Int J CARS
DOI 10.1007/s11548-012-0671-z

ORIGINAL ARTICLE

Voxel classification and graph cuts for automated segmentation of pathological periprosthetic hip anatomy

Daniel F. Malan · Charl P. Botha · Edward R. Valstar

Received: 28 September 2011 / Accepted: 6 January 2012

Abstract
Purpose Automated patient-specific image-based segmentation of tissues surrounding aseptically loose hip prostheses is desired. For this we present an automated segmentation pipeline that labels periprosthetic tissues in computed tomography (CT). The intended application of this pipeline is in pre-operative planning.
Methods Individual voxels were classified based on a set of automatically extracted image features. Minimum-cost graph cuts were computed on the classification results. The graph-cut step enabled us to enforce geometrical containment constraints, such as cortical bone sheathing the femur's interior. The solution's novelty lies in the combination of voxel classification with multilabel graph cuts and in the way label costs were defined to enforce containment constraints.
Results The segmentation pipeline was tested on a set of twelve manually segmented clinical CT volumes. The distribution of healthy tissue and bone cement was automatically determined with sensitivities greater than 82% and pathological fibrous interface tissue with a sensitivity exceeding 73%. Specificity exceeded 96% for all tissues.
Conclusions The addition of a graph-cut step improved segmentation compared to voxel classification alone. The pipeline described in this paper represents a practical approach to segmenting multitissue regions from CT.

Keywords Segmentation · Graph cut · Voxel classification · Osteolysis · Computed tomography

D. F. Malan (✉) · E. R. Valstar
Department of Orthopaedics, Leiden University Medical Center, J11-R, Albinusdreef 2, 2333 ZA Leiden, The Netherlands
e-mail: fmalan@medvis.org

D. F. Malan · C. P. Botha
Department of Mediamatics, EEMCS, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands

C. P. Botha
Department of Radiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands
e-mail: c.p.botha@tudelft.nl

E. R. Valstar
Department of Biomechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
e-mail: e.r.valstar@lumc.nl

Introduction

Periprosthetic osteolysis leading to aseptic loosening is one of the foremost problems limiting the survival of hip prostheses [1]. Loosening caused by osteolysis is characterized by extensive resorption of bone and its replacement by soft fibrous interface tissue that offers little mechanical stability. Surgical treatment becomes necessary when prosthesis loosening ensues. During open revision surgery, the old prosthesis and its cement mantle, along with surrounding fibrous interface tissue, are removed, after which a new prosthesis is placed.

Revision surgery is very demanding on the patient; therefore, experimental techniques substitute open surgery with minimally invasive cement injection to fixate the loosened prosthesis [2,3]. At the time of writing these procedures are annually only performed on a handful of patients, with a much larger potential target group if proven successful. During minimally invasive refixation the surgeon is limited to working under fluoroscopic guidance and can only apply cement to two or three injection sites. Proper pre-operative planning is therefore essential. Femoral strength and stability may be simulated using finite element (FEM) modelling [4,5]

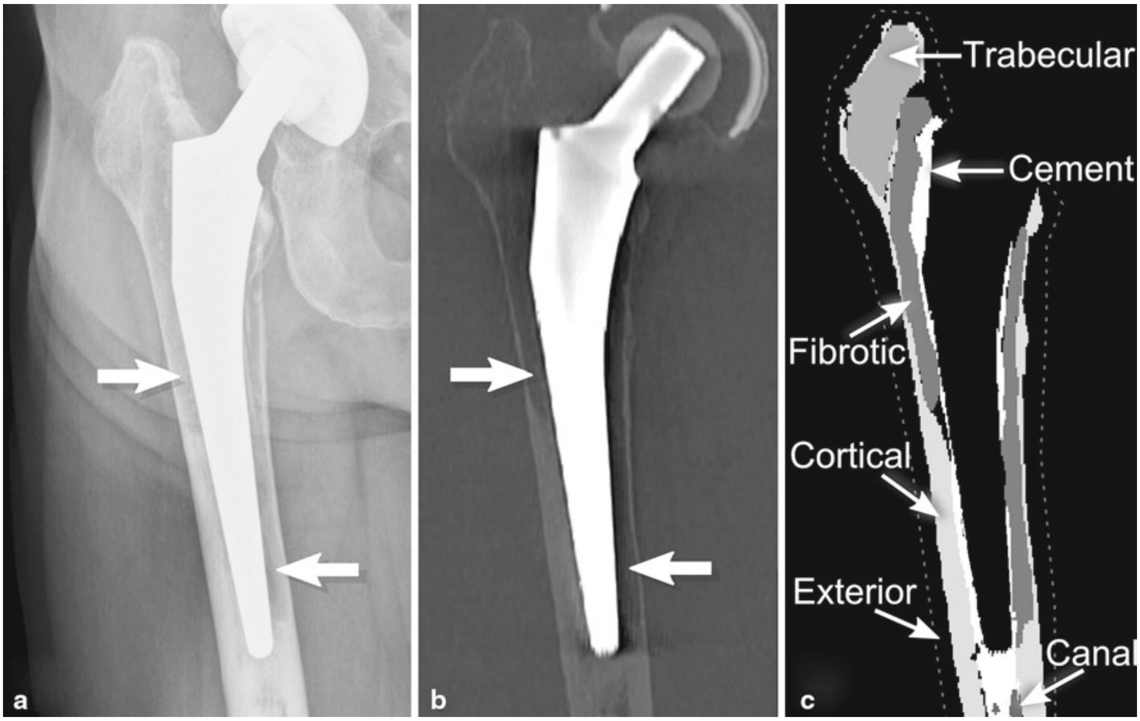

Fig. 1 **a** Coronal X-ray radiograph of femur with osteolysis (*arrows*), **b** coronal CT slice of the same hip, **c** manual segmentation showing periprosthetic tissues. The boundary of the region of interest (ROI) is indicated by the *dotted line*

but requires three-dimensional (3D) tissue segmentation for creation of patient-specific models.

Plain radiographs such as in Fig. 1a are the default imaging modality for diagnosing osteolysis [6]. While sufficient for diagnosis, radiographs do not capture the 3D distribution of periprosthetic tissues, where computed tomography (CT) remains the imaging modality of choice [7,8]. Unfortunately, CT suffers from image degradation in the vicinity of metal prostheses [9,10]. Image degradation makes the 3D classification of periprosthetic tissues a difficult task—especially for low-contrast tissues other than cortical bone. Patients suffering from osteolysis generally have very low bone quality, which exacerbates segmentation problems. In some cases cortical bone, normally thick and easy to discern, is reduced to a thin shell of its former extent and may show regions of very low image intensity.

Manual segmentation of this kind of volume is difficult and too labour intensive for routine use. As an alternative, automatic or semiautomatic techniques have been developed. To segment skeletal structures, Zoroofi et al. [11] use histogram-based thresholding and binary morphological steps. Kang et al. [12] use an automatic region-growing technique augmented by manual correction. Yokota et al. [13] segment the boundary of diseased hip bones with a hybrid statistical shape model. Statistical models based on principal component analysis require well-defined natural shape and tissue distributions [14] and are therefore, similar to atlas-based methods [15], ill-suited to sporadic lesions and surgically modified joints fitted with prostheses.

We set out to develop an automatic 3D CT segmentation pipeline that can segment all mechanically distinct tissues in hip CT volumes, including periprosthetic osteolytic lesions. The envisioned application was to assist with pre-operative planning and the creation of patient-specific finite element models to analyze prosthesis stability.

The main contributions of this work are the following:

- We extend the prototype voxel classification scheme of Malan et al. [16] while simultaneously reducing the feature set to an optimized subset. Using this reduced feature set, we implement a k-centres + k-nearest neighbours voxel classifier.
- We use s/t graph cuts [17] to obtain a "hard" multilabel tissue segmentation from the probabilistic tissue map computed by the aforementioned voxel classifier. To our knowledge this is the first medical image segmentation application of multilabel graph cuts to the output of a probabilistic voxel classifier.
- Following the example of Delong and Boykov [18], we incorporate geometric containment constraints as part of the graph-cut segmentation. The novelty of our approach lies in our definition of the data cost term, which enables

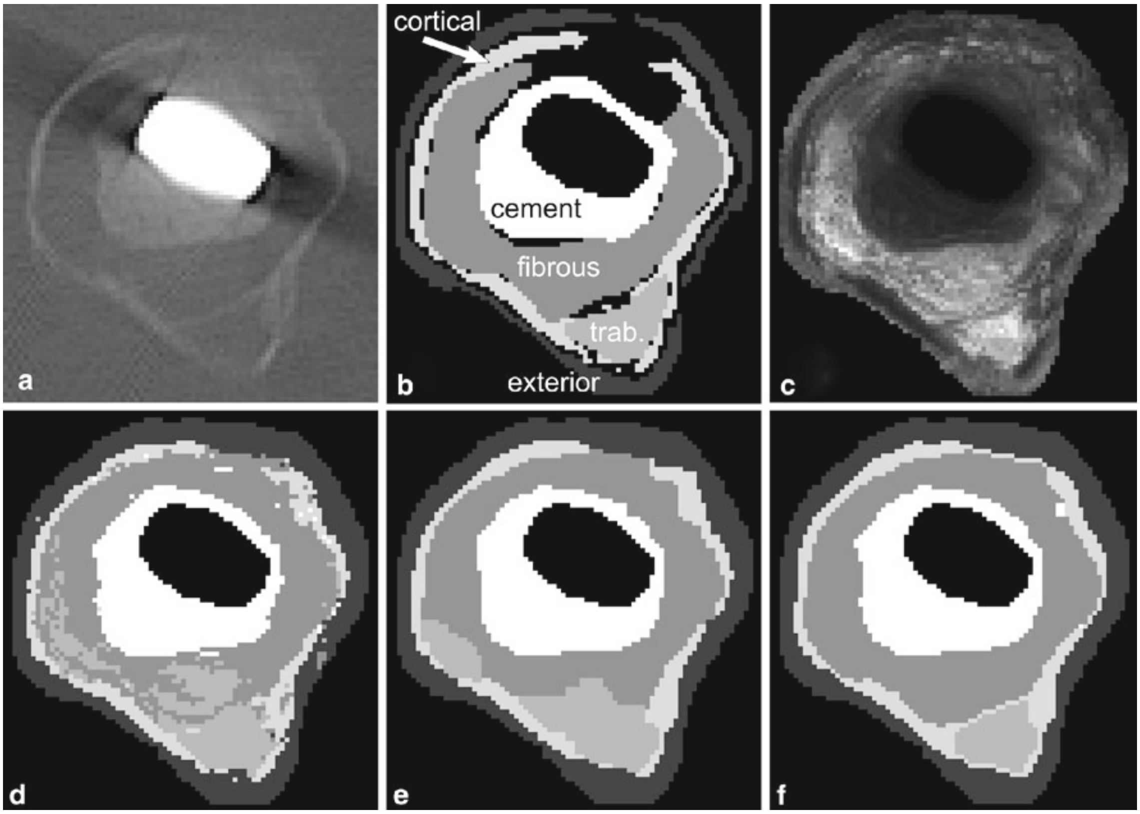

Fig. 2 Cross-section through the proximal femur. **a** Original CT, **b** manually designated tissue regions. *Black areas* were left unassigned, **c** classifier probability map for trabecular bone. Each tissue class has such a map. **d** Maximum posterior probability, **e** segmentation by multilabel graph cut without containment, **f** segmentation by multilabel graph cut with containment

us to define per-node costs. This enables us to use a publically available multilabel graph-cut software library [19] for solving either the unconstrained or constrained case.

Materials and methods

Imaging

We retrieved twelve anonymized clinical CT data sets from twelve patients suffering aseptically loose femoral prostheses. All scans were made with Toshiba Aquilion (Toshiba Medical Systems, Japan) scanners using the FC30 “bone kernel”. We retroactively obtained clinical data and therefore had to accept inter-scan variation, most notably the tube current (150–400 mA), peak voltage (120 or 135 kV) and in-plane voxel spacing (0.44–0.59 mm).

Manual segmentation

An experienced operator (DM) manually segmented each CT volume using the Medical Imaging Interaction Toolkit (MITK) version 0.12.2 [20]. Segmentation was performed using free-hand drawing on intermittent slices with intermediate contours completed by interpolation.

First, a region of interest (ROI) was delimited. The metal prosthesis was removed by thresholding at 5,000 Hounsfield Units (HU). The ROI was then segmented into separate regions of cement, cortical bone, trabecular bone, fibrous interface tissue, intramedullary canal, and exterior muscle tissue. Figures 1c and 2b show examples of segmentations thus obtained. It proved difficult to manually distinguish some regions in the low-image-quality CT volumes. We left these regions unclassified to prevent false positives in the classifier training sets. An example of a difficult to segment region is shown in the upper part of Fig. 2a. The manual segmentations’ correctness was verified by an experienced orthopaedic surgeon (RN). Inter-observer tissue segmentation was evaluated by having a second independent human operator re-segment four randomly selected femurs.

Performance metrics

We used the manually segmented CT image volumes to train voxel classifiers and secondly to serve as ground truth for evaluating the performance of our automatic tissue

Table 1 The thirteen candidate image features

1	CT volume at original resolution ($\sim$0.5 mm $\times$ 0.5 mm $\times$ 0.5 mm)
2	Metal artefact reduced CT volume. Computed using sinogram interpolation
3	CT smoothed with $\sigma = 0.5$ mm isotropic Gaussian filter
4	CT smoothed with $\sigma = 1.0$ mm isotropic Gaussian filter
5	CT smoothed with $\sigma = 2.0$ mm isotropic Gaussian filter
6	Image Gradient magnitude of feature nr 3
7	Image Gradient magnitude of feature nr 4
8	Image Gradient magnitude of feature nr 5
9	Distance (in mm) to the threshold-segmented prosthesis surface
10	Signed distance from prosthesis head, along prosthesis long axis
11	Signed distance from estimated convex hull of the femur's cortical bone
12	Cosine of angle to prosthesis neck in plane perpendicular to long axis
13	Signed radial gradient of MAR volume smoothed by σ = 0.5 mm Gaussian filter

segmentation. A rotating per-patient leave-one-out scheme was used. Voxel classifier performance was evaluated by counting the percentage of correctly classified voxels per tissue class, shown as a confusion matrix in Table 6. Similar to van der Lijn et al. [21] we evaluated the final segmented volumes by their Dice similarity coefficients compared to ground truth. The Dice coefficient is a value between 0 and 1, where 1.0 represents perfect agreement and 0.0 represents completely disjoint segmentations. The coefficient is defined as $2(|A \cap B|)/(|A| + |B|)$, where $|A|$ denotes the volume of region A and $A \cap B$ is the intersection of regions A and B [22].

Computation of image features

To serve as input for voxel classification, thirteen candidate image feature volumes, listed in Table 1, were computed from every original CT volume.

We used the method of Kalender et al. [23], implemented in MATLAB R2009b (Mathworks Inc., MA, USA), to compute the metal artefact reduced (MAR) volume. All other image features were computed using proprietary software developed in the DeVIDE Runtime Environment [24].

Similar to previous studies [25–27] we used multiscale image and image gradient values as features. Isotropic Gaussian low-pass (blurring) with standard deviations of 0.5, 1 and 2 mm, along with their image gradient magnitudes, were computed from the original CT volume.

The next feature consisted of the shortest Euclidian distance to the prosthesis surface. The prosthesis was automatically detected as the largest object exceeding a threshold of 5,000 HU. This value was appropriate for all cobalt-chromium and steel prostheses.

The tenth feature consisted of the signed distance from the prosthesis head's centroid, measured parallel to the femur's long axis. The prosthesis's long axis was automatically computed as the first principal spatial component of the prosthesis' distal half. This direction closely corresponds to the alignment of the femur's long axis.

The next feature was the signed distance from the femur's outer surface. The geometry of the femur is initially unknown to us. By thresholding the ROI between 800 and 3,000 HU and excluding all voxels within 3.5 mm of metal components, we capture the majority of voxels representing cortical bone. Recall that CT image slices are always approximately perpendicular to the femur's long axis. While a human femur is not convex in three dimensions, it is approximately convex in any cross-section perpendicular to its long axis. For every image slice we therefore approximated the femur's outline by the two-dimensional convex hull of the thresholded voxels.

The twelfth candidate feature was the cosine of the angle of each voxel relative to that of the femur head. This planar angle cosine was measured around the central axis of the prosthesis stem.

The last candidate feature was the image gradient of the smoothed MAR image, using a 0.5 mm Gaussian kernel, computed in the radial direction perpendicular to the prosthesis' long axis.

Classifier construction and feature selection

We constructed statistical classifiers using PRSD Studio (PR Sys Design, Delft, The Netherlands) [28], a toolbox that offers implementations of various classification algorithms. Our training data consisted of all twelve manually segmented CT volumes, where individual image voxels were regarded as separate objects. All classifiers were trained and evaluated using a rotating per-patient leave-one-out scheme.

We defined six tissue classes: cortical bone, trabecular bone, bone cement, fibrotic tissue, intramedullary canal, and tissue exterior to the femur. Equal per-class priors were used to prevent infrequent but important tissues like fibrotic zones being suppressed during optimization. Features were scaled to have unit variance. Both forward and backward selection processes were then used to determine an optimal subset of the thirteen candidate image features.

Each classifier computed a "soft" probabilistic classification, that is, probabilities of belonging to the six tissue classes instead of "hard" unambiguous labels. Parametric classifiers that construct internal probability density functions directly output soft classifications. Others like the kNN classifier output feature-space distances that were subsequently converted to unit-sum probabilities by using the normalized inverse of their class separation distances.

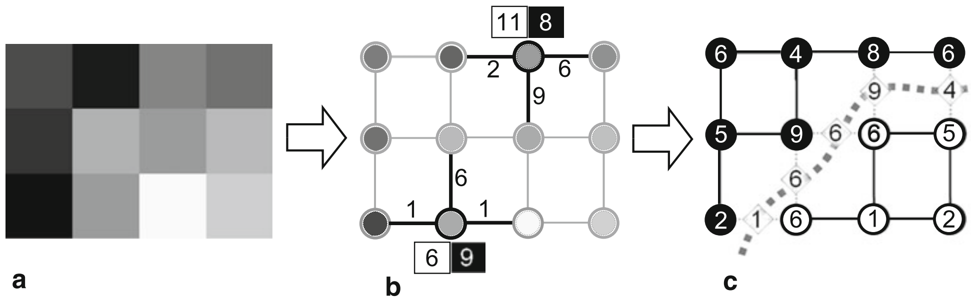

Fig. 3 An example 2D graph cut. **a** 3 × 4 bitmap image. **b** Graph with two highlighted nodes to show data costs for being assigned either "white" or "black". Edge weights were chosen inversely proportional to image gradient. **c** A cut of the graph with resultant data cost for every node and the smoothness cost for every severed edge

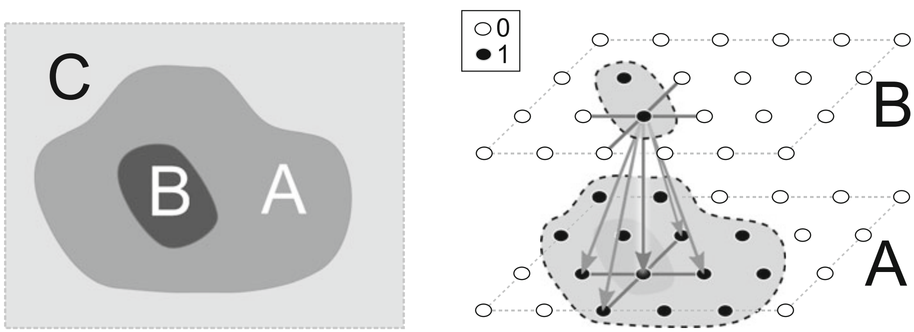

Fig. 4 *Left*: a 2D image with three regions. *Right*: each pixel p is represented by a node pair x_p^A and x_p^B. Edges are shown for a single node's edge pair. Smoothness costs are represented by undirected in-plane edges. The "A contains B" relationship is enforced by an infinity-weight directed edge from plane B to A. "B sheathed in at least 1 pixel width of A" is enforced by directed edges to its 4-connected neighbourhood

Classification performance was compared between the following classification algorithms available in PRSD Studio: linear, quadratic, Gaussian mixture model, kNN with k equal to 1, 3 and 10, a Parzen classifier, neural net, naive Bayes and a decision tree. A *k-centres algorithm was used as* kNN *pre-processing step*. Each classifier was trained and tested identically.

Segmentation by maximum posterior probability

The most straightforward approach for converting the "soft" classifier output to a "hard" segmentation was by independently assigning, for every voxel, the class with highest posterior probability. An example of the classification results obtained with this scheme is shown in Fig. 2d.

Segmentation by graph cuts

Maximum posterior probability classification is prone to noise and irregular segmentation geometry as seen in Fig. 2d. To counter this we instead used multilabel graph cuts to transform the soft classifier output into a multilabel segmentation. For this we used and adapted the publically available gco-v3.0 multilabel graph-cut library [19].

An image may be expressed as a graph by representing individual image pixels or voxels as graph nodes and representing their neighbourhood relationships with edges. Graph-cut algorithms, also known as "minimum cut/maximum flow" algorithms, offer a computationally efficient way of minimizing certain energy functions defined on a graph [29,30]. Figure 3 shows how a 2D image may be segmented by a graph cut.

A cut is minimal if its associated energy is smaller or equal to the energy of any other cut of the same graph. This energy typically consists of two parts: a data cost and a smoothness cost [31]. The data cost of each node is computed from the mismatch between its observed properties and those of its assigned label. Following the example of Boykov et al. [17] we defined each voxel's data costs as the negative logarithm of its per-class membership probability. The smoothness cost of each severed edge may be defined inversely proportional to local image gradient.

Delong and Boykov [18] show how a directed binary graph with two nodes for every image pixel may be constructed to enforce geometric containment, attraction or exclusion. A 2D example of a containment enforcing graph is shown in Fig. 4. Every image pixel is represented by two graph nodes—these node pairs define two separate planes. Within each plane, every node is connected to its neighbours using 4-connected

Table 2 The pairwise data costs of the nodes in Fig. 4 as defined by Delong and Boykov [18]

x_p^A value	x_p^B value	Label of p defined by (x_p^A, x_p^B)	Data cost D_p
0	0	C	$-\log(Pr(C))$
0	1	n/a	K
1	0	A	$-\log(Pr(A))$
1	1	B	$-\log(Pr(B))$

K is an arbitrary constant since the infinity-weight edge prevents a (0, 1) value pair from occurring

Table 3 Our data costs depend only on each node's own binary label

Variable	Value	Data cost D_p
x_p^A	0	$-\log(Pr(C)) + \max A$
x_p^A	1	$-\log(Pr(A)) + \max A$
x_p^B	0	$\max A$
x_p^B	1	$\log(Pr(A)) - \log(Pr(B)) + \max A$

Compared to Table 2 the global energy differs by a constant and has the same labelling as solution

undirected (or bidirectional) edges. Directed edges link node pairs between the planes to enforce containment and attraction relationships.

The graph for a 3D voxel grid is analogous to the 2D model, the difference being the replacement of the 4-connected pixel-node planes A and B with two 8-connected voxel-node grids. A binary cut on this graph assigns every node a value of one or zero. The label of every pixel is defined by the assignment of binary values to its corresponding node pair, as in Table 2.

The difference of our approach compared to that of Delong and Boykov is that we assigned data costs to individual nodes in the graph—dependent only on the binary values which the nodes may individually assume, not on pairwise labelling. The advantage of this is that it allows computation of the minimum graph cut using standard graph-cut libraries such as gco-v3.0 [19].

Starting with Table 2 we replaced the costs of pairwise node assignments by the summed costs of each pair's two separate components, leading to Table 3. This new definition yields the same pairwise costs, up to a constant. For example: the data cost of assigning (0, 0) to x_p^A and x_p^B becomes $[-\log(Pr(C)) + \max A] + [\max A] = -\log(Pr(C)) + 2 * \max A$. We defined $\max A$ as $\max(-\log(A))$, computed across all voxels. This constant term prevents any individual data costs from being negative—a requirement of the gco-v3.0 library. The additional global data cost is therefore $\max A$ times the number of nodes, a known constant. Since the same labelling minimizes all equivalent energy functions that differ by a constant, the solution is an equivalent labelling.

Table 4 Symmetric smoothness cost matrix used for the standard multi-label α-expansion graph cut algorithm

	Outside	Canal	Fibrous	Trabecular	Cortical	Cement
Outside	0	2	2	2	1	2
Canal	–	0	1	2	1	1
Fibrous	–	–	0	1	1	1
Trabecular	–	–	–	0	1	1
Cortical	–	–	–	–	0	1
Cement	–	–	–	–	–	0

Segmentation by single multilabel graph cut on classifier output

We started with smoothness costs similar to the Potts model of Boykov et al. [17,31], but with larger values assigned to tissues that are expected to be non-adjacent, such as between exterior muscle tissue and either the intramedullary canal or fibrous interface tissue. These costs, shown in Table 4, satisfy the definition of a metric as required for convergence of the α-expansion algorithm [32], that is, $V(\beta, \alpha) \leq V(\alpha, \gamma) + V(\gamma, \beta)$ and $V(\alpha, \alpha) = 0$ for all labels α, β, γ.

The costs of Table 4 were additionally scaled with a spatially varying term that depended on the image intensity gradient at each voxel's location. This factor was defined as $S_d = \exp(-c * G_d)$, having a value of 1.0 in areas with zero gradient and exponentially decaying with increasing gradient. The subscript "d" refers to the orthogonal direction, that is, "x", "y" or "z". G_d is the CT image gradient magnitude in the given direction, expressed in HU/mm, and "exp" refers to the exponential function. The parameter c is a scaling parameter that we set to 0.008 to provide the desired falloff rate.

Segmentation by stepwise multilabel graph cut with geometrical containment

The gco-v3.0 library used for solving the unconstrained multilabel problem does not support directed graphs or labels defined on pairs of nodes. Since graph directionality is essential to the containment relationships in Figs. 4 and 5, we modified the code to enable edge directionality. The first modification consisted of allowing asymmetric neighbourhood relationships; that is, node p being a neighbour of node q does not imply that node q must be a neighbour of p. The second modification was for allowing the asymmetric smoothness costs of Table 5 that were subsequently scaled to also be inversely proportional to image gradient as described in the previous section.

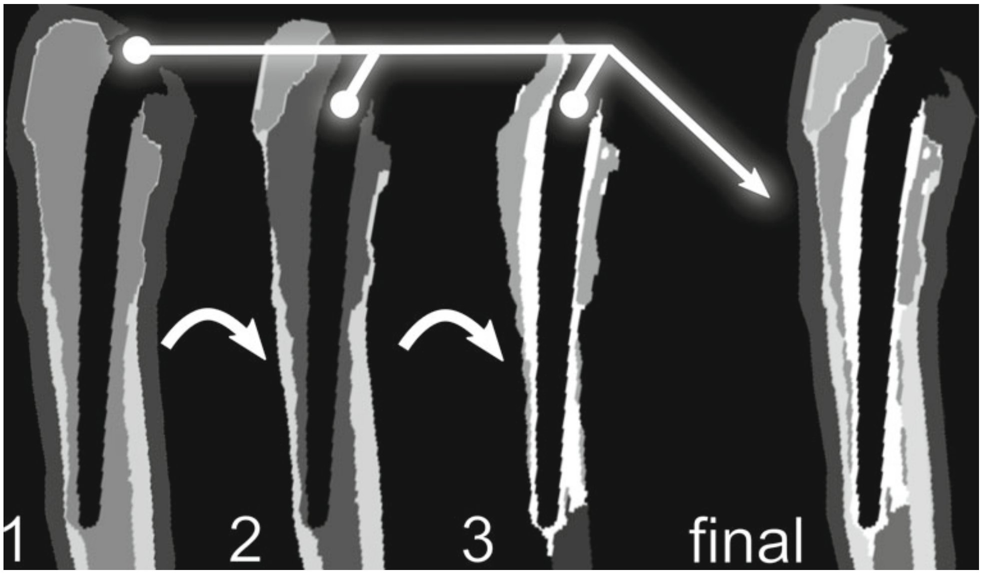

Fig. 5 A coronal slice showing the three-step segmentation process. Steps 1 and 2 enforce geometrical containment relationships

Table 5 Smoothness cost matrix used for the containment-enforcing two-layer binary-valued graph of Fig. 4

	Value 0	Value 1
Value 0	0	1
Value 1	0	0

Looking at Fig. 4 and Table 2 we see that the asymmetry of Table 5 ensures that the "infinity" cost of a containment-rule violation is correctly enforced. Assigning (0, 1) to the (x_p^A, x_p^B) node pair is illegal and does not correspond to any tissue label. This transgression is prohibited by the smoothness cost of 1 multiplied by the graph's "infinity"-weight edge, resulting in "infinite" cost. Conversely, the $(1, 0)$ assignment corresponding to the label "A" is allowed. Multiplying the zero smoothness cost and "infinite" edge weight yields zero cost. Spatially neighbouring voxels are connected by bidirectional edges equivalent to the undirected edges of Fig. 4. Here, label discontinuities are penalized as before, since exactly one of the bidirectional edges will cross a (0, 1) transition.

We used containment relationships to force the region defined by the union of intramedullary canal, cement, fibrous interface tissue and trabecular bone to be enclosed in a layer of cortical bone with a thickness of at least one voxel. This was motivated by the fact that the whole femur is enclosed in cortical bone, even though this layer can be very thin and difficult to discern in the proximal femur. Likewise, we created a penalty term to discourage trabecular bone from not being enclosed in a layer of cortical bone of at least a single-voxel thickness. An example of this is shown in Fig. 2e where an unconstrained graph-cut solution allows holes in the encompassing cortical bone. In Fig. 2f we see that the containment constraint enforces a continuous cortical sheath.

Including multiple containment relationships in a single graph cut complicates graph construction and data cost terms. We instead opted for the data cost structure of Table 3 that only allows three tissue zones at a time. We desired two containment constraints—the femur having an uninterrupted cortical shell and trabecular bone being enclosed in cortical bone. Each of these constraints required a separate graph-cut step. The remaining tissues were separated using a standard α-expansion multilabel graph cut as described by Boykov et al. [17].

The resulting three-step segmentation process is shown in Fig. 5.

1. The ROI is segmented into three classes: exterior, cortical, and "interior" using a containment relationship that forces the interior to be enclosed in a cortical shell of at least single-voxel thickness.
2. The "interior" and "cortical" regions are re-segmented into trabecular, cortical and "rest". A containment relationship similar to step 1 specifies trabecular bone to be enclosed in cortical bone. The final cortical region consists of the union of the cortical regions from steps 2 and 3.
3. Finally, the "rest" tissue from step 2 is segmented into cement, fibrous and canal using a three-label α-expansion graph cut without containment restriction.

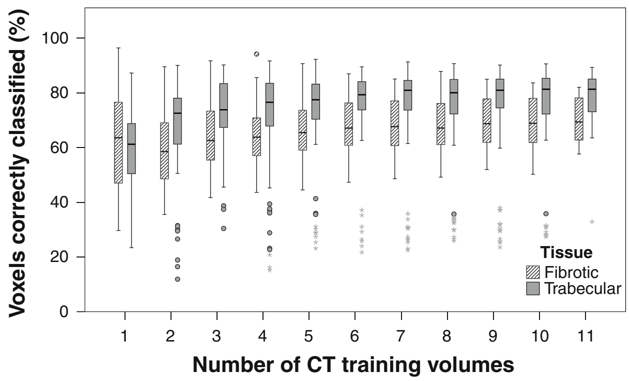

Fig. 6 Voxel classifier sensitivities for the two most difficult-to-classify tissues versus number of CT training volumes

Results

Feature and classifier selection

Using both forward and backward feature selection on disjoint training and test sets, we identified an optimal subset of nine features. Removing any of these or adding any of the remaining features increased the classification error. Referring to Table 1, these nine features, ranked from most to least important, were numbers 10, 11, 2, 4, 9, 13, 7, 1, 5. It is important to note that these features were not necessarily the best individual discriminators, but instead provided the best combined classification power when used as a set.

Using these nine features, we evaluated different classifier algorithms. We found the best overall classification performance using a kNN classifier with $k = 3$, after pre-processing the input data with a k-centres algorithm with $k = 2,000$. We evaluated the classifier's learning curve as it was trained on successively larger patient sets. Testing was performed in a per-patient leave-one-out manner. The results for the difficult-to-classify fibrous and trabecular tissues are plotted in Fig. 6.

We see that classification performance tended to stabilize once a training set size of at least five CT volumes was reached. Further increasing the number of training sets did not significantly improve median classification performance but did reduce the occurrence of negative outliers.

Tissue segmentation

In Table 6 we see that the classifier using the maximum probability criterion managed to correctly classify the large majority of labelled voxels. The most difficult tissues were fibrous interface tissue (66.7%) and trabecular bone (81.1%)—both being low in CT image contrast and located close to the prosthesis in the zone most affected by metal artefacts. Graph-cut post-processing improved classification performance relative to maximum posterior classification—most notably for the fibrotic and trabecular tissue classes. Sensitivity and specificity of the final constrained graph-cut segmentation are shown in Table 7.

In the example slice of Fig. 2 we see that the geometrically unconstrained graph-cut procedure considerably smoothed the segmentation result, with many of the isolated noisy misclassifications removed. The same general tissue distribution was maintained, including a gap in the outer shell of cortical bone. With the constrained graph-cut approach we note that the gap in the outer cortical shell has been closed, and the region of trabecular bone is fully enclosed and shielded from fibrous tissue by a layer of cortical bone as was required by our containment rule. We note from Fig. 7 that there is still

Table 6 Confusion matrix showing median classification performance of our kNN2k classifier when labelling each voxel with its maximum posterior probability (MPP, upper rows) and constrained graph cuts (CGC, lower rows)

Label	Automatically segmented						
	Method	Exterior	Canal	Fibrotic	Trabecular	Cortical	Cement
Manually segmented							
Exterior	MPP	**85.71**	0.00	2.18	8.65	2.18	0.47
	CGC	**89.47**	0.00	1.53	4.76	2.09	0.70
Canal	MPP	0.00	**97.46**	0.31	0.00	0.52	0.96
	CGC	0.00	**98.50**	0.00	0.00	0.59	0.55
Fibrotic	MPP	2.04	0.51	**66.67**	11.70	4.45	3.83
	CGC	0.11	0.22	**73.04**	3.49	5.14	3.43
Trabecular	MPP	5.72	0.00	8.18	**81.08**	2.71	1.20
	CGC	2.76	0.00	8.28	**82.90**	3.86	0.22
Cortical	MPP	0.96	0.96	3.90	2.92	**86.79**	5.79
	CGC	4.08	0.71	3.43	3.90	**86.40**	3.27
Cement	MPP	0.42	1.24	5.80	1.17	5.19	**84.74**
	CGC	0.03	1.29	5.33	0.42	4.41	**86.64**

Bold values, which represent the sensitivities, i.e., the percentages of each tissue type that was classified correctly, form the main diagonal of the confusion matrix

Table 7 Sensitivity and specificity of the constrained graph-cut solution, computed from Table 6

Label	Sensitivity (%)	Specificity (%)
Exterior	89.47	98.60
Canal	95.50	99.56
Fibrotic	73.04	96.29
Trabecular	82.90	97.49
Cortical	86.40	96.78
Cement	86.64	98.37

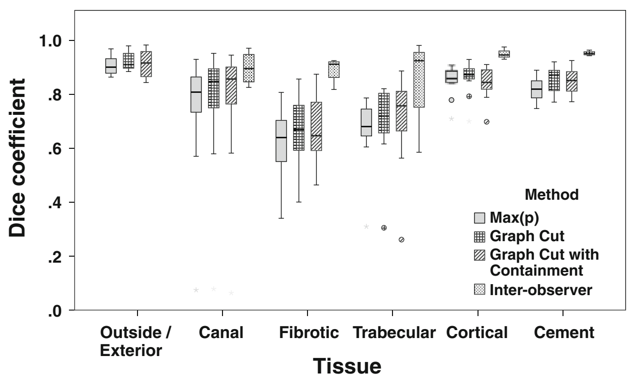

Fig. 7 Comparison of manually segmented ground truth with maximum posterior probability, graph cut and constrained graph cut. Inter-observer variability, computed with a second human segmenter, is also shown

an obvious difference between our automated segmentation methods and a second human segmenter but are encouraged by the overlapping Dice coefficients ranges.

Across all tissues we find that the graph-cut segmentations, both with and without geometric constraints, have significantly higher Dice coefficients than using maximum probability classification. This was confirmed using a Wilcoxon signed rank test ($p < 0.001$ in both cases). Compared to the graph-cut method without containment, however, geometric containment shows no statistically significant improvement in Dice coefficient ($p = 0.466$). Geometric containment had its greatest value in qualitative segmentation improvement—for example, the closing of holes in the cortical shell, as seen in Fig. 2. The lack of statistically significant improvement in Dice coefficient therefore tells only part of the tale.

Computational cost

For a typical CT volume of interest consisting of $200 \times 200 \times 300$ voxels, the total running time of our segmentation pipeline was approximately 15 min. This time was recorded on a 3.0 GHz Intel Core-i7 desktop computer running Microsoft Windows7 64-bit with 12 GB of working memory. Computation of the nine image features took approximately 10 min. Subsequent soft classification by the trained voxel classifier took 3 min. The post-processing of the classification output by the graph-cut algorithm took approximately 40 s regardless of whether containment relationships were specified or not. Since our algorithms currently rely on single-threaded operation, we envisage a substantial possible speed increase should we modify our code to harness processor cores simultaneously.

The most memory intensive operation was the graph-cut step where, in addition to MATLAB's base footprint of 200 MB, a peak amount of 600 MB without containment or 1,100 MB with containment was required.

We note that the execution time and memory requirements of feature generation, voxel classification and the graph-cut algorithm all show linear dependence on the number of classified voxels. This is in accordance with the findings of Delong and Boykov [18], and we experimentally affirmed it.

Discussion

We constructed and optimized a voxel classifier that uses a diverse set of automatically computable image features. The retention of several derived distance metrics in the optimal feature subset showed that image features other than intensities and gradients are beneficial. This was emphasized by observing that the original CT volume ranked as only the eighth most important feature.

The most straightforward way to convert statistical classifier output to a final labelling is by assigning, for each voxel, the label with maximum posterior probability. This classification method leads to noisy results with an excessive number of label transitions. We know that biological tissues tend to form contiguous regions. The maximum posterior probability classification does not incorporate this prior knowledge.

We showed that the α-expansion graph-cut algorithm of Boykov et al. [17] can be applied to obtain an improved segmentation result. The required data cost terms are easily computed from the voxel classifier's output probabilities, and the obtained results exhibit the contiguousness we expect. The resulting segmentations are a qualitative and quantitative improvement over standalone voxel classification.

Additional containment relationships, implemented as modifications to the method of Delong and Boykov [18], have their biggest effect on the segmentation of fibrotic tissue and trabecular bone. This desired result is as expected, since we specifically enforce the containment of these two tissues in a shell of cortical bone. The containment relationships simultaneously enable us to close unwanted holes in the segmentation of thin shell regions of cortical bone.

In Fig. 7, just as in Tables 6 and 7, we note that we had the least success with the softer, irregularly shaped, low-contrast fibrous tissue and trabecular bone. There is a notable outlier in each of the “Canal”, Trabecular and “Cortical” tissue classes, but given the small sample size and large inter-scan variation this not completely unexpected. Indeed the “canal” outlier occurred for a data set where almost no intramedullary canal was included in the defined ROI, thereby leading to a negligibly small volume and low Dice coefficient between successive segmentations. The outliers for trabecular and cortical bone both occurred for the same data set, which had an unusually small femur diameter. This highlights the importance of having a sufficiently large training set to cover the expected inter-patient variation—an assumption which failed for this single data set.

In Fig. 7 we saw that human inter-observer variability has a similar order of magnitude, albeit smaller, than differences between automated and human segmentation. Since we can only evaluate classifier performance relative to the manually segmented ground truth which is itself subject to error and simplification, subtle improvements due to geometrical containment may be obscured in our measurements.

Limitations to this study include the small number of CT volumes used for training and evaluation. Figure 6 suggests, however, that the twelve CT data sets used in this study were sufficiently representative for the goals of this paper.

Since all automatic segmentation algorithms can occasionally fail, it will be good to allow manual segmentation corrections in future work. This could be approached similar to the method of Egger et al. [33]. As in previous literature we used the Dice similarity coefficient to evaluate segmentation accuracy. In future work it will be important to examine the relationship between volumetric segmentation accuracy and derived modelling accuracy, such as when using finite elements.

A further limitation of this study is that all CT image volumes were obtained from the same make of scanner and from the same hospital. By pooling the available data sets we obtained a collection of image volumes containing real clinical data with heterogeneous scan parameters. We foresee the presented algorithm to work similarly well on data from other centres but did not have the opportunity to verify this claim.

Despite CT’s proven diagnostic superiority over standard radiographs [8,34], it is still not routinely performed on patients suffering osteolysis and therefore limited our access to clinical data. Traditional open revision surgery is performed under visual guidance and therefore does not require accurate 3D-image-based tissue segmentation. However, we foresee this situation changing. Minimally invasive refixation is already performed as an alternative to refixation in frail patients [2,3], and here, the surgeon is much more dependent on image-based pre-operative planning. Finite element modelling is a powerful tool for computing mechanical stability of the femur [4,5] but requires 3D tissue distribution models. These advances may lead to validation and wider application of minimally invasive cement injection. This will, in turn, fuel the demand for automated segmentation techniques such as the one described in this paper.

Conclusions

We presented a complete pipeline for segmenting periprosthetic tissues in clinical CT volumes of patients with hip prostheses. Due to low tissue contrast and the presence of beam hardening artefacts, these image volumes are extremely difficult to segment—even manually by a trained human operator.

We applied our algorithm to tissues that pertain to aseptically loose hip prostheses, namely cortical bone, trabecular bone, fibrous interface tissue, bone cement, intramedullary canal, and tissue exterior to the femur. Voxel classification offers a way of combining the strengths of several complementary image features, including metal artefact reduced image data that involve data loss if used on its own [35]. We showed how tissue classifiers’ results may be improved by coupling them with a graph-cut post-processing step. We further showed how an adaptation to the algorithm of Delong and Boykov [18] can be used to incorporate geometrical assumptions into the graph-cut segmentation process. Using this, we enforced the requirement that the femur be enclosed in a layer of cortical bone and trabecular bone to be sheathed in cortical bone. Compared to before, these restrictions helped to close segmentation holes caused by low contrast and partial volume effects in the input CT data.

To our knowledge, in the field of medical image segmentation, this is the first use of graph cuts on voxel classifier output since the pioneering work of van der Lijn et al. [21]. In contrast to their study we extend our solution to a multilabel, as opposed to binary, problem. Not only are the graph-cut solutions qualitatively better than voxel classification on its own, but we show that they quantitatively better represent the manually segmented ground truth in terms of their Dice coefficients, our measure for geometric similarity.

The pipeline described in this paper represents a practical approach to segmenting multitissue regions from CT. The demonstrated approach to containment relationships improves the solution wherever such a priori knowledge is available. We demonstrated our solution using clinical CT images of tissues surrounding metal hip prostheses suffering from low contrast and beam hardening artefacts.

Our proposed solution succeeds for this specific application and may in future also be applied to other anatomical regions and imaging environments that are subject to similar constraints.

Acknowledgments This research is supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs (project number LKG 7943). The authors sincerely thank Professor Rob Nelissen for verifying the correctness of the manually segmented ground truth and Noeska Smit for the re-segmentation of several of the data sets for evaluating inter-observer variability. We furthermore thank David Tax and Marco Loog from the pattern recognition group at Delft University of Technology for their helpful input.

Conflict of interest None.

References

1. Agarwal S (2004) Osteolysis—basic science, incidence and diagnosis. Curr Orthop 18:220–231. doi:10.1016/j.cuor.2004.03.002
2. de Poorter JJ, Hoeben RC, Hogendoorn S, Mautner V, Ellis J, Obermann WR, Huizinga TWJ (2008) Gene therapy and cement injection for restabilization of loosened hip prostheses. Hum Gene Ther 19:83–95. doi:10.1089/hum.2007.111
3. Raaijmaakers M, Mulier M (2010) Percutaneous in situ cementation of a loose femoral stem. J Arthroplast 25:1169.e21–1169.e24. doi:10.1016/j.arth.2009.03.027
4. Cody DD, Gross GJ, Hou J, Spencer HJ, Goldstein SA, Fyhrie DP (1999) Femoral strength is better predicted by finite element models than QCT and DXA. J Biomech 32:1013–1020. doi:10.1016/S0021-9290(99)00099-8
5. Schileo E, Taddei F, Malandrino A, Cristofolini L, Viceconti M (2007) Subject-specific finite element models can accurately predict strain levels in long bones. J Biomech 40:2982–2989. doi:10.1016/j.jbiomech.2007.02.010
6. Garcia-Cimbrelo E, Tapia M, Martin-Hervas C (2007) Multislice computed tomography for evaluating acetabular defects in revision THA. Clin Orthop Relat Res 463:138–143. doi:10.1097/BLO.0b013e3181566320
7. Walde TA, Weiland DE, Leung SB, Kitamura N, Sychterz CJ, Engh CA, Claus AM, Potter HG (2005) Comparison of CT, MRI, and radiographs in assessing pelvic osteolysis: a cadaveric study. Clin Orthop Relat Res 437:138–144. doi:10.1097/01.blo.0000164028.14504.46
8. Cahir JG, Toms AP, Marshall TJ, Wimhurst J (2007) CT and MRI of hip arthroplasty. Clin Radiol 62:1163–1171. doi:10.1016/j.crad.2007.04.018
9. Watzke O, Kalender W (2004) A pragmatic approach to metal artifact reduction in CT: merging of metal artifact reduced images. Eur Radiol 14:849–856. doi:10.1007/s00330-004-2263-y
10. Liu P, Pavlicek W, Peter M, Spangehl M, Roberts C, Paden R (2009) Metal artifact reduction image reconstruction algorithm for CT of implanted metal orthopedic devices: a work in progress. Skeletal Radiol 38:797–802. doi:10.1007/s00256-008-0630-5
11. Zoroofi RA, Sato Y, Sasama T, Nishii T, Sugano N, Yonenobu K, Yoshikawa H, Ochi T, Tamura S (2003) Automated segmentation of acetabulum and femoral head from 3-D CT images. IEEE Trans Inf Technol Biomed 7:329–343. doi:10.1109/TITB.2003.813791
12. Kang Y, Engelke K, Kalender WA (2003) A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data. IEEE Trans Med Imaging 22:586–598. doi:10.1109/TMI.2003.812265
13. Yokota F, Okada T, Takao M, Sugano N, Tada Y, Sato Y (2009) Automated segmentation of the femur and pelvis from 3D CT data of diseased hip using hierarchical statistical shape model of joint structure. Med Image Comput Comput Assist Interv 12:811–818. doi:10.1007/978-3-642-04271-3_98
14. Shlens J (2005) A tutorial on principal component analysis. Systems Neurobiology Laboratory, Salk Institute for Biological Studies. http://www.snl.salk.edu/~shlens/pca.pdf. Accessed 16 Nov 2011
15. Sharma N, Aggarwal LM (2010) Automated medical image segmentation techniques. J Med Phys 35:3–14. doi:10.4103/0971-6203.58777
16. Malan DF, Botha CP, Nelissen RGHH, Valstar ER (2010) Voxel classification of periprosthetic tissues in clinical computed tomography of loosened hip prostheses. In: Proceedings of the IEEE international symposium on biomedical imaging: from nano to macro, Rotterdam, The Netherlands, pp 1341–1344
17. Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell 23:1222–1239. doi:10.1109/34.969114
18. Delong A, Boykov Y (2009) Globally optimal segmentation of multi-region objects. In: Proceedings of the IEEE 12th international computer vision conference, Kyoto, Japan, pp 285–292
19. Veksler O (2010) Code: multi-label optimization. University of Western Ontario. http://vision.csd.uwo.ca/code/. Accessed 22 Sept 2010
20. Maleike D, Nolden M, Meinzer HP (2009) Interactive segmentation framework of the medical imaging interaction toolkit. Comput Methods Programs Biomed 96:72–83. doi:10.1016/j.cmpb.2009.04.004
21. van der Lijn F, den Heijer T, Breteler MMB (2008) Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 43:708–720. doi:10.1016/j.neuroimage.2008.07.058
22. Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, Wells WM, Jolesz FA, Kikinis R (2004) Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 11:178–189. doi:10.1016/S1076-6332(03)00671-8
23. Kalender WA, Hebel R (1987) Reduction of CT artifacts caused by metallic implants. Radiology 164:576–577
24. Botha CP (2008) Hybrid scheduling in the DeVIDE dataflow visualisation environment. In: Hauser H, Strassburger S, Theisel H (eds) Proceedings of simulation and visualization, pp 309–322
25. Loog M, van Ginneken B (2006) Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. IEEE Trans Med Imaging 25: 602–611. doi:10.1109/TMI.2006.872747
26. Folkesson J, Dam EB, Olsen OF, Pettersen PC (2007) Segmenting articular cartilage automatically using a voxel classification approach. IEEE Trans Med Imaging 26:106–115
27. van Rikxoort EM, de Hoop B, van de Vorst S, Prokop M (2009) Automatic segmentation of pulmonary segments from volumetric chest CT scans. IEEE Trans Med Imaging 28:621–630. doi:10.1109/TMI.2008.2008968
28. Paclik P, Lai C (2011) PRSD Studio. PR Sys Design, Delft, The Netherlands. http://www.prsdstudio.com/. Accessed 16 Nov 2011
29. Greig DM, Porteous BT (1989) Exact maximum a posteriori estimation for binary images. J R Stat Soc Ser B Stat Methodol 51:271–279

30. Ahuja RK, Magnanti TL, Orlin JB (1993) Maximum flows: polynomial algorithms. In: Janzow P, Peterson M (eds) Network flows. Prentice Hall, Englewood Cliffs, New Jersey, p 240
31. Boykov Y, Kolmogorov V (2004) An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans Pattern Anal Mach Intell 26:1124–1137. doi:10.1109/TPAMI.2004.60
32. Kolmogorov V, Zabin R (2004) What energy functions can be minimized via graph cuts? IEEE Trans Pattern Anal Mach Intell 26:147–159. doi:10.1109/TPAMI.2004.1262177
33. Egger J, Colen RR, Freisleben B, Nimsky C (2011) Manual refinement system for graph-based segmentation results in the medical domain. J Med Syst. doi:10.1007/s10916-011-9761-7
34. Schwarz EM, Campbell D, Totterman S, Boyd A, O'Keefe RJ (2003) Use of volumetric computerized tomography as a primary outcome measure to evaluate drug efficacy in the prevention of peri-prosthetic osteolysis: a 1-year clinical pilot of etanercept vs. placebo. J Orthop Res 21:1049–1055. doi:10.1016/S0736-0266(03)00093-7
35. Malan DF, Botha CP, Kraaij G, Joemai RM, van der Heide HJL, Nelissen RGHH, Valstar ER (2011) Measuring femoral lesions despite CT metal artefacts: a cadaveric study. Skeletal Radiol. doi:10.1007/s00256-011-1223-2