Simultaneous segmentation of multiple closed surfaces using optimal graph searching," in *Information Processing in Medical Imaging: 19th International Conference*. New York: Springer, 2005, vol. 3565, Lecture Notes in Computer Science.

[22] S. Millington, K. Li, X. Wu, S. Hurwitz, and M. Sonka, "Automated simultaneous 3-D segmentation of multiple cartilage surfaces using optimal graph searching on mri images," *Osteoarthritis Cartilage*, vol. 13, 2005.

[23] T. Dunn, Y. Lu, H. Jin, M. Ries, and S. Majumdar, "T2 relaxation time of cartilage at mr imaging: Comparison with severity of knee osteoarthritis," *Radiology*, vol. 232, no. 2, pp. 592–598, 2004.

[24] Y. Xia, "The total volume and the complete thickness of cartilage determined by mri," *Osteoarthritis Cartilage*, vol. 14, pp. 1781–1786, 2004.

[25] J. Kellgren and J. Lawrence, "Radiological assessment of osteoarthrosis," *Ann. Rheumatic Diseases*, vol. 16, no. 4, 1957.

[26] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," *ACM-SIAM. Discrete Algorithms*, no. 5, pp. 573–582, 1994.

[27] J. Folkesson, O. F. Olsen, P. Pettersen, E. Dam, and C. Christiansen, "Combining binary classifiers for automatic cartilage segmentation in knee mri," in *ICCV 1st Int. Workshop: Comput. Vision Biomed. Imag. Appl.*, 2005, pp. 230–239.

[28] J. Folkesson, E. Dam, O. F. Olsen, P. Pettersen, and C. Christiansen, "Automatic segmentation of the articular cartilage in knee mri using a hierarchical multi-class classification scheme," in *8th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention (MICCAI'05)*, Palm Springs, CA, 2005, pp. 327–334.

[29] C.-H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub, "Molecular classification of multiple tumor types," *Bioinformatics*, vol. 1, no. 1, pp. 316–322, 2001.

[30] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. W. III, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Acad. Radiol.*, vol. 11, pp. 178–189, 2004.

[31] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[32] J. J. Koenderink, "The structure of images," *Biol. Cybern.*, vol. 50, pp. 363–370, 1984.

[33] L. Florack, "The syntactical structure of scalar images," Ph.D. dissertation, Univ. Utrecht, Utrecht, The Netherlands, Oct. 1993.

[34] J. Blom, "Topological and geometrical aspects of image structure," Ph.D. dissertation, Utrecht Univ., Utrecht, The Netherlands, 1992.

[35] M. Descoteaux, L. Collins, and K. Siddiqi, "Geometric flows for segmenting vasulature in mri: Theory and validation," in *7th Int. Conf. Med. Image Comput. Comput.-Assisted Int. (MICCAI'04)*, St Malo, France, 2004, vol. 3216, pp. 500–507.

[36] J. Weickert, *Anisotropic Diffusion in Image Processing*. Stuttgart, Germany: Teubner-Verlag, 1998.

[37] E. B. Dam, J. Folkesson, M. Loog, P. C. Pettersen, and C. Christiansen, "Efficient automatic cartilage segmentation," in *MICCAI Joint Disease Workshop*, Copenhagen, Denmark, 2006 [Online]. Available: http://www.itu.dk/image/joint

[38] J. Folkesson, E. B. Dam, O. F. Olsen, P. C. Pettersen, and C. Christiansen, "Position normalization in automatic cartilage segmentation," in *MICCAI Joint Disease Workshop*, Copenhagen, Denmark, 2006 [Online]. Available: http://www.itu.dk/image/joint

[39] 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*, vol. 44, no. 9, pp. 2072–2077, Sept. 2001.

[40] T. Stammberger, F. Eckstein, K.-H. Englmeier, and M. Reiser, "Determination of 3-D cartilage thickness data from mr imaging: Compputational method and reproducibility in the living," *Magnetic Resonance Med.*, vol. 41, pp. 529–536, 1999.

[41] J. Hohe, G. Ateshian, M. Reiser, K.-H. Englmeier, and F. Eckstein, "Surface size, curvature analysis, and assessment of knee joint incongruity with mri in vivo," *Magnetic Resonance in Medicine*, no. 47, pp. 554–561, 2002.

[42] E. B. Dam, J. Folkesson, P. C. Pettersen, and C. Christiansen, "Automatic cartilage thickness quantification using a statistical shape model," in *MICCAI Joint Disease Workshop*, Copenhagen, Denmark, 2006 [Online]. Available: http://www.itu.dk/image/joint