Learning New Parts for Landmark Localization in Whole-Body CT Scans

Vaclav Potesil*, Timor Kadir, and Sir Michael Brady

***Abstract*—The goal of this work is to reliably and accurately localize anatomical landmarks in 3-D computed tomography scans, particularly for the deformable registration of whole-body scans, which show huge variation in posture, and the spatial distribution of anatomical features. Parts-based graphical models (GM) have shown attractive properties for this task because they capture naturally anatomical relationships between landmarks. Unfortunately, standard GMs are learned from manually annotated training images and the quantity of landmarks is limited by the high cost of expert annotation. We propose a novel method that automatically learns new corresponding landmarks from a database of 3-D whole-body CT scans, using a limited initial set of expert-labeled ground-truth landmarks. The newly learned landmarks, called B-landmarks, are used to build enriched GMs. We compare our method of deformable registration based on such GM landmarks to a conventional deformable registration method and to a "baseline" state-of-the-art GM. The results show our method finds new relevant anatomical correspondences and improves by up to 35% the matching accuracy of highly variable skeletal and soft-tissue landmarks of clinical interest.**

***Index Terms*— Computed tomography (CT), graphical models, image registration, oncology, semi-supervised learning.**

I. Introduction

THE GOAL of this work is to accurately and reliably localize anatomical landmarks in 3-D computed tomography (CT) scans of a large part of the patient's anatomy, for example the upper body in the case of lung cancer. Our method automatically discovers new landmark correspondences (*B-landmarks*), by making use of a limited set of landmarks of clinical interest (*primary landmarks*) whose ground-truth positions have been manually annotated by an expert in the training data.

The automatic localization of anatomical landmarks in medical images can facilitate image reading and reporting, for example automated placement and referencing of regions of interest (ROIs). It can also serve as the basis for more complex processing steps, for example the construction of atlases, organ segmentation, or the planning of radiation therapy. Such clinical applications are particularly important in 3-D whole-body CT images that contain many different anatomical structures which can be highly variable between patients, especially when pathologies such as tumors are present. This variability makes the automatic localization of anatomical landmarks a challenging task [24]. Nevertheless, the relative positions of organs in human body are not random but show common structural patterns. A natural way to model such objects is to use spatial priors based on parts-based graphical models (GM) which have been shown to exhibit attractive properties for substructure labeling in single organs [3], [7], [20]–[22], [25]. Typically, GMs are learned from *manually* annotated training images, and as a result the quantity of landmarks and images is limited by the high cost of annotation. Moreover, annotated primary landmarks are not always placed at the best positions to optimize learning due to the absence of relevant appearance features and/or low spatial correlation with other annotated landmarks. The latter is particularly true in our case where only a sparse set of primary landmarks is used in whole-body volumes. This results in both spatial and appearance models with excessive variance relative to the required matching accuracy. This problem could be resolved by including additional landmarks that have common appearance features in the database and stronger spatial correlations to the individual primary landmarks than other primary landmarks. Unfortunately, manually determining which additional landmarks may be relevant and annotating them in a large database of full-body scans are tedious tasks.

In this paper, we propose to determine automatically such additional landmarks. Our method is weakly supervised in that an initial set of supervised landmarks is used to guide the discovery of new, different landmarks. We refer to the new landmarks as *B-landmarks*. B stands for "Beacon," as we search for confidently localizable reference points that will facilitate the localization of the primary landmarks of clinical interest. To achieve this, we extend recent work based on dense matching of a particular type of parts-based GM, personalized pictorial structures (PS) [18]. Our main contribution is a method that automatically finds the B-landmarks, by searching for image patches that are consistently appearing and localized across the training set. It then builds enriched GMs that combine the initial primary landmarks and the B-landmarks in the same framework. The method is illustrated in Fig. 1, which shows a sagittal section of a CT image, centered on the L5 vertebra. On the left [Fig. 1(a)], we show a detail of the original GM that makes use only of the primary landmarks. In the center [Fig. 1(b)], we show a number of B-landmarks learned using our method from the training data.

Manuscript received July 31, 2013; revised October 25, 2013; accepted October 30, 2013. Date of publication November 20, 2013; date of current version March 31, 2014. *Asterisk indicates corresponding author.*

*V. Potesil is with the Department of Engineering Science, University of Oxford, OX1 3PJ Oxford, U.K., and also with Mirada Medical Ltd., OX1 1BY Oxford, U.K. (e-mail: vaclav@robots.ox.ac.uk).

T. Kadir is with Mirada Medical Ltd., OX1 1BY Oxford, U.K.

M. Brady is with Department of Oncology, University of Oxford, OX1 3PJ Oxford, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2291711

0278-0062 © 2013 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

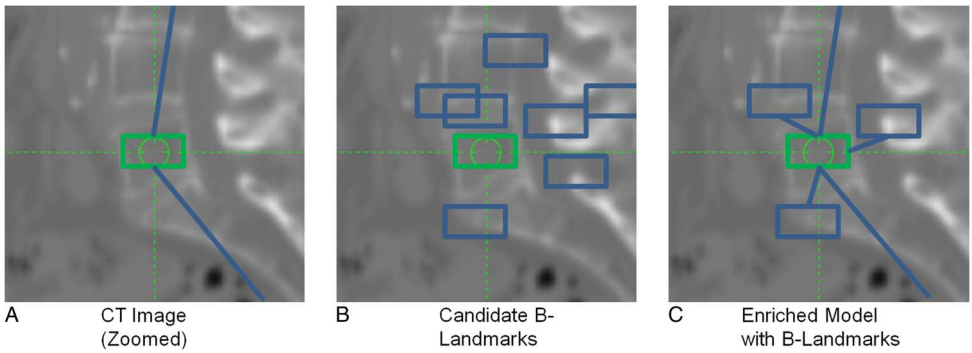
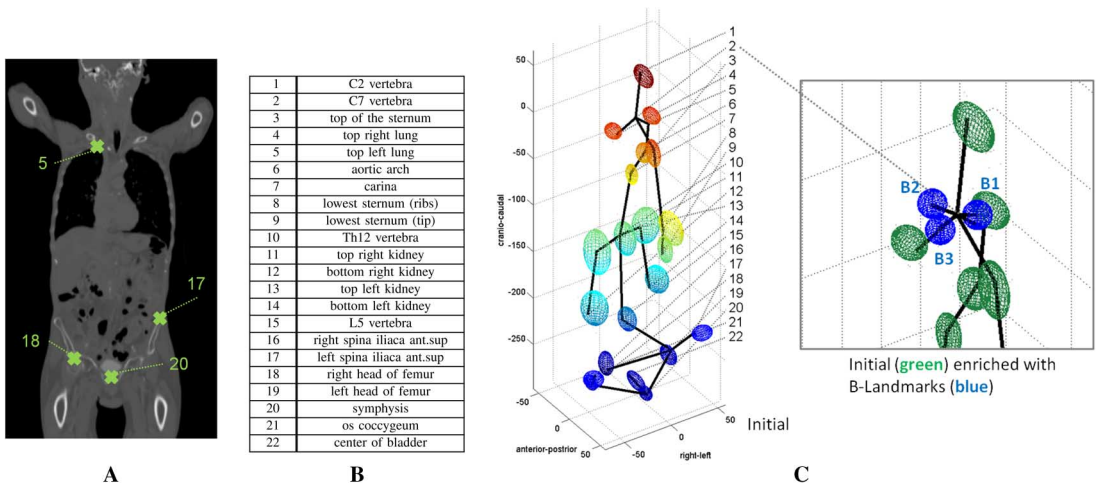


Fig. 1. Illustration of learning new landmarks to enrich a GM. A: Sagittal slice of a CT. Primary landmark (L5 vertebra) is shown in the green box. Blue lines show spatial constraints in the GM to other (distant) primary landmarks. B: New automatically learned landmarks (B-landmarks) are shown in the blue boxes. C: Enriched model built using the primary landmarks and the best subset of B-landmarks.



1	C2 vertebra
2	C7 vertebra
3	top of the sternum
4	top right lung
5	top left lung
6	aortic arch
7	carina
8	lowest sternum (ribs)
9	lowest sternum (tip)
10	Th12 vertebra
11	top right kidney
12	bottom right kidney
13	top left kidney
14	bottom left kidney
15	L5 vertebra
16	right spina iliaca ant.sup
17	left spina iliaca ant.sup
18	right head of femur
19	left head of femur
20	symphysis
21	os coccygeum
22	center of bladder

Fig. 2. Landmarks and anatomical constraints. Shapes show valid population mean constraint ranges, plotted at 2σ from the mean. Stick-man shows the pair-wise relations in a tree-structured GM. Covariances are plotted with respect to the nearest landmark in the graph. Nodes and leafs of the GM correspond to the positions of landmarks (C) in the patient image (A). Figure C shows the initial model learned from expert-annotated data (left) and the enriched GM with B-landmarks. For clarity, only three B-landmarks related to C7 Vertebra are shown.

On the right [Fig. 1(c)], we show a detail of an enriched GM, which makes use of the primary landmarks and the best subset of B-landmarks.

This paper is organized as follows. After a brief introduction to the PS representation in Section III, and related work in Section II, we detail our new method for learning new corresponding landmarks in Section IV. The proposed method finds tens of new anatomical landmarks, and builds spatial and appearance models that combine the primary landmarks and the automatically learned landmarks into enriched models. A quantitative evaluation is performed on clinical 3-D whole-body CT of cancer patients showing an improvement of landmark-based registration across all pairs of training images and an improved localization accuracy, occlusion handling for highly variable skeletal and soft-tissue landmarks (Section V) using the GM matched to unseen test images. Section VII concludes and discusses directions for future work.

II. Related Work

In computer vision, weakly supervised learning of parts-based GM, which foregoes the need for explicit labelling of part locations and scales, has been explored for various tasks. We summarize here only the most relevant work.

Fergus *et al.* [11], [12] built on the so called constellation model studied in previous work [23], where assignments between candidate features proposed by a feature detector and the GM parts are treated as latent variables to be optimized. The joint nature of the constellation shape model is highly computationally expensive in the learning phase, which restricts the number of parts and candidate locations that can be handled. To give good performance, the algorithm needs to use a highly sensitive, specific and geometrically accurate interest point detector [14]. Moreover, the large number of free parameters in the joint model means that hundreds of training examples of each class are needed in order to prevent over-fitting. In subsequent work [11], the authors address some of the practical shortcomings of the constellation model. Nevertheless, the computational requirements remain prohibitive for our purposes, where there are thousands of possible placements for dozens of parts in 3-D medical scans. The method also cannot learn parts which are not detected reliably by one of the generic feature detectors, which is a particular problem if such methods are applied to medical

images, which are noisy, have low contrast and spatial resolution, and are often dominated by irregular image structures

Bourdev *et al.* proposed Poselets [2] as a regression model for detecting people and localizing body parts in 2-D images. Poselets are parts that are tightly clustered both in the 3-D pose space as well as the 2-D image patch appearances. In follow-up work [1], the authors used only 2-D annotations and proposed to consider the spatial context of Poselet activations in order to re-rank noisy Poselet activation scores according to their mutual consistency. Poselets are similar to our work, in that the 3-D annotation guides a greedy search for "good" parts, which are tightly clustered in configuration space and effective at joint localization. However, we directly search for new landmarks in fully 3-D volumetric data, instead of 2-D images. Our proposed approach is also optimized for the localization accuracy of individual landmarks, whereas the Poselet approach is designed for the challenging task of approximate person pose estimation in 2-D images with large pose variation.

Crandall *et al.* [4] proposed techniques for weakly supervised learning of PS models for object recognition and localization, with only a class label for each training image. In a later paper [5], the authors applied their weak learning technique to learn a hierarchical model, composed of a simple star-structured object model and a coarse star-structured model which represents the scene context. The locations of the parts are learned automatically with respect to a labeled bounding box. Crandall's weakly supervised search learns both consistent appearance and spatial models, and has been shown to provide very good results for certain object categories, particularly when augmented with the scene model. However, the redundant representations result in computational costs that are even higher than those of Fergus's algorithm.

In medical imaging, weakly supervised learning for parts-based GMs has been explored recently in the context of group-wise registration of 2-D images, whereas we address a far more complex application by using 3-D upper-body CT scans with widely different patterns of variability between landmarks (skeletal and soft-tissue landmarks) and patients (abnormalities due to cancer). Zhang [25], [26] proposed to learn new correspondences from the data and then to use such correspondences to initialize group-wise deformable registration of 2-D hand radiographs and lumbar spine images. The approach relies on a population-based genetic algorithm or combinatorial optimization strategy to select a subset of the best parts from a large pool of densely sampled candidates. However, their cost function is based on explaining the whole of an image region using nonrigid deformations based on the centers of the part models. This may not be suitable in 3-D scans of cancer patients, which often have substantial variability in tissue appearances and pathologies. Also, they seek a sparse set of parts which can be used to initialize a local optimization based group-wise registration scheme for the entire image, whereas our work seeks to discover new landmarks which can then be detected accurately across the dataset.

Langs [15] described a method of constructing sparse shape models from unlabeled images, by finding multiple interest points and using the minimum description length (MDL) principle to choose the best model and to determine optimal correspondences. However, that approach represents shape correspondences using only sparse interest points and a global active shape model, which cannot deal efficiently with multiple candidates and is more easily affected by local perturbations. Donner *et al.* [8] learn a GM of shape and appearance, which relies on a single manually annotated training image to define the structure of interest, and iteratively include high confidence matches in the emerging model. In subsequent work [7], the authors introduced a method that is similar to our baseline model in that it does not rely on predefined interest point detector. However, unlike our approach, their method entirely relies on manually labelled training data and the corresponding structures in the neighborhood of the labelled landmarks (captured by a learned Random Forest + Haar wavelets classifier) are not identified explicitly as new landmarks.

For completeness, we include the approach by Criminisi, who proposed a discriminatively-trained approach based on random forest regression (i.e., not a parts-based GM). Their computationally efficient method predicts the sides of organ bounding boxes [6] or crudely segments entire organs [13]. Consistent clusters of features in the organ context can be interpreted as anatomical landmarks, although the examples which the authors have presented to date only show bulk reference structures e.g., the lung or the (entire) iliac crest, which do not permit accurate anatomical point correspondences.

III. Parts-Based Models for Anatomical Landmark Localization

Our work is built on a type of parts-based graphical model, called PS. We recall that model briefly, referring the reader to [10], [18] for more detail. Moreover, we build on our previous work which improved localization accuracy and reliability of the standard PS by model personalization. We also briefly describe that here, but refer the interested reader to [18] for detailed description.

We start by introducing the function which assigns a cost to a given landmark configuration (i.e., set of landmark locations) in a upper body CT scan. The following section (Section IV) will describe how the new landmarks are discovered.

A. Cost Function and Basic Notation

All primary landmarks and B-landmarks share the same cost given a landmark configuration. Their cost comprises a unary energy term for local part appearance and a set of pair-wise terms for spatial compatibility. This reflects naturally the relational anatomical constraints between internal structures in the human body. Formally, given a rectangular image volume I that contains the upper body of the patient and background (air), we aim to find the locations $\boldsymbol{l_k}$ of n landmarks. Each landmark and B-landmark is defined by its [x,y,z] coordinates.

The model takes the form of a tree-structured undirected graphical model (GM), named G, with associated parameters $\boldsymbol{\Theta}$. At training time, n landmarks $v_1, \ldots, v_j, v_k, \ldots, v_n$ are placed manually by the expert in a set of images of different patients at locations $\boldsymbol{l_1}, \ldots, \boldsymbol{l_j}, \boldsymbol{l_k}, \ldots, \boldsymbol{l_n}$, defined by their [x,y,z] coordinates. The appearance u_k of each landmark v_k and relational spatial constraints $c_{k,j}$ between the landmarks v_k and v_j are learned automatically from the annotated training

set. The B-landmarks, once learned automatically using the method proposed in this paper, are represented in the same way as the initial hand-annotated landmarks.

We define the probability of a given configuration L conditioned on the image I as

$$p(L|I,\Theta) \propto \prod_{k=1}^{n} \underbrace{p(I|\boldsymbol{l_k}, u_k)}_{\text{appearance terms}} \prod_{v_{k,j}\in G} \underbrace{p(\boldsymbol{l_k}, \boldsymbol{l_j}|c_{kj})}_{\text{spatial terms}}. \quad (1)$$

L is the set of locations for all n landmarks in the image I and Θ is a set of parameters of the GM that comprises the landmark appearance, the relational spatial constraints and the model connectivity. Given the quality of match of patch appearance model u_k to the image I, the first term in (1) represents the cost of placing the landmark v_k at location $\boldsymbol{l_k}$ (unary data terms). The second term is the spatial compatibility cost of connected pairs of landmarks v_k, v_j placed at $\boldsymbol{l_k}, \boldsymbol{l_j}$, given a set of prior anatomical constraints c_{kj} that restrict their relative positions (pair-wise terms). Equivalently, the above equation can be rewritten as an energy minimization problem by taking its negative logarithm.

In the explanation of the learning method (Section IV), we will refer to B-landmarks pertaining to a primary landmark as $\{v_k^{b+}\}$, in order to clearly distinguish it from the manually labeled primary landmark v_k. However, in terms of model parameters and fitting of a learned model, B-landmarks are identical to primary landmarks (that is, equivalently to fitting a larger GM with more manually annotated landmarks).

B. Appearance Models

Local tissue appearance is modelled in terms of local rectangular image patches centered on landmark locations $\boldsymbol{l_k}$ at a scale $s_k(x,y,z)$. Positive and negative examples from $i = 1 \ldots N_I$ training images are used. We do not use intensity normalization before training, since in our application the raw intensity values carry important information. First, for positives, patches are projected onto a set of n basis functions obtained from the ground-truth patches. In the initial model, patches are projected onto a set of n_f features comprising the mean template and the top $(n_f - 1)$ eigen-patches obtained from the ground-truth training patches using Principal Component Analysis. In the personalized model, the generic features are replaced by a subset of training patches adapted to each test image at run-time [18].

The patch scale $s_k(x,y,z)$ for each part is optimized for landmark localization using the training approach described in [17]. In more detail, our patch-size optimization uses a combination of complementary measures of appearance model quality, which favor not only accurate localization, but also spatially compact and repeatable posterior distributions measured in the local regions of interest around the ground-truth locations of the primary landmarks. This leads to reduced over-fitting to (noisy) ground-truth annotations, and more reliable appearance models.

The variability of true part appearance is represented by the distribution of the feature responses, $H_k^+ \sim N(\boldsymbol{M_k^+}, \boldsymbol{\Sigma_k^+})$ for each part v_k, where $\boldsymbol{M}, \boldsymbol{\Sigma}$ are diagonal matrices. $H_k^- \sim N(\boldsymbol{M_k^-}, \boldsymbol{\Sigma_k^-})$ is the model for the local "background" i.e., nonlandmark tissue. The parameters of the Gaussian model (i.e., the means and the covariances) are learned directly from the training data. In the personalized model, the generic models are replaced by a set of weights for the subset of training patients adapted at run-time [18].

C. Spatial Models

Pair-wise spatial compatibility terms $c_{kj}(\boldsymbol{l_k}, \boldsymbol{l_j})$ penalize part placements outside of the range of anatomical variability exhibited within the training database. The functional form of this spatial model is approximated as a distribution over $\vec{l}_{kj,i} = \boldsymbol{l_{k,i}} - \boldsymbol{l_{j,i}}$ (the relative position vector of two landmarks v_k, v_j in training patient i, assuming that the patient's pose has been normalized for bulk-body rotations). $p(\boldsymbol{l_k}|\boldsymbol{l_j}) \sim N(\vec{l}_{kj}, \boldsymbol{M_{kj}^c}, \boldsymbol{\Sigma_{kj}^c})$ is a Gaussian model with diagonal covariance that models the relative position of two landmarks v_k, v_j.

The parameters of the GM are learned following the approach of [10]. The parameters include the model connectivity (i.e., a minimum spanning tree) and the spatial model $c_{kj}(\boldsymbol{l_k}, \boldsymbol{l_j})$ for each edge in the tree. In brief, the connectivity between nodes (landmarks) in a graph G is obtained from the n-by-n matrix Q of edge compatibilities, estimated using all training patients D

$$Q_{kj}(G|D) \sim \sum_{\boldsymbol{l_k}, \boldsymbol{l_j} \in D} -\log p(\boldsymbol{l_k}, \boldsymbol{l_j}|c_{kj}). \quad (2)$$

The set of edges (constraints) connecting all nodes in a tree is obtained by finding the minimum spanning tree of the complete graph such that the sum of the edge costs is at a minimum. Intuitively, this can be seen as searching for edges with the most "rigid" pair-wise constraints.

In the initial population mean model, distributions are calculated from all examples and the pair-wise terms penalize deviation from the mean. The parameters of the pair-wise spatial models are fitted by directly calculating the maximum likelihood estimate (MLE) for the Gaussian. In the personalized model, the population means are replaced with local weighted subsets of spatial vectors from labeled training images, similar to Gaussian radial basis functions [18], and the pair-wise term penalize deviation from any subset of training patients.

D. Model Fitting

Population models in both the spatial and the appearance models introduce a bias towards the population mean. The so-called personalization method replaces population models (i.e., uni-modal Gaussians) by jointly leveraging weighted combinations of labeled exemplars to obtain personalized models for landmark localization, while preserving the advantages of the PS. Once the B-landmarks are learned from the database using our proposed method (Section IV), they are included in the training set for the updated model. During model fitting, B-landmarks are treated identically to primary landmarks.

Here, we provide the reader with a basic overview of the key steps in model fitting. See [18] for full description of our matching approach.

1) **Find Local Reference Frame (offline)**. During training, we automatically find local reference frames (LRFs) for each of the landmarks. The LRFs consist of local subsets of landmarks, which are spatially predictive of each others' locations. During testing, the (fixed) LRFs are used as an index to search in the training database and pull out labeled training images which are locally anatomically similar to

the test images, in order to produce personalized spatial models. In more detail, the LRF for each primary landmark is found offline by taking the three landmarks that have the lowest spatial variance with respect to the primary landmark in the training database.

2) **Initial Fit of the Population Model**. First, a population model built from the entire training set is fitted to the test image. The appearance models are applied to the CT volume using a sliding window approach. We use a Naive Bayes classifier to obtain a dense posterior probability map for each landmark. Min-sum belief propagation is used for exact, globally optimal fitting of the entire GM, using the dense appearance maps and the pair-wise compatibility terms. The efficient computation of the pair-wise messages in the belief propagation framework is implemented by directly extending to 3-D the Fast Generalized Distance Transforms approach of [10].
3) **Spatial Personalization**. Here, we produce spatial maps personalized to a particular test patient that assign a cost of placing a landmark at each voxel in the test patient image (personalized spatial prior). In more detail, this is done at run-time for each landmark, by estimating local anatomical similarities between the test patient and the training database, We use a robust voting procedure, based on pair-wise L2 distances on the LRF between the test patient and all training patients. The pair-wise spatial vectors from the most similar training patients are used to replace the spatial means in the Pictorial Structure model, in a Radial Basis Function-like manner.
4) **Appearance Personalization**. Here, we produce posterior probability maps personalized to the test patient. First, we obtain appearance vote maps, by letting all training image patches vote for voxels within a local query region around the MAP (maximum a posteriori probability estimate, i.e., the mode of the posterior distribution) location. Finally, the appearance vote maps are combined with the personalized spatial maps. We record the locations and the subset of training patches that voted for the top mode in that map.
5) **Model Update and Refit**. In the final step, we produce and refit personalized appearance and spatial models. For the personalized appearance model, we replace the generic PCA features with the subset of training patches. For the personalized spatial model, the spatial constraint mean is set to the mode of personalized spatial prior. The resulting personalized model can also be stored and used for any follow-up scans of the same patient.

IV. Learning B-Landmarks for Parts-Based Graphical Models

Our method is an iterative process that automatically searches for a number of new B-landmarks, $\{v_k^{b+}\}$, in the neighborhood of one primary (i.e., initial, expert-annotated) landmark one at a time (v_k) across the entire training image set (N_I images). It starts with an initial set of candidate landmarks, which are then refined, ranked and reduced until a small subset remains. The best are then chosen and added to the original model.

The human body contains millions of anatomical correspondences, and it is computationally infeasible, and unnecessary, to evaluate all of them. At each iteration, we restrict our search to those B-landmarks that are a) confidently detectable and spatially well-localized; b) present in all images in the training set (i.e., few or no occlusions); and c) can improve localization of the primary landmarks (i.e., highly spatially predictive of the primary landmark).

The following steps are repeated for each primary landmark one at a time.

1) **Sample "interesting" B-landmark locations**. First, we sample initial approximate B-landmark locations from a proposal distribution that favors textured image regions.
2) **Build initial models**. For each candidate B-landmark, we initialize the appearance and spatial models using the sampled locations. Since the human body is seldom completely "rigid," the initial model built from patches at a fixed offset in different training patients will contain incorrect correspondences. Those are dealt with in the following model refinement.
3) **Iteratively refine and select the best B-landmark models**. For each candidate B-landmark, we optimize their locations across all training images. The following algorithm proceeds iteratively for a fixed number of iterations.
   a) We select an active subset of confidently localized training image patches.
   b) Exemplars in the active set vote for the B-landmark location across all training images.
   c) The spatial and appearance models are updated using the new location estimates.

   After each iteration, we remove the worst B-landmarks to limit the computational expense.
4) **Build enriched models**. We build an enriched GM, which contains the primary landmark and the subset of B-landmarks that contributes to the primary landmark localization.

The four steps are now explained in more detail.

A. Appearance-Based Sampling of Interesting B-Landmark Locations

In this section, we show how the initial B-landmark correspondences are obtained from the training images. The aim is to sample locations from tissue regions with rich textures as opposed to homogeneous regions.

N_s sets of candidate correspondences are sampled from a proposal distribution derived from the training images in a local region around primary landmark v_k. For computational reasons, we currently limit the sample size to $N_s = 250$ sets per primary landmark. Each sample from the proposal distribution is a spatial vector. We initialize the correspondences in all of the training images in the set with the same spatial vector relative to the primary landmark location.

In our implementation, a proposal distribution is obtained by summing up pseudo-probabilities derived from distance transforms of the edge map of each training image using the monogenic signal [9], across all training images in a local region surrounding primary landmark v_k. The summation is performed at the same scales as the patch sizes that are used to represent each

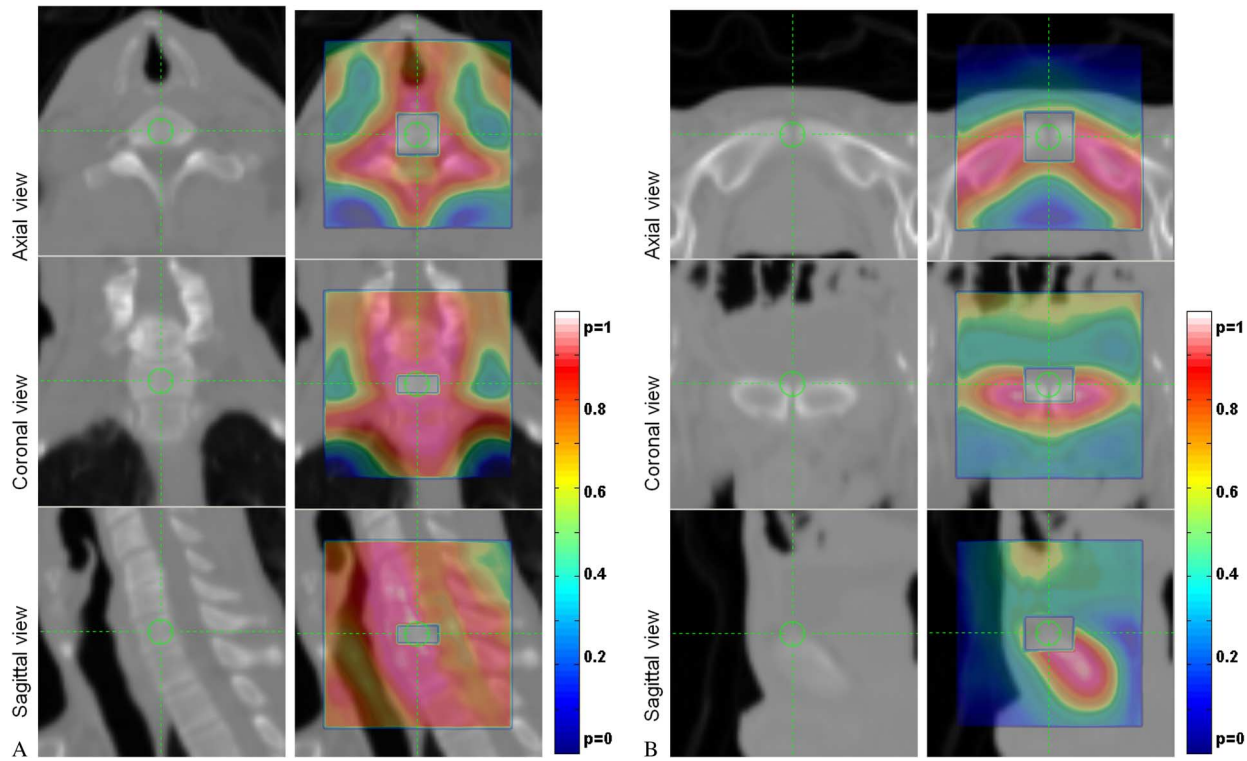


Fig. 3. Example of proposal distributions used to sample candidate B-landmarks. Illustration of the B-landmark proposed distributions around the C7 vertebra (A) and the symphysis (B). Proposal derived from all patients in the database is the colored map (right) overlaid on top of a randomly selected patient CT image (left). Each of the images show a $16 \times 16 \times 16$ cm CT image region centered at the ground-truth location of the respective reference landmarks (in the cross-hair). Red color means locations sampled with high probability ($\mathrm{p} = 1$). Blue color means low probability ($\mathrm{p} = 0$).

B-landmarks. This simple method of generating a proposal distribution aims to sample mostly from tissue regions with rich textures as opposed to homogeneous regions that have no texture at the given scale in the majority of images (e.g., the homogeneous tissues in the abdomen).

In more detail, the pseudo-probabilities for computing a proposal distribution are derived by: 1) running the monogenic filter [9] to obtain edge amplitude maps; 2) converting the maps to pseudo-probabilities (i.e., $\mathrm{edge} = 1$, $\mathrm{nonedge} = 0$) using a sigmoid transform; 3) filtering the pseudo-probabilities with a running average filter at the same patch scale as primary landmark; 4) summing the maps across all training patients and excluding the region immediately surrounding the primary landmark (set to $p = 0$, in order to avoid sampling the same landmark that is already annotated in the training database); and finally 5) renormalizing to obtain a valid pdf. In this way higher values (close to $p = 1$) are accorded to textured regions, while low values (close to $p = 0$) are accorded to regions without texture (i.e., smoothly varying intensity). The method has the additional advantage of being computationally efficient in removing initial correspondences that involve texture-less regions.

An example of the sampling proposed distribution for C7 vertebra is shown in Fig. 3. The proposal down-weights homogeneous regions (lung, muscle) and assigns zero probability to the region under the reference landmark (center) and near the search region boundaries. Note that the role of the proposal distribution is to improve computational efficiency. Any alternative way of generating proposal distributions that up-weights rich textures and down-weights homogeneous regions could be used.

The output of this stage is a large set of initial candidate B-landmark correspondences, located in a neighborhood of each primary landmark (i.e., many candidate B-landmarks for each primary landmark) across all training images. This set contains more correspondences than we need; but we need a large initial number to get a good final few.

B. Building Initial Models

First, we build initial appearance and spatial models for each candidate B-landmark v_k^{b+}, using the spatial vector sampled in Section IV-A (one sampled vector per B-landmark candidates).

The appearance model E is built from patches in all the training images centered around the candidate B-landmark placements using the same scale s_k as the landmark v_k (i.e., a set of patches for each B-landmark candidate). Here, for computational reasons we restrict the search to good landmarks at a single scale. Ideally, the scale would be optimized, for example, by choosing the best landmarks over multiple scales using the technique of [17]. However, this is not done here as it would substantially increase the running time of the current implementation (see Section VI for a discussion of the current running time).

Similarly, we build a spatial model $c_{kb+}(\boldsymbol{l_k}, \boldsymbol{l_{b+}})$. At iteration $t = 0$, we place the initial location of the B-landmark v_k^{b+} at a fixed offset from the primary landmark in each training image. The offset is given by the spatial vector sampled in Section IV-A. Therefore, the initial $\boldsymbol{M^c_{kb+,i}}$ is the same for all $i = 1 \ldots N_I$ training images. $\boldsymbol{\Sigma^c_{kb+,i}}$ is initialized to 10 voxels in all images. This can be seen as a initially 'rigid' spatial

constraint. In the following steps, we allow the B-landmark locations and the models to adjust to the deformations in each training image.

The output of this stage are initial appearance and spatial models for each candidate B-landmark.

C. Iterative Refining and Selection of the Best B-Landmark Models

The initial model is iteratively refined using a four step approach.

1) Estimate Active Subset of Training Images: Next, we remove poorly localized exemplar patches from the initial (noisy) appearance model for each candidate B-landmark. We do this by computing an autocorrelation map in a local neighborhood around the position of the B-landmark in each image. Then we measure how the autocorrelation map is similar to a "Dirac" function, rank all training images and keep the top 50% images.

In more detail, the autocorrelation map is considered to be a three-dimensional spatial histogram and ranked using a quality measure q_i, which is obtained as the Jensen–Shannon divergence (JSD) between the autocorrelation map w_i and an idealized "Dirac" response w_{ref} at the central voxel. This is performed for each B-landmark correspondence and for each image i: $q_{\mathcal{N}_{\boldsymbol{l_0}},\boldsymbol{i}} \sim (1-JSD(w_i|w_{ref}))$. The best localized patches according to q_i are retained in the active set $E_{\rm act}$ (top 50% of the training set, chosen empirically).

The output of this stage is a subset of the most confidently localized training image patches for each candidate B-landmark.

2) Active Images Vote For B-Landmark Locations Across All Other Images: In this step, for each B-landmark and training image, the active subset of training images (Section IV-C1) votes for the most likely B-landmark location in that image.

First, for each training image, all other training images in the active subset vote for the best location of the B-landmark v_k^{b+}. This voting procedure yields exemplar feature response maps $f_{e,i}$ for each image and exemplar template in the active subset. Specifically, we apply all exemplar templates e in $E_{\rm act}$ (except the current training image) to the training image using normalized cross correlation (NCC). For computational efficiency, we evaluate the NCC only in a fixed size region of $s_R = 21$ voxels around the B-landmark mean location given by the spatial model. The region size was motivated by the worst-case localization error of the baseline model formed by primary landmarks. Anatomical structures with higher spatial variability across different patients would be unlikely to form useful B-landmarks that would improve the localization of the primary landmark. Finally, the exemplar feature maps $f_{i,e}$ are turned into a set of exemplar pseudo-probability maps derived from the NCC function $p_{e,i}$: $p_{e,i} \propto \exp(-f_{e,i})$.

For each training image, we combine the strongest peaks from the pseudo-probability maps from all exemplars e into one map p_i per training image I_i

$$p_i(\boldsymbol{l_{b+}}|E_{\rm act}, I_i) \propto \sum_{e \in E_{\rm act}} K_W(p_{e,i}) \tag{3}$$

where K_W is a spatial averaging window that takes account of the intrinsic uncertainty of landmark point annotations and the limited image resolution. The averaging is applied only within windows centered at the strongest peaks in each $p_{e,i}$ (i.e., setting other voxels to 0) and the results are summed into a single map. Finally, we build a set of **consensus maps** $P_C \in \{p_{C,1}, p_{C,2}, \ldots, p_{C,i}\}$ for each image i in the training set by combining the appearance (derived from the pseudo-probability maps) and the spatial model costs

$$p_{C,i}(\boldsymbol{l_{b+}}|I_i, c_{kb+}, E_{\rm act}) \propto \underbrace{p_i(\boldsymbol{l_{b+}}|E_{\rm act}, I_i)}_{\text{appearance}} \times \underbrace{p_{S,i}(\boldsymbol{l_k}, \boldsymbol{l_{b+}}|c_{kb+})}_{\text{spatial}} \tag{4}$$

where $p_{S,i}$ is a spatial prior map obtained using the spatial model c_{kb+}, I_i is the ith training image. The role of the spatial prior is to constrain the search to the neighborhood around the initially sampled location (4). For example, it will penalize spurious matches in the case of a similar structure further away from this location such as the body of another neighboring vertebra. The appearance term comes directly from (6).

The output of this stage is a set of **consensus maps** $P_C \in \{p_{C,1}, p_{C,2}, \ldots, p_{C,N_I}\}$ for each image in the training set which indicate likely locations for correspondences for each candidate B-Landmark.

3) Update the Models: In this step, we update each candidate B-landmark location and the appearance and spatial models based on the consensus maps across the entire set of training images.

First, we aggregate the consensus maps from all training images into a single aggregate consensus map per training set (see Fig. 4 which illustrates the whole process)

$$p_{\rm sum} \propto \sum_{i=1}^{N_I} K_\sigma(p_{C,i}) \tag{5}$$

K_σ is a spatial aggregation kernel that accounts for each B-landmark's intrinsic spatial uncertainty. The kernel is a Gaussian filter applied across the entire search region; the parameter σ is iteratively re-estimated as the spatial variance of the candidate B-landmark correspondence locations across all training patients in the previous iteration.

Finally, the updated candidate B-landmark locations in each training image, $\hat{L}_{b+,i} \in \{\hat{l}_{\mu_1}, \ldots, \hat{l}_{\mu_{N_I}}\}$, are found as the nearest local maximum $l_m \in L_{m,i}$ in each $p_{C,i}$, to the location of the strongest mode $\hat{l}_{\rm sum}$ of the aggregate map $p_{\rm sum}$, according to the L^2 distance. The strongest mode is the voxel with highest value in the aggregate map $p_{\rm sum}$. After each iteration t, the B-landmark locations, appearance models E, spatial models $c_{kb+}(\boldsymbol{l_k}, \boldsymbol{l_{b+}})$, spatial priors $p_{S,i}$, as well as K_σ are also updated using the detections $\hat{L}_{\mu_i}$.

The output of this step are the improved location estimates of each candidate B-landmark in *all* training images; and updated appearance and spatial models.

4) Reduce the Set of Candidate Correspondences: In this step, we reduce the set of candidate B-landmark correspondences, by discarding candidate B-landmarks that are unlikely to improve localization of the labeled primary landmark. This is done after each iteration, in order to reduce the computational requirements on the moderate desktop machine on which the implementation has been developed.

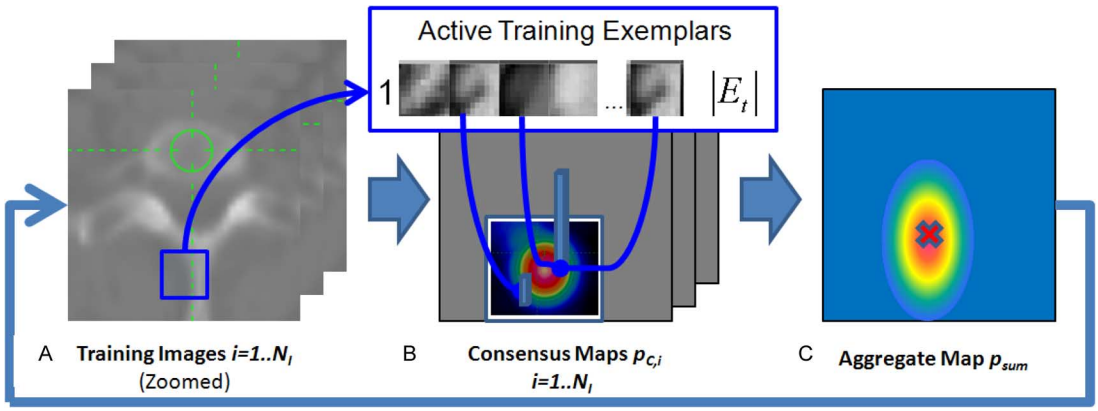


Fig. 4. Training images vote for B-landmark locations across all images. Illustration of the B-landmark search around C7 vertebra. A: Stack of images $i = 1 \ldots N_I$. Blue patches indicate candidate B: Landmark placements from iteration $t-1$. B-all active images vote for B-landmark locations (overlaid on the spatial prior $p_{S,i}$). C Voting results from all images are aggregated into a summary map p_{sum}.

At each iteration t, we first remove poor B-landmark candidates that converge to another landmark or to the landmark v_k. Then, we rank the remaining B-Landmark candidates according to a score combined from three measures. The normalized covariance of the spatial vectors $\vec{l}_{b+k}$ between the B-landmark and the primary landmark position, is a surrogate of how well v_k^{b+} predicts the position of v_k. Pair-wise self-similarity (NCC, allowing for small spatial offsets) between all image patches indicates whether or not the detections capture the same structure. The sum of the votes in the winning modes $p_{C,i}$ in all images is a measure of how well the entire image database supports each B-landmarks hypothesis, i.e., B-landmarks that receive many strong votes are more likely to be genuine correspondences. In a final iteration, B-landmarks are ranked by this combined measure.

In more detail, we look for B-landmark candidate where the following quantity attains its minimum value:

$$\begin{aligned} Q = & Q_{\text{spatial covariance}} \\ & \times Q_{\text{self-similarity}} \\ & \times Q_{\text{consensus votes}} \end{aligned} \tag{6}$$

$Q_{\text{spatial covariance}}$ is the determinant of the covariance matrix of spatial vectors between the primary landmark and the current location estimate of the B-landmark across all training images. We want to make sure we do not penalize B-landmarks which are located further away from the primary landmark (hence naturally have relatively higher covariance), hence we normalize the score by the square of the distance between the primary landmark and the average location of the B-landmark across all training images. Lower value is a better B-landmark. $Q_{\text{consensus votes}}$ is the sum of the values from the maximum mode in the consensus map (4) in all images in the active set of training images. We take $\exp(-X)$ of the sum X, such that lower value is a better B-landmark. $Q_{\text{self-similarity}}$ is the sum of pair-wise similarities between all image patches (at the current location estimates of the B-landmark) in the active training set. The pair-wise similarities are defined as the sum of normalized cross-correlations between all pairs of patches in the active training set. To ensure robustness against very small local offset, we take the minimum value across all shifts by $+-1$ voxels in all dimensions x, y, z. Lower value is a better B-landmark $(0 = \text{perfect correlation})$.

The output of this step is a reduced set of candidate B-landmarks that appear most consistently, are most confidently localized, and contribute to the localization of the primary landmark. This set will be again iteratively optimized using the algorithm described in Sections IV-C1to IV-C4. We stop this process after a fixed number of iterations $(t = 15)$.

D. Building Enriched Graphical Models Including B-Landmarks

The final step is to add new correspondences to the GM. We do this by building star-graphs of all combinations of B-landmarks $\{v_k^{b+}\}$, which are connected to a root primary landmark v_k. Then we choose the combination of B-landmarks that most improves the localization of the primary landmarks. For each combination and training image, we generate max-posteriors, which are spatial maps proportional to the min-marginal of the star GM centered at the primary landmark location v_k, using the approach of [10]. Then, we rank the max-posteriors using quality measures similar to [17]. Good maps are peaked around the ground-truth, have small localization error, and have reliable detection performance. The best combination of B-landmarks for each v_k is included in the updated graph G^{upd}.

The output of this step is an enriched GM, which contains the primary landmarks and the best subsets of new B-landmarks that improves localization of the original primary landmarks.

V. Evaluation on Clinical Data

Our training and test set comprises respectively 63 and 20 patients (split randomly) from a database of lung cancer PET/CT cases. Diagnostic-quality CT scans of the torso for each patient were acquired using a Siemens Biograph 16 and resampled to 2 mm isotropic voxels. 22 anatomical landmarks were selected by an expert nuclear medicine physician, according to their utility as reference points for whole-body PET/CT oncology. Ground-truth landmark positions were annotated by a trainee medic. The expert subsequently validated the annotations using a rigorous protocol.

We apply the method to find B-landmark correspondences across the entire training set and perform a series of experi-

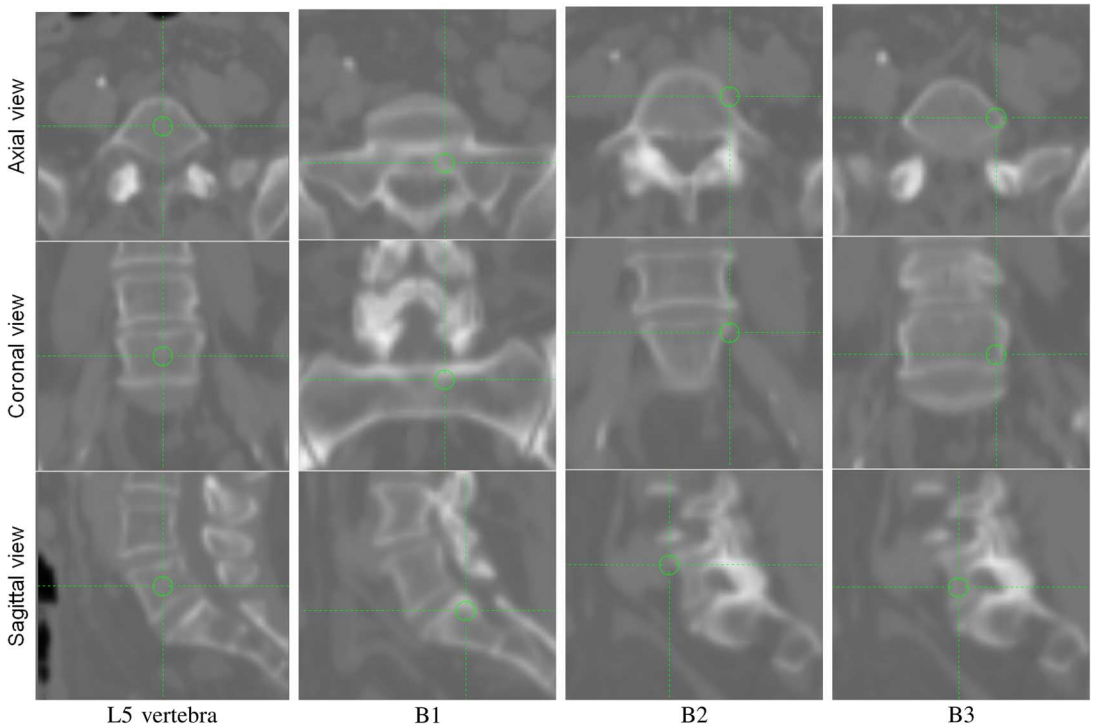


Fig. 5. Example B-landmark correspondences for L5 vertebra. In columns from left, we show the original hand-annotated landmark, and three B-landmarks (green cross-hairs). We show three orthogonal slices of the 3-D volume (rows).

ments. First, we evaluate the newly found B-landmarks both qualitatively and quantitatively, using landmark-based registration between all pairs of training images (Section V-A). Next, we evaluate the enriched GM which include B-landmarks, by matching them to unseen test-images (Section V-B). Finally, we investigate the method performance when occlusions are synthetically introduced to the test images (Section V-C).

A. Anatomic Correspondences Learned From the Training Data

First, we evaluate the quality of the new landmarks outside of the GM framework, to assess their utility as anatomical correspondences in their own right.

As an illustration of the typical B-landmark correspondences found by our proposed method, Fig. 5, shows examples of the top B-landmarks in one patient for vertebra L5. The top B-landmarks are in spatially well localized regions in 3-D. We also show that B-landmarks learned using our method in various body regions are qualitatively valid anatomical correspondences that are highly similar between different patients (Fig. 7).

Next, we quantitatively assess the utility of the learned B-landmarks. To do that, we perform pair-wise registrations of all pairs of training images of 63 different patients (source to target) using a B-splines landmark-based registration [16], [19], using the ground-truth landmark of interest and the top ranked pairs of B-landmarks (Section IV-C4). This is done in turn for each ground-truth landmark of interest. The number of B-landmarks varies from 2 to 10. In total, we perform 952 560 pair-wise registrations. We measure the quality of the alignment by computing the sum of squared differences (SSD) between the warped source and the target image across a local image region ($[35 \times 35 \times 35]$) centered on a landmark of interest. As a baseline, we compute the SSD measure using the same pair of images, manually aligned at the ground-truth landmark correspondence (i.e., a simple translation model, equivalent to uniformly spaced B-splines grid).

The results are shown in Fig. 6. The average improvement in SSD is 13% for 10 B-landmarks. The most improved are the skeletal structures such as the L5 vertebra, where our methods finds many reliable, spatially predictive B-landmarks in the local image neighborhood (up to 27% improvement in SSD). The least improved landmarks are those where our method could not find a sufficient number of reliable B-landmark correspondences across all images: for example the tip of the sternum, defined as the furthermost cartilaginous process connected to the sternum in the mid-line. However, this is a highly variable structure which involves complex patterns of branching, with ambiguous ground-truth placements and low repeatability even for an expert reader. For 18 out of 22 landmarks, an increase in number of B-Landmarks leads to improved SSD measure.

B. Landmark Localization on Unseen Test Images

In this section, we analyze landmark localization accuracy of the learned models on test images of unseen patients. We calculated the mean and the maximum localization error by summarizing the differences (i.e., Euclidean distance) between the detected location and the ground-truth of the original expert-annotated landmarks over the 20 test images.

We compared the enriched GM (M2) against the same GM without the additional B-landmarks (M1, Section III), the latter

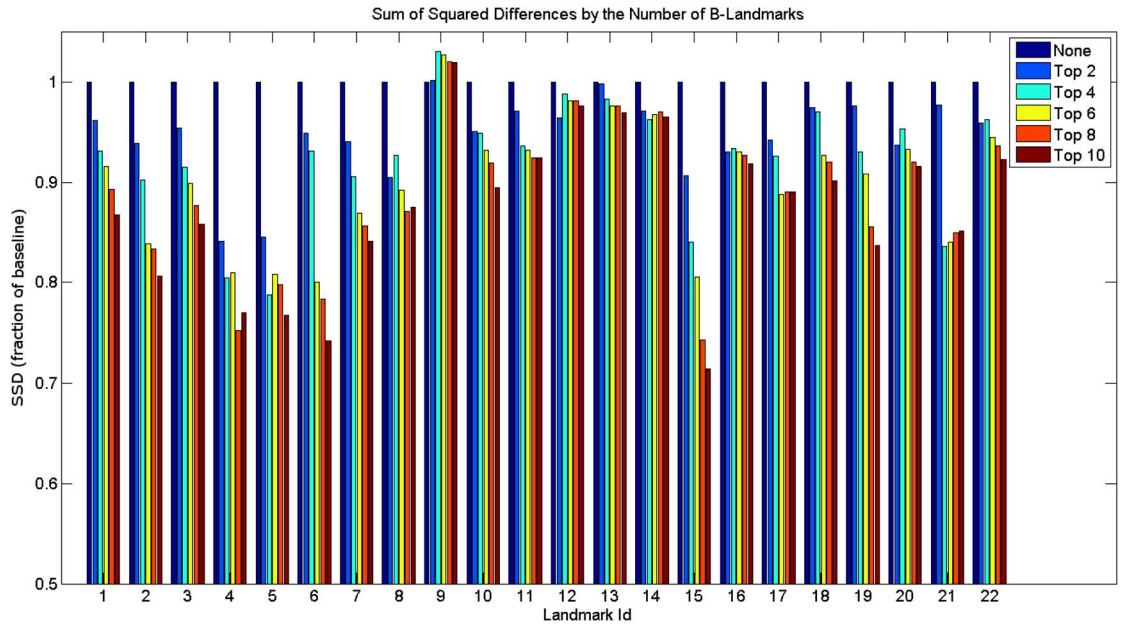


Fig. 6. Sum of squared differences by the number of B-landmarks. Columns (x-axis) show 20 landmarks, the landmark ID correspond to the numbering in Fig. 2. For each landmark we show the SSD measure between registered pairs of images using 2–10 B-landmarks (differently colored bars), expressed as the average improvement (fraction) against the baseline alignment without B-landmarks.

being representative of the state of the art. We also include results obtained using a deformable registration method (M0).[1] Since the proposed method mostly improves large misdetections, we examined in more detail the improvement on those parts which have a baseline model maximum error of at least four voxels, as it does not make sense to train a more complicated model if the baseline accuracy is already high. For completeness, over all 22 landmarks, the mean (2.18 voxels) and average maximum (6.53 voxels) error for the enriched GM improved by 15% and 26% respectively over the original GM. The results are shown in Table I (mean localization error) and Table II (maximum localization error). Compared to the baseline GM, C7 Vertebra has shown the most improved maximum as well as mean error (70% and 36%, respectively).

C. Robustness to Occlusions on Unseen Test Data

We also examined the contribution of the new B-landmarks to the localization of the primary landmarks, irrespective of how well the primary landmark is localized based on its own appearance model. This was done by simulating local occlusions of the primary landmarks, achieved by de-coupling the local appearance term and setting the entire feature map for each primary landmark v_k in turn to zero, for both M1 and M2. Although the main goal of this experiment was to decouple the relative contribution of the B-landmarks from the primary landmark appearance model, it also simulates a real-world occlusion situation, for example due to local image changes due to cancer or an implant.

The results are shown in Table III. On average, the localization accuracies of occluded landmarks improve by 26% with the enriched model and all but one landmark shows improved localization accuracy, compared to the baseline GM. These results provide strong evidence that the new landmarks are valid anatomical correspondences. Interestingly, both the baseline and the enriched GM with occluded data still significantly outperform the deformable registration without any occlusion (Table I).

[1]optimized for whole-body CT-CT registration; from a commercial oncology workstation. Each test image was registered to a reference template, selected by finding the nearest training image to the average anatomy

VI. Discussion

We have validated the contribution of our method in several experiments.

First, we have shown the anatomical relevance of the B-landmarks, both qualitatively and quantitatively using landmark-based registration of entire local image regions that improves the alignment by 15% compared to using the primary landmarks (i.e., translation) alone. Unfortunately, we do not have any ground-truth data for the automatically learned B-landmark correspondences and therefore we had to use indirect measures of registration accuracy. If such an expert-annotated dataset were available, it would be interesting to study the localization error of the B-landmarks directly.

Second, we evaluated the localization performance of the enriched GM that includes the primary as well as B-landmarks on unseen test images. The enriched model improves localization accuracy by 15% and 26% respectively over the original GM across all landmarks. The biggest improvement compared to the original GM for a specific primary landmark is 70% and 36% for the average maximum and mean localization error.

Finally, we investigated the relative contribution of the B-landmarks to the localization of the primary landmark, irrespective of how well localized the primary landmark by itself. This was done by a simulated occlusion experiment, which shows that the localization accuracy of occluded landmarks improve by 26% and all but one landmark shows improved localization accuracy, compared to only using the baseline

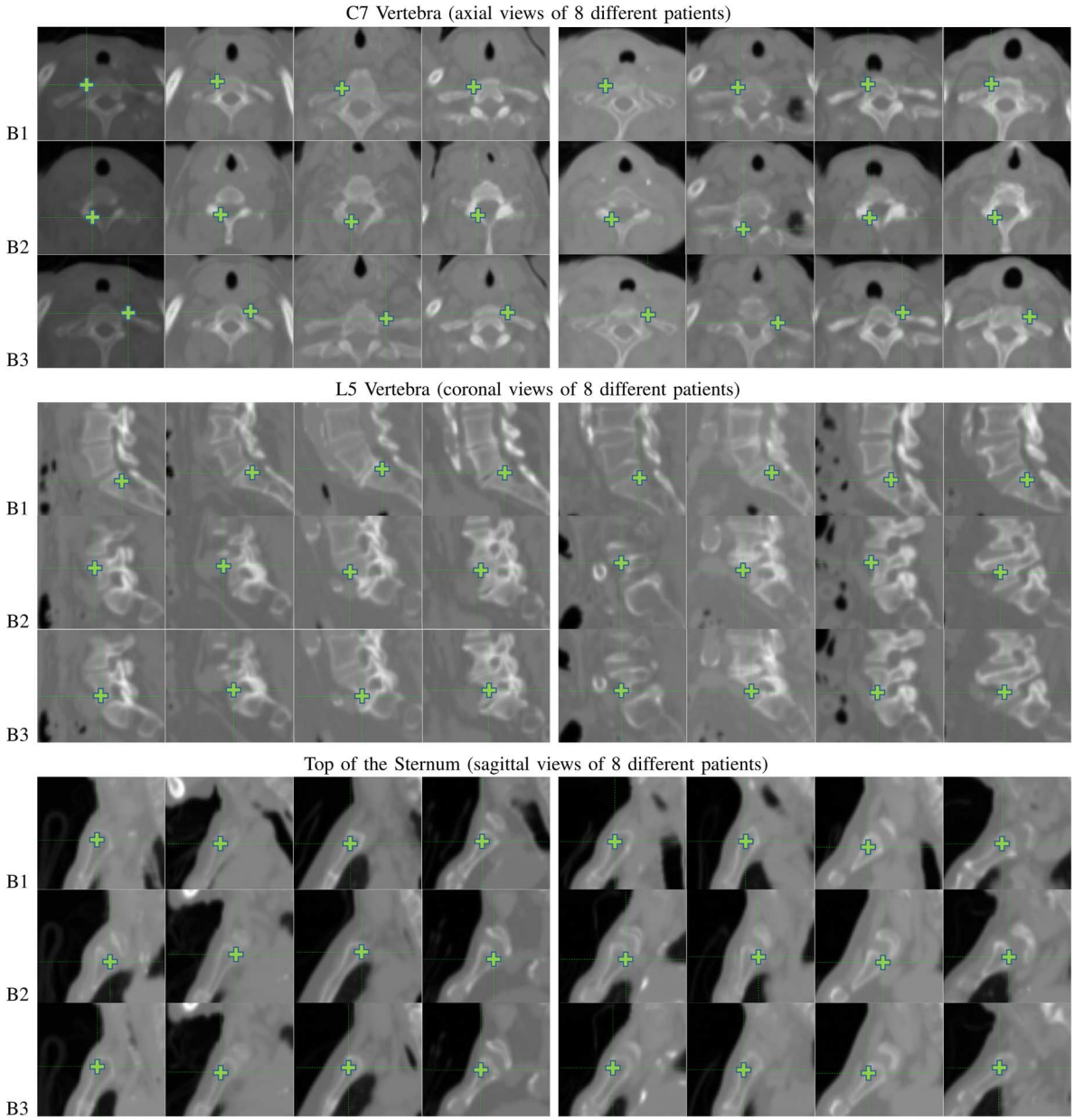


Fig. 7. Example B-landmark correspondences in three different body regions. For space reasons, we show the top three ranked B-landmarks (rows, B1–3; shown in green crosshair) in eight different randomly selected patients (columns) and one slice of the 3-D volume.

GM. The simulated occlusion experiment does not address the fact that while a single landmark is missing, the B-landmarks are unchanged. That may not be the case, for example, after an organ resection. Nevertheless, we clearly show the newly found B-landmarks are valid anatomical correspondences that improve localization of the primary landmark.

The MATLAB implementation of this method, run on our moderate desktop machine, takes on average 25 min for dense matching of the learned model to localize landmarks in an unknown test image. We do not use any search heuristics or efficiency optimizations (e.g., sub-sampling, coarse-to-fine matching, GPU-based patch convolutions . . .). Approximately 80% of the fitting time is spent on computing the appearance costs (unary data terms) and only 20% is spent on the actual model fitting. We expect to achieve a significant speed-up by replacing the existing appearance costs (based on convolutions of patch templates) with more efficient appearance detectors (based e.g., on Random Forests or Histograms of Oriented Gradients). Training time (one-off) is approximately 14 days to learn 220 candidate B-landmarks for 22 primary landmarks using a database of 63 training images. The training-time complexity scales linearly with the number of primary landmarks n and candidate B-landmarks N_b. The most expensive part of training is the voting procedure (Section IV-C3), that scales quadratically with the number of training images N_i and involves applying the NCC measure inside a local search region that contains h voxels ($h = s_R^3$). That results in a combined training-time complexity of $O(n\ N_b\ N_i^2\ h\ \log h)$.

Compared to other state of the art methods that enable localization of landmarks large field-of-view CT scans, typically based on Random Forests (e.g., [6]) and fully manually annotated datasets, our method achieves significantly higher

TABLE I
MEAN LANDMARK LOCALIZATION ERROR. TABLE SHOWS THE AVERAGE LOCALIZATION ERROR IN VOXELS FOR THREE DIFFERENT METHODS (COLUMNS): BASELINE DEFORMABLE REGISTRATION (M0), INITIAL GM (M1), AND THE ENRICHED GM (M2)

MeanRMS	M0	M1	**M2**	Impr.
C7 vertebra	3.57	1.96 ± 1.4	**1.28** ± 0.5	35%
Top of the sternum	5.54	1.57 ± 1.2	**1.35** ± 0.8	14%
Top right lung	5.19	1.88 ± 2.3	**1.29** ± 1.0	31%
Top left lung	5.32	2.13 ± 1.3	**2.01** ± 1.3	5%
Aortic arch	7.22	3.94 ± 2.3	**3.41** ± 2.2	14%
Lowest sternum (ribs)	9.76	4.91 ± 2.9	**4.44** ± 2.3	10%
Th12 vertebra	13.53	3.38 ± 5.7	**2.60** ± 4.6	23%
Top right kidney	15.58	1.95 ± 1.0	**1.88** ± 0.9	4%
Bottom right kidney	7.60	2.24 ± 2.0	**2.19** ± 1.7	3%
Top left kidney	15.25	2.42 ± 1.7	**2.27** ± 2.1	6%
L5 Vertebra	5.95	3.15 ± 4.0	**2.30** ± 3.1	27%
Right head of femur	3.88	1.39 ± 0.9	**1.36** ± 0.8	2%
Os coccygeum	10.03	3.18 ± 2.1	**2.51** ± 0.9	21%
Average	8.34	2.62	**2.22**	**15%**

TABLE II
MAXIMUM LANDMARK LOCALIZATION ERROR. TABLE SHOWS THE MAXIMUM LOCALIZATION ERROR IN VOXELS FOR THREE DIFFERENT METHODS (COLUMNS): BASELINE DEFORMABLE REGISTRATION (M0), INITIAL GM (M1), AND THE ENRICHED GM (M2)

MaxRMS	M0	M1	**M2**	Impr.
C7 vertebra	7.9	8.06	**2.45**	70%
Top of the sternum	11.2	4.00	**3.32**	17%
Top right lung	9.7	10.86	**3.46**	68%
Top left lung	9.3	5.48	**4.12**	25%
Aortic arch	11.3	11.70	**9.43**	19%
Lowest sternum (ribs)	27.0	13.34	**11.18**	16%
Th12 vertebra	30.5	20.12	**16.40**	19%
Top right kidney	30.3	4.58	**3.32**	28%
Bottom right kidney	17.9	8.37	**6.71**	20%
Top left kidney	33.4	14.80	**9.43**	36%
L5 Vertebra	**13.5**	15.23	15.23	0%
Right head of femur	7.2	4.47	**3.61**	19%
Os coccygeum	19.2	8.66	**7.28**	16%
Average	17.57	9.98	**7.38**	**27%**

localization accuracy and obviates the need for large training datasets. This comes at a cost of both increased run-time speed (minutes compared to seconds) and offline training speed (weeks compared to days) of our approach. Nevertheless, we note that we have not performed any speed optimization of our MATLAB implementation and significant improvements are certain to be achieved by more efficient software engineering (e.g., an optimized C++ implementation).

The current study has several limitations. Most importantly, we restrict the search only to point landmarks. Many regions of the body, especially highly variable regions such as the bladder, have few such correspondences. In future work, we intend to investigate higher order geometric features (e.g., symmetries of shapes) in such regions. The computational cost of the current implementation also does not permit search for B-landmarks over multiple scales, which may result in some cases in suboptimal correspondences. Moreover, by restricting the local range of new landmark search, we may miss good correspondences in regions which are not densely covered by the primary landmarks. Finally, we do not explicitly handle structures such as the vertebrae and individual ribs, which repeat at regular intervals. This could increase the complexity of correspondence

TABLE III
MEAN LANDMARK LOCALIZATION ERROR WITH OCCLUDED LANDMARK. TABLE SHOWS THE AVERAGE LOCALIZATION ERROR IN VOXELS FOR TWO DIFFERENT METHODS (COLUMNS): INITIAL GM (M1) AND THE ENRICHED GM (M2). LANDMARK IN QUESTION IS OCCLUDED, BY SETTING ITS FEATURE MAP TO ZERO

MeanRMS	M1	M2	Improvement
C7 vertebra	3.15	**2.27**	28%
Top of the sternum	4.41	**2.77**	37%
Top right lung	4.20	**3.80**	10%
Top left lung	**3.76**	3.85	-3%
Aortic arch	4.83	**4.49**	7%
Lowest sternum (ribs)	6.40	**3.51**	45%
Th12 vertebra	5.68	**3.75**	34%
Top right kidney	7.17	**4.41**	38%
Bottom right kidney	7.27	**4.40**	39%
Top left kidney	6.17	**5.09**	17%
L5 Vertebra	5.39	**3.06**	43%
Right head of femur	3.51	**2.61**	26%
Os coccygeum	5.40	**4.22**	22%
Average (occluded)	5.18	**3.71**	26%

computation if the search region were to be increased appropriately. Such landmarks would likely be down-ranked and simply not selected by the current algorithm, although they may indeed be useful clues for localization. Moreover, we do not assume significant rotation of the patient body. All patients in our training and testing database are aligned in the same way, which is the case for 95% of whole-body PET/CT oncology scans. Our method was shown to deal with limited misalignment and rotation observed in the training database due to the robust nature of the PS model. However, we do not expect it could deal with substantial rotation, e.g., completely different positioning of hands.

Our method has several components, of which only a small number substantially impact performance. Initialization involves many correspondences though the algorithm eventually rejects most of them. We have found that the specific form of the proposal distribution is relatively unimportant, and serves only to reduce computational cost.

In addition, we optimize the initially sampled models, since there are many poor correspondences. Models built from the initially fixed patch positions (i.e., randomly sampled, without ranking and optimization) were not useful, as the resulting appearance posterior probability maps are intrinsically poorly localized, with many false positive high probability landmark placements across the entire image. This is because we use a PCA representation of the patches in our GM. If the image patches used to build the appearance model are not exactly aligned, it results in general features and poor results. Inclusion of such noisy data terms in the GM cannot help improve the already relatively high localization accuracy of the primary landmarks. The optimization overcomes this problem of our representation by improving the alignment of the patches between patients. Two steps at this stage are particularly important. First, selecting an active subset of the training image patches is important, as we guide the model optimization towards intrinsically well localized landmarks. Second, the group-wise voting scheme where exemplars in the active set vote for the B-landmarks locations. That stage is important, as it forces agreement and can overcome initially noisy and misaligned matches and initially non-Gaussian distributions

of appearance models. Most of the improvement happens in the initial iterations of the optimization. The step of building updated models and removing the worst B-landmarks is not very important for accuracy, the main goal is to reduce computational effort. Of course, some landmarks that could become very strong might be removed in the first step thanks to the heuristics we are using. However, we solve that potential problem by redundancy, i.e., sampling many more candidates to start with. Finally, the building of enriched models is important, as we select the best subset of B-landmarks which will help to localize the primary landmark.

An in-depth quantitative investigation of the individual components of the algorithm, and extensive cross-validation, could not be performed due to computational demands of the current implementation, and is a direction for future work. It would also be interesting to investigate larger search regions, since several local regions in our training set, especially in the abdominal area, contain few consistent new landmarks that could be labelled even by a human expert. However, in order to perform this test, we first need to reduce the computational demands of the technique that we have developed.

VII. Conclusion

We have introduced the new concept of B-landmark correspondences, learned automatically from a database of whole-body CT scans and used to augment a GM trained with supervision. This weakly supervised approach may discover reliable spatial correlations that would be difficult to determine manually. Such B-landmarks can then serve to augment a GM to detect the clinician-annotated landmarks more reliably and accurately. We show our method improves both the accuracy and reliability of clinical landmark localization in whole-body CT scans over the state-of-the-art and improve the handling of occlusions. In future work, we plan to extend the method to discover many more new correspondences and to evaluate these clinically. Moreover, if sufficient correspondences can be found, we aim to combine these as constraints for deformable image registration as well as for dense group-wise registration.

Acknowledgment

V. Potesil would like to thank EPSRC, Siemens Molecular Imaging, Mirada Medical, and St. Johns College Oxford for support of his research. M. Brady would like to thank Cancer Research UK and EPSRC for support of his research via the Oxford Cancer Imaging Centre. The authors would like to thank the anonymous reviewers for their diligence and constructive criticisms of the paper that have improved considerably its clarity.

References

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Comput. Vis. ECCV*, 2010, pp. 168–181.

[2] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3-D human pose annotations," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 1365–1372.

[3] J. J. Corso, R. S. Alomari, and V. Chaudhary, "Lumbar disc localization and labeling with a probabilistic model on both pixel and object features," in *Proc. MICCAI*, 2008.

[4] D. Crandall and D. Huttenlocher, "Weakly supervised learning of part-based spatial models for visual object recognition," in *Comput. Vis. ECCV*, 2006, pp. 16–29.

[5] D. J. Crandall and D. P. Huttenlocher, "Composite models of objects and scenes for category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[6] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in *Med. Comput. Vis. Recognit. Tech. Appl. Med. Imag.*, 2011, pp. 106–117.

[7] R. Donner, E. Birngruber, H. Steiner, H. Bischof, and G. Langs, "Localization of 3-D anatomical structures using random forests and discrete optimization," in *Med. Comput. Vis. Recognit. Tech. Appl. Med. Imag.*, 2011, pp. 86–95.

[8] R. Donner, H. Wildenauer, H. Bischof, and G. Langs, "Weakly supervised group-wise model learning based on discrete optimization," in *Proc. MICCAI*, 2009, pp. 860–868.

[9] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Trans. Signal Process.*, vol. 49, no. 12, pp. 3136–3144, Dec. 2001.

[10] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," in *Int. J. Comput. Vis.*, 2005, vol. 61, no. 1, pp. 55–79.

[11] R. Fergus, P. Perona, and A. Zisserman, "A sparse object category model for efficient learning and exhaustive recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, pp. 380–387.

[12] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *Int. J. Comput. Vis.*, vol. 71, no. 3, pp. 273–303, 2007.

[13] J. A. Iglesias, E. Konukoglu, E. Montillo, Z. Tu, and A. Criminisi, "Combining generative and discriminative models for semantic segmentation of CT scans via active learning," in *Proc. 22nd Int. Conf. IPMI 2011*, Kloster Irsee, Germany, Jul. 2011, pp. 25–36.

[14] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, 2001.

[15] G. Langs, R. Donner, P. Peloschek, and H. Bischof, "Robust autonomous model learning from 2-D and 3-D data sets," in *Proc. MICCAI*, 2007, pp. 968–976.

[16] S. Lee, G. Wolberg, and S. Y. Shin, "Scattered data interpolation with multilevel b-splines," *IEEE Trans. Visualizat. Comput. Graph.*, vol. 3, no. 3, pp. 228–244, Jul./Sep. 1997.

[17] V. Potesil, T. Kadir, G. Platsch, and M. Brady, "Improved anatomical landmark localization in medical images using dense matching of graphical models," in *Proc. Br. Mach. Vis. Conf.*, F. Labrosse, R. Zwiggelaar, Y. Liu, and B. Tiddeman, Eds., Sep. 2010, pp. 37.1–37.10.

[18] V. Potesil, T. Kadir, G. Platsch, and M. Brady, "Personalization of pictorial structures for anatomical landmark localization," in *Proc. 22nd Int. Conf. IPMI 2011*, Kloster Irsee, Germany, Jul. 2011, pp. 333–345.

[19] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[20] S. Schmidt, J. Kappes, and M. Bergtholdt, "Spine detection and labeling using a parts-based graphical model," in *Proc. 20th Int. Conf., IPMI 2007*, Kerkrade, The Netherlands, 2007, pp. 122–133.

[21] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu, "Hierarchical parsing and semantic navigation of full body CT data," *Med. Imag.*, vol. 7259, p. 725902, 2009.

[22] M. Toews and T. Arbel, "A statistical parts-based model of anatomical variability," *IEEE Trans. Med. Imag.*, vol. 26, no. 4, pp. 497–508, Apr. 2007.

[23] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Comput. Vis.*, 2000, pp. 18–32.

[24] S. Woerz and K. Rohr, "Localization of anatomical point landmarks in 3-D medical images by fitting 3-D parametric intensity models," in *Med. Image Anal.*, Feb. 2006, vol. 10, no. 1, pp. 41–58.

[25] P. Zhang, S. Adeshina, and T. Cootes, "Automatic learning sparse correspondences for initialising groupwise registration," in *Proc. 13th Int. Conf. MICCAI*, Beijing, China, Sep. 2010, pp. 635–642.

[26] P. Zhang and T. Cootes, "Automatic part selection for groupwise registration," in *Information Processing in Medical Imaging*. New York: Springer, 2011, pp. 636–647.