AUTOMATIC LOCALIZATION OF LOCALLY SIMILAR STRUCTURES BASED ON THE SCALE-WIDENING RANDOM REGRESSION FOREST

Darko Stern[1,2], *Thomas Ebner*[1], *Martin Urschler*[2]

[1]Institute for Computer Graphics and Vision, Graz University of Technology, BioTechMed Graz, Austria
[2]Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria

ABSTRACT

Selection of set of training pixels and feature range show to be critical scale-related parameters with high impact on results in localization methods based on random regression forests (RRF). Trained on pixels randomly selected from images with long range features, RRF captures the variation in landmark location but often without reaching satisfying accuracy. Conversely, training an RRF with short range features in a landmark's close surroundings enables accurate localization, but at the cost of ambiguous localization results in the presence of locally similar structures. We present a scale-widening RRF method that effectively handles such ambiguities. On a challenging hand radiography image data set, we achieve median and 90th percentile localization errors of 0.81 and $2.64mm$, respectively, outperforming related state-of-the-art methods.

Index Terms— anatomical landmark localization, random regression forest, scale range of features, hand X-ray

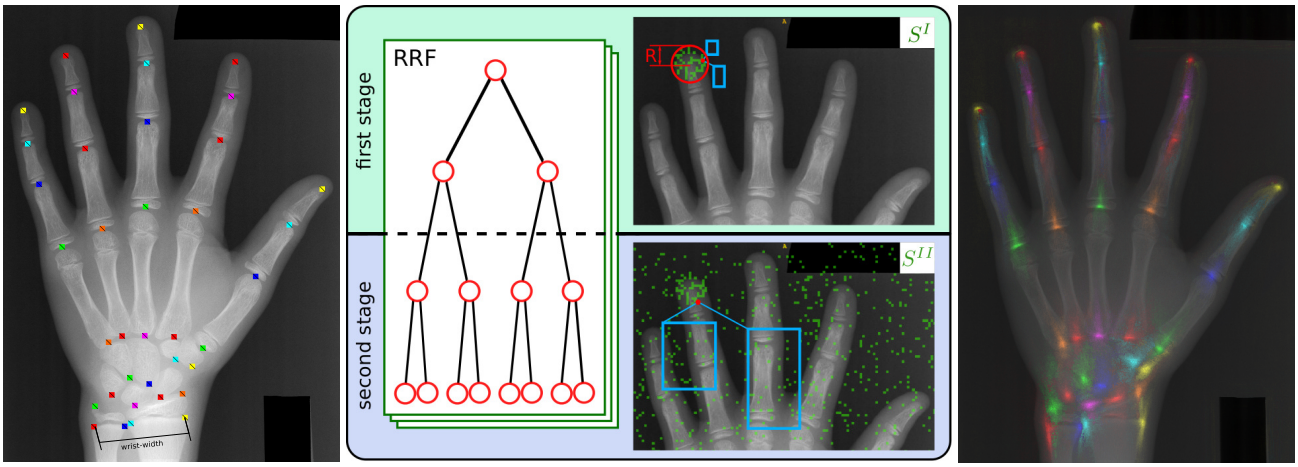
1. INTRODUCTION

Important for both registration and initialization of segmentation, automatic localization of anatomical strucutures or landmarks in medical images is a challenging task. Among different automatic object localization methods, random forest (RF) has recently shown to be the most prominent method for many medical image analysis applications [1].

For both classification and regression RF, pixels in the local and/or global vicinity of the object as well as image features extracted around these training pixels are used to build the forest decision rules voting for object locations. Selection of the scale range of features, i.e. patch size around a pixel and size of area where training pixels are chosen, shows to be a critical, yet currently not fully understood mechanism with high impact on localization results. When introducing regression random forests (RRF) for bounding box localization of anatomical structures from 3D computed tomography (CT) images, Criminisi et al. [2] proposed long range features as the image intensity difference of two cuboids randomly generated at an arbitrary distance from voxels randomly selected from training images. Although their global RRF successfully models relative configurations of anatomical structures, voting can be inaccurate due to variations in pose and location of surrounding anatomical structures. Consequently, when all image voxels contribute to voting, the estimated position tends to converge to the object's mean position as determined from the training dataset. To improve landmark localization accuracy, Ebner et al. [3] introduced a weighting scheme decreasing the weight of votes at testing time according to the length of the voting vector. Thus, by lowering the contribution of votes coming from far away structures, more trust is put into the surroundings of an object, reducing the uncertainty caused by pose changes and anatomical variation. A significant improvement in accuracy is obtained, when cascading a global with an additional RRF restricted to the area estimated by the first stage, as proposed e.g. in [3]. However, a number of implementation-specific parameters have to be empirically defined to choose the restriction area properly according to the first stage estimation. Depending crucially on the result of the global RRF, a too small area might not capture the object location, while a too large area might lead to localization of neighboring, locally similar structures. Conversely, RRFs solely trained on a local area surrounding the object, with features capturing only their local appearance, i.e. short range features, are capable to predict object locations very accurately, at the cost of potential ambiguities due to similar appearance of close-by structures. By using training voxels in the vicinity, but allowing arbitrarily large features, ambiguities can be partially reduced. Instead of the commonly used uniform sampling of the feature range during node split optimization, Loic et al. [4] recently proposed a classification RF automatically adapting to the most informative scale range of image features at which contextual information is extracted. Trained on the area locally surrounding the object, such a localization RRF achieves an effective trade-off between global and local RF, but also shows weaknesses of both approaches.

In this paper an automatic two-stage landmark localization method is presented, that combines the accuracy of locally trained RRFs with the ability to cope with ambiguities caused by locally similar structures. By widening the scale-range of informative features during training, the proposed method outperforms all the above mentioned RF approaches on a challenging hand radiography image dataset.

978-1-4799-2349-6/16/$31.00



(a) (b) (c)

Fig. 1. (a) The location of 37 anatomical landmarks in 2D hand radiographic images are estimated with RRFs. (b) The first RRF stage is trained locally on the area surrounding a landmark location (inside radius R) with short range features. The second RRF stage is trained on the pixels from the whole image with an increased maximal feature range. (c) Landmark locations are estimated by accumulating the predictions of pixels locally surrounding the landmark.

2. METHOD

Generally constrained by all surrounding structures, the location of an anatomical landmark is most accurately defined by the structures in its immediate neighborhood. Following this observation, we train the first RRF stage on the area closely surrounding the landmark with short range features, allowing an accurate estimation, see Fig. 1. However, an efficient differentiation from locally similar structures can only be made when global context is used. If the surrounding area of a landmark is sufficiently defined, i.e. the maximal depth level of the first stage is reached, RRF training enters a second stage that allows generation of features located at an arbitrary distance from training pixels. To make feature selection more efficient in discriminating locally similar areas, pixels randomly chosen from other parts of the image are also introduced into the second RRF stage, see Fig. 1. At testing time, pixels outside of a landmark's local vicinity are recognized and banned from estimating the landmark location in order to maintain the accuracy achieved during the first RRF stage.

2.1. Training the RRF

For each anatomical landmark, an RRF is trained independently. At each node of the T trees of a forest, the set of pixels S_n reaching the node n is pushed to the left ($S_{n,L}$) or the right ($S_{n,R}$) child node according to the splitting decision made by thresholding a feature response for each pixel. Feature response is calculated as difference between the mean intensity of two rectangles with maximal size in each direction s and maximal offset o relative to each 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, that maximize an information gain measure I, computed as:

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

For pixel set S, $\boldsymbol{d}_i$ is the i-th voting vector, defined as distance 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. To prepare 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 stage: Trained on the set of pixels (S^I), randomly selected from the training images at the location inside a circle centered at landmark position and with radius R, the first RRF stage is trained with features whose rectangles have maximal size in each direction s^I and maximal offset o^I, see Fig. 1(b). Training of this stage is finished when the maximal depth D^I is reached.

Second training stage: In addition to the set of pixels S^I introduced in first stage, the second RRF stage is trained on a set of pixels S^{II} randomly selected from all pixels in the training images, see Fig. 1(b). In a manner that will be explained in subsection 2.2, the newly introduced pixels are pushed through the trees till they reach the terminal nodes of the first RRF stage. Thus, at the beginning of second stage training, pixels from the area surrounding the landmark as well as pixels with locally similar appearance are located in the terminal nodes of the first stage. Second stage RRF training is carried out in the same fashion as first stage training, with the exception that the scale range of image features, i.e. the maximal size in each direction s^{II} and maximal offset o^{II}

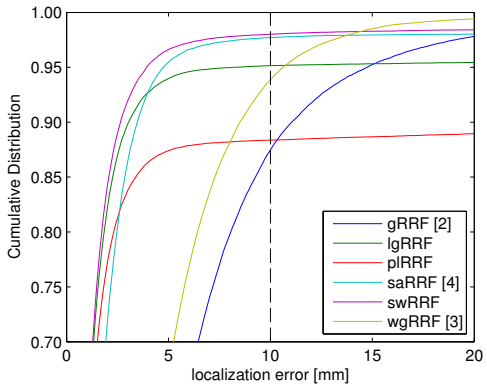


Fig. 2. The cumulative distribution of localization errors.

relative to the pixel position, is increased. When the maximal tree depth D^{II} is reached or there is no improvement in I, the recursive splitting procedure is finished.

2.2. 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 $l_T(\boldsymbol{v})$ 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 accumulated in an image initialized in the same coordinate system as the testing image and the location of the landmark is estimated by the accumulator's maximum.

3. EXPERIMENTAL SETUP AND RESULTS

Experimental setup: We evaluate the performance of our scale-widening regression forest method (swRRF) on 600 hand radiographs from the publicly available Digital Hand Atlas Database System[1]. Since the resolution of the images is not known, but the field of view of the hands is similar, we resampled all images to the same height of 625 pixels to achieve a similar spatial resolution. Normalizing image distances according to the wrist width, defined by the ground-truth annotation of 2 landmarks as shown in Fig. 1(a) and assuming a wrist-width of $50mm$, pixel size is thus provided with a physical interpretation in mm. For evaluation, $N_L = 37$ landmarks, many of which have locally similar structures, e.g. the tip of the fingers or joint between the bones as shown in Fig. 1(a), were manually annotated by three engineers well experienced in medical image analysis.

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

Table 1. Accuracy defined as 50th and 90th percentiles. Reliability defined as ratio of correctly detected landmarks with error below threshold (10 and 20 mm) and number of outliers.

	median (mm)		reliability / # outliers	
	50%	90%	<10mm	<20mm
gRRF [2]	4.42	11.04	0.87 / 5569	0.98 / 971
wgRRF [3]	3.67	8.51	0.94 / 2709	**0.99 / 261**
saRRF [4]	1.24	3.46	0.98 / 1016	0.98 / 881
plRRF	0.87	25.98	0.88 / 5163	0.89 / 4903
lgRRF	0.82	3.05	0.95 / 2155	0.95 / 2024
swRRF	**0.81**	**2.64**	**0.98 / 883**	0.98 / 702

Our swRRF is built separately for each landmark and consists of $N_T = 7$ trees. This setup is identical for all compared approaches. The first stage of the swRRF is trained up to depth level $D^I = 12$ on a set of pixels randomly selected from training images within a circle of radius $R = 10mm$ centered at each landmark position. To greedily optimize the splitting criterion for each node, $N_F = 20$ candidate features and $N_T = 10$ candidate thresholds are generated. The random feature rectangles are defined by maximal size in each direction $s^I = 1mm$ and maximal offset $o^I = R$. In the second RRF stage, new pixels randomly selected from all test image locations are introduced. Feature size is increased to a maximal size $s^{II} = 50mm$ and an offset in each direction of $o^{II} = 50mm$. Training the second RRF stage stops after reaching a maximal tree depth $D^{II} = 17$.

Algorithm performance is evaluated and compared to other methods in a cross-validation setup with N = 3 rounds, splitting 600 input images into 200 for training and 400 for testing. As evaluation measure, we used Euclidean distance between ground truth and estimated landmark positions. Results are compared with the global RRF (gRRF) approach of Criminisi et al. [2], weighted global RRF (wgRRF) of Ebner et al. [3] and scale adaptive RRF (saRRF) of Loic et al. [4]. To ensure a fair comparison, we used the same RRF parameters for all methods, except for the number of candidate features in our implementation of the Loic et al. [4] method, which was set to $N_F = 500$ as suggested by the authors. The results are also compared to a purely local RRF (plRRF) trained by continuing the first stage of our RRF to the maximal depth of $D^M = 17$. Finally, we also trained a local-global RRF (lgRRF) that differs from swRRF by restricting training to the voxels locally surrounding a landmark, i.e. no new pixels outside radius R are selected in the second RRF stage.

Results: Figure 2 shows the cumulative localization error distribution calculated for all landmarks (44400) of the six tested methods. In Table 1, quantitative results are presented as 50% and 90% percentile localization error and reliability given as the percentage of tested cases with a distance less then 10 and $20mm$ from the landmark position, respectively.

4. DISCUSSION AND CONCLUSION

In addition to the range of features used to train RRF [4] and the distance between voting pixel and landmark [3], this paper demonstrates that the area from which voxels are selected to train the RRF also has a high influence on landmark localization results. As shown by median values in Table 1, accuracy of the global methods, that allow arbitrary long voting ranges such as [2] and [3], is significantly smaller than for the methods trained on a local area with short range features, e.g. $4.42mm$ for global RRF and $0.87mm$ for local RRF. Trained on the area locally surrounding the landmarks, meaning that voting range effectively is locally limited, scale-adaptive RRF proposed by Loic et al. [4] shows to improve accuracy ($1.24mm$) compared to global methods but still is not as accurate as methods trained only with short range features. On the other hand, our method that was trained on the same area at first stage and allows pixels to vote for the landmark position only if it belongs to that area, demonstrates the same accuracy as the local RRF methods ($0.81mm$) while substantially improving on reliability in terms of outliers. It is important to note that without restricting voting to the pixels surrounding the landmark at testing time, the accuracy obtained in the first stage would be lost. Thus, the pixels randomly selected from images in the second stage are required only to make feature selection more effective in discriminating locally similar structures, but should not participate in voting due to their distance from the landmark.

As shown in Fig. 2, global RRF methods such as [2] or [3] are capable of capturing pose variation in hand images as their cumulative distribution is larger than 0.98 at $20mm$. Nevertheless, having the landmark prediction widely spread around an inaccurate maximum, cascading the global RRF with local RRF in order to improve accuracy often leads to localization of neighboring, locally similar landmarks. On the other hand, the estimated distribution of a local RRF is spiky and, in the case of locally similar structures as present e.g. in hand images, multiple maxima with equally high estimation responses at the location of e.g. the tip of the fingers or joint between the bones can be observed. To maintain the accuracy of the local RRF and to efficiently cope with multiple estimation responses, global geometrical models like statistical shape models (SSM) [5] or Markov Random Fields (MRF) [6] may be used as an additional step to disambiguate multiple landmarks, i.e. to improve the reliability. As our method follows the same idea of selecting the best among accurately located landmark candidates, similarities can be found between optimizing a global geometrical model and our second stage. However, the best candidate in our single landmark optimization method is defined by all structures in the training data, while in global geometrical model approaches, the best landmark candidate is defined by the location of other, geometrically related landmarks. Anyhow, in this paper, we did not further investigate this option, as the global geometrical model can always be used in addition to the proposed method to further improve reliability in multiple landmark localization. Moreover, it is to be expected that due to suppressing the estimation response at the location of locally similar structures in the accumulator's images, replacement of the local RRF with our method should make such an approach initialized closer to the optimal solution and thus trivial for optimization. Finally, increasing the feature range without including the voxels selected from other parts of the image, will not make the method competitive in reliability, which can be seen by comparing the local-global RRF and the proposed method in Fig. 2. Thus, by including the voxels selected from all other parts of the image into the second stage, our method achieved the highest reliability (0.98) as defined with a distance $< 10mm$ from the true landmark position.

To conclude, we proposed an automatic landmark localization method based on RRF that can cope with locally similar structures while still achieving high accuracy due to our novel scale-widening two-stage RRF architecture.

5. ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF): P 28078-N33 (FAME).

6. REFERENCES

[1] A Criminisi and J Shotton, Eds., *Decision Forests for Computer Vision and Medical Image Analysis*, Springer, 2013.

[2] A Criminisi, D Robertson, E Konukoglu, J Shotton, S Pathak, S White, and K Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[3] T Ebner, D Stern, R Donner, H Bischof, and M Urschler, "Towards Automatic Bone Age Estimation from MRI: Localization of 3D Anatomical Landmarks," in *MICCAI*, 2014, vol. 8674 of *LNCS*, pp. 421–428.

[4] P Loic, O Pauly, P Chatelain, D Mateus, and N Navab, "Scale-Adaptive Forest Training via an Efficient Feature Sampling Scheme," in *MICCAI*, 2015, vol. 9349 of *LNCS*, pp. 637–644.

[5] C Lindner, P A Bromiley, M C Ionita, and T F Cootes, "Robust and Accurate Shape Model Matching using Random Forest Regression-Voting," *IEEE Trans. PAMI*, vol. 37, pp. 1862–1874, 2015.

[6] R Donner, B H Menze, H Bischof, and G Langs, "Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization," *Med. Image Anal.*, vol. 17, no. 8, pp. 1304–1314, 2013.