Localization and Labeling of Posterior Ribs in Chest Radiographs Using a CRF-regularized FCN with Local Refinement

Alexander Oliver Mader[1,2,3(✉)], Jens von Berg[3], Alexander Fabritz[2], Cristian Lorenz[3], and Carsten Meyer[1,2,3]

[1] Institute of Computer Science, Kiel University of Applied Sciences, Kiel, Germany
alexander.o.mader@fh-kiel.de
[2] Department of Computer Science, Faculty of Engineering, Kiel University, Kiel, Germany
[3] Department of Digital Imaging, Philips Research Hamburg, Hamburg, Germany

Abstract. Localization and labeling of posterior ribs in radiographs is an important task and a prerequisite for, e.g., quality assessment, image registration, and automated diagnosis. In this paper, we propose an automatic, general approach for localizing spatially correlated landmarks using a fully convolutional network (FCN) regularized by a conditional random field (CRF) and apply it to rib localization. A reduced CRF state space in form of localization hypotheses (generated by the FCN) is used to make CRF inference feasible, potentially missing correct locations. Thus, we propose a second CRF inference step searching for additional locations. To this end, we introduce a novel "refine" label in the first inference step. For "refine"-labeled nodes, small subgraphs are extracted and a second inference is performed on all image pixels. The approach is thoroughly evaluated on 642 images of the public Indiana chest X-ray collection, achieving a landmark localization rate of 94.6%.

Keywords: Posterior ribs · Localization and labeling
Chest radiography · Fullyconvolutional network
Conditional random field

1 Introduction

Segmenting ribs in chest radiographs is used for the analysis of the lung parenchyma, as the overlaid ribs may obscure important findings. Rib shadows may be either excluded from the automatic analysis [1] or suppressed from the image [2] to minimize their impact. Ribs may also be used as an anatomical reference

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-00934-2_63) contains supplementary material, which is available to authorized users.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 562–570, 2018.
https://doi.org/10.1007/978-3-030-00934-2_63

to automatically locate findings like lung lesions or to establish correspondence between different images (e.g., in follow-up acquisition). Further on, counting the ribs in the lung field is a standard radiological procedure used to assure proper inhalation state in chest X-ray quality assessment. Unlike the previous two, these applications require the ribs not only to be segmented, but also to be anatomically labeled correctly.

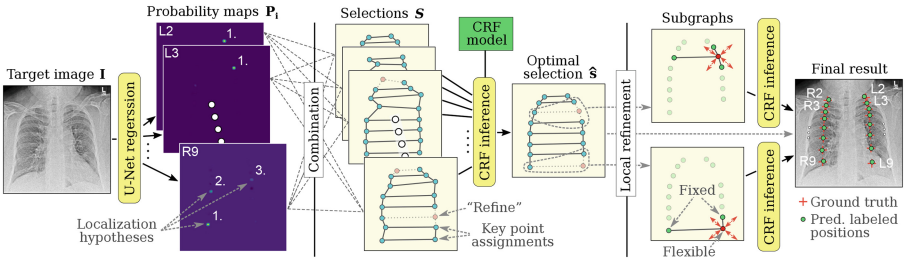
There is a number of methods described in literature to segment ribs in chest radiographs using either pixel classification [3], atlas registration [1], or a mixture of methods [4]. But there is no method described yet that does a robust anatomical labeling of posterior ribs. Even the atlas-based method did not use rib labels. Unlike CT where this task could be solved easier [5,6], the upper ribs are often overlaid in a chest radiograph (by, e.g., clavicles and other ribs) in a way that may prevent an algorithm from identifying and counting all the upper ribs properly. Also using the lung field as reference space appears not to be sufficient to unambiguously assign an anatomic label to a detected rib.

In this paper, we propose an automatic and general approach for localizing and labeling spatially correlated point landmarks. We apply our approach specifically to rib localization and labeling in posterior-anterior chest radiographs, by formulating the problem as finding a key point on each rib near the rib center. Unlike previous methods, it is a general approach and does not make or need any assumption about the task. Instead, all model parameters are automatically learned from annotated training data. First, the fully convolutional network (FCN) U-Net [7] is used to generate localization hypotheses. Then, a conditional random field (CRF) is applied to assess spatial information between landmarks. For feasibility, the CRF state space is combinatorically defined by the U-Net-generated localization hypotheses. Since the CRF has no means to select other than these localization hypotheses, we introduce a novel "refine" label. This allows the CRF to select this label instead of any of the localization hypotheses in case, e.g., none of them presents a viable option w.r.t. the CRF model. A second inference is performed for all "refine"-labeled nodes on a local subgraph over all image pixels rather than the set of localization hypotheses. Applying our approach to 642 images of the publicly available Indiana chest X-ray collection [8], we are able to localize and label 94.6% of the 16 individual landmarks correctly, corresponding to 83.0% fully correct cases. A median distance to the rib centerline of 0.7 mm is achieved.

2 Method

We formulate the problem as predicting $N = 16$ labeled key points for each posterior rib (2nd to 9th) close to its centerline (see Fig. 3a) in posterior-anterior chest radiographs. Our approach to solve this problem is split into three steps (compare Fig. 1): First, a FCN is used to regress heat maps to derive $n = 15$ localization hypotheses $\hat{\mathcal{X}}_i = \{\hat{\mathbf{x}}_{i,1}, \ldots, \hat{\mathbf{x}}_{i,n}\}$ for each key point $i \in [1\,..\,N]$ (Fig. 1a). Second, the unary information of the localizer is combined with binary information assessing spatial features between key point localization hypotheses. Both

are jointly modeled in a CRF with key points being the nodes and the corresponding localization hypotheses the respective states. An additional "refine" label is introduced for each node to be selected if no localization hypotheses is plausible (given the CRF model). CRF A* inference [9] is applied to find the best selection $\hat{\mathbf{S}} = (\hat{s}_1, \ldots, \hat{s}_N) \in [0\,..\,n]^N$ out of all possible selections $|\mathcal{S}| = (n+1)^N$. For each key point i, the inference either selects a localization hypothesis $\hat{\mathbf{x}}_{i,\hat{s}_i}$ if $\hat{s}_i \in [1\,..\,n]$, or the "refine" label if $\hat{s}_i = 0$ (Fig. 1b). In the third step, we derive positions for the "refine"-labeled nodes. We fix all nodes with predicted positions and optimize each "refine" node in a small subgraph over all image pixels rather than over the n localization hypotheses only (Fig. 1c). The following three subsections describe each step in detail.



(a) Key point localization (b) Spatial combination (c) Local refinement

Fig. 1. Schematic illustration of our three-step approach: (a) Generation of localization hypotheses using a U-Net, (b) followed by a CRF modelling spatial relations between key points and (c) a final local refinement based on a subgraph considering the whole image domain.

2.1 Generating Localization Hypotheses Using a U-Net

The goal of the first step is to predict candidate positions for each key point. The basic idea is to transform an image $\mathbf{I} : \mathbb{R}^2 \to \mathbb{R}$ into pseudo (not normalized) probability maps $\widetilde{\mathbf{P}}_i : \mathbb{R}^2 \to \mathbb{R}^+$ (for each target key point i) in which the location of the highest value $\hat{\mathbf{x}}_{i,1} = \arg\max_{\mathbf{x}} \widetilde{\mathbf{P}}_i(\mathbf{x})$ corresponds to the most likely predicted position of key point i.

To do so, we use U-Net [7], which has proven to deliver good results in the medical domain. However, its architecture is designed for pixel-wise segmentation, while we aim at localizing points to combine it with a CRF. Therefore, we directly formulate the problem as a pixel-wise heat map regression. This is achieved by dropping the soft-max classification and extending the final layer to N feature maps. Each feature map corresponds to a key point specific heat map that we want to regress. Assuming that high values in these heat maps correspond to likely positions of the searched key point, we can simply apply non-maximum suppression to each heat map to generate n localization hypotheses $\hat{\mathcal{X}}_i = \{\hat{\mathbf{x}}_{i,1}, \ldots, \hat{\mathbf{x}}_{i,n}\}$ (we use $n = 15$) for each key point i. This setup allows to generate localization hypotheses jointly for all key points in a single network.

Training. The modified U-Net is trained using stochastic gradient descent in the form of the Adam [10] algorithm using standard parameters with a mini-batch size of 8 and a sum-squared-error loss function. The target regression values are defined by a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}_i^*, 1/9\mathbf{1}r^2)$ with its mean located at the key point's true position $\mathbf{x}_i^*$. This provides high values close to the true position and very low values outside a small neighborhood ($r = 6$). As advocated in [7], we also perform data augmentation in form of elastic transformations of the training images, effectively increasing the training set size by the factor 11. We stop the training after 1000 epochs.

2.2 Selecting Reasonable Localization Hypotheses Using a CRF

To compensate for potentially incorrect first best localization hypotheses $\hat{\mathbf{x}}_{i,1}$ for arbitrary key points i, we use a CRF to model geometric relationships between key points. Each key point $i \in [1\,..\,N]$ is represented by a node in the graph with the corresponding localization hypotheses $\hat{\mathcal{X}}_i$ being the respective labels. We introduce an additional "refine" label for the CRF to choose during inference to compensate for cases where none of the localization hypotheses is plausible (and might negatively influence the selection of neighboring nodes). This "refine" label is used in our third step (Sect. 2.3) to still derive an accurate prediction in case CRF inference assigned the "refine" label to any node.

An energy-based formulation is applied where a low energy $E(\mathbf{S})$ of a selection $\mathbf{S} = (s_1, \ldots, s_N)$ implies a large posterior probability. For each node, either a localization hypothesis $\hat{\mathbf{x}}_{i,s_i}$ is assumed if $s_i > 0$, or the "refine" label if $s_i = 0$. The energy $E(\mathbf{S})$ of the CRF is parameterized by a set of J unary and binary potential functions $\Phi = \{\phi_1(\cdot), \ldots, \phi_J(\cdot)\}$ with corresponding weights $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_J)$ and missing potential values $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ for the "refine" label $s_i = 0$:

$$E(\mathbf{S}) = \sum_{j=1}^{J} \lambda_j \cdot \begin{cases} \beta_j & \text{if } s_i = 0 \text{ for any } i \in \text{Scope}(\phi_j) \\ \phi_j(\mathbf{S}) & \text{else} \end{cases} . \tag{1}$$

The inclusion of the missing energy values $\boldsymbol{\beta}$ is necessary, because it is not possible to compute potential values for the "refine" label $s_i = 0$. We use the same four potential types as introduced in [11]: For each key point i, an unary potential $\phi_i^{\text{loc}}(\mathbf{S})$ corresponding to the localizer's respective heat map value is introduced (see Fig. 2a). For each key point pair i and j, a distance 0potential $\phi_{i,j}^{\text{dist}}(\mathbf{S})$, an angle potential $\phi_{i,j}^{\text{ang}}(\mathbf{S})$ and a vector potential $\phi_{i,j}^{\text{vec}}(\mathbf{S})$ are used to model the geometric relations. The probability densities of estimated Gaussian, von Mises and multivariate Gaussian distributions, respectively, are used as potential values (see Fig. 2b). Finally, to efficiently find the best selection $\hat{\mathbf{S}} = \arg\min_{\mathbf{S} \in \mathcal{S}} E(\mathbf{S})$, exact inference in form of the A* algorithm [9] is applied.

Training. The weights $\mathbf{\Lambda}$ and the missing energies $\boldsymbol{\beta}$ are automatically learned in a training phase using a gradient descent scheme minimizing a max-margin hinge loss L over data $\mathcal{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$. The idea is to increase the energy gap between the "correct" selection $\mathbf{S}^+$ and the best (lowest energy) "incorrect" selection $\mathbf{S}^-$ until a certain margin m is satisfied. Let our loss function be defined as

$$L(\mathbf{\Lambda}, \boldsymbol{\beta}) = \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{S}_k^+, \mathbf{S}_k^-) \cdot \max\left(0, m + E(\mathbf{S}_k^+) - E(\mathbf{S}_k^-)\right) + \theta \cdot \|\mathbf{\Lambda}\|_1 \tag{2}$$

subject to $\lambda_j \geq 0$ for $j = 1, \ldots, J$, with

$$\delta(\mathbf{S}_k^+, \mathbf{S}_k^-) = \frac{1}{NR}\left(e(\mathbf{S}_k^-) - e(\mathbf{S}_k^+)\right) \in [0, 1] \tag{3}$$

weighting each training sample k w.r.t. the reduction in error (capped at $R = 100$)

$$e(\mathbf{S}) = \sum_{i=1}^{N} \begin{cases} 0 & \text{if " refine "}(s_i = 0) \text{ predicted and true,} \\ \min\left(R, \|\hat{\mathbf{x}}_{i,s_i} - \mathbf{x}_i^*\|_2\right) & \text{if "non-refine"}(s_i > 0) \text{ pred. and true,} \\ R & \text{else,} \end{cases} \tag{4}$$

going from the incorrect selection $\mathbf{S}_k^-$ to the correct selection $\mathbf{S}_k^+$. The "refine" label ($s_i = 0$) is assumed true, if none of the localization hypotheses ($s_i > 0$) is correct (the localization criterion is defined in Sect. 3). An additional θ-weighted L1 regularization term w.r.t. $\mathbf{\Lambda}$ was added to further accelerate the sparsification of terms. To optimize the loss function from Eq. (2), we apply again the Adam algorithm [10] starting from a grid-structured (Fig. 1b) graph. Once all CRF parameters are estimated, we remove unnecessary potentials where $\lambda_j = 0$, effectively optimizing the graph topology and improving the inference time while simultaneously improving the localization performance.

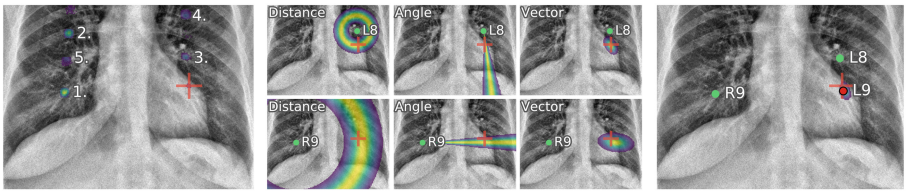
2.3 Going Beyond Potentially Incorrect Localization Hypotheses

After finding the optimal selection $\hat{\mathbf{S}}$ using CRF inference, we look at all key points $\{i \mid s_i = 0\}$ that have the "refine" label $s_i = 0$ instead of a localization hypothesis assigned. In order to assign those nodes a position, we start by fixing all nodes $\{i \mid s_i > 0\}$ with a properly selected localization hypothesis $\hat{\mathbf{x}}_{i,s_i}$. Then, we individually optimize each "refine"-labeled node by considering all connected binary potentials $\Phi_i = \{\phi_j \mid i \in \text{Scope}(\phi_j) \wedge \exists i' : (i' \in \text{Scope}(\phi_j) \wedge s_{i'} > 0)\}$ that are fully specified (except for the current node). Given that this second inference

$$\widetilde{\mathbf{x}}_i = \underset{\mathbf{x} \in \mathbf{I}}{\arg\min} \sum_{\phi'_j \in \Phi_i} \lambda_j \cdot \phi'_j(\mathbf{x}) \tag{5}$$

is performed on a very small subgraph, we can increase the search space to all possible pixel positions $\mathbf{x} \in \mathbf{I}$ for that node, rather than the set of localization

hypotheses, which would be intractable on the full problem. Note that we used some handwavy notation ϕ'_j to indicate that the (binary) potentials are computed by solely altering the position of key point i, since all others are known and fixed. By optimizing all "refine" nodes in decreasing order of the number of connected potentials $|\Phi_i|$, we can use previously refined positions in the next optimization in terms of more usable potentials. This also prevents the case that a node may not have any usable potential. How this approach can overcome the limitation of the fixed state space is illustrated on a test case in Fig. 2. Note that this final step does not require any training.



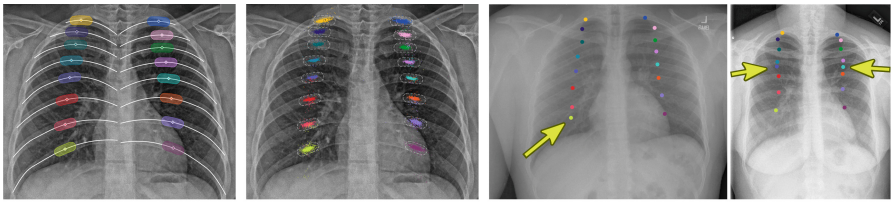
(a) Localizer hypotheses (b) Binary potentials (c) Combined result

Fig. 2. Illustration of the refinement process for L9 (true key point indicated as red cross). The "refine" label was chosen by the CRF inference for L9 because all n localizer hypotheses – shown enumerated in (a), heat map overlaid on cropped original image – yield large total energies $E(\mathbf{S})$. Utilizing the connected (b) binary potentials from L8 and R9, we are still able to (c) predict a correct position (red point) for L9 by evaluating over all image pixels instead of just the (here incorrect) localization hypotheses.

3 Experiments and Results

We evaluated our approach on 1000 consecutive images of various quality of the publicly available anonymized Indiana chest X-ray collection from the U.S. National Library of Medicine [8], downsampled to an isotropic resolution of 1×1mm/px. To derive key points for the unlabeled images for training and evaluation, we started by generating unlabeled posterior rib centerlines using an automatic approach based on [4]. The generated centerlines were then manually checked for quality, i.e., the line should be properly located within the rib, and correctly labeled, potentially discarding images. Following this approach, we generated labeled centerlines for the posterior ribs $\mathrm{L2}, \ldots, \mathrm{L9}, \mathrm{R2}, \ldots, \mathrm{R9}$ for 642 images. The middle points on the centerlines (w.r.t. the x-axis) have been selected as point annotations for each key point (except for the second and third rib where a factor of ± 0.3 and ± 0.4, respectively, instead of 0.5 was chosen). A corresponding predicted point is treated correct (localization and labeling criterion) if it is close to the annotated point (distance $\leq 15\,mm$) and very close to the centerline (distance $\leq 7.5\,mm$). This resembles the test whether the point lays on the rib while allowing for some translation along the rib. An example annotation as well as this localization criterion are depicted in Fig. 3a.

Note that correct point localizations also mean correct labels for the previously generated unlabeled centerlines, which effectively means the automatic generation of labeled centerlines as well.



(a) Annotations (b) Pred. positions (c) Typical errors

Fig. 3. (a) Illustration of the centerline and key point annotations and the resulting localization criterion, i.e., the area where a localization hypothesis is considered correct. Images in (a) and (b) are cropped and have been enhanced using adaptive histogram equalization. (b) Predicted positions in 642 test images visualized in a single image by registering the images using the true positions (affine and b-spline) and warping the predicted positions. Labels are shown color coded. (c) Typical errors involve an incorrect localization in the abdomen (first image) and chain errors caused by intermediate mistakes (second image).

We used a 3-fold cross-validation setup in our experiments, which provided us with 428 training images in each fold. Each training corpus was divided into three non-overlapping subsets $\mathcal{D}_{\text{pot}}$, $\mathcal{D}_{\text{weights}}$ and $\mathcal{D}_{\text{val}}$, containing 50%, 40%,10% randomly selected training images, respectively. $\mathcal{D}_{\text{pot}}$ was used to train the localizer (Sect. 2.1) and to estimate the statistics of the CRF potential functions (means, variances). $\mathcal{D}_{\text{weights}}$ was used to optimize the CRF potential weights $\mathbf{\Lambda}$ and to estimate the missing energies $\boldsymbol{\beta}$. The last subset $\mathcal{D}_{\text{val}}$ was used as validation corpus to select unknown meta parameters like learning rate and regularization parameter θ.

Applying our method, 94.6% of the key points were labeled correctly, corresponding to 83.0% of the images where all 16 key points were correct. The rates for individual key points and for the different steps in our chain are depicted in Fig. 4. First, we see that the CRF improves upon the plain U-Net results, especially in terms of the number of correct cases. Second, we see that the U-Net provides few good alternative localization hypotheses, which is apparent in a bad upper bound of the CRF of just 59.7% and justifies our third step. Third, we see that the additional CRF refinement step improves upon the CRF, where the percentage of correct cases increases dramatically from 57.3% to 83.0%. Fourth, the performance slightly decreases towards the lower ribs, which is probably caused by low contrast, higher variability and fewer meaningful surrounding structures (Fig. 3c). Errors in terms of Euclidean distance to the true position as well as distance to the centerline are listed in Table 1. The resulting median values of 2.8 mm and 0.7 mm, respectively, are in line with the visualization of the prediction results depicted in Fig. 3b. The overall average runtime of our approach

per case comes down to 36 ms U-Net + 61 ms CRF + 73 ms refinement = 170 ms running our unoptimized Python implementation on an Intel Xeon CPU E5-2650 in combination with an NVIDIA Titan X.

See supplement 1 for supporting content.

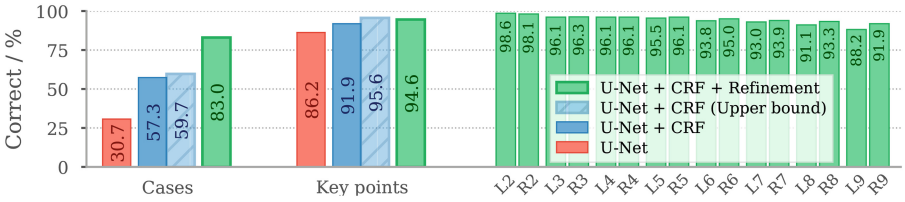


Fig. 4. Rates for correct cases (i.e., all 16 key points localized correctly) and correctly localized key points for our three steps in percent. Upper bound indicates the theoretical maximal performance of the CRF, caused by the limitation of the state space to the set of localization hypotheses. 100% corresponds to 642 cases and individual key points (L2-R9), and $642 \cdot 16 = 10272$ total key points.

Table 1. Median and mean Euclidean distance between true and predicted position (error) as well as median and mean distance between centerline and predicted position (line distance) for individual and all key points in mm.

Metric	L2	R2	L3	R3	L4	R4	L5	R5	L6	R6	L7	R7	L8	R8	L9	R9	All
Error / mm																	
Median	3.3	3.4	3.0	3.0	2.1	2.6	2.0	2.1	2.3	2.3	2.5	2.6	3.4	2.8	4.6	3.8	**2.8**
Mean	4.1	4.3	3.9	4.0	3.4	3.6	3.3	3.5	4.0	3.9	4.6	4.8	6.4	5.6	8.4	6.8	**4.7**
Line distance / mm																	
Median	0.8	0.9	0.8	0.9	0.6	0.6	0.5	0.5	0.5	0.5	0.6	0.6	0.9	0.8	1.2	1.1	**0.7**
Mean	1.2	1.2	1.6	1.6	1.7	1.4	1.6	1.5	2.1	1.9	2.6	2.7	3.9	3.3	4.8	4.0	**2.3**

4 Discussion and Conclusions

We presented a general approach for localization and labeling of spatially correlated key points and applied it to the task of rib localization. The state-of-the-art FCN U-Net has been used as localizer, which was regularized by a CRF incorporating spatial information between key points. The limitation of a reduced CRF state space in form of localization hypotheses imposed by the exact CRF inference in large graphs has been overcome with a novel "refine" node label. After a first CRF inference, a second inference is performed on small subgraphs formed by the marked "refine" nodes to refine the respective key points over all image pixels (rather than the set of localization hypotheses). Applying our approach to 624 images of the publicly available Indiana chest X-ray collection [8], we were

able to correctly localize and label 94.6% of the 16 key points, corresponding to 83.0% fully correct cases. The introduced refinement allowed for an increase of 25.7 percent points in fully correct cases over the global CRF alone. Note that this was achieved without domain-specific assumptions; all CRF model parameters were automatically learned from annotated training data. Our approach is thus directly applicable to other anatomical localization tasks.

In future work, we are going to increase the rotation and scaling invariance by incorporating ternary potentials over the commonly used binary ones, with tractability being the main challenge.

Acknowledgements. This work has been financially supported by the Federal Ministry of Education and Research under the grant 03FH013IX5. The liability for the content of this work lies with the authors.

References

1. Candemir, S., et al.: Atlas-based rib-bone detection in chest X-rays. CMIG **51**, 32–39 (2016)
2. von Berg, J., et al.: A novel bone suppression method that improves lung nodule detection. IJCARS **11**(4), 641–655 (2016)
3. Loog, M., Ginneken, B.: Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. T-MI **25**(5), 602–611 (2006)
4. von Berg, J., et al.: Decomposing the bony thorax in X-ray images. In: ISBI, pp. 1068–1071 (2016)
5. Staal, J., Ginneken, B., Viergever, M.: Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data. MIA **11**(1), 35–46 (2007)
6. Wu, D., et al.: A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images. In: CVPR, pp. 980–987. IEEE (2012)
7. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. U.S. National Library of Medicine (NLM): Open-i Open Access Biomedical Image Search Engine (2017). https://openi.nlm.nih.gov. Accessed 14 Feb 2018
9. Bergtholdt, M., Kappes, J.H., Schnörr, C.: Learning of graphical models and efficient inference for object class recognition. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.) DAGM 2006. LNCS, vol. 4174, pp. 273–283. Springer, Heidelberg (2006). https://doi.org/10.1007/11861898_28
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2014)
11. Mader, A.O., et al.: Detection and localization of landmarks in the lower extremities using an automatically learned conditional random field. In: GRAIL (2017)