Anatomical triangulation: from sparse landmarks to dense annotation of the skeleton in CT images

Marie Bieth[12]
marie.bieth@tum.de
Rene Donner[3]
Georg Langs[3]
Markus Schwaiger[1]
Bjoern Menze[2]

[1] Nuclear Medicin
Klinikum Rechts der Isar
Munich, Germany

[2] Department of Computer Science
TU Munich
Munich, Germany

[3] Department of Biomedical Imaging and Image-guided Therapy, CIR Lab
Medical University of Vienna
Vienna, Austria

Abstract

The automated annotation of bones that are visible in CT images of the skeleton is a challenging task which has, so far, been approached for only certain subregions of the skeleton, such as the spine or hip. In this paper, we propose a novel annotation algorithm for automatically identifying structures and substructures in the whole skeleton.

Our annotation algorithm makes use of recent advances in anatomical landmarks detection and is capable of generalising local information about landmarks to a dense label map of the full skeleton by anatomical triangulation. We follow a recognition approach that combines the use of distance-based features for measuring Euclidean and geodesic distances to a few given landmark locations, a parts-based model that is disambiguating anatomical substructures, and an iterative scheme for considering distances to the previously detected structures and, hence, to a dense set of anatomical reference points.

We propose an annotation protocol for 136 substructures of the skeleton and test our annotation algorithm on 18 CT images. On average, we obtain a Dice score of 90.54.

1 Introduction

Dense skeleton annotation is necessary for a variety of clinical and research applications, in particular in orthopaedics or oncology. In patients with primary or secondary tumours in the bones, for example, numerous lesions have to be mapped and analysed, and also to be reidentified and compared in case several scans are acquired while monitoring treatment. Evaluating those lesions and staging the patient is a very time-consuming task to perform manually, in particular in the context of clinical application, where dozens to hundreds of lesions have to be localised and evaluated repeatedly for each individual patients.

Different semi-automatic and automatic methods exist that perform skeleton annotation for the orthopaedic domain in MR or CT images. For example, in [9] a parts-based model is used and in [6] an improved Adaboost is used to annotate the spine. In [8], a specific shape model is matched to each vertebra. In [11], a pose shape estimation followed by a multi-pass Random Forest and a graph cut regularisation are used to segment cartilage in the knee. In [5], during the training, manual centroids of the vertebrae are converted to rough dense labels which are then used to train a Random Forest classifier for centroid localisation of the vertebrae. In [12], deformable template matching is used to annotate the ribs. Finally, in [4], an atlas is used to annotate the hip region. However, these methods are specific to a particular region of the skeleton and generalising them to the whole skeleton, with a much larger field of view, a higher variability of the anatomy, but also variations in the positioning of the patient remains an open task. It has – very much to our own surprise and in spite of the high demand for such approaches in the staging of bone tumour patients – not been sufficiently addressed yet.

In this article, we propose an approach for fully automatic dense annotation of substructures of the whole skeleton in CT images for staging bone tumour patients, where we want to localise anatomical substructures in order to estimate the local tumour load as visible from PET/CT images. Our approach builds upon the sparse annotation method in [3] and also shares some conceptual similarities with the approach of [5], but uses distance and not appearance-based features. Different from [5], it provides a dense labelling of the skeleton and not only the centroids of the structures and is not limited to the spine. As our approach builds solely on distance-based features that are measuring, for each pixel, the proximity to multiple anatomical landmarks in order to infer the relative positioning with respect to this landmark and, hence, the correct anatomical label, we refer to our approach as *anatomical triangulation*. Features that are measuring Euclidean and geodesic distances to a few automatically localised landmarks are used in a Random Forest that provides a first anatomical classification. A parts-based model then disambiguates the position of the centroid of each anatomical substructure. These centroids are then used by a second Random Forest classifier to provide a final labelling. The final set of dense labels is obtained by iterated anatomical triangulation.

Our approach is tailored to the needs of PET/CT data analysis in oncological staging. At present, because of the time required for manual annotation of PET/CT data, clinicians only use the two or three biggest metastasis to evaluate the evolution of the cancer or the response to a therapy, ignoring both smaller metastasis and information about the global distribution of lesions. Our method will, firstly, allow to process the PET/CT scan in a fully automatic fashion, extracting their location and different values of interest and following them over time. Secondly, it will also provide new means for disambiguating the correspondence problem over different scans that can arise from inexact registration or tumours merging or splitting, as mapping anatomical substructures of the skeleton allows us to calculate the average local tumour load. As both tasks require a sufficient high granularity of the anatomical atlas, we also propose a new annotation scheme that not only segments individual bones, but also substructures at a scale relevant for oncological staging throughout the skeleton: for example, each rib is divided into three substructures and each of the femurs and humerus in five substructures, while each foot is considered a single structure. This is a total of 136 regions.

In the following section, we detail our method for automatically annotating the skeleton. We then evaluate it in Section 3. We end with a conclusion in Section 4.

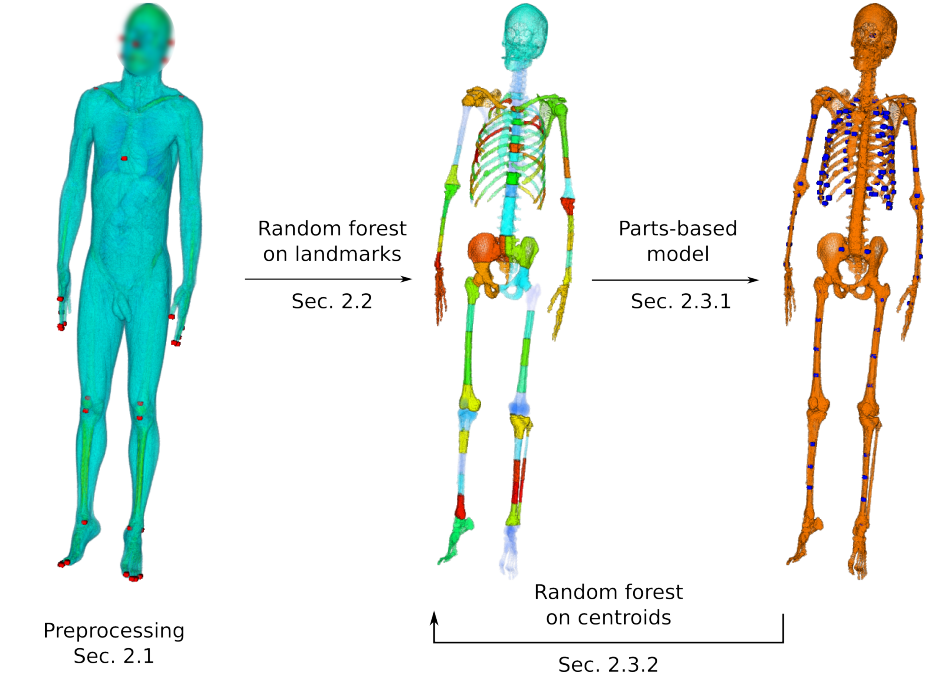

Figure 1: Description of the workflow. On the left, landmarks generated by Hough regression forest [3] are shown as part of the preprocessing. A first labelling (middle) is obtained using a Random Forest classifier (segments are assigned random colors for better visualisation). Centroids of each segment (right) are assigned anatomical labels by the parts-based model resolving ambiguous assignments. Centroids serve as new and evenly distributed landmarks for calculating distance features that serve as input to a another Random Forest classifier. The last two steps can be iterated.

2 Methods

Localising multiple anatomical regions and subregions of the skeleton is a multi-class problem, where each class corresponds to a bone or a segment of bone. In this section, we detail the four main steps of our method. The workflow is summarised in Figure 1. In a first step, we segment bones from the background and compute anatomical landmarks (Sec. 2.1). We then follow a recognition approach: a Random Forest classifier that uses distances to a few anatomical landmarks as input predicts a probabilistic label for each voxel (Sec. 2.2). A parts-based model is then used to disambiguate the position of the centroids of each bone subregion, resolving in particular an incorrect numbering of the ribs. As the newly identified substructures provide a much denser set of anatomical references than the initial landmarks, we now use the distances to the centroid of newly identified regions as input to another Random Forest classifier, predicting the final label (Sec. 2.3). We finally iterate local prediction and global correction using the parts-based model.

2.1 Preprocessing: bone segmentation and landmarks localisation

In a preprocessing step, the bones are segmented from the background in the CT image, and anatomical landmarks are computed.

2.1.1 Bone segmentation

We define bone segmentation as a two class-labelling problem: 0 denotes the background class, and 1 the bone class. To this end, we have adapted the method described in [7], modelling the intensities in the image as a mixture of two Gaussians. We then apply a conservative threshold to define a first segmentation of bones. As a second step, where morphological operations are used in [7], we choose to minimise the following energy on the grid-structured graph of the image:

$$E = \sum_x U(x, l(x)) + \sum_{x \sim y} B(x, y) \tag{1}$$

where $l \in [|0,1|]$ is the label, $\sim$ denotes the neighbouring relation, δ is the indicator function, μ and σ are the mean and the standard deviation of the intensity distribution of bones defined by the thresholding, I_x is the intensity of the image at voxel x, λ_1 and λ_2 are parameters, $d(x,y)$ is the Euclidean distance between the voxels x and y and

- $U(x,l) = \begin{cases} e^{-(I_x-\mu)/\sigma^2} \times \delta_{I_x<T} + \delta_{T<I_x<\lambda_1 T} + \lambda_2 \times \delta_{I_x>\lambda_1 T} & \text{if } l = 0 \\ 1 - U(x,0) & \text{else} \end{cases}$,
- $B(x,y) = e^{-(I_x-I_y)^2/\sigma^2} e^{-d(x,y)^2}$.

The unary potential term treats voxels differently according to their intensity: if the intensity is *much* higher than the threshold T (*much* being defined by the parameter λ_1), the cost of classifying them as background is high (defined by λ_2); if the intensity is *slightly* higher than T, the cost of classifying the voxel as background is unitary because the confidence that it belongs to bone is smaller; if the intensity is lower than T the cost of classifying the voxel as background decreases in a Gaussian fashion with its intensity. The binary potential term is similar to a bilateral filter and enforces spatial consistency for neighbours with similar intensities.

This is a binary min-cut problem and we solve it using the implementation of α-expansion provided in [2].

2.1.2 Anatomical landmarks

Landmarks are obtained in a preprocessing step by the method of Donner et al. [3]. It consists of a classifier for pre-filtering landmarks positions that are then refined through a Hough regression model and a parts-based model of the global landmark topology. N_l landmarks are spread over the whole body at meaningful places (at joints or tips of bones) that can be localised with high accuracy (a few millimetres) even in an automated fashion and each landmark is present in all the datasets (Figure 1: left).

2.2 Anatomical triangulation: dense probabilistic labelling using distance features

In the second step, we compute a probabilistic labelling of the L structures and substructures of the skeleton. This is an L-classes segmentation problem. We calculate the following features that include both local and long-range contextual information and are used as input to a Random Forest classifier:

1. the triangulation features: signed distance to each landmark in each of the three directions (sagittal, coronal and transversal, overall 171 features)
2. geodesic distances to selected landmarks along the skeleton at different scales (original scale, and subsampling to 1 cm and 2 cm resolutions, either by majority or maximum voting, overall 110 features),
3. mean image intensity and bone proportion in a window around the center voxel (8 features),
4. spatial position of the voxel within the field of view (3 features).

Since the calculation of geodesic distances is computationally costly, we do not compute them to all the landmarks but only to a selected number of landmarks that are well distributed over the body. In subsampling of the bone mask using majority voting, a voxel is classified as bone in the new mask if the majority of its parents are bones. In a subsampling using maximum voting, a voxel is classified as bone if at least one of its parents is. The geodesic distances are computed using the method described by Toivanen in [10].

Note that only bone voxels are presented to the Random Forest. We train a Random Forest with 20 trees for each triplet of subjects from the training data, each tree being trained with 10% of the voxels of the training set sampled randomly with replacement and evaluating 146 random subspaces at each split. Applied to a test set, it outputs a probability P of belonging to each structure for each voxel that we can convert to a labelling by majority voting.

2.3 Iterative triangulation using centroids

To account for local anatomy variability among subjects, local information can be extracted from the first labelling and used to refine the classification, in particular in regions where finer structures are present, e.g. the rib cage.

2.3.1 Centroid regularisation using a parts-based model

In the third step, we regularise the position of the centroid of each substructure using a parts-based model. This allows us to disambiguate anatomical substructures in regions where the label probability distribution outputted by the Random Forest is spread among several labels.

We construct a Conditional Random Field (CRF) with L nodes $\mathcal{N}$. L is the number of substructures in the skeleton. Each node l has C_l centroid candidates. The candidates $c(l)$ are obtained by thresholding the distribution probability computed by the Random Forest in the previous step for each label and taking the centroid of each connected component. The task of the CRF is to select the candidates that fit best the local and pairwise distributions of centroids learned from the training data. We have experimented with different configurations

of the model: fully connected, anatomy connected (only structures that anatomically touch each other are neighbours in the graph), and torso connected (the model is fully connected in the torso and anatomy connected elsewhere). The objective function to minimise is therefore:

$$\mathrm{E} = \sum_{l \in \mathcal{N}} \mathrm{U}(l, c(l)) + \sum_{l_1 \sim l_2} \mathrm{B}(l_1, c(l_1), l_2, c(l_2)) \tag{2}$$

The unary term U is based for each label l and candidate $c(l)$ on the distribution of label l modelled for each training subject s as a Gaussian $\mathcal{G}_{s,l}(c|\mu_l, \Sigma_l)$ and the aggregated probability of $CC_{c(l)}$ from the Random Forest, where $CC_{c(l)}$ is the connected component of the thresholded probability distribution of label l in the image to which $c(l)$ belongs, $P_{agg}(c(l)) = \sum_{x \in CC_{c(l)}} P(l|x)$:

$$\mathrm{U}(l, c(l)) = -\log(P_{agg}(c(l))) - \log(\max_s(\mathcal{G}_{s,l}(c(l)|\mu_l, \Sigma_l))) \tag{3}$$

The binary term B is based for each pair of labels l_1, l_2 on the distribution of the distances between centroids of theses classes modelled as a Gaussian $\mathcal{G}_{l_1,l_2}(x|\mu_{l_1,l_2}, \Sigma_{l_1,l_2})$:

$$\mathrm{B}(l_1, c(l_1), l_2, c(l_2)) = -\log(\mathcal{G}_{l_1,l_2}(c(l_1) - c(l_2)|\mu_{l_1,l_2}, \Sigma_{l_1,l_2})) \tag{4}$$

We solve the optimisation problem using the implementation of alpha-expansion-fusion provided in [1].

2.3.2 Anatomical triangulation from centroids

As a final step, a second Random Forest classifier refines the dense labelling. While the first classifier was only relying on distances to a few anatomical landmarks at joints or tips of bones, we can now use distances to the nearby anatomical regions and substructures that we have identified in the previous step. This is again an L-class segmentation problem and we follow the same classification approach, now replacing the distances to the landmarks in Sec. 2.2 by distances to the centroids of the 136 anatomical structures, leading to a total set of 419 features. The centroids play the same role as the landmarks in the first Random Forest, but they are more densely distributed and therefore capture more accurately the local variations in anatomy among subjects. Again, the probabilistic labelling returned by the classifier can either be converted to a hard labelling by majority voting or be used to iterate the anatomical triangulation with a better estimation of the centroids.

3 Experiments

We have evaluated our method on 18 whole body CT scans. The volumes have an average size of 512 × 512 × 1900 voxels with a voxel size of 1.3mm × 1.3mm × 1mm, and we subsampled them by a factor of two. Around 1% of all voxels are bone. 136 bones or bone segments have been manually labelled on each image. An example with shuffled labels can be seen in Figure 1 (middle). We had for each dataset 57 landmarks distributed at meaningful places over the whole body. Examples of landmarks can be seen in Figure 1 (left). All the tests were run using six-fold cross-validation. It means that every prediction was made by aggregating the results of the Random Forests trained on 15 subjects (3 subjects for each forest) and totalising a total of 100 trees.

We evaluate results using Dice scores (DS) weighted by the size of each structure, that we calculate as follows for a labelling S and a ground truth G:

$$\mathrm{DS}(\mathrm{S},\mathrm{G}) = \frac{100}{|\mathrm{G}|}\sum_{l\in\mathrm{G}}|\mathrm{G}==l|\frac{|\mathrm{S}==l\cap\mathrm{G}==l|}{|\mathrm{S}==l|+|\mathrm{G}==l|} \tag{5}$$

For visibility reasons, we show the results for 9 groups of substructures: whole body, skull, left and right arm, left and right leg, pelvis, spine, ribs.

3.1 Anatomical triangulation from landmarks for probabilistic labelling

In a first experiment, we measure the accuracy of the baseline Random Forest (Sec. 2.2). Figure 2 (red) shows the results in terms of DS after converting the output of the Random Forest into a labelling by majority voting. We obtain a mean overall DS of 88.77. The most difficult region to annotate is the rib cage, where the mean DS is 66.88. However, most of the bone metastasis in case of prostate cancer for example are located in the ribs and the spine. It is therefore important to refine the results in these regions.

Note that the feature importances computed from the forests show that the most important features are the distance and geodesic distance features. In particular, the distances to landmarks in the feet-head direction are more important than in the left-right direction, and both are more important than the distances to landmarks in the front-back direction. This was expected, because the depth in the body provides useful location information for only few bones (for example to differentiate between spine and sternum).

3.2 Iterative triangulation and parts-based model

Instead of converting the first probabilistic labelling to a hard labelling, we can apply the iteration scheme described in Sec. 2.3 to refine it.

Using a threshold of 0.1 on the probability distributions, the parts-based model typically had between 1 and 20 centroid candidates depending on the label. The Random Forest was trained using centroids computed from the ground truth data. As an alternative to the parts-based model for centroid regularisation, one can also compute the centroids directly from the segmentation. We show results for both methods in Figure 2 (green and blue).

We observe that iterating, both with and without regularisation through a parts-based model, improves the segmentation accuracy only for selected regions. It is however most effective in the regions that are most important for oncological staging and difficult to label: the rib cage and the spine. For example, the overall mean DS is 87.58 after three iterations with regularisation. However, in the ribs, the mean DS increases from 66.88 for the baseline to 72.67 after three iterations without regularisation and 76.81 after three iterations with regularisation respectively. The iterative triangulation with regularisation therefore performs better than without regularisation and achieves an improvement of 10 percentage points in the ribs region, which makes the method very useful for oncology applications. Figure 3 also shows for one subject that the iterative process improves the accuracy in the ribs and the pelvis regions.

Figure 4 shows that the fully connected and the torso connected models have similar results, which are better than the results from the anatomically connected model. This is

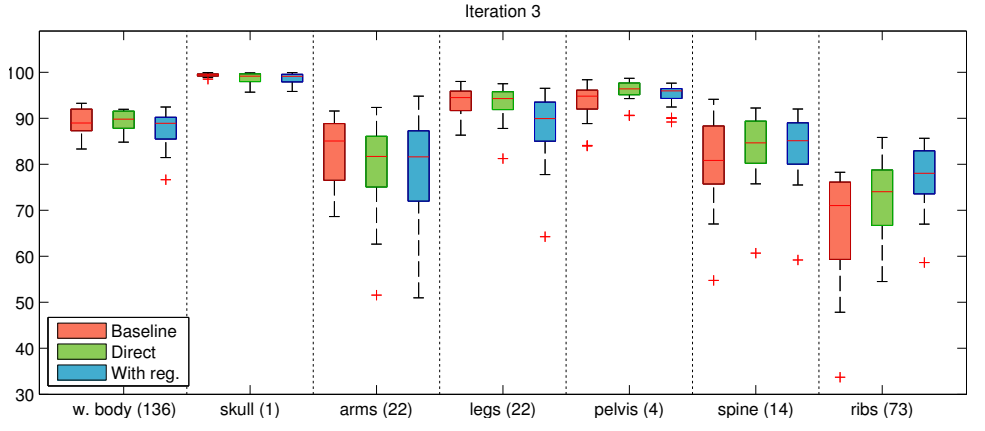

Figure 2: Weighted Dice scores for the segmentation of different groups of substructures by the baseline Random Forest (red), after 3 iterations without centroid regularisation (green), after three iterations with centroid regularisation (blue). The total number of substructures of each group is given in brackets. More detailed results are shown in the supplementary material.

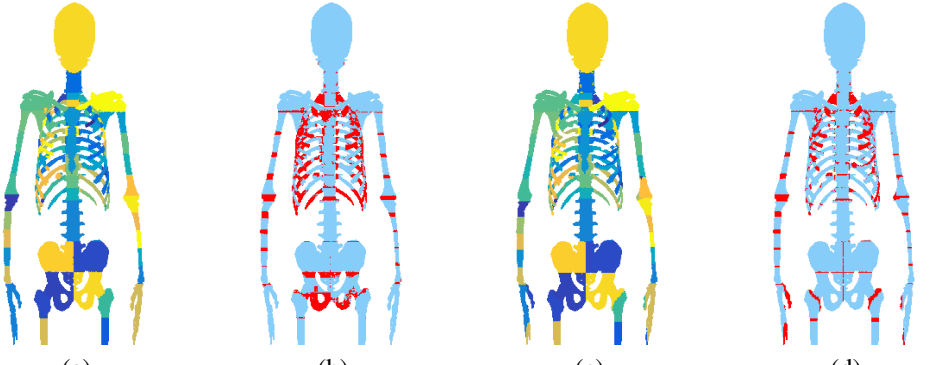

Figure 3: (a) Ground truth labelling. (b) Mislabelled regions (red) in the baseline labelling. (c) Labelling after three iterations with regularisation. (d) Mislabelled regions (red) in the labelling after three iterations with regularisation.

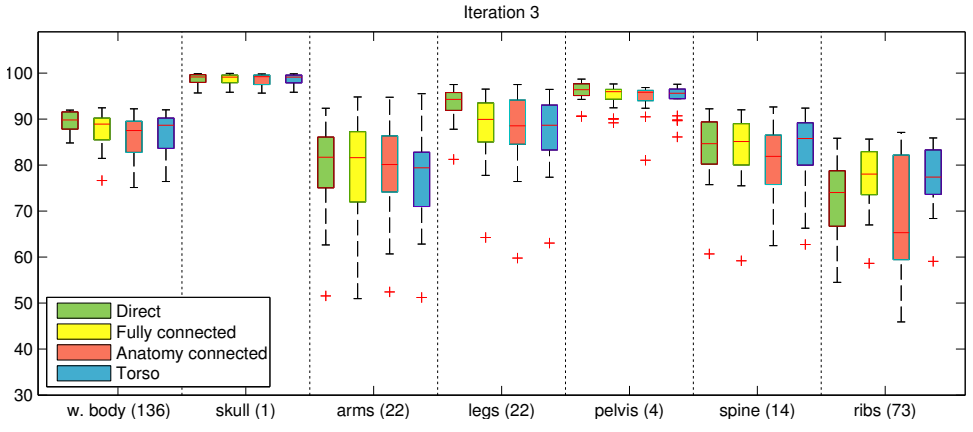

Figure 4: Weighted Dice scores for the segmentation of different groups of substructures with different configurations of the parts-based model after 3 iterations. "Direct" is iteration without regularisation. The three other configurations are detailled in Section 2.3.1.

because even structures that are far away carry global anatomical information, and the more connected models therefore recover better from any error made in the previous step

By selecting for each region the best suited method, we obtain a mean overall DS of 90.54.

We believe that a bigger training dataset and more precise ground truth labels will improve the results of our experiments. In future work, we also want to compare our results with the inter-rater labelling variation, and to evaluate the intra-patient variation for several scans at different time points.

4 Conclusion

We have developed a novel approach for localising substructures in the whole skeleton. To the best of our knowledge, only methods for certain subregions of the skeleton had been published so far. Our method achieves a mean overall weighted Dice Score of 90.54 over 18 subjects, which is sufficient, for example, for measuring local tumour load in the staging of bone tumour patients and to quantify local disease progression in longitudinal PET/CT scans. In future work, we want to improve the iterative scheme by using auto-context, and experiment with different fields of view.

References

[1] B. Andres, T. Beier, and J.H. Kappes. OpenGM: A C++ library for discrete graphical models. *ArXiv e-prints*, 2012.

[2] Andrew Delong, Anton Osokin, Hossam N Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. *International Journal of Computer Vision*, 96(1): 1–27, 2012.

[3] René Donner, Bjoern H Menze, Horst Bischof, and Georg Langs. Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization. *Medical Image Analysis*, 17(8):1304–1314, 2013.

[4] Jan Ehrhardt, Heinz Handels, Thomas Malina, Bernd Strathmann, Werner Plötz, and Siegfried J Pöppl. Atlas-based segmentation of bone structures to support the virtual planning of hip operations. *International Journal of Medical Informatics*, 64(2):439–447, 2001.

[5] Ben Glocker, Darko Zikic, Ender Konukoglu, David R Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In *Proc. MICCAI 2013*, pages 262–270. Springer, 2013.

[6] Szu-Hao Huang, Yi-Hong Chu, Shang-Hong Lai, and Carol L Novak. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. *IEEE Transactions on Medical Imaging*, 28(10):1595–1605, 2009.

[7] Yan Kang, Klaus Engelke, and Willi A Kalender. A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data. *IEEE Transactions on Medical Imaging*, 22(5):586–598, 2003.

[8] Tobias Klinder, Jörn Ostermann, Matthias Ehm, Astrid Franz, Reinhard Kneser, and Cristian Lorenz. Automated model-based vertebra detection, identification, and segmentation in CT images. *Medical Image Analysis*, 13(3):471–482, 2009.

[9] Stefan Schmidt, Jörg Kappes, Martin Bergtholdt, Vladimir Pekar, Sebastian Dries, Daniel Bystrov, and Christoph Schnörr. Spine detection and labeling using a parts-based graphical model. In *Information Processing in Medical Imaging*, pages 122–133. Springer, 2007.

[10] Pekka J Toivanen. New geodosic distance transforms for gray-scale images. *Pattern Recognition Letters*, 17(5):437–450, 1996.

[11] Quan Wang, Dijia Wu, Le Lu, Meizhu Liu, Kim L Boyer, and Shaohua Kevin Zhou. Semantic context forests for learning-based knee cartilage segmentation in 3D MR images. In *Medical Computer Vision. Large Data in Medical Imaging*, pages 105–115. Springer, 2014.

[12] Dijia Wu, David Liu, Zoltan Puskas, Chao Lu, Andreas Wimmer, Christian Tietjen, Grzegorz Soza, and Shaohua Kevin Zhou. A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images. In *Proc. CVPR 2012*, pages 980–987. IEEE, 2012.