Content-based Image Retrieval Applied to Bone Age Assessment

Benedikt Fischer[*a], André Brosig[a], Petra Welter[a], Christoph Grouls[b], Rolf W. Günther[b], Thomas M. Deserno[a]
[a]Department of Medical Informatics, [b]Department of Diagnostic Radiology,
RWTH Aachen University, Pauwelsstr. 30, 52057 Aachen, Germany

ABSTRACT

Radiological bone age assessment is based on local image regions of interest (ROI), such as the epiphysis or the area of carpal bones. These are compared to a standardized reference and scores determining the skeletal maturity are calculated. For computer-aided diagnosis, automatic ROI extraction and analysis is done so far mainly by heuristic approaches. Due to high variations in the imaged biological material and differences in age, gender and ethnic origin, automatic analysis is difficult and frequently requires manual interactions. On the contrary, epiphyseal regions (eROIs) can be compared to previous cases with known age by content-based image retrieval (CBIR). This requires a sufficient number of cases with reliable positioning of the eROI centers. In this first approach to bone age assessment by CBIR, we conduct leaving-one-out experiments on 1,102 left hand radiographs and 15,428 metacarpal and phalangeal eROIs from the USC hand atlas. The similarity of the eROIs is assessed by cross-correlation of 16x16 scaled eROIs. The effects of the number of eROIs, two age computation methods as well as the number of considered CBIR references are analyzed. The best results yield an error rate of 1.16 years and a standard deviation of 0.85 years. As the appearance of the hand varies naturally by up to two years, these results clearly demonstrate the applicability of the CBIR approach for bone age estimation.

Keywords: bone age assessment, epiphyseal region of interest, content-based image retrieval, computer-aided diagnosis, hand radiograph, skeleton

1. INTRODUCTION

Bone age assessment based on hand radiographs is a frequent and time-consuming task for radiologists. The two most commonly used methods are based on image comparison. In the method of Greulich and Pyle[1], the radiologist compares all bones of the hand to radiographs in the standard atlas. In the method of Tanner and Whitehouse (TW3)[2], a certain subset of bones is examined. Different approaches have been taken to fully or partially automate bone age assessment[3,4,5,6]. Recently, commercial software has been introduced that is based on the active appearance model[7]. Critical points of existing approaches are: (i) reliable region of interest extraction, (ii) registration to the references, and (iii) combination of matching results to obtain a final age suggestion, and (iv) need of individual parameterizations dependant on age, gender, and ethnic origin of the patient. Content-based image retrieval (CBIR) promises comparisons of the images and regions of interest (ROI) to earlier cases without the need to specify many separate heuristics. While our previous efforts have been focused on the automatic ROI extraction[8], the presented work serves as a proof-of-concept for CBIR as an instrument for bone-age assessment.

2. METHODOLOGY

In order to assess the bone age for a given hand radiograph, the epiphyseal regions (eROIs) between metacarpals and distal phalanges are compared to the corresponding regions of all hand radiographs with known age in the retrieval database. From these similarities the unknown age is estimated. For faster response times, the retrieval database already contains the center coordinates and the images of the eROIs for all hand radiographs. The age assessment for a new radiograph by CBIR is then performed in four stages which are subsequently explained in more detail (Fig. 1):

1. 14 eROIs are extracted from the new image.
2. Each eROI is subjected to an individual content-based query to the retrieval engine.

[*]Corresponding author: Dipl.-Inform. Benedikt Fischer. Department of Medical Informatics, RWTH Aachen University, Pauwelsstr. 30, 52057 Aachen, Germany, email: bfischer@mi.rwth-aachen.de; web: http://irma-project.org, phone: +49 241 80 85174, fax: +49 241 80 33 85174

Medical Imaging 2010: Computer-Aided Diagnosis, edited by Nico Karssemeijer, Ronald M. Summers,
Proc. of SPIE Vol. 7624, 762412 · © 2010 SPIE · CCC code: 1605-7422/10/$18 · doi: 10.1117/12.844392

3. For each eROI a list of the K best matching eROIs in the CBIR database is retrieved along with their age and similarity score.
4. From the list of retrieved eROIs, their respective ages are accumulated for the age prediction.

2.1 eROI Extraction

The first stage, i.e. the extraction of the eROIs, can be either performed fully automatically as demonstrated by our previous works[8], or semi-automatic. In the latter case, the user first defines the centers of the eROIs by clicking into them. Afterwards the rest of the eROI extraction is performed automatically by extracting oriented bounding boxes around the center coordinates. The orientation is determined by the mean of the orientation of the connecting lines between the eROI center and the center above (if applicable) and below of the same finger. For an image of 256 pixels height, an eROI size of 25x30 pixels has proven suitable. The eROI size for larger images is chosen appropriately. The similarity comparison as described in the next section is computed on 16x16 downscaled versions of the extracted eROIs.

After extraction, each eROI p is automatically assigned a label number $\lambda(p) \in \{1,\ldots,19\}$ according to its anatomical position (Fig. 2). In the case of automatic eROI extraction, the labeling is directly available from the labels of the

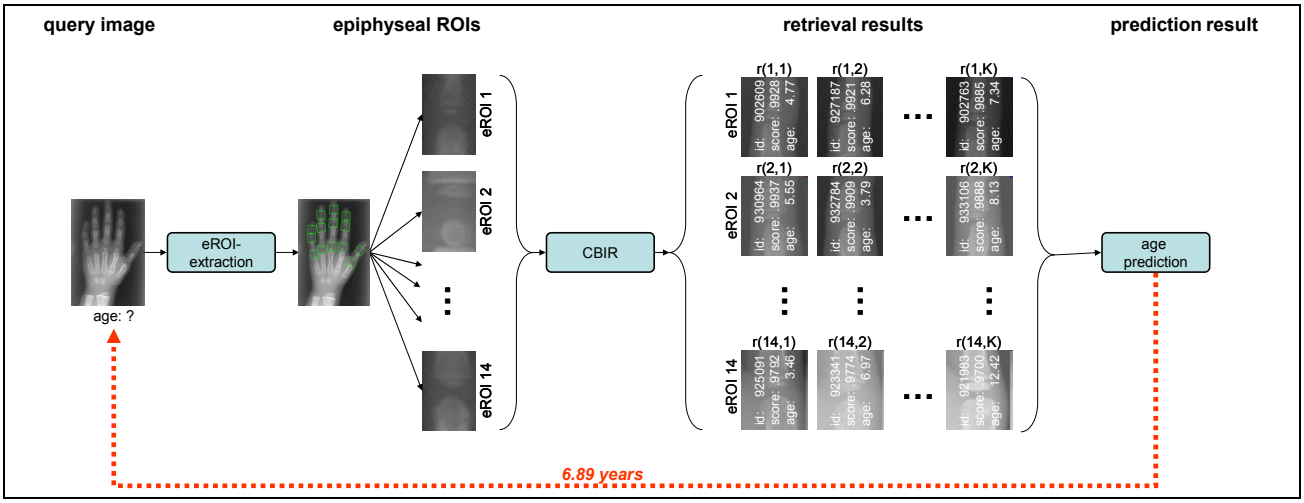

Figure 1: CBIR approach to bone-age assessment: For a hand of unknown age, the epiphyseal ROIs (eROIs) are

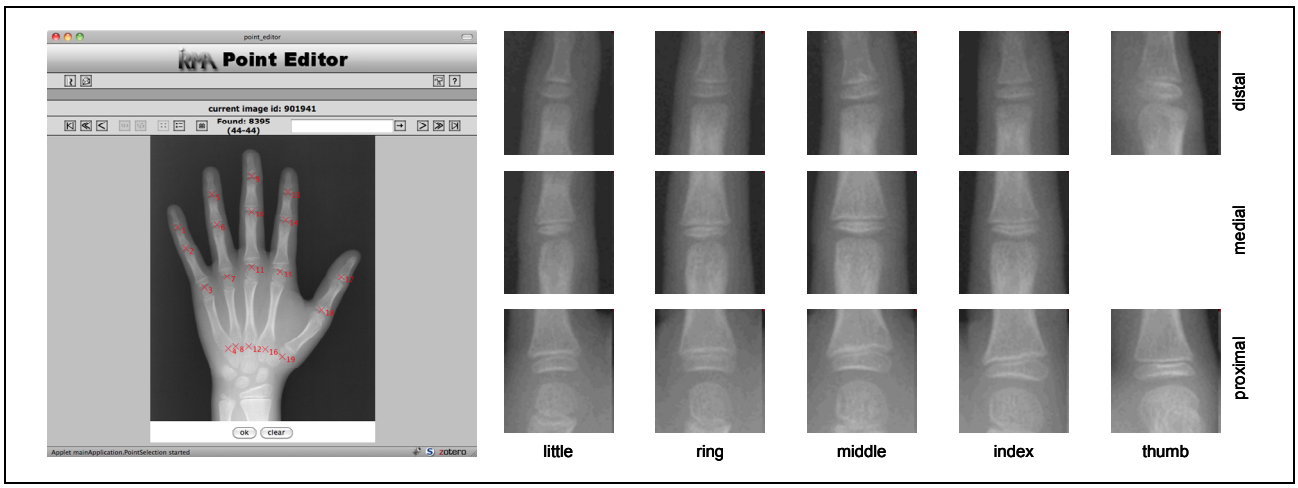

Figure 2: Web interface for the manual positioning of epiphyseal centers (left) and corresponding extracted eROIs (right)

surrounding bones. In the case of manual positioning of the eROI centers, the labels are assigned in the order of the center definitions, although an automatic assignment by the relative positions of the center coordinates would also be possible. In this work, the eROIs were set manually and stored in the retrieval database to establish a ground truth for the conducted experiments.

2.2 eROI Similarity

In the second stage, the extracted eROIs of a new image are to be compared by similarity to the ones stored in the retrieval database. For this purpose, the cross-correlation function (CCF) is chosen, as it is easy to compute and as it normalizes image brightness, which is a common cause of variability found in medical images. The eROI extraction as described above ensures that orientation and scaling are not an issue.

The similarity of a query eROI q to an eROI p in the retrieval database is therefore computed by:

$$D_{\mathrm{CCF}}(p,q) = \max_{|m|,|n|\le d} \left\{ \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(p(x-m,y-n)-\overline{p}\right)\left(q(x,y)-\overline{q}\right)}{\sqrt{\left(\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(p(x-m,y-n)-\overline{p}\right)^2\right)\cdot\left(\sum_{x=1}^{X}\sum_{y=1}^{Y}\left(q(x,y)-\overline{q}\right)^2\right)}} \right. \tag{1}$$

The variables $\overline{p}$ and $\overline{q}$ denote the respective mean gray values of p and q. In our experiments, we use 16x16 scaled versions of the eROIs, i.e. X=Y=16, and a warp range of d=2 to determine the maximum correlation.

2.3 Retrieval Result

In order to restrict comparisons to anatomically corresponding regions, only eROIs with identical labels are compared, i.e., $\lambda(p)=\lambda(q)$, $\forall p,q$ in (1). For each query eROI q of an image of unknown age, the similarity to all eROIs of the same label in the retrieval database is computed in the CBIR process. The result of the third stage in the age assessment process provides a sorted similarity list $\partial(q,k)$ for each query eROI q, where k constitutes the k^{th}-similar eROI in the database according to (1).

2.4 Age Estimation

In the final stage, the retrieved eROIs and their corresponding ages need to be combined to approximate likely ages for the single query eROIs and for the complete hand. Two approaches to build the age estimate from the results of the previous stages are investigated: The first method weights each eROI identically, independent of its anatomical position:

$$a_{\mathrm{local}}^{\mathrm{predict}}(h) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{\sum_{k=1}^{K}\partial(r,k)}\sum_{k=1}^{K}a^{\mathrm{known}}(r,k)\partial(r,k)\right) \tag{2}$$

where $a^{\mathrm{known}}(r,k)$ denotes the known age of the k^{th}-similar eROI to the current eROI r of the query image of a hand h. The similarity scoring for the k^{th} similar eROI to eROI r is provided by the sorted list $\partial(r,k)$. K and R denote the number of considered best-ranked retrieval results and the number of considered eROIs, respectively. Although strictly speaking, $a^{\mathrm{known}}(r,k)$ and $\partial(r,k)$ both depend on h, the h has been omitted in the notation in favor of improved readability.

In (2), the age is predicted for each eROI separately and the resulting age is determined by summing up the predicted ages and normalizing the sum by the number of eROIs, i.e. each eROI receives the same weight. Yet the second method of global similarity normalization does not consider the age for the eROIs separately and then concludes the hand age, but instead it weights each retrieval result on a global similarity scale, in order to avoid a high influence by eROIs with relatively low cross-correlation:

$$a_{\mathrm{global}}^{\mathrm{predict}}(h) = \frac{1}{\sum_{r=1}^{R}\sum_{k=1}^{K}\partial(r,k)}\sum_{r=1}^{R}\sum_{k=1}^{K}a^{\mathrm{known}}(r,k)\partial(r,k) \tag{3}$$

2.5 Quality Measures

The overall quality of the experiments is assessed by the mean and standard deviation of the absolute error of all age assessments per hand, despite using the mean absolute error on a per eROI basis:

$$\mu_{err} = \frac{1}{H}\sum_{h=1}^{H}\left|a^{actual}(h) - a^{predict}_{\{local,global\}}(h)\right|$$
$$\sigma_{err} = \sqrt{\frac{1}{H}\sum_{h=1}^{H}\left(\left|a^{actual}(h) - a^{predict}_{\{local,global\}}(h)\right| - \mu_{err}\right)^2} \quad (4)$$

H denotes the number of hands, i.e. the number of age estimations.

2.6 Experimental Setup

The retrieval engine is part of the IRMA system[9,10], which easily allows the configuration, execution, and analysis of large scale image retrieval experiments. All components used for the presented experiments are realized as a module. Experiments are set up as modular networks in combination with their parameters allowing easy re-use of the methods and of the results for future experiments.

The retrieval database contains the hand radiographs of the first cycle of data collection from the USC hand atlas along with the patients' ages[6]. One duplicate image has been removed from the image, therefore leaving 1,102 hand radiographs for our evaluation. The retrieval database is enriched by the centers and extracted image regions for the 14 eROIs between metacarpals and distal phalanges for each radiograph.

The major goal of all experiments is the proof of concept for assessing the bone-age by a relatively simple CBIR approach rather than by highly sophisticated and possibly over-specialized solutions with poor generalization abilities on untrained data.

In the experiments, the influences of the following aspects have been considered:

1. the number *R* of eROIs used for comparison,
2. the number of considered most-similar neighbors *K* of a query eROI, and
3. the normalization methods for age prediction.

For each aspect, leaving-one-out experiments on all radiographs are performed. Concerning the number of eROIs, all individual eROIs are used at first individually to determine their prediction potential. Afterwards, sets of the (according to the absolute error) best six and best seven eROIs, i.e., about half of all eROIs, and subsequently the full set of all 14 eROIs is examined. For the number of retrieved references, the results are compared for $K = 1$ to 10, 20, 30, and 100. All experiments are conducted for both normalizations. Finally, a conclusion on the overall performance of the CBIR approach is drawn.

3. RESULTS

The resulting output for an individual query of one hand is illustrated in Figures 3.a) and 3.b). At first, the selected hand to be analyzed is shown together with the corresponding query eROIs. These are fed to the CBIR engine which returns the list of the (in this example) five most similar matches for each eROI and their known ages. The age estimation according to equations (2) and (3) is displayed together with the (in this case) known age of the hand. For better relevance feedback, all similarity scores according to (1) and all known ages of the returned eROIs in the database are provided as well as the eROI images themselves and the corresponding hand images (Figure 3.b). A click into one of the displayed thumbnails opens the corresponding full resolution image for further analysis. As the retrieval is part of the very modular IRMA system[10], the involved distance and age computation functions as well as all involved parameter settings can be easily exchanged for future experiments.

3.1 Number of eROIs

If only a single eROI is used for the CBIR bone age estimation, mean absolute error rates range between 1.21 and 1.99 years and standard deviations occur between 0.851 and 1.595 years. The best results for each eROI are listed in Tab. 1. The best individual eROI is the proximal eROI of the middle finger (11). As evident in Fig. 4 (middle), the other

proximal eROIs (IDs 15, 7, 3, and 5) are next in the ranking. In general, the ranking degrades in the distal and less central directions. This also holds for the ranking of the standard deviation (Fig. 4, right).

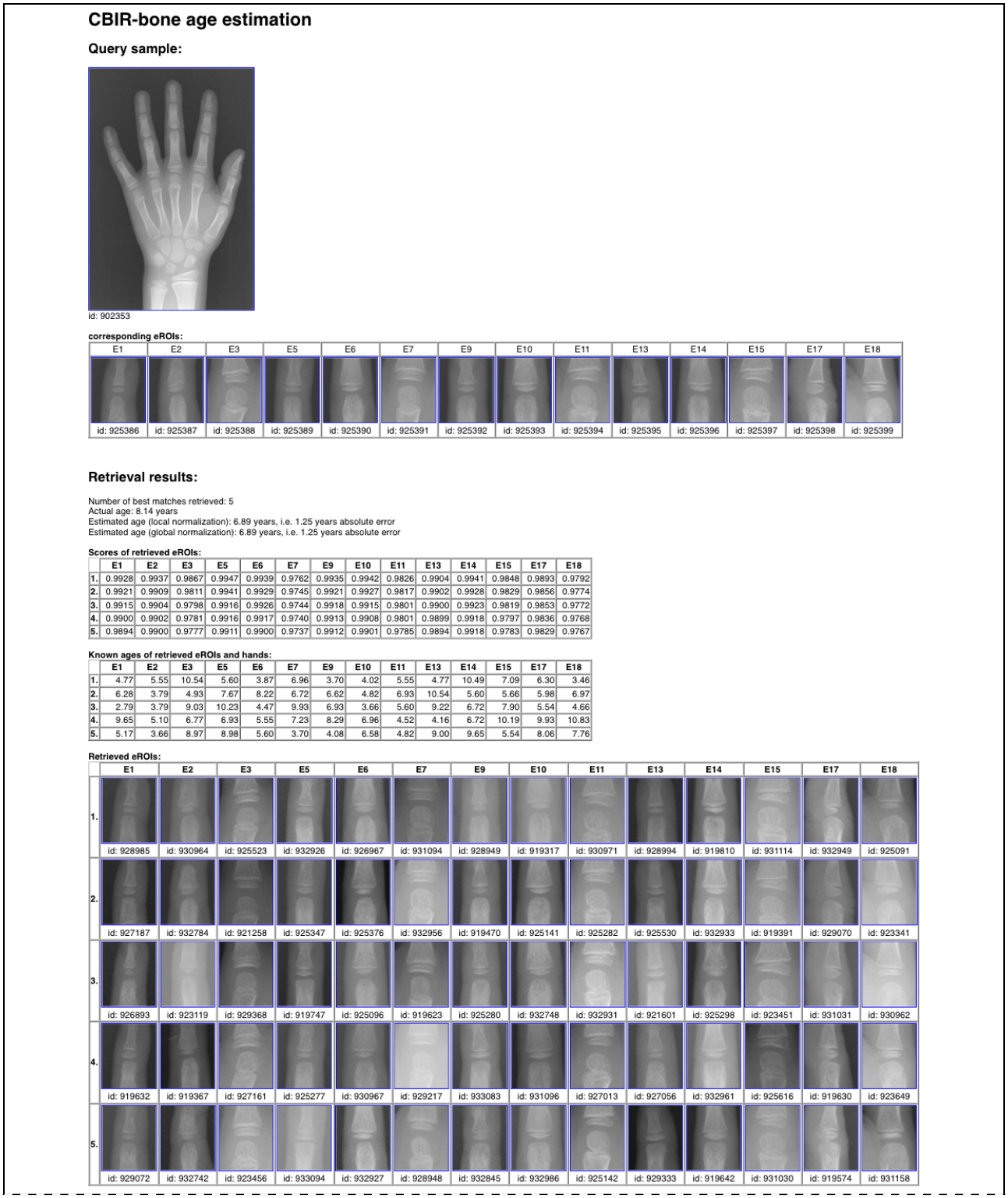

Figure 3.a): Example result with relevance feedback for one hand radiograph. The output is continued in Figure 3.b).

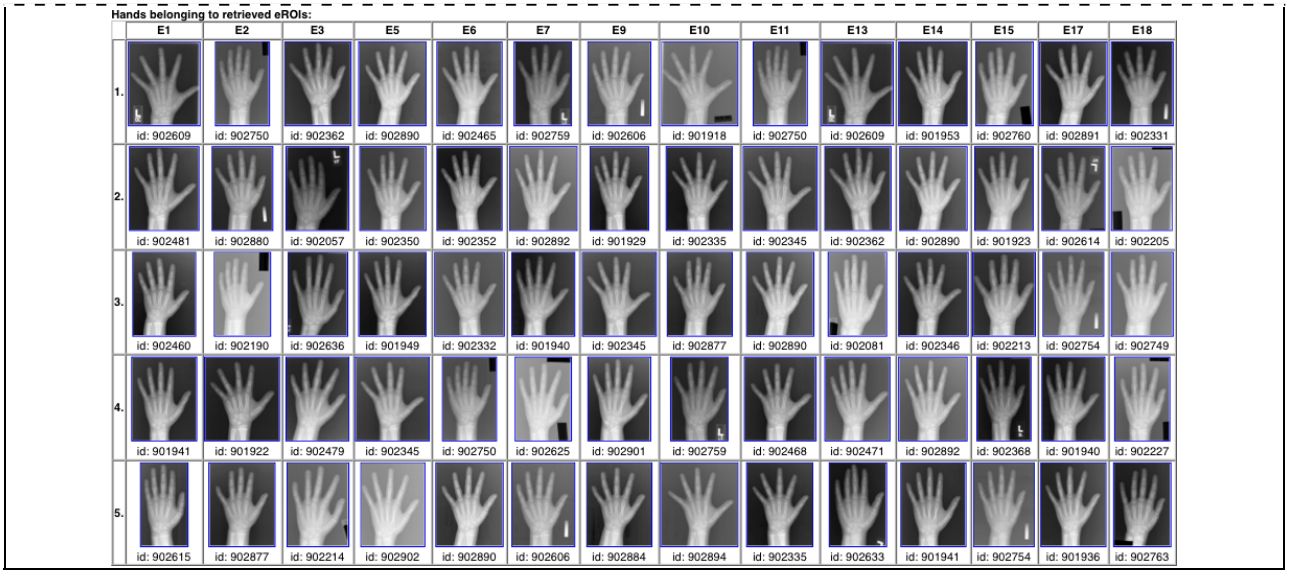

Figure 3.b): Example result with relevance feedback for one hand radiograph (continued). These are the hands belonging to the five most similar matches for each query epiphyseal region (Figure 3.a)).

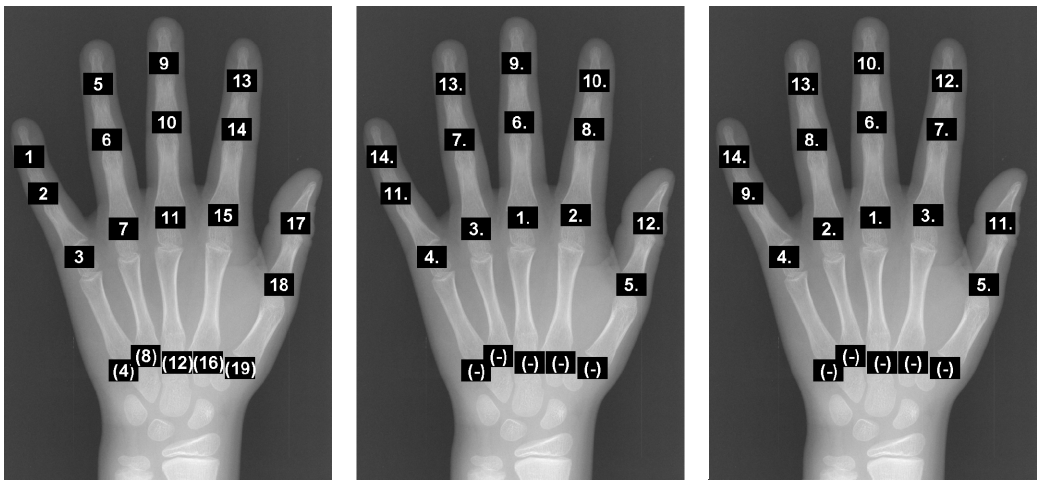

Figure 4: eROI-IDs (left) and ranking according to minimum mean (middle) and standard deviation (right) of the absolute error for each eROI as obtained from Table 1. IDs in brackets are not used in the experiments.

To assess the benefit of an increasing number of eROIs, the ranking is used to build subsets of the individually best six and best seven eROIs of lowest mean absolute error and compare their quality measures to the single eROIs and the full set of eROIs. This comparison as well as the influence of the number of neighbors is provided in Tables 2 and 3 for the local and global normalization. In general, the best six eROIs outperform the other sets. The best result of the seven eROIs is also better than the best individual eROI, while the best result using all eROIs is slightly below the best two individual eROIs. Of all experiments, the best mean absolute error is as low as 1.156 independent of the normalization and is obtained with the set established from the best six individual eROIs.

Table 1: Best results for individual eROIs: for each eROI, the minimum absolute prediction error in years (left) and the minimum standard deviation of the absolute error (right) are given along with the number of considered best retrieval references *K (K-NN)*. By definition, local (2) and global normalization (3) are identical in this case since *R*=1.

eROI	min. error	*K*
1	1.990	7
2	1.743	5
3	1.438	9
5	1.786	8
6	1.545	10
7	1.316	9
9	1.655	10
10	1.501	6
11	**1.209**	10
13	1.739	5
14	1.573	9
15	1.243	9
17	1.757	7
18	1.441	9
min	1.209	
max	1.990	

eROI	min. std.deviation	*K*
1	1.595	7
2	1.314	8
3	1.014	10
5	1.439	7
6	1.206	9
7	0.926	20
9	1.380	20
10	1.146	6
11	**0.851**	30
13	1.439	5
14	1.196	9
15	0.930	20
17	1.425	10
18	1.097	10
min	0.851	
max	1.595	

Table 2: eROI-set performances for increasing k-NN and local normalization:
Shown in bold are the minimum values for each set. The overall minimum is highlighted.

	Local Normalization					
	6 best eROIs		7 best eROIs		all eROIs	
K	mean	std.dev.	mean	std.dev.	mean	std.dev.
1	1.2038	0.9482	1.1887	0.9346	1.2674	0.9307
2	1.1631	0.8806	**1.1604**	0.8750	**1.2649**	**0.9130**
3	**1.1563**	0.8873	1.1624	0.8872	1.2875	0.9316
4	1.1752	0.8707	1.1845	0.8714	1.3122	0.9368
5	1.1820	0.8677	1.1945	0.8731	1.3269	0.9481
6	1.1908	0.8581	1.2034	**0.8656**	1.3446	0.9546
7	1.1960	**0.8570**	1.2111	0.8692	1.3576	0.9652
8	1.2047	0.8609	1.2194	0.8723	1.3675	0.9736
9	1.2101	0.8650	1.2265	0.8755	1.3787	0.9801
10	1.2196	0.8631	1.2355	0.8739	1.3880	0.9869
20	1.2788	0.8797	1.3040	0.8995	1.4814	1.0463
30	1.3294	0.8959	1.3583	0.9228	1.5524	1.0991
50	1.4086	0.9316	1.4436	0.9676	1.6689	1.1792
100	1.5692	1.0293	1.6125	1.0748	1.8859	1.3396
min	1.1563	0.8570	1.1604	0.8656	1.2649	0.9130
max	1.5692	1.0293	1.6125	1.0748	1.8859	1.3396

Table 3: eROI-set performances for increasing k-NN and global normalization:
Shown in bold are the minimum values for each set. The overall minimum is highlighted.

	Global Normalization					
	6 best eROIs		7 best eROIs		all eROIs	
K	mean	std.dev.	mean	std.dev.	mean	std.dev.
1	1.2040	0.9513	1.1890	0.9380	1.2685	0.9439
2	1.1630	0.8783	**1.1612**	0.8800	**1.2655**	**0.9160**
3	**1.1564**	0.8849	1.1631	0.8897	1.2880	0.9326
4	1.1753	0.8683	1.1854	0.8750	1.3128	0.9387
5	1.1820	0.8633	1.1954	0.8758	1.3272	0.9470
6	1.1905	0.8525	1.2043	**0.8685**	1.3448	0.9524
7	1.1957	**0.8502**	1.2121	0.8717	1.3577	0.9626
8	1.2044	0.8535	1.2202	0.8734	1.3675	0.9709
9	1.2099	0.8576	1.2275	0.8767	1.3787	0.9768
10	1.2193	0.8560	1.2366	0.8766	1.3880	0.9845
20	1.2785	0.8734	1.3054	0.9035	1.4816	1.0449
30	1.3294	0.8905	1.3600	0.9271	1.5527	1.0977
50	1.4088	0.9274	1.4456	0.9722	1.6698	1.1788
100	1.5702	1.0282	1.6153	1.0807	1.8876	1.3415
min	1.1564	0.8502	1.1612	0.8685	1.2655	0.9160
max	1.5702	1.0282	1.6153	1.0807	1.8876	1.3415

3.2 Number of Neighbors

As apparent in Fig. 5, an increasing number *K* of considered retrieved references leads to an increase in the mean error and standard deviations. The details are provided in Tables 2 and 3. The mean absolute error is minimal on all sets for *K*=2 or *K*=3. The standard deviation for the sets of the six or seven best eROIs is the lowest for *K*=6 or *K*=7, respectively. For the full set, *K*=2 is optimal. Concerning the results of individual eROIs, the best result for the mean error is observed for *K*=10, and for the standard deviation for *K*=30. The mean of *K* for the best error results of the individual eROIs is 8, for the best standard deviations it is 12.

3.3 Normalization

For the experiments with single eROIs, local and global normalization are identical by definition, as (2) and (3) are equivalent for *R*=1. For the other experiments the differences between the two normalizations are neglectable. Concerning the mean, the differences occur only in the third decimal, the standard deviations differ at most in the second decimal.

4. DISCUSSION

Even with just a single eROI, a fair prediction is possible. The best result is achieved by the epiphyseal area between metacarpal and proximal phalanx of the middle finger. The best result with six eROIs features a mean absolute age prediction error of only 1.156 years with a standard deviation of 0.885. Since (i) natural variations of the appearance have a range of approximately two years[11], (ii) inter-observer differences up to 2.5 years may occur[11], (iii) for the one to five year old patients, only five images are contained in the retrieval database per group, and (iv) the number of references for the newborns is even less, these results are very promising.

Concerning the number of considered closest retrieval references, more than ten references are not beneficial to the prediction error. This is to be expected since the USC hand atlas contains groups of 5 or 10 images for each age, gender and ethnic origin. Therefore, this observation is consistent to the fact that as more retrieval results are considered, these must belong to different ages and ethnic groups and hence are likely to distort the age estimation.

The selection of eROIs is dependent on the desired application. If only one predicted age is to be estimated, the "ideal" combination is sought. While the best result in our experiments has been obtained by the set of the six best eROIs of

Table 1 (left), this has been only a first rough attempt. While a brute force experiment of all 2^{14}=16,384 sets would be possible, this would deny the option of weighting the eROIs differently, for example to strengthen the influence of the eROIs of known good performance (Figure 4). Therefore, other methods such as support vector machines will be applied for optimization in future experiments[12].

The difference between the two normalization methods is negligible. This indicates that the retrieved correlation values for the separate eROIs are very similar and that good matches are found for all eROIs. If the similarity function (1) would lead to lower values for certain eROIs, their ages would be less regarded in the global normalization than the weighting of $1/R$ in the local case. In future experiments, the weighting of each eROI is likely to change according to prior training and e.g. depend on the occurred variances of the similarity function.

As far as the similarity or distance function for the retrieval engine is regarded, future experiments will include additional features and functions which have proven favorable in earlier work on medical image retrieval[9,10,13].

5. CONCLUSION

While other recently published approaches to automated bone age estimation rely on heuristics for the segmentation and measurements of (inter-)bone segments and need distinct heuristics for different ages, gender, and ethnic origin[6], our first approach by CBIR is able to make direct use of visual similarities without the need to explicitly formulate heuristics.

The presented results of the first prototype are promising. Besides further analysis on the distinct performance for different age, gender and ethnic groups, the eROI set combination needs comprehensive analysis. In addition, future work will focus on the automatic localization of the eROIs[8] and on user interfaces allowing quick input to the retrieval mechanism. If copyright questions can be settled, we will also set up a public web interface, where uploaded hand radiographs may be used to test the retrieval and age estimations for research (not diagnostic) purposes.

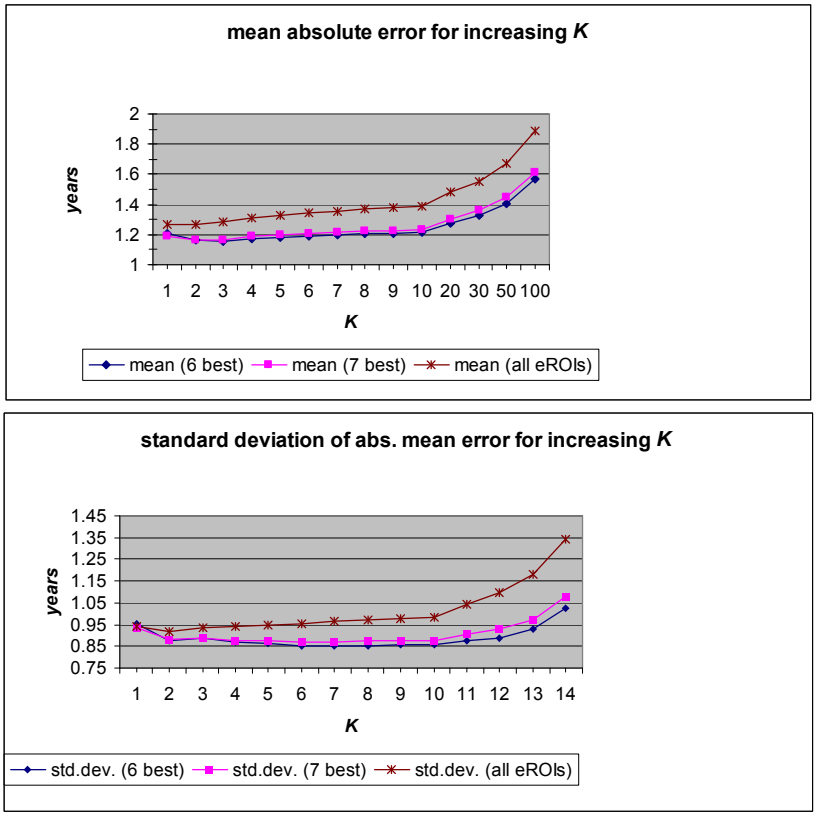

Figure 5: Influence of increasing k-NN on the quality measures for the combined sets and global normalization.

ACKNOWLEDGEMENT

This research was supported (in part) by the German Research Foundation (DFG), grant no. Le 1108/9.

REFERENCES

[1] Greulich WW, Pyle SI. Radiographic atlas of skeletal development of hand wrist. Stanford University Press, Stanford CA. 1971.
[2] Tanner JM, Healy MRJ, Goldstein H, Cameron N. Assessment of skeletal maturity and prediction of adult height (TW3). WB Saunders, London. 2001.
[3] Martin-Fernandez MA, Martin-Fernandez M, Alberola-Lopez C. Automatic bone age assessment: A registration approach. Proc. SPIE 5032: 1765–76 (2003).
[4] Pietka E, Gertych A, Pospiech S, et al. Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal ROI extraction. IEEE Trans Med Imaging; 20(8): 715-29 (2001).
[5] Park KH, Lee JM, Kim WY. Robust epiphyseal extraction method based on horizontal profile analysis of finger images. Proc ISITC; 278–82 (2007).
[6] Gertych A, Zhang A, Sayre J, et al. Bone age assessment of children using a digital hand atlas. Computerized Medical Imaging and Graphics; 31(4-5):322–31 (2007).
[7] Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging;28(1):52-66 (2009).
[8] Fischer B, Brosig A, Deserno TM, Ott B, Günther RW. Structural scene analysis and content-based image retrieval applied to bone age assessment. Proc. SPIE 7260: 041-11 (2009).
[9] Lehmann TM, Güld MO, Thies C, et al. Content-based image retrieval in medical applications. Methods Inf Med; 43(4): 354-61 (2004).
[10] Güld MO, Thies C, Fischer B, Lehmann TM. A generic concept for the implementation of medical image retrieval systems. Int J Med Inform; 76(2-3): 252-9 (2007)
[11] Zhang A. A computer-aided-diagnosis (CAD) method combining phalangeal and carpal bone features for bone age assessment of children. Dissertation at University of Southern California; 2007
[12] Thies C, Schmidt-Borreda M, Seidl T, Lehmann TM. A classification framework for content-based extraction of biomedical objects from hierarchically decomposed images. Proc. SPIE 6144:559-68 (2006).
[13] Deselaers T, Keysers D, Ney H. Features for image retrieval: An experimental comparison. Information Retrieval; 11(2):77-107 (2008).