From Local to Global Random Regression Forests: Exploring Anatomical Landmark Localization

Darko Štern[1], Thomas Ebner[2], and Martin Urschler[1,2,3(✉)]

[1] Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
martin.urschler@cfi.lbf.ac.at
[2] Institute for Computer Graphics and Vision,
Graz University of Technology, Graz, Austria
[3] BioTechMed-Graz, Graz, Austria

Abstract. State of the art anatomical landmark localization algorithms pair local Random Forest (RF) detection with disambiguation of locally similar structures by including high level knowledge about relative landmark locations. In this work we pursue the question, how much high-level knowledge is needed in addition to a single landmark localization RF to implicitly model the global configuration of multiple, potentially ambiguous landmarks. We further propose a novel RF localization algorithm that distinguishes locally similar structures by automatically identifying them, exploring the back-projection of the response from accurate local RF predictions. In our experiments we show that this approach achieves competitive results in single and multi-landmark localization when applied to 2D hand radiographic and 3D teeth MRI data sets. Additionally, when combined with a simple Markov Random Field model, we are able to outperform state of the art methods.

1 Introduction

Automatic localization of anatomical structures consisting of potentially ambiguous (i.e. locally similar) landmarks is a crucial step in medical image analysis applications like registration or segmentation. Lindner et al. [5] propose a state of the art localization algorithm, which is composed of a sophisticated statistical shape model (SSM) that locally detects landmark candidates by three step optimization over a random forest (RF) response function. Similarly, Donner et al. [2] use locally restricted classification RFs to generate landmark candidates, followed by a Markov Random Field (MRF) optimizing their configuration. Thus, in both approaches good RF localization accuracy is paired with disambiguation of landmarks by including high-level knowledge about their relative location. A different concept for localizing anatomical structures is from Criminisi et al. [1],

D. Štern—This work was supported by the province of Styria (HTI:Tech_for_Med ABT08-22-T-7/2013-13) and the Austrian Science Fund (FWF): P 28078-N33.

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 221–229, 2016.
DOI: 10.1007/978-3-319-46723-8_26

suggesting that the RF framework itself is able to learn global structure configuration. This was achieved with random regression forests (RRF) using arbitrary long range features and allowing pixels from all over the training image to globally vote for anatomical structures. Although roughly capturing global structure configuration, their long range voting is inaccurate when pose variations are present, which led to extending this concept with a graphical model [4]. Ebner et al. [3] adapted the work of [1] for multiple landmark localization without the need for an additional model and improved it by introducing a weighting of voting range at testing time and by adding a second RRF stage restricted to the local area estimated by the global RRF. Despite putting more trust into the surroundings of a landmark, their results crucially depend on empirically tuned parameters defining the restricted area according to first stage estimation.

In this work we pursue the question, how much high-level knowledge is needed in addition to a single landmark localization RRF to implicitly model the global configuration of multiple, potentially ambiguous landmarks [6]. Investigating different RRF architectures, we propose a novel single landmark localization RRF algorithm, robust to ambiguous, locally similar structures. When extended with a simple MRF model, our RRF outperforms the current state of the art method of Lindner et al. [5] on a challenging multi-landmark 2D hand radiographs data set, while at the same time performing best in localizing single wisdom teeth landmarks from 3D head MRI.

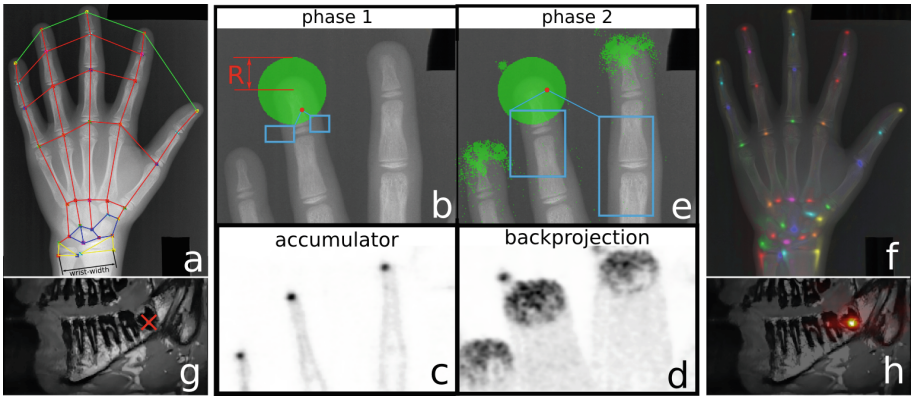

Fig. 1. Overview of our RRF based localization strategy. (a) 37 anatomical landmarks in 2D hand X-ray images and differently colored MRF configurations. (b) In phase 1, RRF is trained locally on an area surrounding a landmark (radius R) with short range features, resulting in accurate but ambiguous landmark predictions (c). (d) Backprojection is applied to select pixels for training the RRF in phase 2 with larger feature range (e). (f) Estimated landmarks by accumulating predictions of pixels in local neighbourhood. (g, h) One of two independently predicted wisdom teeth from 3D MRI.

2 Method

Although being constrained by all surrounding objects, the location of an anatomical landmark is most accurately defined by its neighboring structures. While increasing the feature range leads to more surrounding objects being seen for defining a landmark, enlarging the area from which training pixels are drawn leads to the surrounding objects being able to participate in voting for a landmark location. We explore these observations and investigate the influence of different feature and voting ranges, by proposing several RRF strategies for single landmark localization. Following the ideas of Lindner et al. [5] and Donner et al. [2], in the first phase of the proposed RRF architectures, the local surroundings of a landmark are accurately defined. The second RRF phase establishes different algorithm variants by exploring distinct feature and voting ranges to discriminate ambiguous, locally similar structures. In order to maintain the accuracy achieved during the first RRF phase, locations outside of a landmark's local vicinity are recognized and banned from estimating the landmark location.

2.1 Training the RRF

We independently train an RRF for each anatomical landmark. Similar to [1,3], at each node of the T trees of a forest, the set of pixels S_n reaching node n is pushed to left ($S_{n,L}$) or right ($S_{n,R}$) child node according to the splitting decision made by thresholding a feature response for each pixel. Feature responses are calculated as differences between mean image intensity of two rectangles with maximal size s and maximal offset o relative to a pixel position $\boldsymbol{v}_i$; $i \in S_n$. Each node stores a feature and threshold selected from a pool of N_F randomly generated features and N_T thresholds, maximizing the objective function I:

$$I = \sum_{i \in S_n} \left\| \boldsymbol{d}_i - \overline{\boldsymbol{d}}(S_n) \right\|^2 - \sum_{c \in \{L,R\}} \sum_{i \in S_{n,c}} \left\| \boldsymbol{d}_i - \overline{\boldsymbol{d}}(S_{n,c}) \right\|^2 . \tag{1}$$

For pixel set S, $\boldsymbol{d}_i$ is the i-th voting vector, defined as the vector between landmark position $\boldsymbol{l}$ and pixel position $\boldsymbol{v}_i$, while $\overline{\boldsymbol{d}}(S)$ is the mean voting vector of pixels in S. For later testing, we store at each leaf node l the mean value of relative voting vectors $\overline{\boldsymbol{d}}_l$ of all pixels reaching l.

First training phase: Based on a set of pixels S^I, selected from the training images at the location inside a circle of radius R centered at the landmark position, the RRF is first trained locally with features whose rectangles have maximal size in each direction s^I and maximal offset o^I, see Fig. 1b. Training of this phase is finished when a maximal depth D^I is reached.

Second training phase: Here, our novel algorithm variants are designed by implementing different strategies how to deal with feature ranges and selection of the area from which pixels are drawn during training. By pursuing the same local strategy as in the first phase for continuing training of the trees up to a maximal depth D^{II}, we establish the *localRRF* similar to the RF part in [2,5].

If we continue training to depth D^{II} with a restriction to pixels S^I but additionally allow long range features with maximal offset $o^{II}>o^I$ and maximal size $s^{II}>s^I$, we get *fAdaptRRF*. Another way of introducing long range features, but still keeping the same set of pixels S^I, was proposed for segmentation in Peter et al. [7]. They optimize for each forest node the feature size and offset instead of the traditional greedy RF node training strategy. For later comparison, we have adapted the strategy from [7] for our localization task by training trees from root node to a maximal depth D^{II} using this optimization. We denote it as *PeterRRF*. Finally, we propose two strategies where feature range *and* area from which to select pixels are increased in the second training phase. By continuing training to depth D^{II}, allowing in the second phase large scale features (o^{II}, s^{II}) and simultaneously extending the training pixels (set of pixels S^{II}) to the whole image, we get the *fpAdaptRRF*. Here S^{II} is determined by randomly sampling from pixels uniformly distributed in the image. The second strategy uses a different set of pixels S^{II}, selected according to back-projection images computed from the first training phase. This concept is a main contribution of our work, therefore the next paragraph describes it in more detail.

2.2 Pixel Selection by Back-Projection Images

In the second training phase, pixels S^{II} from locally similar structures are explicitly introduced, since they provide information that may help in disambiguation. We automatically identify similar structures by applying the RRF from the first phase on all training images in a testing step as described in Sect. 2.3. Thus, pixels from the area surrounding the landmark as well as pixels with locally similar appearance to the landmark end up in the first phase RRFs terminal nodes, since the newly introduced pixels are pushed through the first phase trees. The obtained accumulators show a high response on structures with a similar appearance compared to the landmark's local appearance (see Fig. 1c). To identify pixels voting for a high response, we calculate for each accumulator a back-projection image (see Fig. 1d), obtained by summing for each pixel $\boldsymbol{v}$ all accumulator values at the target voting positions $\boldsymbol{v} + \boldsymbol{d}_l$ of all trees. We finalize our *backProjRRF* strategy by selecting for each tree training pixels S^{II} as N_{px} randomly sampled pixels according to a probability proportional to the back-projection image (see Fig. 1e).

2.3 Testing the RRF

During testing, all pixels of a previously unseen image are pushed through the RRF. Starting at the root node, pixels are passed recursively to the left or right child node according to the feature tests stored at the nodes until a leaf node is reached. The estimated location of the landmark $L(\boldsymbol{v})$ is calculated based on the pixels position $\boldsymbol{v}$ and the relative voting vector $\boldsymbol{d}_l$ stored in the leaf node l. However, if the length of voting vector $|\boldsymbol{d}_l|$ is larger than radius R, i.e. pixel $\boldsymbol{v}$ is not in the area closely surrounding the landmark, the estimated location is

omitted from the accumulation of the landmark location predictions. Separately for each landmark, the pixel's estimations are stored in an accumulator image.

2.4 MRF Model

For multi-landmark localization, high-level knowledge about landmark configuration may be used to further improve disambiguation between locally similar structures. An MRF selects the best candidate for each landmark according to the RRF accumulator values and a geometric model of the relative distances between landmarks, see Fig. 1a. In the MRF model, each landmark L_i corresponds to one variable while candidate locations selected as the N_c strongest maxima in the landmark's accumulator determine the possible states of a variable. The landmark configuration is obtained by optimizing energy function

$$E(\mathbf{L}) = \sum_{i=1}^{N_L} U_i(L_i) + \sum_{\{i,j\} \in C} P_{i,j}(L_i, L_j), \tag{2}$$

where unary term U_i is set to the RRF accumulator value of candidate L_i and the relative distances of two landmarks from the training annotations define pairwise term $P_{i,j}$, modeled as normal distributions for landmark pairs in set C.

3 Experimental Setup and Results

We evaluate the performance of our landmark localization RRF variants on data sets of 2D hand X-ray images and 3D MR images of human teeth. As evaluation measure, we use the Euclidean distance between ground truth and estimated landmark position. To measure reliability, the number of outliers, defined as localization errors larger than 10 mm for hand landmarks and 7 mm for teeth, are calculated. For both data sets, which were normalized in intensities by performing histogram matching, we perform a three-fold cross-validation, splitting the data into 66 % training and 33 % testing data, respectively.

Hand Dataset consists of 895 2D X-ray hand images publicly available at Digital Hand Atlas Database[1]. Due to their lacking physical pixel resolution, we assume a wrist width of 50 mm, resample the images to a height of 1250 pixels and normalize image distances according to the wrist width as defined by the ground-truth annotation of two landmarks (see Fig. 1a). For evaluation, $N_L = 37$ landmarks, many of them showing locally similar structures, e.g. finger tips or joints between the bones, were manually annotated by three experts.

Teeth Dataset consists of 280 3D proton density weighted MR images of left or right side of the head. In the latter case, images were mirrored to create a consistent data set of images with $208 \times 256 \times 30$ voxels and a physical resolution of $0.59 \times 0.59 \times 1$ mm per voxel. Specifying their center locations, two wisdom teeth per data set were annotated by a dentist. Localization of wisdom teeth is challenging due to the presence of other locally similar molars (see Fig. 1g).

[1] Available from http://www.ipilab.org/BAAweb/, as of Jan. 2016.

Experimental setup: For each method described in Sect. 2, an RRF consisting of $N_T = 7$ trees is built separately for every landmark. The first RRF phase is trained using pixels from training images within a range of $R = 10\,\text{mm}$ around each landmark position. The splitting criterion for each node is greedily optimized with $N_F = 20$ candidate features and $N_T = 10$ candidate thresholds except for *PeterRRF*. The random feature rectangles are defined by maximal size in each direction $s^I = 1\,\text{mm}$ and maximal offset $o^I = R$. In the second RRF phase, $N_{px} = 10000$ pixels are introduced and feature range is increased to a maximal feature size $s^{II} = 50\,\text{mm}$ and offset in each direction $o^{II} = 50\,\text{mm}$.

Treating each landmark independently on both 2D hands and 3D teeth dataset, **the single-landmark experiments** show the performance of the methods in case it is not feasible (due to lack of annotation) or semantically meaningful (e.g. third vs. other molars) to define all available locally similar structures. We compare our algorithms that start with local feature scale ranges and increase to more global scale ranges (*localRRF, fAdaptRRF, PeterRRF, fpAdaptRRF, backProjRRF*) with reimplementations of two related works that start from global feature scale ranges (*CriminisiRRF* [1], with maximal feature size s^{II} and offset o^{II} from pixels uniformly distributed over the image) and optionally decrease to more local ranges (*EbnerRRF* [3]). First training phases stop for all methods at $D^I = 13$, while the second phase continues training within the same trees until $D^{II} = 25$. To ensure fair comparison, we use the same RRF parameters for all methods, except for the number of candidate features in *PeterRRF*, which was set to $N_F = 500$ as suggested in [7]. Cumulative error distribution results of the single-landmark experiments can be found in Fig. 2. Table 1 shows quantitative localization results regarding reliability for all hand landmarks and for subset configurations (fingertips, carpals, radius/ulna).

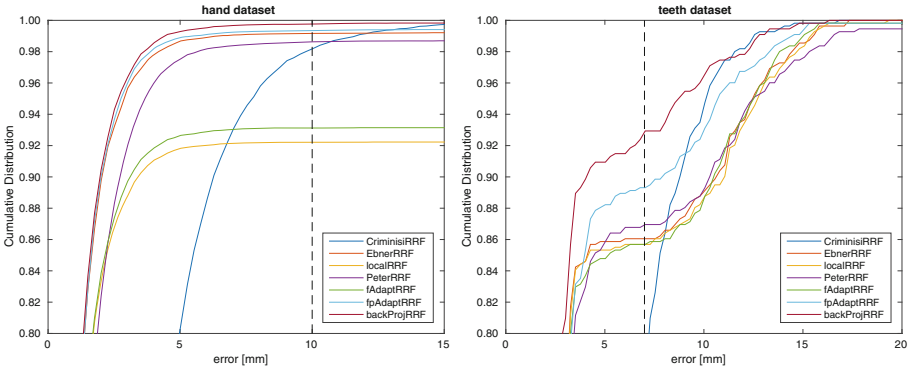

Fig. 2. Cumulative localization error distributions for hand and teeth data sets.

The multi-landmark experiments allow us to investigate the benefits of adding high level knowledge about landmark configuration via an MRF to the prediction. In addition to our reimplementation of the related works [1,3],

Table 1. Multi-landmark localization reliability results on hand radiographs for all landmarks and subset configurations (compare Fig. 1 for configuration colors).

method	mean ± std.	outliers
EbnerRRF	0.97 ± 2.45	228 (6.89‰)
Lindner et al. [5]	0.85 ± 1.01	20 (0.60‰)
localRRF+MRF	**0.80** ± **0.91**	**14** (0.42‰)
backProj	0.84 ± 1.58	57 (1.72‰)
backProj+MRF	**0.80** ± **0.91**	15 (0.45‰)

landmark subset configuration	localRRF +MRF	backProj +MRF	backProj
full ● ● ● ●	**14** (0.4‰)	15 (0.5‰)	57 (1.7‰)
fingertips ●	14 (3.1‰)	**5** (1.1‰)	17 (3.8‰)
radius,ulna ●	495 (92.2‰)	**6** (1.1‰)	11 (2.0‰)
carpals ●	17 (2.7‰)	**13** (2.1‰)	14 (2.2‰)

Lindner et al. [5] applied their code onto our hand data set using $D^I = 25$ in their implementation of the local RF stage. To allow a fair comparison with Lindner et al. [5], we modify our two training phases by training two separate forests for both stages until maximum depths $D^I = D^{II} = 25$, instead of continuing training trees of a single forest. Thus, we investigate our presented *backProjRRF*, the combination of *backProjRRF* with an MRF, *localRRF* combined with an MRF, and the two state of the art methods from Ebner et al. [3] (*EbnerRRF*) and Lindner et al. [5]. The MRF, which is solved by a message passing algorithm, uses $N_c = 75$ candidate locations (i.e. local accumulator maxima) per landmark as possible states of the MRF variables. Quantitative results on multi-landmark localization reliability for the 2D hand data set can be found in Table 1. Since all our methods including *EbnerRRF* are based on the same local RRFs, accuracy is the same with a median error of $\mu_E^{hand} = 0.51\,\text{mm}$, which is slightly better than accuracy of Lindner et al. [5] ($\mu_E^{hand} = 0.64\,\text{mm}$).

4 Discussion and Conclusion

Single landmark RRF localization performance is highly influenced by both, selection of the area from which training pixels are drawn and range of hand-crafted features used to construct its forest decision rules, yet exact influence is currently not fully understood. As shown in Fig. 2, the global *CriminisiRRF* method, is not giving accurate localization results (median error $\mu_E^{hand} = 2.98\,\text{mm}$), although it shows the capability to discriminate ambiguous structures due to the use of long range features and training pixels from all over the image. As a reason for low accuracy we identified greedy node optimization, that favors long range features even at deep tree levels when no ambiguity among training pixels is present anymore. Our implementation of *PeterRRF* [7], which overcomes greedy node optimization by selecting optimal feature range in each node, shows a strong improvement in localization accuracy ($\mu_E^{hand} = 0.89\,\text{mm}$). Still it is not as accurate as the method of Ebner et al. [3], which uses a local RRF with short range features in the second stage ($\mu_E^{hand} = 0.51\,\text{mm}$), while also requiring a significantly larger number (around 25 times) of feature candidates per node. The drawback of *EbnerRRF* is essentially the same as for *localRRF* if the area, from which local RRF training pixels are drawn, despite being reduced by the global RRF of the first stage, still contains neighboring, locally similar structures. To investigate RRFs capability to discriminate ambiguous structures

reliably while preserving high accuracy of locally trained RRFs, we switch the order of *EbnerRRF* stages, thus inverting their logic in the spirit of [2,5]. Therefore, we extended *localRRF* by adding a second training phase that uses long range features for accurate localization and differently selects areas from which training pixels are drawn. While increasing the feature range in *fAdaptRRF* shows the same accuracy compared to *localRRF* ($\mu_E^{hand} = 0.51\,\text{mm}$), reliability is improved, but not as strong as when introducing novel pixels into the second training phase. Training on novel pixels is required to make feature selection more effective in discriminating locally similar structures, but it is important to note that they do not participate in voting at testing time since the accuracy obtained in the first phase would be lost. With our proposed *backProjRRF* we force the algorithm to explicitly learn from examples which are hard to discriminate, i.e. pixels belonging to locally similar structures, as opposed to *fpAdaptRRF*, where pixels are randomly drawn from the image. Results in Fig. 2 reveal that highest reliability (0.172 % and 7.07 % outliers on 2D hand and 3D teeth data sets, respectively) is obtained by *backProjRRF*, while still achieving the same accuracy as *localRRF*.

In a multi-landmark setting, RRF based localization can be combined with high level knowledge from an MRF or SSM as in [2,5]. Method comparison results from Table 1 show that our *backProjRRF* combined with an MRF model outperforms the state-of-the-art method of [5] on the hand data set in terms of accuracy and reliability. However, compared to *localRRF* our *backProjRRF* shows no benefit when both are combined with a strong graphical MRF model. In cases where such a strong graphical model is unaffordable, e.g. if expert annotations are limited (see subset configurations in Table 1), combining *backProjRRF* with an MRF shows much better results in terms of reliability compared to *localRRF+MRF*. This is especially prominent in the results for radius and ulna landmarks. Moreover, Table 1 shows that even without incorporating an MRF model, the results of our *backProjRRF* are competitive to the state of the art methods when limited high level knowledge is available (fingertips, radius/ulna, carpals). Thus, in conclusion, we have shown the capability of RRF to successfully model locally similar structures by implicitly encoding global landmark configuration while still maintaining high localization accuracy.

References

1. Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K.: Regression forests for efficient anatomy detection and localization in computed tomography scans. Med. Image Anal. **17**(8), 1293–1303 (2013)
2. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global localization of 3D anatomical structures by pre-filtered hough forests and discrete optimization. Med. Image Anal. **17**(8), 1304–1314 (2013)
3. Ebner, T., Stern, D., Donner, R., Bischof, H., Urschler, M.: Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 421–428. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10470-6_53

4. Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A.: Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 262–270. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40763-5_33
5. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. PAMI **37**, 1862–1874 (2015)
6. Lindner, C., Thomson, J., Consortium, T.O.G.E.N., Cootes, T.F.: Learning-based shape model matching: training accurate models with minimal manual input. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 580–587. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_69
7. Peter, L., Pauly, O., Chatelain, P., Mateus, D., Navab, N.: Scale-adaptive forest training via an efficient feature sampling scheme. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 637–644. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_78