Automated 3-dimensional segmentation of pelvic lymph nodes in magnetic resonance images

O. A. Debats,[a)] G. J. S. Litjens, J. O. Barentsz, N. Karssemeijer, and H. J. Huisman
Radboud University Nijmegen Medical Centre, Geert Grooteplein-Zuid 10—Radiologie 767, Nijmegen, Gelderland 6525 GA, The Netherlands

(Received 21 April 2011; revised 15 August 2011; accepted for publication 30 September 2011; published 26 October 2011)

Purpose: Computer aided diagnosis (CAD) of lymph node metastases may help reduce reading time and improve interpretation of the large amount of image data in a 3-D pelvic MRI exam. The purpose of this study was to develop an algorithm for automated segmentation of pelvic lymph nodes from a single seed point, as part of a CAD system for the classification of normal vs metastatic lymph nodes, and to evaluate its performance compared to other algorithms.
Methods: The authors' database consisted of pelvic MR images of 146 consecutive patients, acquired between January 2008 and April 2010. Each dataset included four different MR sequences, acquired after infusion of a lymph node specific contrast medium based on ultrasmall superparamagnetic particles of iron oxide. All data sets were analyzed by two expert readers who, reading in consensus, annotated and manually segmented the lymph nodes. The authors compared four segmentation algorithms: confidence connected region growing (CCRG), extended CCRG (ECC), graph cut segmentation (GCS), and a segmentation method based on a parametric shape and appearance model (PSAM). The methods were ranked based on spatial overlap with the manual segmentations, and based on diagnostic accuracy in a CAD system, with the experts' annotations as reference standard.
Results: A total of 2347 manually annotated lymph nodes were included in the analysis, of which 566 contained a metastasis. The mean spatial overlap (Dice similarity coefficient) was: 0.35 (CCRG), 0.57 (ECC), 0.44 (GCS), and 0.46 (PSAM). When combined with the classification system, the area under the ROC curve was: 0.805 (CCRG), 0.890 (ECC), 0.807 (GCS), 0.891 (PSAM), and 0.935 (manual segmentation).
Conclusions: We identified two segmentation methods, ECC and PSAM, that achieve a high diagnostic accuracy when used in conjunction with a CAD system for classification of normal vs metastatic lymph nodes. The manual segmentations still achieve the highest diagnostic accuracy.
© *2011 American Association of Physicists in Medicine*. [DOI: 10.1118/1.3654162]

Key words: CAD, segmentation, lymph node, USPIO, MR-lymphography

I. INTRODUCTION

Prostate cancer (PCa) is the second leading cause of cancer mortality in men, with 218 000 new cancer cases and 32 000 deaths in the United States in 2010.[1] For patients diagnosed with PCa, treatment options depend on whether the cancer is contained within the prostate, or has broken through its capsule and formed metastases in pelvic lymph nodes. In the latter case, radical prostatectomy or radical radiotherapy is not considered a curative treatment.

MR lymphography (MRL)—MRI with a lymph node specific contrast agent—is currently the most accurate imaging modality for assessing metastatic involvement of pelvic lymph nodes, with reported sensitivities up to 91% at 98% specificity.[2] With MRL, lymph node metastases can be found that are not detected by routine pelvic lymph node dissection.[3] MRL uses ferumoxtran-10, a contrast agent based on ultra-small superparamagnetic particles of iron oxide (USPIO), which results in signal intensity differences between metastatic and normal lymph node tissue. Other modalities, such as CT or gadolinium-enhanced MRI cannot distinguish metastatic tissue from normal lymph node tissue and have to rely on size criteria instead, resulting in a sensitivity of only 34% at 97% specificity.[4]

The interpretation of MRL images is time-consuming, with reported average interpretation times up to 80 min,[5] and is highly dependent on the experience of the reader. MRL interpretation time can be reduced and the accuracy further improved by using a computer aided diagnosis (CAD) system. Such a CAD system would, ideally, automatically detect all lymph nodes in the MRL images, and subsequently determine which of them are likely to contain metastasis. This would allow the radiologist to efficiently display and interpret suspicious nodes on a dedicated viewing station.

An important part of a CAD system for MRL is segmentation of the lymph nodes. Lymph node segmentation has been subject of research for roughly 15 years,[6] recent publications being,[7,8] mostly directed to segmentation in CT images. Unal *et al*.[9] segmented lymph nodes in MRL images by fitting an ellipse in a 2-D slice, and propagating it to the next slices in 3-D. To our knowledge, no other studies have been published which focus on lymph node segmentation in MR images.

In this study, we developed a parametric shape and appearance model-based segmentation method and an extended confidence connected region growing method, and compared them to two existing methods: graph cut segmentation and nonextended confidence connected region growing. We evaluated the four methods with respect to a database of 2347 lymph nodes, annotated in MRL images.

II. MATERIALS AND METHODS

II.A. Segmentation methods

Sections II A 1 and II A 2 describe confidence connected region growing (CCRG) and our extended CCRG method (ECC). In subsection II A 3, graph cut segmentation (GCS) is briefly reviewed, and in subsection II A 4, our parametric shape and appearance model-based segmentation method (PSAM) is described.

II.A.1. Confidence connected region growing

Thresholding with either one or two fixed thresholds is the simplest possible region-based segmentation method. Given an image volume $\mathcal{S}$ represented as a set of voxels $v \in \mathcal{S}$, a thresholded segmentation $\mathcal{S}_{\text{th}} \subset \mathcal{S}$ is the subset of voxels whose intensities I_v are within the threshold range,

$$\mathcal{S}_{\text{th}} = \{v \in \mathcal{S} \,|\, t_{\text{lower}} < I_v < t_{\text{upper}}\}. \tag{1}$$

When the segmentation task is preceded by a detection step, from which a seed location is available for the object to be segmented, seeded region growing can be performed. We can define a connected region $\mathcal{S}_{\text{cr}}$ as the union of a series of subsequent layers $\mathcal{L}_i$. Except for the first layer, which contains only the seed voxel, each layer consists of all voxels $v_p \in \mathcal{S}_{\text{th}}$ that are neighbor-connected to the previous layer $\mathcal{L}_{i-1}$, according to a standard 6-, 18-, or 26-connected 3-dimensional neighborhood system H,

$$\mathcal{S}_{\text{cr}} = \bigcup_i \mathcal{L}_i \tag{2}$$

$$\mathcal{L}_i = \begin{cases} \{v_{\text{seed}}\} & \text{if } i = 1 \\ \mathcal{S}_{\text{th}} \cap \left\{ v_p \notin \bigcup_{j=1}^{i-1} \mathcal{L}_i \,|\, H(v_p) \cap \mathcal{L}_{i-1} \neq \emptyset \right\} & \text{otherwise,} \end{cases} \tag{3}$$

using a voxel set based notation for region growing, which follows the notation of Dawant and Zijdenbos[10] and Liu *et al.*[11]

CCRG segments a region by taking a seed region as an initial segmentation and then iteratively adding all connected voxels that have a signal intensity within a dynamically defined threshold range.[12] It has been shown to produce accurate results when used for tumor volume segmentation on FDG-PET images.[13] With an initial segmentation $\mathcal{S}_{\text{cc},0} = \mathcal{S}_{\text{seed}}$ and N iterations, the confidence connected segmentation $\mathcal{S}_{\text{cc},N}$ is defined as

$$\mathcal{S}_{\text{cc},n} = \mathcal{S}_{\text{cc},n-1} \cup \mathcal{S}_{\text{cr},n} \quad 1 \leq n \leq N, \tag{4}$$

where $\mathcal{S}_{\text{cr},n}$ depends on the two threshold values $t_{\text{lower},n}$ and $t_{\text{upper},n}$, which are determined by the mean signal intensity μ and standard deviation σ of the voxels segmented in the previous iteration, and the bandwidth factor b, as follows:

$$t_{\text{lower},n} = \mu_{n-1} - b \cdot \sigma_{n-1} \tag{5}$$

$$t_{\text{upper},n} = \mu_{n-1} + b \cdot \sigma_{n-1}. \tag{6}$$

Because convergence is not guaranteed, the number of iterations N is specified beforehand.

II.A.2. Extended confidence connected region growing

In the original CCRG method, the bandwidth factor b is set to a fixed value. If it is set too high, the segmented volume will "leak" out of the lymph node into surrounding structures with similar signal intensities, but when set too low, it will lead to undersegmentation. When segmenting lymph nodes in MR images, the optimal value of b may be very different between lymph nodes. Therefore, segmentation results might be greatly improved if an optimal setting of b could be computed for each lymph node separately. We developed an extended CCRG segmentation method, which includes selection of the optimal bandwidth factor by a leakage detection mechanism based on the expansion rate, defined as

$$e(b) = \frac{|\mathcal{S}_{\text{cc}}(b + \Delta b)|}{|\mathcal{S}_{\text{cc}}(b)|}. \tag{7}$$

Vertical bars denote the cardinality of a set, which is in this case the size of a segmented region expressed as the number of voxels it contains. The expansion rate is a measure of the relative increase in size with increasing values of b. When an increase of b results in leakage, there is usually a sudden expansion of the segmented region, resulting in a high value of $e(b)$. By increasing b until $e(b)$ reaches a predefined maximum, an optimal b is selected.

For both the CCRG and ECC methods, we apply seed point optimization: in the neighborhood of each seed point, the voxel with the lowest signal intensity is selected as seed voxel. This increases robustness against small variations in seed location, e.g., when the user defines the seeds by clicking on the lymph nodes.

II.A.3. Graph cut segmentation

In graph cut segmentation,[14] the image to be segmented is regarded as a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, defined as a set of nodes or vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$, where each edge $e = \{p, q\} \in \mathcal{E}$ connects two "neighboring" nodes p and q. The nodes represent image pixels or voxels. There are also two special nodes s (source terminal) and t (sink terminal), representing "object" and "background" labels, which are connected to every voxel. Thus, $\mathcal{E}$ can be divided into three subsets, $\mathcal{E}_n$ containing edges between two voxels (called neighbor-links or *n-links*), $\mathcal{E}_s$ and $\mathcal{E}_t$ containing edges between a voxel and the source or sink, respectively (both called terminal-links or *t-links*).

All edges e have an associated weight or cost w_e. A graph cut is then defined as a subset of edges $\mathcal{C} \subset \mathcal{E}$ such that in $\mathcal{G}' = \langle \mathcal{V}, \mathcal{E} \backslash \mathcal{C} \rangle$ the terminals s and t are separated. The cost or energy of a cut is defined as the sum of the costs of the edges that it contains,

$$E(\mathcal{C}) = \sum_{e\in\mathcal{C}} w_e = \sum_{e\in\mathcal{E}_n\cap\mathcal{C}} w_e + \sum_{e\in\mathcal{E}_s\cap\mathcal{C}} w_e + \sum_{e\in\mathcal{E}_t\cap\mathcal{C}} w_e. \tag{8}$$

$$w_e = \begin{cases} \exp\left(-\frac{(I_p-I_q)^2}{2\sigma^2}\right) & \text{if } e\in\mathcal{E}_n \\ \lambda\cdot -\ln L(I_p\,|\,p \text{ is a lymph node voxel}) & \text{if } e\in\mathcal{E}_s \\ \lambda\cdot -\ln L(I_p\,|\,p \text{ is a background voxel}) & \text{if } e\in\mathcal{E}_t \end{cases}, \tag{9}$$

where I_p is the intensity of voxel p and $\lambda \geq 0$ is a coefficient specifying the relative cost of t-links versus n-links. Assuming Gaussian distributions, the likelihood function L is defined by the means and standard deviations of voxel intensity: μ_{lymph}, μ_{bg}, σ_{lymph}, and σ_{bg}.

The cost function is chosen such that the costs of cutting an n-link are high if it connects two voxels with similar intensities, for which $(I_p - I_q)^2 < \sigma^2$, but the costs are low when the intensities I_p and I_q are very different. The costs of cutting a t-link are proportional to the negative log-likelihood of the concerning voxel's intensity occurring in a lymph node or in the background.

To optimize the cost function of a graph cut representation, one can choose from a number of different combinatorial min-cut/max-flow algorithms. All these algorithms guarantee convergence to the cost function's global minima, and yield identical segmentations; they only differ in execution speed.[15] In a minimum-cost $\mathcal{G}'$, each voxel p is either connected to s or to t, but not both. The corresponding segmentation is

$$\mathcal{S}_{\text{gr}} = \{p \in \mathcal{G}' \,|\, p \text{ connected to } s\}. \tag{10}$$

II.A.4. Shape and appearance-based segmentation

A novel PSAM was developed to segment lymph nodes. The rationale was to have a method that intuitively integrates prior knowledge about an objects' shape and appearance in a general way, instead of the often ad hoc integration into dedicated algorithms. The PSAM method includes a volume model which is a synthetic representation of a lymph node within a background, characterized by shape and appearance parameters and initially centered at a seed location. The volume model defines a lymph node likelihood L_v for each voxel v, based on location and gray value parameters. For a given set of model parameters $\vec{p}$, a lymph node segmentation $\mathcal{S}_{\text{psam}}$ results from the application of a threshold t on the voxel likelihoods,

$$\mathcal{S}_{\text{psam}} = \{v \in \mathcal{S} \,|\, L_v > t\}. \tag{11}$$

The volume model is fitted to a lymph node by finding the parameter values that minimize a cost function, which comprises an internal term, E_{int}, and an external term, E_{ext},

$$\vec{p}_{\text{optimal}} = \arg\min_{\vec{p}}(E_{\text{int}}(\vec{p}) + E_{\text{ext}}(\vec{p})). \tag{12}$$

The internal cost term penalizes deviations from a population PSAM model using the population distribution of the model parameters. The population distribution is obtained in a separate training session. The model is matched to the image by an external cost term that penalizes deviations of the current segmentation estimates and model parameters.

The PSAM volume model is characterized by the following parameters: *distance* of the model center to the seed location, the volume by the *radius* (shape), and the *gray value difference* between the volume and its background (appearance). The population model distribution is defined as an independent, multi-variate Gaussian which is characterized by each parameters' mean and standard deviation μ_{distance}, σ_{distance}, μ_{radius}, σ_{radius}, $\mu_{\text{grayvalue}}$, and $\sigma_{\text{grayvalue}}$. The mean and standard deviation were computed from a training set of parameters obtained by segmenting a subset of lymph nodes using PSAM with manual parameter optimization by an expert observer (OD). The internal term of the cost function is the logarithm of the population model distribution.

The external term of the cost function is the sum of the squared differences of the actual and estimated parameters. During optimization for each segmentation, an estimate of each parameter is obtained as follows. The estimated distance is the actual center (center of mass) minus the seed location. The estimated lymph node radius is obtained from the voxel position standard deviation (with correction factor). The mean gray values are computed from the segmented voxel gray value statistics. Optimization was performed using the L-BFGS-B optimization method. L-BFGS-B is a limited-memory, quasi-Newton, bound-constrained optimization method. The required cost function derivative was computed from the analytical derivative of the above cost function. For each parameter the optimization bounds are set to the 95% range (population mean ±2 × standard deviation).

II.B. Experiments

Two experiments were performed to evaluate the performance of the different segmentation methods. The first experiment quantitatively analyzed the segmentation performance of all segmentation methods using an overlap criterion. In the second experiment, the ability to classify lymph nodes as either metastasized or normal was studied. All methods were implemented in open source programming environments, the VISUALIZATION TOOLKIT and the INSIGHT TOOLKIT, using the tool command language (TCL) and C++.

II.B.1. Imaging

All imaging was performed on a Siemens 3.0 Tesla MR scanner. Images were acquired in the coronal plane, covering the whole pelvis. Four MR series were used in this study:

- T1-weighted "volumetric interpolated breath hold examination" (VIBE), with 0.78 mm isotropic voxel size
- T2*-weighted "multi echo data image combination" (MEDIC), with 0.78 mm isotropic voxel size
- Apparent diffusion coefficient (ADC), with 2.3 × 2.3 mm in-plane resolution and 3.0 mm slice distance
- "Fast low angle shot" (FLASH), with 1.0 mm isotropic voxel size

Data were collected from a consecutive set of patients who had biopsy-proven prostate cancer and underwent MR lymphography using the USPIO-based lymph node specific contrast agent ferumoxtran-10 (Sinerem®, Guerbet, France) in the Radboud University Nijmegen Medical Centre, as part of their clinical evaluation. The patients received a ferumoxtran-10 drip infusion, 36 to 24 h before MR imaging was performed. Buscopan (20 mg i.v. and 20 mg i.m.) and glucagon (20 mg i.m.) were administered immediately before the MR examination in order to suppress bowel peristalsis. All patients provided informed consent for the use of their images for research purposes.

To be included in the analysis, the images of a patient had to fulfill the following inclusion criteria:

- Scan date between January 2008 and April 2010
- All four above mentioned MR series available
- One or more lymph nodes visible in the VIBE image

Between January 2008 and April 2010, 289 patients underwent USPIO imaging. For 146 patients, all MR series were available with at least one lymph node visible in the VIBE image; these patients were included in the analysis.

II.B.2. Reference standard

Two types of reference standard were used in this study: a segmentation reference standard, which was used in experiment 1, and a classification reference standard, used in experiment 2. The images of each patient were assessed by two expert readers in consensus: an MD researcher (reader 1) and an experienced radiologist (reader 2). As a common first step, all lymph nodes visible in the pelvic area were identified and numbered by reader 1.

II.B.2.a. Segmentation reference standard. Each lymph node was interactively segmented by reader 1, using the application Lymph Node Task Card, developed by Siemens, Malvern, PA (USA). Segmentation was performed in the VIBE image, which was also used as the input for the automatic segmentation algorithms. The location of each lymph node was listed as a seed point, defined as the center of the manual segmentation of the node. Subsequently, the two readers discussed each lymph node and changed its segmentation where needed.

II.B.2.b. Classification reference standard. Each lymph node was assigned a level of suspicion (LOS), on a five-point scale where LOS 1 = "definitely not metastatic," LOS 2 = "probably not metastatic," LOS 3 = "equivocal," LOS 4 = "probably metastatic," and LOS 5 = "definitely metastatic." The two readers discussed each lymph node until they reached consensus and defined its level of suspicion. For the classification experiment described below, the lymph nodes had to be defined as either negative or positive; the cut-off was placed between LOS 3 and LOS 4.

II.B.3. Segmentation

All segmentation methods used the VIBE series as input image. Because this MR series, as described above, has isotropic voxels, no resampling or interpolation was needed. To reduce noise but retain sharp edges, we preprocessed the VIBE images prior to segmentation by application of a median smoothing filter using a 3-dimensional, cubic smoothing kernel with a width of 2.34 mm (3 voxel widths).

Long-axis lymph node diameter did not exceed 2.5 cm in our database. We selected a 3 × 3 × 3 cm region of interest (ROI) around each seed point to reduce computation time.

A subset of the MR lymphographies that were excluded because no FLASH or no ADC series was available was used for the purpose of trying out the segmentation methods and determine the most suitable parameter settings. For this subset, manual segmentations were created in the same way as the reference standard segmentations.

The CCRG method has two parameters: the number of iterations N and the bandwidth factor b. For segmentation of lymph nodes, which are relatively small with respect to voxel size, N can be set to a low value. In this study, it was set to 4 iterations. Any value of b will either result in oversegmentation or undersegmentation or both, of part of the data set. We used $b = 1.8$, which gave the best balance between over- and under-segmentation.

The ECC method extends CCRG by dynamically selecting a value of b for each lymph node, based on maximum expansion rate $e_{\max}$ and step size Δb. The value of these parameters was set to 2 and 0.2, respectively, indicating that the value of b will be selected at which a further increase of 0.2 would lead to a more than twofold increase in segmented volume.

For GCS, we calculated the mean and standard deviation of signal intensity in the try-out set, using the manually segmented lymph nodes, and set the parameters accordingly: $\mu_{\text{lymph}} = 219$, $\sigma_{\text{lymph}} = 97$, $\mu_{\text{bg}} = 391$, and $\sigma_{\text{bg}} = 143$.

For PSAM, the means and standard deviations of the population model distribution were: $\mu_{\text{distance}} = 0.0$, $\sigma_{\text{distance}} = 1.5$, $\mu_{\text{radius}} = 3.3$, $\sigma_{\text{radius}} = 0.9$, $\mu_{\text{grayvalue}} = 150$, and $\sigma_{\text{grayvalue}} = 50$.

Because of the inhomogeneity of lymph node tissue with respect to signal intensity in the VIBE image, all tested segmentation methods frequently miss one or more bright or dark patches inside the lymph nodes. To solve this, a morphological closing operation is applied to obtain the final segmentation result, using a spherical kernel with a 2.34 mm diameter (3 voxel widths).

In experiment 1, the performance of the four segmentation methods was assessed in terms of spatial similarity between the resulting segmentations and the reference segmentations. Two spatial overlap metrics are often used in segmentation studies: the Dice similarity coefficient (DSC),[16] derived as a special case of the kappa statistic by Zijdenbos *et al.*,[17]

$$\text{DSC}_{\text{x}} = \frac{|\mathcal{S}_{\text{x}} \cap \mathcal{S}_{\text{h}}|}{\frac{1}{2}(|\mathcal{S}_{\text{x}}| + |\mathcal{S}_{\text{h}}|)} \tag{13}$$

and the Jaccard index (JI),[18]

$$\text{JI}_{\text{x}} = \frac{|\mathcal{S}_{\text{x}} \cap \mathcal{S}_{\text{h}}|}{|\mathcal{S}_{\text{x}} \cup \mathcal{S}_{\text{h}}|}, \tag{14}$$

where $\mathcal{S}_{\text{x}}$ is the segmentation result of method x, and $\mathcal{S}_{\text{h}}$ is the segmentation made by the human experts. For both DSC and JI, a value of 1 means a perfect match between the two regions, whereas 0 means no overlap at all. In case of a partial match, the JI is always lower than the DSC value. Both metrics describe similarity in terms of overlap, and they have a one-to-one correspondence,

$$\mathrm{JI} = \mathrm{DSC}/(2 - \mathrm{DSC}). \tag{15}$$

Note that this equation holds for single measurements but *not* for mean values. Following the majority of recent publications on lymph node segmentation, we use the DSC to report overlap results, but to enable comparison with studies that report overlap as JI, we will include also JI values.

II.B.4. Normal vs metastasis classification

The four segmentation methods were also evaluated by studying classification performance of a classifier that used features computed from their segmentations. For each segmentation, 25 features were calculated: volumetric lymph node size and 24 features based on voxel intensity values. The intensity features were the 24 combinations of 6 descriptive statistics: mean, standard deviation, median, 25th and 75th percentile, and interquartile range (IQR) of signal intensity inside the segmented region, in 4 MR sequences (VIBE, MEDIC, ADC, and FLASH).

No smoothing was applied to the MR images when used as input for the feature extraction. Because the ADC and FLASH had a different spatial resolution than the VIBE series, the surface of the segmented lymph nodes did not follow the ADC and FLASH voxel boundaries, i.e., some voxels were not completely inside or outside the segmented lymph node, but were intersected by the surface of the segmentation. For feature computation, both these surface voxels and the voxels inside were included in the analysis.

Classification was performed for each of the four sets of segmentations (one from each method) using linear discriminant analysis with leave-one-out cross-validation. Additionally, the same classification system was also applied to the set of manual segmentations, and to the population-averaged shape model (PAS) described below. The complete classification experiment was done twice, once with and once without the lymph node size feature, to evaluate how much this feature contributed to the performance. The likelihood of malignancy provided by the classifier was used as a discrimination score to classify the lymph nodes as normal or metastatic. The discrimination scores were analyzed with ROC methodology, and the diagnostic performance of the classification was calculated as the area under the ROC curve (AUC).

II.B.5. Population-averaged lymph node shape

In order to demonstrate the need for individual segmentation of each lymph node when classifying normal vs metastasized lymph nodes, we constructed a population-averaged shape model. In MR lymphography images, lymph nodes have varying shapes. What their shapes do have in common is that they are more or less blob-like, as opposed to tubular or sheet-like. Therefore, we used a generalized blob shape—i.e., a sphere with a fixed diameter—as a population average. For each lymph node, the center of the sphere was defined by the corresponding seed point; the diameter was estimated by the mean diameter of a set of manual segmentations. Although, strictly, it is not a segmentation method, because it does not use any image information, this model does result in a set of voxels defining a "segmented" structure for each seed point. As such, it can be used as a basis for feature computation. Our features currently do not include lymph node shape. We performed classification (as described in the subsection II B 4) based on this fixed lymph node model to judge the additional value of segmentation.

II.B.6. Statistical analysis

The statistical significance of differences in Dice similarity was tested using Wilcoxon's signed rank test for paired nonparametric data.

The ROC analysis was performed using the R Project for Statistical Computing and the ROCKIT software package (Kurt Rossmann Laboratories, University of Chicago).

As we performed multiple significance tests, Bonferroni correction was included.[19]

III. RESULTS

III.A. Segmentation overlap results

The automated segmentation result of each of the described methods for three example lymph nodes is visually presented in Fig. 2. For each lymph node in the data set, the Dice similarity between the automated and the reference segmentations was calculated. To enable comparison with studies that use the JI as similarity metric, we also calculated the JI values for each segmentation. The results are summarized in Table I. The mean DSC value was 0.35 for the original CCRG method, 0.57 for the ECC method, 0.44 for graph cut segmentation, and 0.46 for the parametric shape and appearance model-based segmentation method. The difference in Dice similarity between ECC and PSAM was statistically significant ($p < 0.004$ with Bonferroni correction).

III.B. Normal vs metastasis classification

The second experiment evaluated how the ability of the classifier to discriminate between normal and metastasized lymph nodes was influenced by the choice of the segmentation method. This evaluation was done with the four segmentation methods, the PAS model, and the manual segmentations. Classification accuracy was determined by calculating the AUC. The ROC curves are displayed in Fig. 3. Note that the experts'

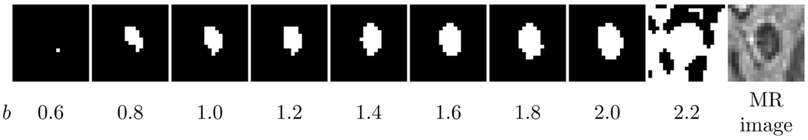

FIG. 1. Example of a CCRG segmentation leaking out of the lymph node at a certain (unpredictable) value of b.

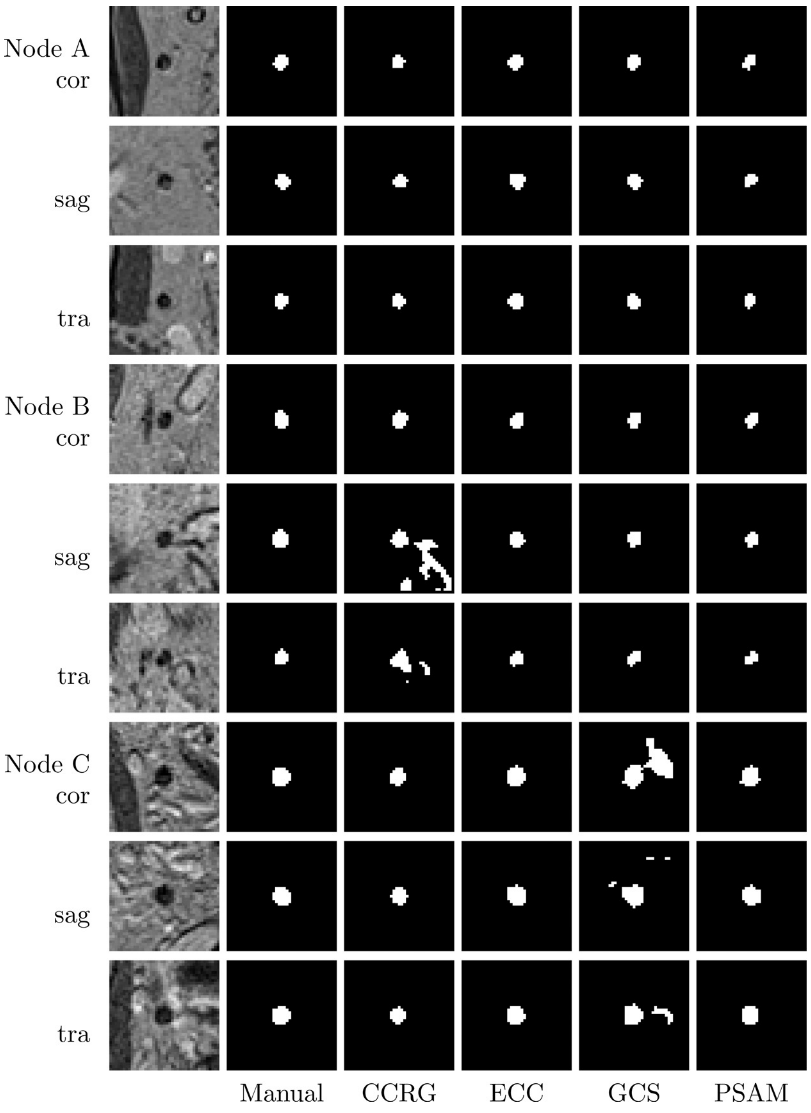

FIG. 2. Segmentation results of three example lymph nodes. For each node, a coronal, sagittal, and axial slice through the seed point are shown. Lymph node A is segmented well by all four automated segmentation methods. Leakage out of lymph nodes B and C is seen in the segmentations by the CCRG and GCS methods, respectively. Note that, while in these 2-D slices, some leaked-into regions seem disjoint, they are connected in the 3-D image.

lymph node classifications, but not their manual segmentations, were used as reference standard in this experiment. That is why when evaluating classification with the manual segmentations the AUC does not equal the ideal value of 1. The mean AUC values were 0.935 for the manual segmentations, 0.890 for the ECC segmentations, 0.891 for PSAM, 0.807 for graph cut segmentation, 0.805 for CCRG, and 0.728 for the population averaged shape method (Fig. 4). When classification was done without lymph node size as a feature, performance changed only very slightly, as shown in Fig. 4.

While PSAM had the highest AUC performance of the four automated segmentation methods, the difference between PSAM and ECC was not statistically significant. However, both PSAM ($p < 0.0004$ with Bonferroni correction) and ECC ($p < 0.0004$ with Bonferroni correction) had a significantly higher performance than GCS, which had the third best performance.

IV. DISCUSSION

This study provides baseline results for automated segmentation of lymph nodes in MR images. We evaluated four segmentation methods on a large number of lymph nodes, for all of which a reference segmentation was available,

TABLE I. Summary statistics (minimum, first quartile, median, mean, third quartile, and maximum) for the segmentation performance in terms of DSC and JI.

	DSC						JI					
Method	Min	Q1	Med	Mean	Q3	Max	Min	Q1	Med	Mean	Q3	Max
CCRG	0.00	0.02	0.37	0.35	0.64	0.91	0.00	0.01	0.23	0.26	0.47	0.84
ECC	0.00	0.43	0.64	0.57	0.77	0.93	0.00	0.27	0.47	0.44	0.63	0.88
GCS	0.00	0.17	0.42	0.44	0.70	0.91	0.00	0.09	0.27	0.32	0.54	0.84
PSAM	0.00	0.25	0.49	0.46	0.68	0.90	0.00	0.14	0.32	0.33	0.51	0.82

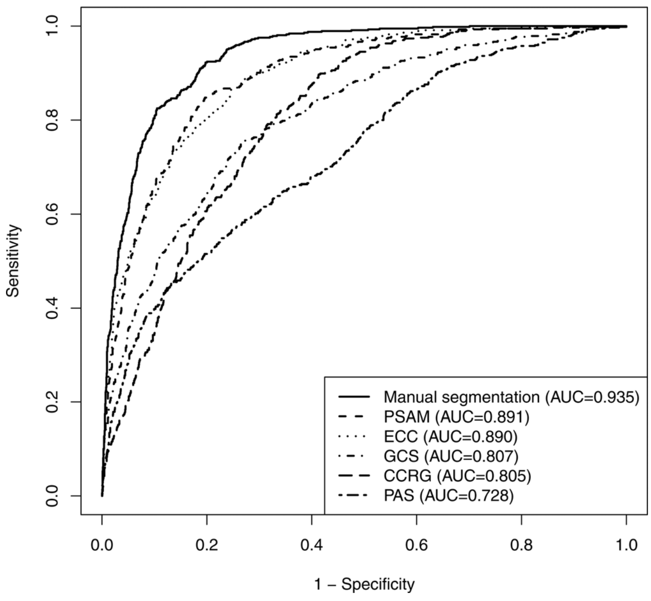

FIG. 3. ROC curves of all segmentation methods.

created by two expert readers who annotated all MR lymphographies that met the inclusion criteria in a consecutive set of patients. All four methods are automated segmentation methods, i.e., they require only a seed point as initialization. The seed points can be provided beforehand, e.g., by an automated lymph node detection algorithm (which is outside the scope of this study), or can be defined through a mouse click by a human reader, as all four methods are fast enough for interactive use on a state-of-the-art desktop PC.

The presented results show that the ECC method scores significantly better on segmentation performance than the other segmentation methods that were compared in this study. This might seem surprising at first sight because of its relative simplicity, but can be understood better when one considers the fact that it has two properties that are very advantageous for this particular application. ECC does not depend on predefined gray values, which is important because lymph nodes in MRI—as opposed to CT—vary widely in image intensity, even within one MR image. Second, it is capable of segmenting lymph nodes within a wide size range, which is needed because with MR lymphography, metastases can be found in very small lymph nodes as well as in large ones. Another advantage is that ECC has an explicit mechanism to prevent leakage into surrounding structures. The idea that for segmentation of small blob-like objects, methods based on region growing perform well, is supported by research in related fields such as lung nodule segmentation[20] and segmentation of liver metastases.[21]

The results of the second experiment show that with the segmentations from either the ECC or the PSAM method, the classification system achieves a good diagnostic accuracy, with AUC values around 0.89. However, using the manual segmentations still yields the highest accuracy.

When lymph node size was omitted as a feature for the classifier, performance changed only very slightly, as shown in Fig. 4.

Measurement of radiological lymph node size is generally accepted as a method to detect metastatic lymph nodes in CT and MR imaging, and is recommended in the response evaluation criteria in solid tumors (RECIST) guideline.[22] However, in the case of prostate cancer, a meta-analysis showed that while specificity was 0.82, sensitivity was only 0.42, indicating that 82% of normal lymph nodes but also 58% of metastatic nodes were normal-sized.[23] Another study analyzing 980 prostatectomy patients concluded that in normal-sized nodes, size did not correlate with the presence of metastasis.[24] Seen in this light, the very modest

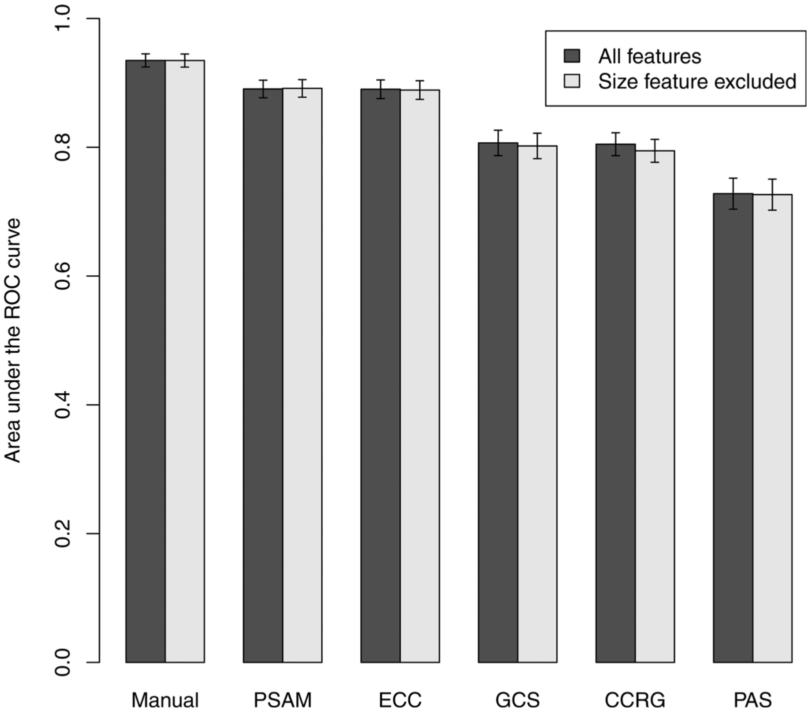

FIG. 4. The effect on the diagnostic accuracy of the normal vs metastasis classification using six types of segmentation. Error bars represent 95% confidence intervals.

contribution of lymph node size to classification performance is not surprising, because in our study population almost all metastatic lymph nodes are normal-sized.

To date, a limited number of studies have been published on lymph node segmentation and most of them focus on CT data, exceptions being Zhang *et al.*,[25] who used ultrasound images, and Unal *et al.*[9] who focused on lymph nodes in MR images. However, Unal's method is semi-automatic, 2D-oriented, and forces lymph node shape to be elliptical. While both CT and MR lymphography are cross-sectional imaging techniques, radiodensity in CT is a quantitative measurement of x-ray attenuation, expressed in Hounsfield units (HU), and each tissue type has a specific range of HU values, whereas signal intensity in most MR sequences is nonquantitative by nature. The intensities of lymph node tissue in MR therefore cannot be defined *a priori*. They can vary between scanners but also between patients. Moreover, because bias fields and other image artifacts are ubiquitous in MR imaging, lymph nodes can have very different intensity ranges even within one image. Another difference is that MRI distinguishes different types of soft tissue that have the same HU value, which causes lymph nodes to have a more heterogeneous appearance than in CT. These differences between CT and MR images hamper the interpretation of differences in segmentation result between studies.

Yan *et al.* were the first to evaluate semi-automatic segmentation on real lymph node images, after a few studies that evaluated their methods on synthetic images.[6,26] They proposed a fast marching method,[27] and a marker-controlled watershed method[28] to segment lymph nodes in CT scans of lymphoma patients, which achieved a JI value (called "overlap ratio" in their paper) of 0.832. This high accuracy was made possible by extensive input from a human reader, who selected an optimal seed point and drew a circle around the lymph node to prevent boundary leaking. In a separate study,[29] they developed an algorithm for lymph node segmentation in follow-up scans, based on registration with manually segmented baseline scans. The reported JI of 0.73 was the mean of only the *successfully* segmented nodes.

Lu *et al.*[30] proposed a method called "single-click live wire." They evaluated their method on central-chest lymph nodes in an interactive setting where they permitted up to three single-click attempts, and achieved a mean DSC (called "accuracy" in their paper) of 0.79, but they did not report what DSC would result if only one single-click were permitted.

For the segmentation of pelvic lymph nodes, Young *et al.*[31] evaluated a commercially available software package that used atlas registration to generate "autocontours," which were corrected by human readers. The DSCs that they report, 0.78, 0.86, and 0.78, are from the *corrected* autocontours.

A number of publications address segmentation of lymph nodes in head and neck CT images. Chen *et al.*,[32] for example, report a DSC of 0.698 and Stapleford *et al.*[33] reach a DSC of 0.76. It is important to realize that these studies did not segment individual lymph nodes nor separate neck lymph node levels. Instead, all included lymph nodes on either side of the body were taken together and regarded as one single volume. Segmenting lymph nodes individually, as we did, is much more challenging in terms of overlap accuracy, as small segmentation errors on the boundaries of each lymph node add up together, which decreases DSC values. Dornheim *et al.*[34] achieved in 2007 already a remarkably good overlap with their method based on "stable mass-spring models." The mean JI of the first version was 0.56, and with the improved method this increased to 0.721.[8] Although this is a very good result, it is unlikely that the method can be easily adapted for MR images, because it relies on the specific HU range of lymph nodes. Moreover, it must be noted that the evaluation was done on 40 manually selected lymph nodes, and that the mean overlap might be less if the method were evaluated on a larger set of nodes from a consecutive set of patients.

Other studies that do distinguish at least between the different neck lymph node levels report much lower DSCs. For example Gorthi *et al.*, who segmented levels Ia, Ib, IIa, IIb, III, IV, Va, Vb, and VI separately, reported in 2009 DSCs of 0.42 an 0.46 with and without leave-one-out cross-validation, respectively.[35] In their most recent study, they improved the method and achieved a DSC of 0.503.[36] Our best performing method, with a mean DSC of 0.57, compares favorably to those results.

One limitation of this study is that we use one fixed seed point per lymph node in the experiments. Therefore, we cannot draw conclusions concerning the effect of variations in seed point location. However, all methods have the ability to deal with a seed point that deviates from the lymph node's center of mass. Our CCRG and ECC implementations apply seed point optimization, PSAM directly optimizes lymph node center, so it uses the seed point only for initialization, and GCS assumes only that the seed point is inside the lymph node but not necessarily in its center.

In imaging studies it is, in general, important to include images of both healthy and diseased patients, or else the conclusions drawn may not apply to both groups of patients. While it is true that all patients included in our study are prostate cancer patients, our study specifically addresses the lymph nodes, not the prostate gland. Therefore, in our case the important issue is to include both nonmetastasized and metastasized patients. Indeed, more than half of the patients in our database did not have metastases (as defined by the reference standard). Moreover, only 566 of the 2347 analyzed lymph nodes (24%) contained metastases.

The ferumoxtran-10 contrast agent used for MRI acquisition, which is needed for the normal vs metastasis classification, is currently not available. While imaging without contrast agent, or with another type of USPIO, might have a big impact on classification performance, it has little or no influence on the segmentation results, because the input for the segmentation algorithms is the VIBE image, which is insensitive to USPIO contrast. Therefore the performance, in terms of similarity between manual and automated methods, would not be affected. Furthermore, it is expected that new USPIO agents with imaging characteristics very similar to ferumoxtran-10 will be introduced in the near future.

Future improvements in segmentation and classification performance can probably be made by including a segmentation of the main pelvic anatomical structures, such as the pelvic bones and the major blood vessels. By doing so, erroneous inclusion of parts of those structures in segmented lymph nodes can be further eliminated. As pelvic lymph nodes are mainly located near big blood vessels, a segmentation of those can also serve as basis for a fully automatic lymph node segmentation system that includes automated seed point detection. Classification performance may be further improved by including a registration step: although in our images, the misalignments are small (in the order of 1 mm) registration of the MEDIC, ADC, and FLASH images to the VIBE image, on which the segmentation was performed, may increase accuracy especially for small lymph nodes.

V. CONCLUSION

In this paper, we compared four automated segmentation methods applied to MR lymphography images. We evaluated the performance of the methods by two experiments, with a database of 2347 lymph nodes, including manual segmentations created by two expert readers in consensus. The first experiment, in which each method's results were compared to a set of reference segmentations using the Dice similarity coefficient, showed that our ECC method is significantly more accurate than the other segmentation methods ($p < 0.004$). The second experiment, in which the methods were compared on the basis of their diagnostic accuracy when used as input for a lymph node metastasis CAD system, showed that both ECC and PSAM had a good diagnostic accuracy and were significantly better than GCS and CCRG ($p < 0.0004$).

ACKNOWLEDGMENTS

This project was funded by Grant KUN2007-3971 from the Dutch Cancer Society. The content is solely the responsibility of the authors and does not necessarily represent the official views of this institution.

[a] Author to whom correspondence should be addressed. Electronic mail: o.debats@rad.umcn.nl

[1] A. Jemal, R. Siegel, J. Xu, and E. Ward, "Cancer statistics, 2010," CA Cancer J. Clin. **60**, 277–300 (2010).

[2] M. G. Harisinghani, J. Barentsz, P. F. Hahn, W. M. Deserno, S. Tabatabaei, C. Hulsbergen van de Kaa, Jean de la Rosette, and R. Weissleder, "Noninvasive detection of clinically occult lymph-node metastases in prostate cancer," N. Engl. J. Med. **348**, 2491–2499 (2003).

[3] R. A. M. Heesakkers, G. J. Jager, A. M. Hövels, B. de Hoop, H. C. M. van den Bosch, F. Raat, J. A. Witjes, P. F. A. Mulders, C. Hulsbergen van der Kaa, and J. O. Barentsz, "Prostate cancer: detection of lymph node metastases outside the routine surgical area with ferumoxtran-10-enhanced MR imaging," Radiology **251**, 408–414 (2009).

[4] R. A. M. Heesakkers, A. M. Hövels, G. J. Jager, H. C. M. van den Bosch, J. A. Witjes, H. P. J. Raat, J. L. Severens, E. M. M. Adang, C. Hulsbergen van der Kaa, J. J. Fütterer, and J. Barentsz, "MRI with a lymph-node-specific contrast agent as an alternative to CT scan and lymph-node dissection in patients with prostate cancer: A prospective multicohort study," Lancet Oncol. **9**, 850–856 (2008).

[5] H. C. Thoeny, M. Triantafyllou, F. D. Birkhaeuser, J. M. Froehlich, D. W. Tshering, T. Binser, A. Fleischmann, P. Vermathen, and U. E. Studer, "Combined ultrasmall superparamagnetic particles of iron oxide-enhanced and diffusion-weighted magnetic resonance imaging reliably detect pelvic lymph node metastases in normal-sized nodes of bladder and prostate cancer patients," Eur. Urol. **55**, 761–769 (2009).

[6] J. Rogowska, K. Batchelder, G. S. Gazelle, E. F. Halpern, W. Connor, and G. L. Wolf, "Evaluation of selected two-dimensional segmentation techniques for computed tomography quantitation of lymph nodes," Invest. Radiol. **31**, 138–145 (1996).

[7] J. H. Moltz, L. Bornemann, J.-M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, M. Fabel, H.-C. Bauknecht, M. Hittinger, A. Kiessling, M. Pusken, and H.-O. Peitgen, "Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans," IEEE J. Sel. Top. Signal Process. **3**, 122–134 (2009).

[8] L. Dornheim, J. Dornheim, and I. Rossling, "Complete fully automatic model-based segmentation of normal and pathological lymph nodes in CT data," Int. J. Comput. Assist. Radiol. Surg. **5**, 565–581 (2010).

[9] G. Unal, G. Slabaugh, A. Ess, A. Yezzi, T. Fang, J. Tyan, M. Requardt, R. Krieg, R. Seethamraju, M. Harisinghani, and R. Weissleder, "Semiautomatic lymph node segmentation in LN-MRI," *Proceedings of International Conference on Image Processing* IEEE, (2006), pp. 77–80.

[10] B. M. Dawant and A. P. Zijdenbos, "Image segmentation" in *Handbook of Medical Imaging*, edited by M. Sonka and J. M. Fitzpatrick (SPIE, Bellingham, Washington, 2004), Chap. 2, pp. 71–127.

[11] J. Liu, S. Huang, V. Ihar, W. Ambrosius, L. C. Lee, and W. L. Nowinski, "Automatic model-guided segmentation of the human brain ventricular system from CT images," Acad. Radiol. **17**, 718–726 (2010).

[12] K. Martin, L. Ibáñez, L. Avila, S. Barré, and J. H. Kaspersen, "Integrating segmentation methods from the Insight Toolkit into a visualization application," Med. Image Anal. **9**, 579–593 (2005).

[13] E. Day, J. Betler, D. Parda, B. Reitz, A. Kirichenko, S. Mohammadi, and M. Miften, "A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients," Med. Phys. **36**, 4349–4358 (2009).

[14] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," Int. J. Comput. Vis. **70**, 109–131 (2006).

[15] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," IEEE Trans. Pattern Anal. Mach. Intell. **26**, 1124–1137 (2004).

[16] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology **26**, 297–302 (1945).

[17] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: method and validation," IEEE Trans. Med. Imaging **13**, 716–724 (1994).

[18] P. Jaccard, "The distribution of the flora in the alpine zone," New Phytol. **11**(2), 37–50 (1912).

[19] J. M. Bland and D. G. Altman, "Multiple significance tests: the bonferroni method," Br. Med. J. **310**, 170 (1995).

[20] B. van Ginneken, "Supervised probabilistic segmentation of pulmonary nodules in CT scans," Med. Image Comput. Comput. Assist. Interv., Volume 4191 of Lecture Notes Computer Science (2006), pp. 912–919.

[21] J. H. Moltz, L. Bornemann, V. Dicken, and H. O. Peitgen, "Segmentation of liver metastases in CT scans by adaptive thresholding and morphological processing," *Proceedings of MICCAI Workshop on 3-D Segmentation in the Clinic*, 2008.

[22] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: revised recist guideline (version 1.1)," Eur. J. Cancer **45**(2), 228–247 (2009).

[23] A. M. Hövels, R. A. M. Heesakkers, E. M. Adang, G. J. Jager, S. Strum, Y. L. Hoogeveen, J. L. Severens, and J. O. Barentsz, "The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: A meta-analysis," Clin. Radiol. **63**, 387–395 (2008).

[24] R. Tiguert, E. L. Gheiler, M. V. Tefilli, P. Oskanian, M. Banerjee, D. J. Grignon, W. Sakr, J. E. Pontes, and D. P. Wood, "Lymph node size does not correlate with the presence of prostate cancer metastasis," Urology **53**, 367–371 (1999).

[25] J. Zhang, Y. Wang, and X. Shi, "An improved graph cut segmentation method for cervical lymph nodes on sonograms and its relationship with node's shape assessment," Comput. Med. Imaging Graph. **33**, 602–607 (2009).

[26] D. M. Honea and W. E. Snyder, "Three-dimensional active surface approach to lymph node segmentation," Proc. SPIE, **3661**, 1003 (1999).

[27] J. Yan, Tian ge Zhuang, B. Zhao, and L. H. Schwartz, "Lymph node segmentation from CT images using fast marching method," Comput. Med. Imaging Graph. **28**(1–2), 33–38 (2004).

[28] J. Yan, B. Zhao, L. Wang, A. Zelenetz, and L. H. Schwartz, "Marker-controlled watershed for lymphoma segmentation in sequential CT images," Med. Phys. **33**, 2452–2460 (2006).

[29] J. Yan, B. Zhao, S. Curran, A. Zelenetz, and L. H. Schwartz, "Automated matching and segmentation of lymphoma on serial CT examinations," Med. Phys. **34**, 55–62 (2007).

[30] C. Lu, S. Chelikani, X. Papademetris, J. P. Knisely, M. F. Milosevic, Z. Chen, D. A. Jaffray, L. H. Staib, and J. S. Duncan, "An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy," Med. Image Anal., **15**, 772–785 (2011).

[31] A. V. Young, A. Wortham, I. Wernick, A. Evans, and R. D. Ennis, "Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes," Int. J. Radiat. Oncol. Biol. Phys. **79**(3), 943–947 (2011).

[32] A. Chen, M. A. Deeley, K. J. Niermann, L. Moretti, and B. M. Dawant, "Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images," Med. Phys. **37**, 6338–6346 (2010).

[33] L. J. Stapleford, J. D. Lawson, C. Perkins, S. Edelman, L. Davis, M. W. McDonald, A. Waller, E. Schreibmann, and T. Fox, "Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer," Int. J. Radiat. Oncol. Biol. Phys. **77**, 959–966 (2010).

[34] J. Dornheim, H. Seim, B. Preim, I. Hertel, and G. Strauss, "Segmentation of neck lymph nodes in CT datasets with stable 3D mass-spring models," Acad. Radiol. **14**, 1389–1399 (2007).

[35] S. Gorthi, V. Duay, N. Houhou, M. Bach Cuadra, U. Schick, M. Becker, A. S. Allal, and J.-P. Thiran, "Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration," IEEE J. Sel. Top. Signal Process. **3**(1), 135–147 (2009).

[36] S. Gorthi, V. Duay, X. Bresson, M. B. Cuadra, F. J. Sánchez Castro, C. Pollo, A. S. Allal, and J.-P. Thiran, "Active deformation fields: Dense deformation field estimation for atlas-based segmentation using the active contour framework," Med. Image Anal., **15**, 787–800 (2011).