Towards Automatic Bone Age Estimation from MRI: Localization of 3D Anatomical Landmarks

Thomas Ebner[1,3,⋆], Darko Stern[1], Rene Donner[2,1], Horst Bischof[1], and Martin Urschler[1,3,4]

[1] Institute for Computer Graphics and Vision, BioTechMed, Graz University of Technology, Austria
[2] Computational Image Analysis and Radiology Lab, Department of Radiology, Medical University Vienna, Austria
[3] Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
[4] Department of Legal Medicine, Medical University of Graz, Austria

Abstract. Bone age estimation (BAE) is an important procedure in forensic practice which recently has seen a shift in attention from X-ray to MRI based imaging. To automate BAE from MRI, localization of the joints between hand bones is a crucial first step, which is challenging due to anatomical variations, different poses and repeating structures within the hand. We propose a landmark localization algorithm using multiple random regression forests, first analyzing the shape of the hand from information of the whole image, thus implicitly modeling the global landmark configuration, followed by a refinement based on more local information to increase prediction accuracy. We are able to clearly outperform related approaches on our dataset of 60 T1-weighted MR images, achieving a mean landmark localization error of 1.4±1.5mm, while having only 0.25% outliers with an error greater than 10mm.

1 Introduction

Skeletal bone age estimation (BAE) of adolescents based on 2D hand radiographs has applications in clinical and legal medicine, like growth predictions, diagnosis of endocrinological diseases [1], assessing asylum seekers without proper identification documents, or preventing age manipulation in junior-level sports competitions [2]. Recently, non-invasive 3D MRI methods have gained in importance [1,2], especially in legal medicine, since the use of ionizing radiation is prohibited in many countries for non-diagnostic reasons. To provide an objective, repeatable and radiation-free measure of chronological age, a fully automatic BAE method from hand MRI may significantly advance the use of age estimation in legal medicine. Automated localization of individual hand bone landmarks is a mandatory and crucial first step in a BAE pipeline to analyze bone ossification

⋆ This work was partly supported by the city of Graz (A16-21628/2013) and a European Community FP7 Marie Curie Intra European Fellowship (331239).

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 421–428, 2014.

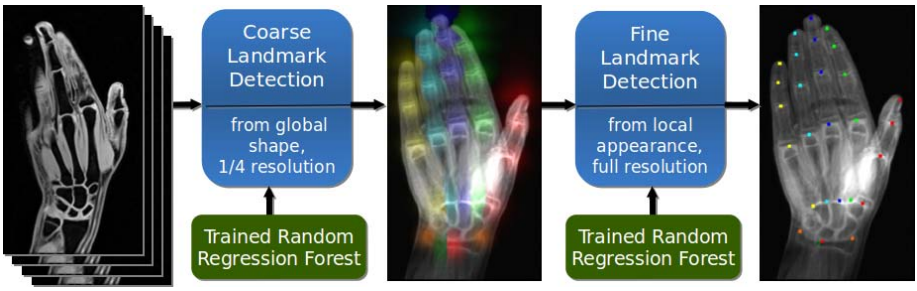

Fig. 1. Overview of the proposed multiscale localization method

stages. Anatomical landmark localization may be performed based on low-level interest point detection [3], that requires a reasoning on the high-level semantics of localized points in a subsequent step. A more specific low-level algorithm for hand bone localization was presented in [4], where ROIs were extracted from X-ray images by detecting finger tips and bone longitudinal axes using gradient information. This approach is not robust to the presence of variations in typical clinical images. BoneXpert [5] uses a statistical shape/appearance model to detect 2D bone contours from X-ray images, however, such models require a high effort to extend to 3D surfaces and their generative modelling strategy requires a large number of training data [6]. Recently discriminative machine learning approaches have received a lot of attention for anatomical structure localization, see [7] for a survey. Criminisi et al. apply random regression forests (RRF) to estimate distances to the planes of bounding boxes containing anatomical structures [8]. They are able to coarsely locate a large number of different organs, but especially in the presence of flexible anatomical structures like the fingers, their approach lacks in precision, presumably due to their axis-aligned bounding box design. Donner et al. introduce a three-step procedure consisting of a coarse, generic landmark localization without global knowledge of the landmark configuration, followed by a per-landmark refinement and finally imposing a global structure using a Markov Random Field (MRF) [7]. In [9] the same authors have proposed an alternative localization approach using dictionaries of multi-scale image patches, which jointly predicts landmarks visible within each patch using nearest-neighbor dictionary lookups. They have shown similar localization accuracy compared to [7], but at a fraction of the runtime.

In this paper, we present a novel 3D anatomical landmark localization approach for 3D hand MR images (see Fig. 1). We propose a two-step multiscale RRF based approach, that first makes a prediction of the coarse bone landmark positions analyzing the whole shape of the hand and using feature information from all over the image. This step finds the area, where the landmark locations are expected, and implicitly models the global landmark configuration. Based on these locations, the second step uses more localized information for accurate landmark prediction. We see this idea of gradually decreasing the area of interest of an RRF as our main contribution, which resembles a generic localization

strategy. We apply our method and related approaches to a database containing youths in an age range, where BAE is relevant for forensic applications. This data set is challenging due to its age range, presence of repeating structures, and variation due to the non-fixed configuration of fingers (see Fig. 2a,b).

2 Method

The location of anatomical landmarks is constrained by all of their surrounding structures. However, coarse localization of landmarks is supported by global information from all over the image, while closer structures provide the information to increase the accuracy of landmark localization. We realize this concept by a weighting scheme, that lets local structures have a higher contribution to the estimation of landmark positions. To implement this idea the RRF framework [8] is perfectly suitable, since it selects proper image structures that vote for landmark distances in a probabilistic fashion, where position estimates can be weighted by the distance to the estimate. Additional information about the landmark position can subsequently be obtained by connecting multiple estimation steps, where the output of individual steps restricts the area for estimating landmarks in the following step. This connection is made by using several RRF stages, that gradually decrease the areas around landmarks, where structural information is taken from. Together with the weighting scheme, we regard this idea as our main contribution compared to related work [7, 8].

For our application of landmark detection from hand MR images, we propose using two RRF steps according to the strategy above, as shown in Fig. 1. The first RRF coarsely locates the landmarks, and due to its multi-class architecture where each voxel in the image votes for all landmark positions, it implicitly models spatial relations between the landmarks. The second RRF learns from the restricted areas around landmarks given by the first step, thus improving localization accuracy. In the following we describe our generic RRF and focus on its use for the two proposed landmark detection steps.

2.1 Random Regression Forest

RRF Training: Our regression forest models the distances in x, y, and z of voxels in training images to multiple individual landmark positions $\boldsymbol{l}_c$ simultaneously. At each node of the T independently constructed trees, the set of voxels (S) reaching the node is split into voxels reaching the left (S_L) and the right child node (S_R). The splitting decision is made by thresholding for each voxel a feature response, calculated by taking the mean intensity difference between two cuboids with arbitrary size and offset relative to each voxel position $\boldsymbol{v} = (v_x, v_y, v_z) \in S$. From a pool of randomly generated features and thresholds, one feature and threshold is selected in order to maximize an information gain measure IG, computed according to

$$IG(S, S_L, S_R) = H(S) - \sum_{i \in \{L,R\}} \frac{|S_i|}{|S|} H(S_i), \tag{1}$$

where entropy $H(S) = \sum_c q(c;S) \cdot log|\Lambda_c(S)|$, and $q(c;S)$ is the ratio between the number of voxels that vote for landmark c and the total number of voxels in S. Entropy is computed from per-landmark variances $\Lambda_c(S)$:

$$\Lambda_c(S) = \frac{1}{|S|} \sum_{i \in S} ||\boldsymbol{d}_c(\boldsymbol{v}_i) - \frac{1}{|S|} \sum_{j \in S} \boldsymbol{d}_c(\boldsymbol{v}_j)||^2. \tag{2}$$

The maximization of the information gain aims to minimize the uncertainty $\Lambda_c(S_{\{L,R\}})$ of the distance estimates $\boldsymbol{d}_c(\boldsymbol{v})$ to all landmarks from the voxels in left and right child node. Each voxel votes for the landmark positions $\boldsymbol{l}_c$ relative to its position $\boldsymbol{v}$, with the relative voting vector $\boldsymbol{d}_c(\boldsymbol{v}) = \boldsymbol{l}_c - \boldsymbol{v}$ for a landmark c. Node splitting is done recursively and stops, when the maximum tree depth D is reached. For testing we store at each leaf node for each landmark the 1D histograms of the x, y and z components of $\boldsymbol{d}_c(\boldsymbol{v})$ of all voxels reaching the node.

RRF Testing: During testing voxels are pushed through all of the T trained trees. Starting at the root node, voxels are passed recursively to the left or right child according to the binary feature tests stored at the split nodes until a leaf node $l_t(\boldsymbol{v})$ is reached. We apply the distance estimates given by the histograms at the leaf nodes $\boldsymbol{h}_{\{x,y,z\},c}(l_t(\boldsymbol{v}))$ relative to the voxel positions $\boldsymbol{v}$ and sum them up with a weight $w(\boldsymbol{v})$, according to (3), to get for each landmark three histograms $\boldsymbol{h}_{\{x,y,z\},c}$, representing the probabilities of a landmark being located at a certain position separately for x, y, and z.

$$\boldsymbol{h}_{\{x,y,z\},c} = \frac{1}{T \cdot \sum_v w(\boldsymbol{v})} \sum_{t=1}^{T} \sum_{\boldsymbol{v}} w(\boldsymbol{v}) \boldsymbol{h}_{\{x,y,z\},c}(l_t(\boldsymbol{v})) \tag{3}$$

The final probability estimate $p(\boldsymbol{l}_c)$ is obtained by the product of the three histograms $\boldsymbol{h}_{\{x,y,z\},c}$, and the final landmark positions by the maxima of $p(\boldsymbol{l}_c)$.

Our main contribution is the introduced weighting factor $w(\boldsymbol{v})$ in (3), which lets local structures contribute more, by decreasing the weight of the voting vectors according to their length $||\boldsymbol{d_c}||$. The weighting factor (4) also incorporates the goal of reducing the area for estimating landmarks during the second detection step, by plugging the outcome of the first detection step into the prior probability $p_c(\boldsymbol{v}) = p(\boldsymbol{l}_c)$.

$$w(\boldsymbol{v}) = e^{-||\boldsymbol{d_c}|| \cdot \alpha} \cdot p_c(\boldsymbol{v}) \tag{4}$$

The parameter α allows adjusting the steepness and it is set to $1/cm$ in all experiments. With the lack of prior knowledge about landmark positions in the first detection step, we use the same prior probability for all voxels, i.e. $p_c(\boldsymbol{v}) = 1$.

2.2 First Detection Step: Coarse RRF (CRRF) Estimation

We train an RRF according to Sec. 2.1 by letting all voxels within the training images vote for all landmark positions simultaneously. Input images are resampled to a quarter of the original resolution, since this first step only requires a coarse localization, and experiments on full resolution did not show any benefit.

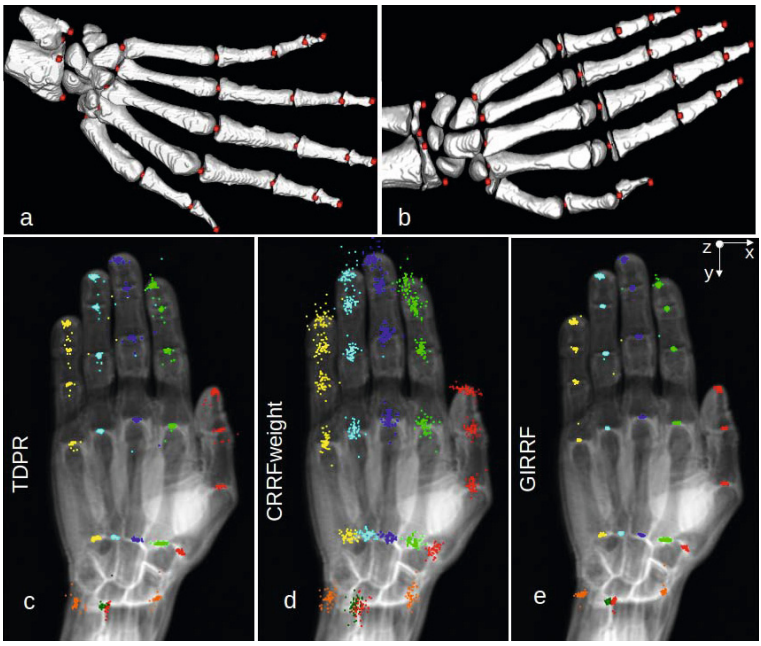

Fig. 2. Hand bone segmentation with landmark annotation (**a**) and different subject with GIRRF localization result (**b**). Results of compared algorithms (**c-e**) presented on a 2D projection of a selected MR volume, with error vectors from cross-validation drawn relative to ground truth position of the specific MR volume

2.3 Second Detection Step: Refinement

In the second detection step another regression forest is trained by considering only a small region around the landmarks retrieved by the CRRF step, resulting in voting for neighboring landmarks only. For training we apply CRRF to all training images to get the probability $p(\boldsymbol{l}_c)$ of the landmark c being at position $\boldsymbol{l}_c$. We use this probability to focus on local areas by randomly selecting voxels for training according to the distribution $p(\boldsymbol{l}_c)$. Additionally we apply a threshold τ to eliminate voxels with low probability, representative of non-local structures. All selected voxels are put into one single forest, and the same kind of features as in the CRRF step are used. This makes effective use of feature sharing, since a lot of landmarks share similar local appearance, an idea that was presented in [10]. When going down to deeper levels of the tree, voxels of landmarks with a different local appearance will be passed to different branches of the tree. During the IG calculation and in the voting aggregation in the leaf nodes, voxels are voting only for those landmark positions where $p_c(\boldsymbol{l}_c) \geq \tau$.

In testing, we accumulate the leaf histograms of all voxels in a range r around the estimation of the landmark position from the first detection step. Due to a higher prediction accuracy when moving close to the actual landmark position, this process is repeated $n_{iter} = 3$ times, initialized by the prediction of the CRRF

or the previous iteration step. This resembles a greedy optimization scheme for landmark localization, where we experienced convergence after a few steps. We refer to the combination of first and second detection step as our gradually improving random regression forest (GIRRF) localization method.

3 Materials and Experimental Setup

Materials: Our dataset of left hand T1-weighted 3D gradient echo MR images consisted of scans from 60 caucasian male subjects between 13 and 23 years. The average dimension of the volumes was $294 \times 512 \times 72$ with a voxel size of $0.45 \times 0.45 \times 0.9mm^3$. Hands are located roughly in the center and rotation about the z-axis is varying in the range of about $\pm 15°$ (see Fig. 2e). In each volume, 28 landmarks were manually annotated by a scientist, who selected characteristic locations within the hand at the ends of each of the metacarpal and phalanx bones, three points at the radius and one at the ulna bone (see Fig. 2a).

Experimental Setup: We evaluated our algorithm and compared it to the Top-Down Patch Regression (TDPR) [9] method with the parameters proposed by the authors in a cross-validation setup with $N = 5$ rounds. In each round we randomly split the 60 input images into 43 training and 17 testing images. The measure we used for evaluating the performance is the Euclidean distance between the ground truth and the estimated landmark position.

First Detection Step: We built $T = 8$ trees with maximum depth $D = 14$, where for each node split 100 candidate features and 10 candidate thresholds were generated. The maximum size and range of the random feature cuboids was limited to 50mm and 25mm in each dimension, respectively. Further, to show the benefit of the introduced weighting scheme, we made an experiment on the first detection step with and without the use of the weighting function $w(\boldsymbol{v})$, denoted as CRRFweight and CRRF, respectively. Note that our first detection step without the weighting function resembles an implementation of the method in [8], but focusing on landmark localization instead of bounding boxes, since we aim for accurate localization independent of bounding box orientation.

Second Detection Step: The threshold used for selecting the voxels for training was set to $\tau = 0.4 \cdot max\{p(\boldsymbol{l}_c)\}$. Using the selected voxels, we built $T = 8$ trees with maximum depth $D = 15$. At each node split 20 random candidate features and 10 candidate thresholds are generated. The maximum size in each dimension and distance of the feature cuboids is 7mm. To iteratively estimate final landmark positions, we used the voxels in the ranges $r = \{30mm, 10mm, 5mm\}$ around the previous estimation starting with the CRRF result.

4 Results

Figure 2 shows a visualization of the cross-validation results of the TDPR and our proposed two detection steps. For all landmarks we achieve a localization error ($\pm$ standard deviation) of 1.44 ± 1.51mm. In x, y and z direction we achieve

Table 1. Comparison of localization errors from cross validation on hand bone landmarks, radius/ulna (R/U), carpometacarpal (CMP), metacarpal (MCP), distal and proximal interphalangeal joints (DIP,PIP), finger tips (FT)

Method	Localization Error [mm]: Mean ± Std.						
	R/U	CMC	MCP	PIP	DIP	FT	overall
TDPR [9]	2.8±2.8	2.0±1.1	2.0±2.4	2.0±3.1	1.8±3.9	2.7±4.2	2.2±3.1
CRRF [8]	7.9±5.1	7.4±5.1	6.7±3.1	6.5±3.2	6.5±3.3	8.1±5.3	7.2±4.4
CRRFweight	4.8±2.4	3.7±1.5	4.0±2.1	4.1±2.0	4.5±2.5	5.5±3.1	4.4±2.4
GIRRF	**1.8±1.3**	**1.5±0.7**	**1.2±0.6**	**1.3±2.2**	**1.3±2.4**	**1.5±0.8**	**1.4±1.5**

a mean error of $0.68mm$, $0.57mm$ and $0.84mm$, respectively. A more detailed quantitative comparison of the evaluated methods can be found in Table 1. From the $5 \cdot 17 \cdot 28 = 2380$ detected landmark positions, only six outliers (0.25%) had a localization error larger than 10mm. One outlier was on the radius bone, the others occurred on the distal interphalangeal (DIP) and proximal interphalangeal (PIP) joints. The TDPR approach showed 35 (1.5%) outliers.

Runtime of our C++ algorithm, which was implemented on top of the open-source Sherwood library from Microsoft Research, is about 400s per volume on an 8-core Intel(R) Core(TM) i7 CPU. Non-parallelized forest training for one round of cross validation takes 24 hours on the same PC. Runtimes for training and testing of TDPR are around 2 hours and 10s, respectively.

5 Discussion

As can be seen in Table 1 and Fig. 2, our proposed algorithm achieves superior overall and individual localization accuracy in terms of mean error and standard deviation among the compared algorithms. A detailed analysis of the outliers shows that for TFPR and GIRRF they occur in hands with a finger pose that is not covered in the training set during cross validation, however, more often these situations occur in the TDPR approach. In case something went wrong during the detection in the TDPR approach, almost all landmarks located on the phalanges were detected wrong in the same image. TDPR seems to be even more constrained by the variability in the training data through the explicit use of a PCA-based point distribution model (PDM). An experiment showed us that adding this PDM to GIRRF does not fix the remaining outliers, but rather introduces new errors on already well detected landmarks. In GIRRF, there were at most three outliers in one single image, compared to 12 for PFHR.

All evaluated algorithms achieved the worst mean error on radius and ulna bone, which can be explained by the large anatomical variation especially at the ulna bone and because the landmarks had to be chosen at locations, that were hard to define in manual annotation due to lack of proper anatomical structures near the bone. On our dataset CRRF is able to achieve a much better accuracy when including the weighting function according to (4), compared to a weighting equal to one as proposed in [8]. The reason for this improvement is, that local information around each landmark provides a more accurate estimation, since

there is a large pose variation of the fingers in our database. This fact is exactly what has driven the development of our proposed approach. Since automatic BAE relies on a very accurate bone localization, we find that we can improve by using GIRRF compared to related work, due to its capability to extract age related features to learn an age regression model based on located bone landmarks. A drawback of our approach is higher runtime compared to e.g. TDPR. Our major bottleneck is leaf histogram summation, which could be sped up by a GPU implementation.

6 Conclusion and Outlook

We have shown a novel hand bone landmark detection approach based on random regression forests at multiple scales, which outperforms other methods regarding localization accuracy on our hand MRI data. First experiments have demonstrated that GIRRF is able to initalize an automatic skeletal bone age estimation algorithm that requires extraction of age related features like ossification stages of the bone. In future work we plan to investigate GIRRF on different data sets as well, to show its generalization capabilities.

References

1. Terada, Y., Kono, S., Tamada, D., Uchiumi, T., Kose, K., Miyagi, R., Yamabe, E., Yoshioka, H.: Skeletal age assessment in children using an open compact MRI system. Magnet. Reson. Med. 69(6), 1697–1702 (2013)
2. Dvorak, J., George, J., Junge, A., Hodler, J.: Age determination by magnetic resonance imaging of the wrist in adolescent male football players. Brit. J. Sport Med. 41(1), 45–52 (2007)
3. Wörz, S., Rohr, K.: Localization of anatomical point landmarks in 3D medical images by fitting 3D parametric intensity models. Med. Image Anal. 10(1) (2006)
4. Pietka, E., Gertych, A., Pospiech, S., Cao, F., Huang, H., Gilsanz, V.: Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal roi extraction. IEEE Trans. Med. Imag. 20(8), 715–729 (2001)
5. Thodberg, H.H., Kreiborg, S., Juul, A., Pedersen, K.D.: The BoneXpert method for automated determination of skeletal maturity. IEEE Trans. Med. Imag. 28(1), 52–66 (2009)
6. Heimann, T., Meinzer, H.P.: Statistical shape models for 3D medical image segmentation: A review. Medical Image Analysis 13, 543–563 (2009)
7. 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)
8. 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)
9. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Fast anatomical structure localization using top-down image patch regression. In: Menze, B.H., Langs, G., Lu, L., Montillo, A., Tu, Z., Criminisi, A. (eds.) MCV 2012. LNCS, vol. 7766, pp. 133–141. Springer, Heidelberg (2013)
10. Razavi, N., Gall, J., van Gool, L.: Scalable multi-class object detection. In: IEEE Computer Vision and Pattern Recognition (CVPR), pp. 1505–1512 (2011)