Fast Anatomical Structure Localization Using Top-Down Image Patch Regression

René Donner[1,2,⋆], Bjoern H. Menze[3,4,5], Horst Bischof[2], and Georg Langs[1,3]

[1] Computational Image Analysis and Radiology Lab, Department of Radiology, Medical University Vienna, Austria
[2] Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[3] CSAIL, MIT, Cambridge MA, USA
[4] Asclepios Project, INRIA Sophia-Antipolis, France
[5] Computer Vision Laboratory, ETH Zurich, Switzerland
rene.donner@meduniwien.ac.at

Abstract. Fully automatic localization of anatomical structures in 2D and 3D radiological data sets is important in both computer aided diagnosis, and the rapid automatic processing of large amounts of data. We present a simple, accurate and fast approach with low computational complexity to find anatomical landmarks, based on a multi-scale regression codebook of informative image patches and encoded landmark contexts.

From a set of annotated training volumes the method captures the appearance of landmarks over several scales together with relative positions of neighboring landmarks and a spatial distribution model. During multi-scale search in a target volume, starting from the coarsest level, each landmark model predicts all landmark positions it has encoded, with the median of all predictions yielding the final prediction for each scale.

We present results on two challenging data sets (hand radiographs and hand CTs), where our method achieves comparable accuracy to the state of the art with substantially improved run-time.

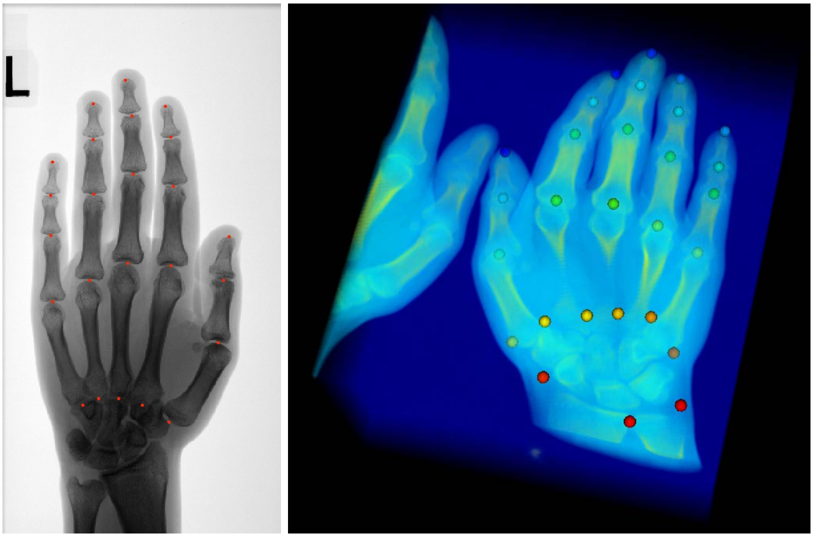
Keywords: Anatomical structure localization, nearest neighbor regression, image patch codebooks.

1 Introduction

The accurate localization of anatomical landmarks in medical imaging data is a challenging problem, due to rich variability and frequent ambiguity of their appearance. Among the reasons for the difficulties are noise (including local

[⋆] This work was partly supported by the European Union FP7 Project KHRESMOI (FP7-257528), by the Austrian National Bank grants BIOBONE (13468) and AORTAMOTION (13497) and the Austrian Sciences Fund grant PULMARCH (P 22578-B19).

B.H. Menze et al. (Eds.): MCV 2012, LNCS 7766, pp. 133–141, 2013.



(a) (b)

Fig. 1. Examples from the two data sets employed in this paper. a) Hand radiographs and b) high resolution hand CTs. The objective of the proposed method is to localize the depicted anatomical landmarks in an unseen target image or volume.

and global intensity changes), cluttered image data (overlapping structures in 2D projections, highly structured background in 3D organ segmentation), and anatomical structures that exhibit a high degree of similarity (e.g., fingers or vertebrae). We propose an algorithm that copes with these challenges and offers a general approach to accurately localize landmarks without initialization or subsequent refinement. The method constructs a multi-level regression codebook which associates image patches with the corresponding positions of anatomical landmarks depicted in the patch. During search the scale-pyramid is traversed, finding the most similar patch for each landmark using k-nearest neighbor search.

The localization of anatomical structures is crucial for several areas of medical imaging analysis: Segmentation approaches such as Level-Sets [4] and Appearance Models [3], typically require at least a coarse initial localization, while registration approaches can exploit spatial initialization to avoid local minima. The automatic localization of anatomical structures is fundamental for the field of Computer Aided Diagnosis [7] and for structuring image information in image retrieval, since it allows the algorithms to focus on target regions in the data and subsequently invoke more specialized analysis stages. Landmark localization can also be regarded as a form of semantic parsing [13] when point-wise rather than regional information is required.

State of the art. Several approaches to anatomical structure localization exist in recent literature. They mainly differ in the type of semantic representation that is obtained to describe the image data. We thus distinguish between approaches that either 1) indicate the *positions* of individual landmarks, 2) provide *bounding boxes* for entire organs, 3) result in *model parameters* which describe the position and shape of the object or 4) provide *voxel-wise labels* for different organs.

Localizing anatomical landmarks using the *positions* of selected interest points has been the objective of [8,1]. The methods learn interest point detectors on training data, estimate positions of landmark candidates in the target volume and finally disambiguate these candidates through a model matching step. Both methods rely on the classification of the entire volume. [9] reduces this computational burden by performing a low-resolution step and a refinement step using Hough regressors. Reducing the complexity by working on axial slices, [13] parse whole body CT data in a hierarchical fashion, but are concerned with finding larger organs. While substantially speeding up the localization this only works for objects which are rather large in respect to the overall volume size, since the objects have to be visible in at least one of the three central orthogonal slices. Using Random Forests for the localization of organs in thorax CTs through *bounding boxes* has been been proposed in [5]. An extension using Hough ferns was presented in [12] to predict the bounding boxes of multiple organs at once in full-body MR data. Relying on stochastic optimization instead of ensemble classification or regression, Marginal Space Learning [15] tries to find the parameters of a bounding box or a parametric and data-driven *shape model* [2] to localize and segment anatomical structures. This allows for fast localization, but instead of representing a global search algorithm, iterative approaches have to be used to cope with repetitive structures [10]. The task of assigning *voxel-wise labels* to segment entire organs or organ structures has been approached by [6] and [11] using Random Forest classification.

Contribution. We present a simple, fast method for the global, accurate localization of anatomical structures in 2D/3D data based on an appearance codebook, and location predictors that capture sub-configurations of a landmark set. It demonstrates that a top-down nearest neighbor matching strategy of image patches drastically reduces the number of required feature computations and yields localization results comparable to the state of the art.

Paper structure. The paper is structured as follows: Sec. 2.1 details the construction of the codebook, with the localization on a target volume described in Sec. 2.2. Sec. 3 introduces the experiments, with the results presented in Sec. 3.3. A discussion and an outlook can be found in Sec. 3.4 and Sec. 4.

2 Methods

The approach is divided into a training phase and a localization phase as shown in Fig. 2 and Fig. 3. During localization a multi-scale codebook of image patches and landmark positions is constructed, which is traversed during the localization phase to obtain increasingly accurate landmark estimates at each scale.

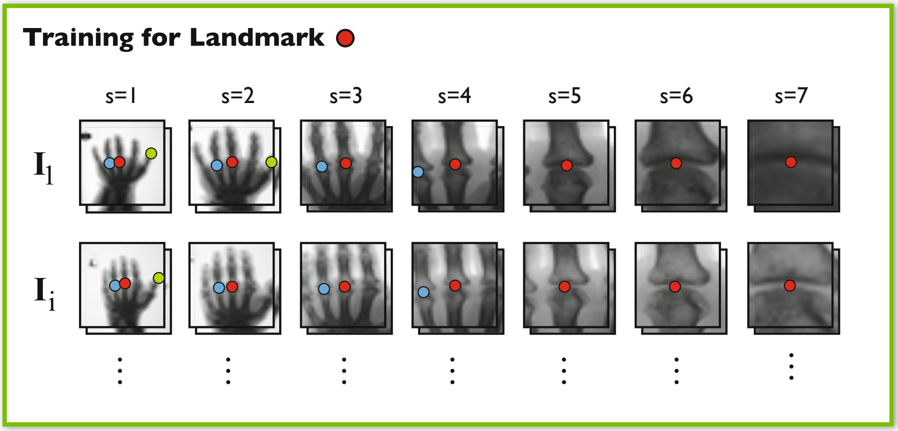


Fig. 2. Construction of the regression codebooks during training. For each landmark and scale patches at various offsets and the corresponding relative landmark positions are recorded, using all training images/volumes.

2.1 Training – Constructing the Landmark Regression Codebook

The training phase requires a set of N training images or volumes $\mathbf{I}_i$ with corresponding annotations. The annotations represent the coordinates $\mathbf{x}_x^i$ of the $x \in \{1, \ldots, L\}$ landmarks of the anatomical structure in question. Each landmark is present in each of the training volumes.

Codebook Construction to Connect Local Appearance and Landmark Information. Our aim is to build multi-scale regression codebooks $\mathcal{C}$ of image patches and corresponding relative landmark positions – one codebook per scale $s \in 1, \ldots, S$ and landmark x. The patches stored in the codebook are extracted around the landmarks with varying offsets and scaling, capturing the typical visual appearance around each landmark. For each patch the positions of all landmarks visible in the patch are recorded, relative to the patch's center. Each of the PN entries in the codebook $\mathcal{C}_{s,x}$ consists of the tuple $\langle \mathbf{P}^p, \mathbf{L}^p \rangle$ of the patch $\mathbf{P}^p$ and the corresponding relative $D \times L$ landmark coordinates $\mathbf{L}^p$ which are visible in the patch. $\mathbf{L}^p$ specifies the coordinates of the landmarks $x \in 1 \ldots L$ relative to the center of the given patch[1]. Landmarks which are outside of the patch are denoted as not visible.

The construction of the codebook proceeds as follows: At the top-most scale $s = 1$ each image or volume is represented by an an-isotropically downscaled miniature of size $m \times m \times m$ (similarly $m \times m$ for images). At each scale s the volume is considered to possess an edge length of $\sqrt{2}(s-1)m$. This re-sampling of the entire image is never actually computed, it simply forms the reference frame for each scale of the codebook generation.

[1] The necessary transformations between image coordinates and patch coordinates are omitted for clarity throughout the text.

At each scale s, patches $\mathbf{P}$ are extracted from the image or volume data using linear interpolation for each landmark x from all training volumes N. The patches are of size $m \times m \times m$, i. e. at scale $s = 1$ they correspond to the entire image, and for scales $s > 1$ the patches *zoom in* on the landmark, as illustrated in Fig. 2. Parts of patches which would be sampled from outside of the volume are set equal to the closest voxel on the volume's border. The gray values of each patch is normalized to zero mean and unit variance.

To explore the image information in the vicinity of a landmark the entries in the codebook $\mathcal{C}_{s,x}$ at a certain scale s and landmark x, are constructed by extracting several patches around the landmark with, empirically chosen, 7 offsets in the range of $[-6, 6]$ voxels for each dimension, along with scaling factors of $\{0.9, 1, 1.1\}$, resulting in $P = 1029$ patches for one landmark in one training volume at one scale ($P = 147$ for images). To considerably reduce the memory requirements and computational complexity for the codebook lookup, dimensionality reduction of each codebook is performed using PCA, retaining 90% of variance, resulting in PCA coefficients $\mathbf{P}_{PCA}$ and final codebook tuples $\langle \mathbf{P}^p_{PCA}, \mathbf{L}^p \rangle$. This training scheme results in the $S \times L$ regression codebooks $\mathcal{C}_{s,x}$.

Shape model to regularize the localization. To be able to regularize the intermediate solutions during the prediction phase, a model of the spatial distribution of the landmarks $\mathbf{s} = \langle \mathbf{x}^i_1, \ldots, \mathbf{x}^i_L \rangle$ in the training data is learned. We compute a point distribution model $\mathcal{S} = \langle \bar{\mathbf{s}}, \mathbf{S} \rangle$ using an eigen-decomposition of the covariance matrix of the training landmarks $\mathbf{x}_x$ as proposed in [2], retaining all eigenvectors and thus the entire shape variance observable in the training set, where the shapes $\mathbf{s}$ in the model can be constructed through a parameter vector $\mathbf{b}$ such that:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{S}\mathbf{b}$$

2.2 Localization – Regularized Top-Down Matching

Similar to the training phase the localization is performed in a multi-scale fashion, shown in Fig. 3. The $D \times L$ landmark localization matrix $\mathbf{L}^*_{s=1}$ is initialized with all landmarks starting at the center of the test volume $\mathbf{I}_{target}$. Starting with scale $s = 1$, a patch $\mathbf{P}^x$ for each landmark x is extracted (without additional offsets or scaling variations). The patch is normalized and projected onto the patch PCA model of $\mathcal{C}_{s,x}$, resulting in $\mathbf{P}^x_{PCA}$. The most similar patch p^{x*} in the codebook is found using euclidean nearest neighbor search – leading to the tuple $\langle \mathbf{P}^{x*}_{PCA}, \mathbf{L}^{x*}_p \rangle$ and thus the landmark coordinate predictions $\mathbf{L}^{x*}_p$ as estimated by landmark x. Repeating this codebook lookup for all landmarks yields the $D \times L \times L$ prediction tensor $\mathbf{M}_{\mathbf{d},\mathbf{i},\mathbf{j}}$ with position estimates from each landmark i to all landmarks that are visible in the same patch. The median over all predictions j which are not marked as not-visible yields the updated landmark localization matrix $\mathbf{L}^*_s$. This procedure is repeated through all scales, resulting in the final localization result $\mathbf{L}^*_S$.

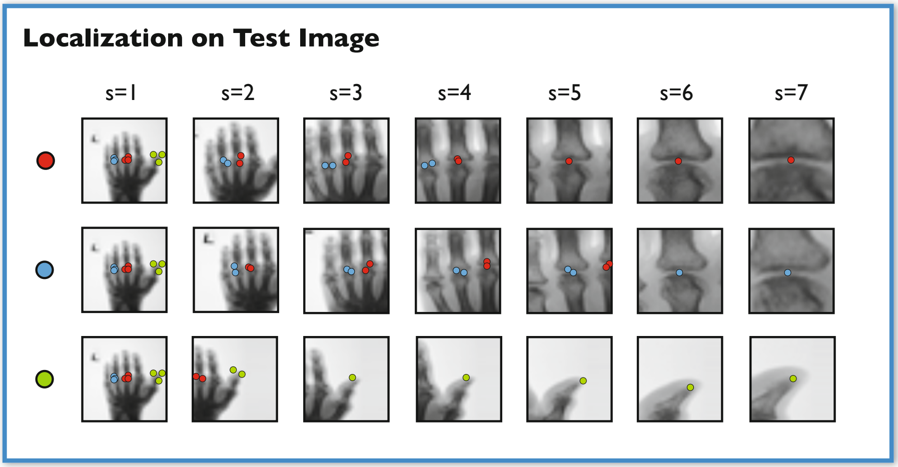


Fig. 3. The localization of three landmarks on a test image/volume descends the scale pyramid. At each level regression based on the image patch generates not only a position estimate for the primary landmak, but also for other landmarks visible in the patch. When progressing to a finer scale, for each landmark these estimates vote for the next estimate and center of the finer patch.

Shape regularization. The position estimates $\mathbf{L}_s^*$ are regularized by projecting them onto the shape PCA model $\mathcal{S}$ and reconstructing them again thereafter. This enforces landmark positions which can be modeled by a linear combination of the shapes observed in the training data. This regularization is performed for scales $s \leq S - 3$, to allow for landmark positions which can not be modeled though the shape model at scales $s > S - 3$.

3 Experiments

3.1 Data Sets

We evaluated the proposed approach on the two separate data sets shown in Fig. 1: 20 hand radiographs and 12 high resolution hand CTs.

Data set 1: Hand Radiographs $N = 20$ hand radiographs with an average size of 460×260 pixels with a resolution of 0.423mm/pixel were annotated with $L = 24$ landmarks. The landmarks include the five finger tips, as well as the distal interphalangeal (DIP), proximal interphalangeal (PIP), metacarpophalangeal (MCP) and carpometacarpal (CMC) joints for each finger.

Data set 2: Hand CTs The 3D hand CTs have a voxel size of $0.5mm \times 0.5mm \times 0.66mm$ resulting in an average size of $256 \times 384 \times 330$ voxels. They are annotated with the same 24 landmarks as the hand radiographs, with three additional landmarks placed around the carpus at the radiocarpal, radioulnar, and ulnocarpal joints, totaling in $L = 27$.

Table 1. Experimental results, localization accuracy in mm: Residual distances of the localization result to the ground truth annotation for the proposed method, in comparison with a state of the art approach

Residual in mm	MRF-based graph-matching			Proposed Patch-Regression Method		
	Median	Mean	Std	Median	Mean	Std
Hand Radiographs	0.80	0.99	0.82	0.63	0.77	0.64
Hand CTs	1.19	1.45	1.13	1.43	1.96	1.80

3.2 Setup

The experiments were run using four-fold cross validation, learning the landmark regression codebook on 75% of the N images / volumes and performing the localization on the remaining images / volumes. The main measure of interest for each landmark is the residual distance between the position of the predicted landmark position and the corresponding ground truth. The parameter settings are identical for the experiments on the two data sets, except for the size of the patches: 32×32 in the 2D case and $32 \times 32 \times 32$ for the 3D data. The results are compared with the recently proposed pre-filtered Hough regression Random forests [9], which in turn showed to outperform alternative approaches such as classification-based landmark candidate estimation with graph-based optimization [1] and classification + mean-shift based approaches [14].

3.3 Results

The results of the evaluation of the landmark localization are presented in Tab. 1, which shows the aggregated localization performance for the two data sets. The accuracy on the 2D radiograph data set is very high with a median residual of 0.63 mm and a mean/std of 0.77/0.64 mm. This result compares favorably with the results reported and methods tested on the same data in [9]. The result on the 3D hand CT data set show a median residual of 1.43 mm and a mean/std of 1.96/1.80 mm. It can be seen that despite a similar median residual, the proportion of localizations with higher error is slighty larger in this case. The run-times of the proposed approach were in the order of 0.6sec for the 2D data set and 4.5sec for the 3D data set on a single core of a 2009 Xeon MacPro. The method was entirely implemented in Matlab - we expect a potential speed-up by a factor of 10 to 100 through a more optimized implementation.

3.4 Discussion - Feature Computation Complexity

The main contribution of this work is the demonstration of a feature computation scheme which requires significantly less memory accesses then existing methods.

Voxel-wise classification / prediction approaches such as those proposed in [1,11] scale with the number of voxels, while pre-filtered Hough regression [9] reduces

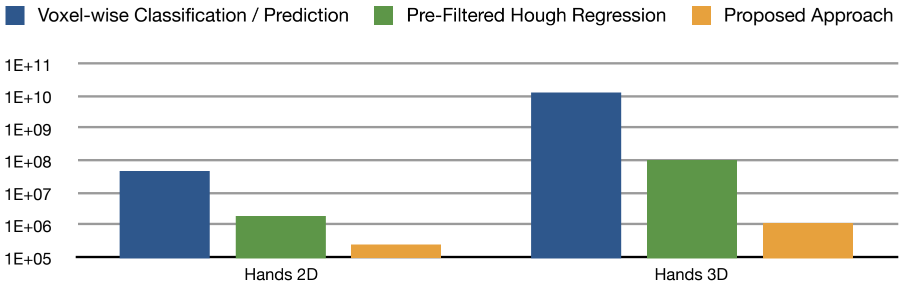


Fig. 4. Number of image/volume accesses necessary to compute the features required during the localization phase. Voxel-wise classification / prediction approaches [1,11] scale with the number of voxels, while pre-filtered Hough regression [9] works on strongly downsampled volumes. In constrast to this, the proposed approach is indepedent of the number of voxels and scales with the number of landmarks.

computational complexity by working on strongly down-sampled volumes. A typical number of 400 memory accesses to compute the classification for a single voxel was assumed in the calculation, corresponding to e. g. 20 individual features in an ensemble of 20 individual classifiers.

In contrast to this, the proposed approach is independent of the number of voxels and only depends on the number of landmarks, with $m \times m \times m$ voxels sampled for the patch at each landmark and scale. The proposed approach thus requires one to four orders of magnitude less image/volume accesses, allowing for fast localization even in unoptimized implementations or cheap commodity hardware.

4 Conclusion and Outlook

We present an approach for localizing complex, partly repetitive anatomical structures in 2D and 3D data. We demonstrate that a top-down nearest neighbor matching strategy of image patches drastically reduces the number of required feature computations and that the prediction of relative landmark positions using codebook regression is feasible.

The results on the two data sets clearly demonstrate the ability of the proposed approach to find the landmark positions in the target volume with accuracy comparable to the state of the art, with the consistent localization of detailed anatomical structures with a median residual of 1.7 to 2.7 pixels/voxels.

We consider the results to be very promising for such a simple method, and will focus on several topics in upcoming work: A detailed analysis of the parameters involved, namely the patch size and the perturbation strategy during codebook generation, as well as approximations of the nearest neighbor search through random subspaces.

References

1. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A Study of Parts-Based Object Class Detection Using Complete Graphs. IJCV 87(1-2), 93–117 (2010)
2. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graha, J.: Active Shape Models - Their Training and Application. CVIU 61(1), 38–59 (1995)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. TPAMI 23(6), 681–685 (2001)
4. Cremers, D., Rousson, M., Deriche, R.: A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape. IJCV 72(2), 195–215 (2007)
5. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in ct studies. In: Medical Computer Vision 2010: Recognition Techniques and Applications in Medical Imaging, MICCAI Workshop (2010)
6. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision Forests with Long-Range Spatial Context for Organ Localization in CT Volumes. In: Proc. of MICCAI Workshop on Probabilistic Models for Medical Image Analysis, MICCAI-PMMIA (2009)
7. Doi, K.: Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics 31, 198–211 (2007)
8. Donner, R., Birngruber, E., Steiner, H., Bischof, H., Langs, G.: Localization of 3D Anatomical Structures Using Random Forests and Discrete Optimization. In: Proc. MICCAI Workshop on Medical Computer Vision (2010)
9. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global Localization of 3D Anatomical Structures by Pre-filtered Hough Forests and Discrete Optimization. Medical Image Analysis (accepted, 2013)
10. Kelm, B.M., Zhou, S.K., Suehling, M., Zheng, Y., Wels, M., Comaniciu, D.: Detection of 3D Spinal Geometry Using Iterated Marginal Space Learning. In: Proc. MICCAI Workshop on Medical Computer Vision (2010)
11. Montillo, A., Shotton, J., Winn, J., Iglesias, J.E., Metaxas, D., Criminisi, A.: Entangled Decision Forests and Their Application for Semantic Segmentation of CT Images. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 184–196. Springer, Heidelberg (2011)
12. Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Möller, A.M., Nekolla, S., Navab, N.: Fast Multiple Organ Detection and Localization in Whole-Body MR Dixon Sequences. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 239–247. Springer, Heidelberg (2011)
13. Seifert, S., Barbu, A., Zhou, S., Liu, D., Feulner, J., Huber, M., Suehling, M., Cavallaro, A., Comaniciu, D.: Hierarchical Parsing and Semantic Navigation of Full Body CT Data. In: SPIE Medical Imaging (2009)
14. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moorea, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image. In: Proc. CVPR (2011)
15. Zheng, Y., Georgescu, B., Comaniciu, D.: Marginal Space Learning for Efficient Detection of 2D/3D Anatomical Structures in Medical Images. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 411–422. Springer, Heidelberg (2009)