Support Vector Machine Classification based on Correlation Prototypes applied to Bone Age Assessment

Markus Harmsen, Benedikt Fischer, Hauke Schramm, Thomas Seidl,
and Thomas M Deserno, *Senior Member, IEEE*

***Abstract*—Bone age assessment (BAA) on hand radiographs is a frequent and time consuming task in radiology. We present a method for (semi)automatic BAA which is done in several steps: (i) extract 14 epiphyseal regions from the radiographs, (ii) for each region, retain image features using the IRMA framework, (iii) use these features to build a classifier model (training phase), (iv) evaluate performance on cross validation schemes (testing phase), (v) classify unknown hand images (application phase). In this paper, we combine a support vector machine (SVM) with cross-correlation to a prototype image for each class. These prototypes are obtained choosing one random hand per class. A systematic evaluation is presented comparing nominal- and real-valued SVM with k nearest neighbor (kNN) classification on 1,097 hand radiographs of 30 diagnostic classes (0 – 19 years). Mean error in age prediction is 1.0 and 0.83 years for 5-NN and SVM, respectively. Accuracy of nominal- and real-valued SVM based on 6 prominent regions (prototypes) is 91.57% and 96.16%, respectively, for accepting about two years age range.**

***Index Terms*—Bone Age Assessment, Support Vector Machine, Classification, Cross Correlation, Prototypes**

I. Introduction

BONE AGE ASSESSMENT (BAA) usually is based on hand radiographs and constitutes a frequent as well as time consuming task in diagnostic radiology. The bone age reflects the skeletal maturity and indicates disease when differing significantly from the chronological age. For BAA, two conventional methods are common. In the method developed by Greulich & Pyle (GP) [1], the radiologist visually compares all bones of the left hand with a standard atlas and assesses the bone age according to his perception. Applying the method of Tanner & Whitehouse (TW) [2], only a certain subset of bones is considered and described individually with respect to the epiphyseal gap and shape. The radiologist classifies regions into several stages, which are described literally and do not rely on visual comparison with an atlas. The bone age is calculated by scoring and adding up the scores of classified regions. Both conventional methods have the drawback of being highly subjective, like especially GP, or being relatively complex as TW. Furthermore, the gender has to be treated with care, since the growth spurts differ significantly for girls and boys [3], resulting in an even more complex process. For these reasons, an automated BAA is definitely preferable.

Already in 1996, Al-Taani et al. [4] reported an automatic BAA approach based on a point distribution model of 130 feature points. During training, examples of bones from each class are collected so that the allowable shape deformations are learned. The system was tested classifying two bones of the third finger, the distal and the middle phalanx. A set of 120 images of nine age classes was used for evaluation, and the classification rates were 70.5% and 73.7%. About 50% of the errors were only one stage off.

In 2001, Pietka et al. comprehensively reviewed early attempts on BAA and – following a promising strategy – developed semantic (heuristic) features for BAA by measuring the gap between metaphyses and diaphyses [5]. A multiple-step processing pipeline was suggested: (i) preprocessing for orientation correction and background removal, (ii) localization of phalangeal tips by superimposing wedge functions over the hand image, (iii) detection of phalangeal long axis, (iv) extraction of epiphyseal regions of interest (eROI)s, and (v) determination of global size and distance measures (epiphyseal gap). Discrimination power was proven based on 200 hand radiographs of limited age ranges (male $<$ 14y, female $<$ 12y). Providing rather a solid view on fundamental principals in BAA, age computation was not performed.

An approach that seems to be inspired by Pietka et al. has been presented by Martin-Fernandez et al. [6]. First, the authors locate regions of interest (ROIs) as landmarks in relevant hand radiographs and describe the finger positioning with a wire model. This model is matched with a reference model built from a template hand and is therefore directly used for comparison. For registration, several affine transforms are applied to the entire hand as well as to individual fingers, and mutual information is used in a second stage of registration. Experiments on age prediction, however, were not performed.

An attempt using fuzzy methodology has been introduced by Aja-Fernandez et al. [7]. A decision tree is used as a straightforward representation of rules given from the TW method. Six computational features are derived. For the rule-based system, large training data is avoided and experiments are reported on 85 diagnosed radiographs from girls. Accuracy of classification is 97.6% and 95.3% for ulna and proximal phalange I, respectively. A similar technique based on artificial

M. Harmsen, B. Fischer, and T. M. Deserno are with the Department of Medical Informatics, RWTH Aachen University, 52057 Aachen, Germany.

H. Schramm is with the University of Applied Sciences, Kiel, Germany.

T. Seidl is with the Data Management and Data Exploration Group, RWTH Aachen University, Germany.

Corresponding author: T. M. Deserno, Institut für Medizinische Informatik, Universitätsklinikum Aachen, Pauwelsstr. 30, 52074 Aachen, Germany, fon: +49 241 80 88793, fax: +49 241 80 33 88793, mail: deserno@ieee.org

Manuscript received April 20, 2012

neural networks was published by Bocci et al. [8], where 20 individual bones and regions of TW2 method were classified and based on 120 images for training and 40 images for testing, a maximum error of 1.4 years is reported.

Such approaches are based on more or less direct use of GP or TW method. In contrast, Hsieh et al. [9] and Chang et el. [10] extract radiographic features of phalanges or carpal bones and analyze them by computerized shape and area description using a classifier. Evaluation is performed on larger data. The authors use a private database of 909 radiographs, 465 male pattern of 12 groups aged from 2 to 8 years, and 444 female patterns of 14 groups, aged from 2.5 to 10 years. The lowest mean error was 0.5 years in females. Based on pure radiolucency analysis, a mean error of 1.5 years is reported [10].

The idea of using eROIs in a pattern-based approach was suggested by Kim & Kim [11]. After segmenting nine relevant eROIs automatically, discrete cosine transform and linear discriminant analysis are applied for BAA. In contrast to Pietka et al., this approach does not require heuristic feature extraction. A private dataset of 396 radiographs (93 male, 303 female) was collected to report an average error of 0.6 years.

Further standardization of experiments and improved comparability of approaches was achieved by Gertych et al. [12] when publishing a reference database for BAA computation. This digital hand atlas has been established at the University of Southern California (USC) and therefore is referred to as USC hand atlas. It is composed of 19 age classes, four ethnic groups and both genders, with ten to forty images carefully selected into each individual class, summing up to a total of 1,097 digitized radiographs that still are publicly available[1].

A method based on content-based image retrieval (CBIR) of eROI patches extracted from USC data has been presented recently by Fischer et al. [13]. Using all 19 eROIs of the query image, similar patches are retrieved from the database using the Image Retrieval in Medical Applications (IRMA[2]) framework [14], [15]. The retrieval approach is based on the k nearest neighbor (kNN) method and – as a novelty due to the metric nature of age – BAA is calculated algebraically from a weighted sum of reference ages linked to the most similar patterns. An error rate of 0.97 years is reported.

Currently, the leading commercial product for BAA is BoneXpert[3]. The BoneXpert approach uses an active shape model to estimate bone structures and directly follows the methods of GW and TW to compute the bone age with a mean error of 0.72 years computed on an extract of USC data [16], [17].

Although a reference database is available, it is still recognized insufficiently, and – if experiments have been published at all – some groups investigated only a specific age range. For instance, BoneXpert is focused on 2 – 17 years. Depending on the gender, the method of Hsieh et al. considers ranges of 1 – 8 years or 2.5 – 10 years.

The results of Fischer are based on the entire age range of USC data but non-commercial approaches might not be optimized completely. For instance, some weak points have been identified in the method of Fischer et al. [13]: (i) a fixed amount of k neighbors is used for classification, which – depending on the dataset – may not be optimal; (ii) the classification considers the eROIs to be independent and uses only a (weighted) age average of these regions for age determination; (iii) the gender is disregarded completely, although male and female growth spurts differ significantly; (iv) BAA is performed strictly data-driven disregarding any medical knowledge of growth spurts; and (v) computation is expensive, since the cross-correlations between the test image and all existing references are determined.

In the past few years, the support vector machine (SVM) has been introduced into many classification fields and has demonstrated state-of-the-art performance. For example, a combination of SVM with CBIR has been successfully applied to detect malign structures in mammography [18]. Despite of their broad applicability, some essential problems have to be addressed when using SVM. Besides the fundamental choices of features, attributes and parameters, the SVM only applies to binary problems, i.e., a classification into more than two classes requires several SVMs, and the class size has to be chosen carefully.

In this work, our method on automatic CBIR-based BAA [13] is extended to class prototypes with SVM classification. It is evaluated critically with respect to the standard kNN classifier.

II. Materials and methods

Within the IRMA framework, global, local, and structural features are supported to describe the image, an eROI, or a constellation of eROIs, respectively [19]. In the following, we describe the eROI and feature extraction, prototype generation, classification using kNN and SVN, age computation, and our design of experimental validation.

A. eROI extraction

Automatic extraction of eROIs has been presented previously [19]–[21]. Essentially, a structural prototype is trained, where the phalanges and metacarpal bones are represented by nodes, and location, shape as well as texture parameters are modeled with Gaussians. In a recently published web interface [22], a manual procedure is also offered, where the user quickly hits the centers of relevant epiphyses. Thereafter, 14 eROIs are extracted (Fig. 1), rotated and geometrically aligned into an upright position, and inserted into the IRMA database with reference to the according hand radiography. Hence, all patterns are in upright position and uniformly scaled, disregarding individual finger positioning in the original radiograph (Fig. 2).

B. Similarity measure

The cross-correlation function (CCF) is (i) easy to compute, (ii) robust regarding the radiation dose, (iii) robust regarding translation for a given range, (iv) normalizes intensity, and (v) has already been used successfully in BAA tasks [13]. Disadvantages of CCF, such as sensitivity to rotation and scaling,

[1] http://www.ipilab.org/BAAweb/
[2] http://irma-project.org
[3] http://www.bonexpert.com

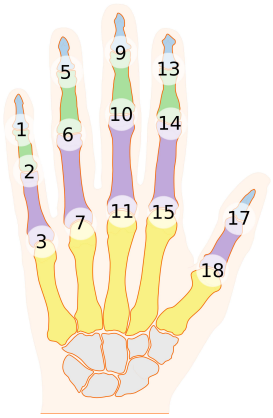

Fig. 1. eROIs and corresponding numbers as used in this paper.

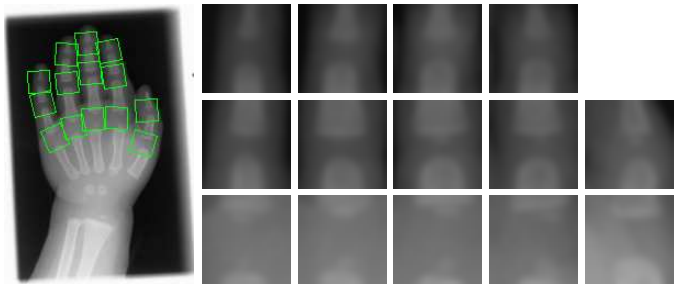

Fig. 2. eROI extraction. From distances of eROIs and orientations of eROI interconnections, a geometric model is derived and used to normalize the eROIs regarding rotations and scalings.

are less important in our framework since such alternations are corrected in the extraction process based on the constellation of epiphyses centers. The similarity between two images – more specifically two eROIs – a and b is hence computed by:

$$S_{\mathrm{CCF}}(a,b) = \max_{|m|,|n|\leq d} \left\{ \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y} A \cdot B}{\sqrt{\sum_{x=1}^{X}\sum_{y=1}^{Y} A^2 \cdot \sum_{x=1}^{X}\sum_{y=1}^{Y} B^2}} \right\} \quad (1)$$

with

$$A = a(x-m, y-n) - \overline{a}$$
$$B = b(x,y) - \overline{b}$$

where $\overline{a}$ and $\overline{b}$ denote the mean gray values of a and b, respectively. The position ranges m and n of correlation are limited to d, i.e., $m, n \leq d$. According to [13], we set $d = 2$ and use a scaled version of the eROIs with 32 $\times$ 32 pixels.

C. Class prototypes

To address the problem of class size, the data is grouped according to the growth spurts. Using the ontology defined by Gilsanz & Ratib [23], reference ages are quantized in steps of 2m, 4m, 6m, and 12m for the intervals [8m ...20m), [20m ...28m), [2.5y ...6y), and [6y ...18y], respectively, where m and y denote month and year, respectively. This creates a set of 29 classes with four different ranges. A 30$^{\mathrm{th}}$ class for bone ages > 18 years was added (Table I). Notice that gender information is not used so far for class building. In Figure 3, a subset of prototypes for a specific region (no. 15) is shown.

TABLE I
AGE CLASSES AND THEIR CORRESPONDING AGE RANGE.

Class	Age range in years	Class	Age range in years
00	0.00 - 0.66	15	05.00 - 05.50
01	0.66 - 0.83	16	05.50 - 06.00
02	0.83 - 1.00	17	06.00 - 07.00
03	1.00 - 1.16	18	07.00 - 08.00
04	1.16 - 1.33	19	08.00 - 09.00
05	1.33 - 1.50	20	09.00 - 10.00
06	1.50 - 1.66	21	10.00 - 11.00
07	1.66 - 2.00	22	11.00 - 12.00
08	2.00 - 2.33	23	12.00 - 13.00
09	2.33 - 2.50	24	13.00 - 14.00
10	2.50 - 3.00	25	14.00 - 15.00
11	3.00 - 3.50	26	15.00 - 16.00
12	3.50 - 4.00	27	16.00 - 17.00
13	4.00 - 4.50	28	17.00 - 18.00
14	4.50 - 5.00	29	18.00 - 99.00

D. Feature extraction

For each hand h and each region r, the CCF similarities are computed between the eROI image $I(h,r)$ and all corresponding prototypes $P(r,c)$, where $c \in \{0, 1, \cdots, 29\}$ represents the class label. The prototypes were chosen randomly. This yields:

$$\vec{F}_{\mathrm{CCF}}(h,r) = \begin{pmatrix} S_{\mathrm{CCF}}\left(I(h,r), P(r,0)\right) \\ S_{\mathrm{CCF}}\left(I(h,r), P(r,1)\right) \\ \vdots \\ S_{\mathrm{CCF}}\left(I(h,r), P(r,30)\right) \end{pmatrix} \quad (2)$$

For each region r of the hand radiograph h, a vector is obtained. The resulting feature vector:

$$\vec{F}(h) = \begin{pmatrix} g_f(h) \\ g_m(h) \\ \vec{F}_{\mathrm{CCF}}(h,1) \\ \vec{F}_{\mathrm{CCF}}(h,2) \\ \cdots \\ \vec{F}_{\mathrm{CCF}}(h,r) \end{pmatrix} \quad (3)$$

is composed of gender information g and all region-specific feature vectors $\vec{F}_{\mathrm{CCF}}(h,r)$. For female and male, we set $(g_f = 1, g_m = 0)$ and $(g_m = 1, g_f = 0)$, respectively. All other values are scaled to the range $[-1, +1]$ avoiding attributes in greater numeric ranges that dominate those in smaller ranges. A database is used to store $\vec{F}(h)$ and index all extracted features.

E. K-Nearest-Neighbor Algorithm

kNN is a simple method for classifying objects based on the k closest training examples in feature space. Since all computation is only done at classification time, kNN is known as a lazy learning classifier. A feature vector $x \in \mathbb{R}^n$, where n denotes the number of features, is associated to the class that is set by the majority of k most similar feature vectors, according to a distance function.

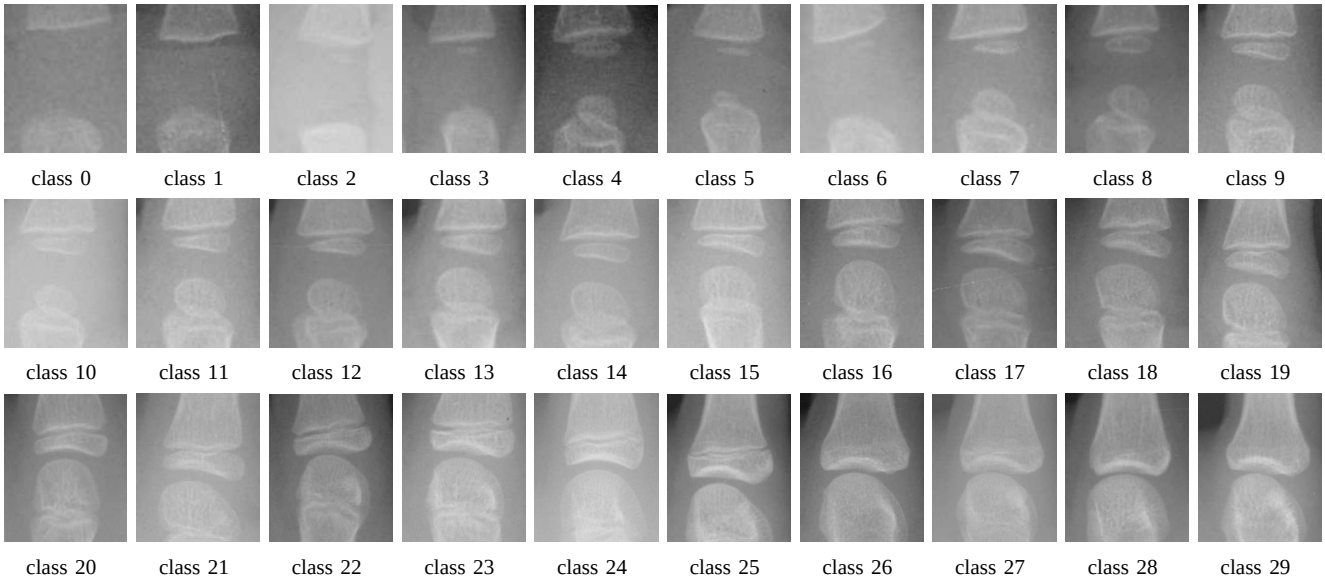

Fig. 3. Prototypes for region 15. The development of the epiphysis is clearly visualized.

In this paper, we use the Euclidean distance, which is defined for given vectors $\vec{p} = (p_1, \ldots, p_n)^T$ and $\vec{q} = (q_1, \ldots, q_n)^T$, where $\vec{p}, \vec{q} \in \mathbb{R}^n$, by:

$$d(\vec{p}, \vec{q}) = d(\vec{q}, \vec{p}) = \frac{1}{n}\sqrt{\sum_n (p_n - q_n)^2} \tag{4}$$

The kNN algorithm is easy to implement and by its nature a multi-class classifier. Especially for a large set of training data, however, it is recommended to integrate indexing structures to prevent a time consuming distance calculation for each training vector. Furthermore, the only parameter k has to be chosen wisely, since it may depend on the used dataset and has to balance a good discrimination with over-fitting.

F. Support Vector Machine

In contrast to kNN, the SVM is a binary classifier and uses a given set of training examples to create a model that can be used to classify new examples [24]. Basically, a hyperplane is calculated, which separates the samples towards a maximum margin. Using appropriate kernels, SVM copes with the classification of non-linear data by mapping the input space into a higher dimensional feature space.

In our method, the radial basis function is used as kernel [25] as we assume a non-linear relation between classes and features, as well as a small number of features. Since we have 30 classes, SVM is extended according to the "one-against-one" approach instead of the traditional "one-against-all" approach, due to its good performance and short training time [26].

G. Age computation

For a hand radiograph with unknown bone age x, the eROIs and features are extracted and the data instance vector $\vec{F}(x)$ is built. Both, kNN and SVM with the trained classificatory model are used to determine the bone age of the new radiograph. Based on the complex feature vectors, both classifiers return the suggested class c. The estimated bone age a for the class c is calculated as the arithmetic mean of upper and lower bound B of the age range, the prototype of c is corresponding to:

$$a = \frac{1}{2}\Big(B^u(c) + B^l(c)\Big) \tag{5}$$

H. Validation experiments

Using 1,097 images from the USC hand atlas [27], the class prototypes were selected randomly but kept fixed for all experiments. They represent the class samples. Afterwards, for each hand, the feature vector is generated by measuring the similarity to corresponding region class-prototypes, with and without considering the gender.

We then apply the same feature vectors for kNN and SVM, using k-fold cross-validation for SVM with $k = 5$ and the leaving-one-out cross-validation scheme for kNN. For SVM we use a fixed seed for the random function to ensure an equal class distribution. For each experiment, 5-fold cross-validation is applied. The optimal SVM parameters are computed by grid search.

To determine the optimal set of eROIs included in $\vec{F}(h)$, we apply a brute force method of simply building all possible subsets of all 14 relevant regions (no. 1 – 3, 5 – 7, 9 – 11, 13 – 15, 17, and 18 in Fig. 1) and run experiments on the $2^{14} = 16,384$ sets.

Based on the nature of experimental data, some fuzziness of adjacent classes is expected, which has already been reported by other researchers. Pietka et al. [5], for example, have used diameter-based features to show this effect, related to intrinsic class prototypes. For quantification purposes, this paper compares SVM with ordinal ranking (ordered classes) to SVM for regression (rSVM) providing real-valued output.

III. Implementation

Our implementation is completely written in C++, using the Qt framework version 4.7.3 [28] and a SQLite database

version 2.8.17 as data-backend [29]. The Qt framework allows easy image file reading and database access. SQLite has been chosen since it entirely runs from memory providing radiographs and features by standard SQL queries. The SVM is implemented by using the libSVM library version 3.11 [30], which offers a rich set of features like build-in cross-validation and multi-class handling. The implementation consists of mainly four parts:

1) *DatabaseLibrary* connects the database and abstracts from SQL to a hand-related layer. This library is used as a data provider to retrieve radiographic features. Furthermore, a simple feature storage is implemented.
2) *ImageLibrary* implements a fast image representation using arrays and pointer access including basic bitmap manipulation and CCF.
3) *PrototypeFeatures* executes the ImageLibrary to load all 30 prototype images into the memory and to calculate the CCF for all remaining hand radiographs and eROIs to each of the prototypes. The double values obtained are stored using DatabaseLibrary.
4) *BoneAge* is a console executable to perform the experiments. A set of parameters is used to configure the options: "Classificator (kNN/SVM); Regions; Include-Subsets; Use-Gender" and some output specific options. If "Include-Subsets" is used, each possible subset of "Regions" is used for classification and results are printed as comma separated values (CSV) lines. The output for an experiment contains "Regions; Hands; Correct classes; Accuracy; Error in [-2, 2]; Mean age error; Error variance (s^2)". The application also implements kNN and performs scaling on attributes before using libSVM. Furthermore, any experiment can be exported in a libSVM specific data-format, allowing grid search with the libSVM tool for parameter selection.

The framework is currently being merged with the IRMA BAA tool, that can be accessed at http://irma-project.org/onlinedemos_en.php (Fig. 4).

IV. Results

In terms of classification error rate (age class accuracy), the results for a single region for kNN range from $15\% - 25\%$ and $11\% - 28\%$ for $k = 1$ and $k = 5$, respectively. In terms of mean age error (age assessment accuracy), they range for individual regions from $1.24 - 2.26$ years and $1.01 - 2.47$ years, respectively. The SVM achieves an age class accuracy of $19\% - 34\%$ and an age assessment accuracy of $0.95 - 2.15$ years (Table II).

Both perform best only on a subset of regions, whereas the kNN uses a lower number of regions for the ten best sets – 3.5 regions on average. Region 11 is used always and if it is used exclusively, it yields the 3rd lowest age assessment error of 1.01 years (Table III). In contrast to kNN classification, SVM performs best on a larger number of 7.5 regions on average for the ten best results. Like the kNN classifier, the SVM tends to use region 11 (Table IV). The 2nd best SVM result uses a subset of 10 regions, reaching a mean age error of 0.83 years. kNN and SVM classification perform best on an age range of

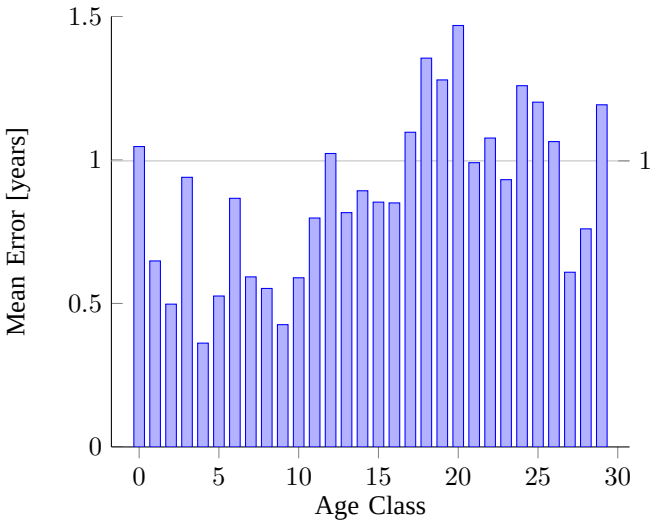

Fig. 5. kNN: Mean error itemized by class for regions $3, 6, 7, 11, 15$ and age in $[0, 18]$ years, distribution seems to be likewise SVM.

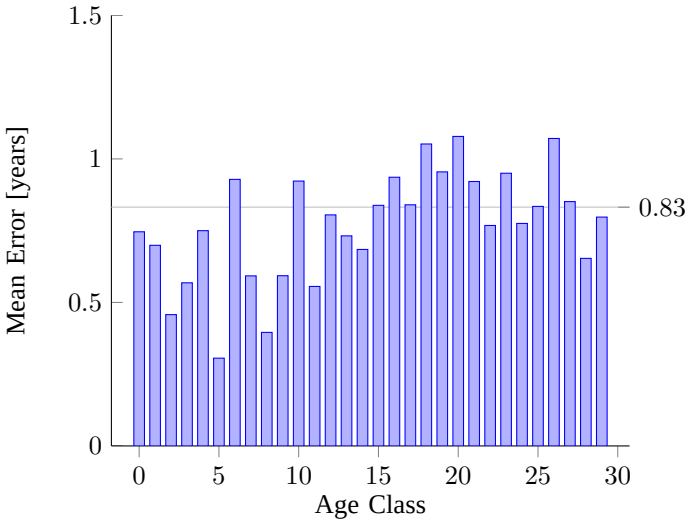

Fig. 6. SVM: Mean error itemized by class for regions $2, 6, 11, 13, 15, 18$ and age in $[0, 18]$ years.

$0 - 8$ years and getting worse especially in the range $8 - 16$ years. The mean class-specific errors are shown in Figs. 5 and 6 for kNN and SVM, respectively. Mean and variance are decreased clearly when using the SVM classifier, indicating its better performance. Also, one can note an imbalance according to the age classes. According to that larger differences in development of bones, the infant age classes (0 – 7 years) are easier to recognize.

The best experimental result for kNN (age class accuracy 26.71%, age assessment error 1.00 years) is obtained by a subset of 5 regions and $k = 5$. For SVM, the best result (accuracy 36.93%, mean error 0.83 years) is obtained using a subset of six regions. It outperforms both, kNN as well as the method by Fischer et al., who has reported a mean error of 0.97 years on the same data. However, correctness rates of 36% might be considered as inapplicable for routine use. Our age classes span 4 – 12 months. Allowing two classes of

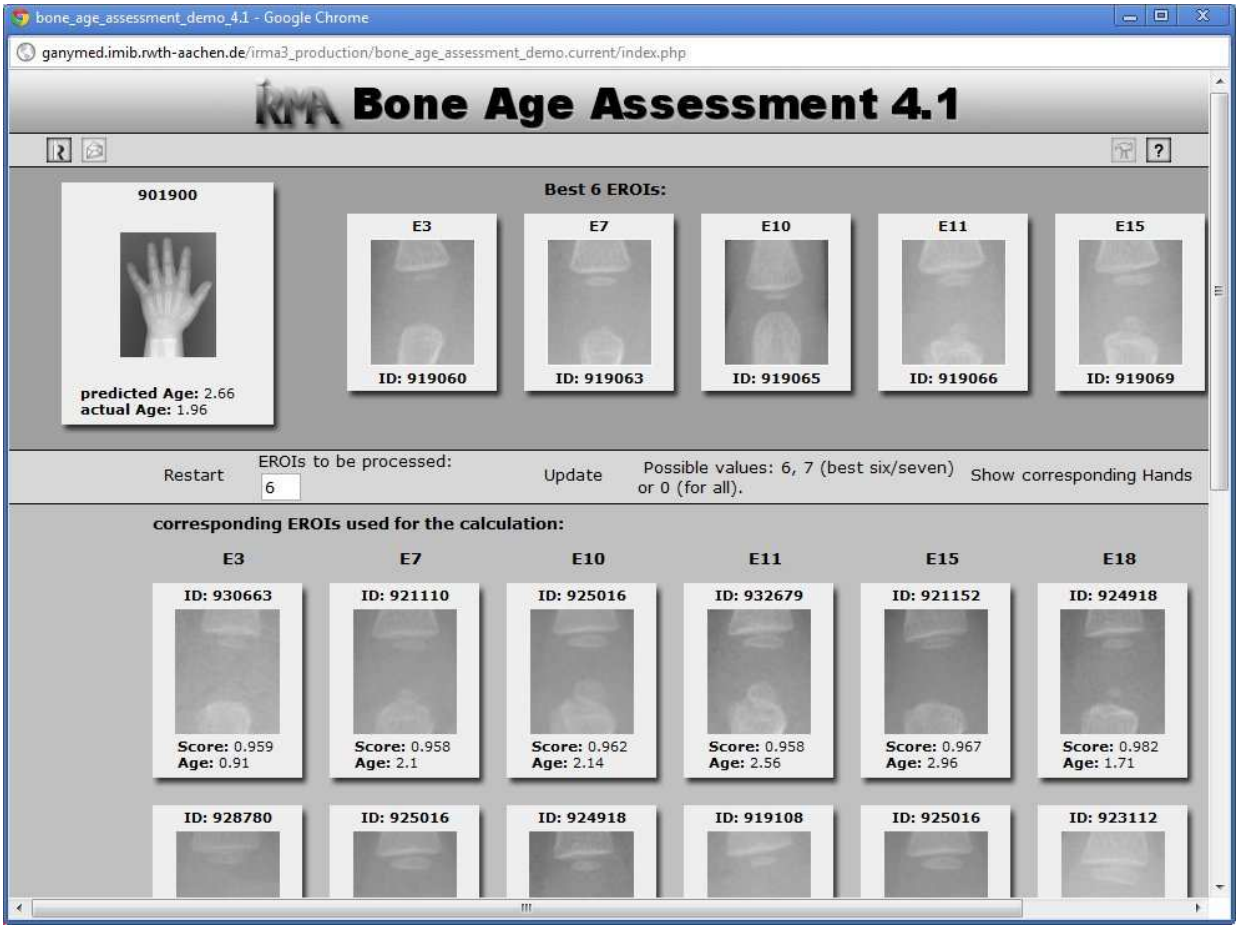

Fig. 4. Screenshot of the current IRMA Bone Age Assessment interface [22].

difference yields a range of up to two years, which corresponds to the human observer agreement that is manifested in the USC reference reading and that has been reported by others [5], [16]. Within this interval, kNN, SVM, and rSVM reach a performance of 89.79%, 91.57%, and 96.16%, respectively (Table V). Furthermore, rSVM increases correctness of class labeling to 58.86% as compared to 36.93% and 26.71% for SVM and kNN, respectively.

In comparison, BoneXpert reaches a mean error of 0.72 years on an age range from $2-17$ years. Projected on the same age range, this is still superior to our method with a mean error of 0.82 years (Table VI). However, the result with BoneXpert is obtained with 14 rejections due to several reasons such as bad image quality or abnormal bone morphology [17]. When disregarding the 14 worst radiographs from our experiments, the mean error yields 0.796 years. A direct comparison to the methods by Hsieh et al. [9], [10] is not possible, since they used private data.

V. Discussion

We have presented a novel method using CBIR from an atlas for computer-aided diagnosis (CAD) such as BAA. Improving previous concepts, we applied nominal- and real-valued and SVM using correlation-based features on class prototypes characterizing semantically defined age groups. The advantages of this method are full automation, robustness, and generality, since all a-priori knowledge is hosted in the reference database but not modeled into the image processing algorithms. In contrast to other approaches, our method neither needs semantic atlases such as the GP and TW methods nor semantic features as suggested by, for instance, Pietka et al. It is purely data-based using an image repository enriched with annotated readings. If such readings are considered as ground truth, the approach is applicable directly to other tasks of CAD, such as screening mammography, skin lesions, or tumor staging in general.

Region 11 was found most reliable. This finding corresponds to the medical TW and GP methods, where region 11 is also used. The middle finger is the largest and best developed bone and, hence, best contrasted in x-ray imaging. The overall classification accuracy seems to be low, but most mislabeled classes only have a class distance of 1 or 2 (Table V).

Using correlation prototypes, the number of comparisons needed for a single hand in the application phase is reduced from 1,097 [13] down to 30 (i.e., the number of prototypes), significantly improving the classification performance.

Further speedup is obtained from SQLite-based implementation. For instance, a complete 5-fold cross validation cycle is computed in 1.775 seconds (s) and 3.577 s for 6 and all 14 relevant regions, respectively. In comparison, a single kNN

TABLE II
EXPERIMENT OUTCOME FOR KNN/SVM AND SINGLE REGIONS

	kNN $k=1$			kNN $k=5$			SVM		
Region	Accuracy	Mean error	Error variance(s^2)	Accuracy	Mean error	Error variance(s^2)	Accuracy	Mean error	Error variance(s^2)
1	15.56	**2.90**	**7.81**	**11.06**	2.47	4.40	19.59	2.15	4.08
2	18.28	2.26	4.59	12.46	1.96	2.52	23.15	1.81	2.83
3	20.71	1.64	2.26	19.40	1.37	1.22	29.05	1.15	1.06
5	16.03	2.47	5.19	15.37	1.99	3.08	24.84	1.68	2.90
6	18.09	1.89	3.05	17.24	1.59	1.66	27.18	1.47	1.96
7	24.09	1.44	1.73	23.15	1.16	1.04	33.18	1.01	0.75
9	17.15	2.20	4.07	13.68	1.87	2.60	24.27	1.66	2.86
10	19.31	1.78	2.56	17.62	1.49	1.44	30.08	1.22	1.31
11	25.02	1.24	1.15	27.55	1.01	0.67	**34.40**	**0.95**	**0.65**
13	14.62	2.37	4.53	14.25	1.94	2.70	23.43	1.67	2.48
14	16.87	2.00	3.45	17.71	1.60	1.86	25.87	1.47	1.82
15	21.93	1.36	1.38	24.18	1.11	0.84	32.52	1.00	0.83
17	18.84	2.24	5.06	15.18	2.04	3.40	27.18	1.54	2.56
18	16.68	1.92	2.97	18.93	1.62	1.72	27.74	1.24	1.31

All experiments use the gender, SVM parameters $C = 8,192$ and $\gamma = 0.0078125$ have been copied from our best set experiment and therefore may not be optimal. Max and min values are bold.

TABLE III
EXPERIMENT OUTCOME FOR KNN BEST REGIONS

Regions	Correct Classes	Accuracy[%]	Error in $[-2,2]$y[%]	Mean Error[y]	Error variance(s^2)
3, 6, 7, 11, 15	285	26.710	90.91	**0.997**	0.63
7, 11, 15	282	26.429	91.00	0.997	0.65
3, 10, 11, 15	292	27.366	90.35	1.00	0.68
11, 15	292	27.366	90.35	1.00	0.72
11	294	27.553	90.63	1.01	0.69
3, 6, 11, 15	290	27.179	90.07	1.01	0.69
7, 10, 11, 15	294	27.553	90.25	1.01	0.71
3, 7, 10, 11, 15	284	26.616	89.69	1.01	0.65
3, 7, 11, 15	287	26.897	90.16	1.01	0.65
6, 11, 15	289	27.085	90.25	1.02	0.72

Best set of regions for kNN classification. Notice the much lower amount of regions used than in SVM. The best value is bold.

TABLE IV
EXPERIMENT OUTCOME FOR SVM BEST REGIONS

Regions	Correct Classes	Accuracy[%]	Error in $[-2,2]$y[%]	Mean Error[y]	Error variance(s^2)
2, 6, 11, 13, 15, 18	394	36.93	93.81	**0.83**	0.50
2, 3, 6, 9, 10, 11, 13, 14, 15, 18	397	37.21	93.63	0.83	0.51
2, 6, 9, 10, 11, 13, 15, 18	381	35.71	93.72	0.85	0.51
2, 3, 6, 9, 10, 11, 13, 15, 18	390	36.55	93.06	0.85	0.51
2, 6, 10, 11, 13, 15, 18	381	35.71	93.63	0.85	0.51
2, 3, 6, 9, 11, 13, 14, 15, 18	392	36.74	93.16	0.85	0.53
2, 9, 11, 13, 15, 18	379	35.52	94.19	0.85	0.51
2, 10, 11, 14, 15, 18	383	35.90	94.10	0.85	0.49
2, 3, 6, 9, 10, 14, 15, 18	391	36.64	93.06	0.85	0.56
5, 9, 10, 11, 15, 18	377	35.33	93.81	0.85	0.51

Best set of regions for SVM classification. SVM parameters $C = 8,192$ and $\gamma = 0.0078125$ have been computed via grid search. Our overall best value is bold.

TABLE V
CLASSIFICATION RESULTS FOR MULTIPLE REGIONS

Distance	0	1	2	3	4	5	6	7	8	9	10	11	…	29
kNN hits(%)	26.71	43.21	19.87	6.19	2.06	1.22	0.28	0.37	0.09	0	0	0	…	0
		⟵ 89.79 ⟶						⟵ 10.21 ⟶						
SVM hits(%)	36.93	39.55	15.09	4.31	2.53	0.66	0.47	0.37	0	0.09	0	0	…	0
		⟵ **91.57** ⟶						⟵ 8.43 ⟶						
rSVM hits(%)	58.86	27.65	9.65	2.06	0.94	0.28	0.19	0.19	0	0.09	0.09	0	…	0
		⟵ **96.16** ⟶						⟵ 3.84 ⟶						

A distance of 0 denotes a correctly labeled class, whereas a distance of 1 indicates the classifier has labeled a wrong class directly one before or after the actual age class.

TABLE VI
COMPARISON TO PUBLISHED RESULTS - MEAN ERROR

Age range	Use gender	SVM [2 6 11 13 15 18]	SVM best	Region numbers	BoneXpert	Fischer et al.
$[0, 18]$	Yes	0.8320	0.8320	6	–	–
$[0, 18]$	No	0.9917	0.9637	8	–	0.97
$[2, 17]$	Yes	0.8426	0.8265	7	–	–
			0.7958[1]		0.72	–
$[2, 17]$	No	1.0588	0.9850	9	–	–

[1] Removal of worst 14 hands.

leaving-one-out cycle needs 1.112 s and 2.273 s, respectively. These figures indicate real time performance in the application phase, where the SVM model is built already. Here, SVM is even faster than kNN.

It is noticeable that our SVM only uses gender and CCF as features and may easily be enriched by further features. In other words, gender is used only as an attribute for classification and not for prototype building, since the classifier should implicit model the fact of different growth spurts for male and female subjects. An experiment using the gender to build twice as many prototypes – and therefore increasing the feature space – verifies this hypothesis. The mean error was even slightly higher (about 0.06 years). Here, no grid search was used to optimize SVM parameters so a small improvement should be possible.

As a limitation of this study, we investigated classification quality based on extracted eROIs. For routine application, errors in epiphysis detection must be analyzed, too. However, if inaccurate regions are extracted, the prototype-based correlation will yield poor similarity, and the related measures will not contribute in the weighted summation of age computation. So far, the age computation is based on the nominal SVM output. As we have shown, rSVM improves accuracy but still needs to be integrated in the age computation. Another drawback might be the random prototype selection. Experiments have shown that the mean error decreases about 0.03 years, when selecting the prototypes optimal according to their CFF value in the ground truth. This could be improved by using prototypes from a standard reference atlas – which is also used in radiology and therefore perfectly match our age classes.

Further evaluation needs larger data sets that are not selected carefully for the purpose of atlas generation but taken from daily routine including all artifacts and problems radiologists have to face when diagnosing bone age. Such data will, however, lack reliable ground truth, which is considered as the general limitation of automatic BAA. Without such a ground truth, systems cannot be tuned optimally, and comprehensive evaluations cannot be performed. It is worth mention that differences between both expert readings in the USC data reach up to 2.5 years.

ACKNOWLEDGEMENT

This work was partly supported by the Image Retrieval in Medical Applications (IRMA) project, funded by the German Research Foundation (DFG), grants no. Le 1108/6 and Le 1108/9.

REFERENCES

[1] W. W. Greulich and S. I. Pyle, Radiographic atlas of skeletal development of hand wrist. Stanford CA.: Stanford University Press, 1971.

[2] J. M. Tanner, M. J. R. Healy, H. Goldstein, and N. Cameron, Assessment of skeletal maturity and prediction of adult height (TW3) Method. London: WBSaunders, 2001.

[3] P. Hindmarsh and K. Geertsma, "Growth and bone age," http://www.cahisus.co.uk, 2011.

[4] A. T. Al-Taani, I. W. Ricketts, and A. Y. Cairns, "Classification of hand bones for bone age assessment," Proc ICECS, vol. 2, pp. 1088–1091, 1996.

[5] E. Pietka, A. Gertych, S. Pospiech, F. Cao, and H. K. Huang, "Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal ROI extraction," IEEE Transactions on Medical Imaging, vol. 20, no. 8, pp. 715–729, 2001.

[6] M. A. Martin-Fernandez, M. Martin-Fernandez, and C. Alberola-Lopez, "Automatic bone age assessment: A registration approach," Proc SPIE, vol. 5032, pp. 1765–1776, 2003.

[7] S. Aja-Fernande, R. D. Luis-Garcia, M. A. Martin-Fernandez, and C. Alberola-Lopez, "A computational TW3 classifier for skeletal maturity assessment. a computing with words approach," Journal of Biomedical Informatics, vol. 37, no. 2, pp. 99–107, 2004.

[8] L. Bocchi, F. Ferrara, I. Nicoletti, and G. Valli, "An artificial neural network architecture for skeletal age assessment," Proc ICIP, vol. 1, pp. 1077–1080, 2003.

[9] C.-W. Hsieh, T.-L. Jong, Y.-H. Chou, and C.-M. Tiu, "Computerized geometric features of carpal bone for bone age estimation," Chinese Medical Journal, vol. 120, no. 9, pp. 767–770, 2007.

[10] C.-H. Chang, C.-W. Hsieh, T.-L. Jong, and C.-M. Tiu, "A fully automatic computerized bone age assessment procedure based on phalange ossification analysis," Proc IPPR, vol. 16, pp. 463–468, 2003.

[11] H.-J. Kim and W.-Y. Kim, "Computerized bone age assessment using dct and lda," Lecture Notes in Computer Science, vol. 4418, pp. 440–448, 2007.

[12] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H. Huang, "Bone age assessment of children using a digital hand atlas," Computerized Medical Imaging and Graphics, vol. 31, no. 4-5, pp. 322–331, 2007.

[13] B. Fischer, P. Welter, C. Grouls, R. Guenther, and T. M. Deserno, "Bone age assessment by content-based image retrieval and case-based reasoning," Proc SPIE, vol. 7963, 2011.

[14] T. M. Lehmann, M. O. Gld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. B. Wein, "Content-based image retrieval in medical applications," Methods of Information in Medicine, vol. 43, no. 4, pp. 354–361, 2004.

[15] M. Gueld, C. Thies, B. Fischer, and T. M. Lehmann, "A generic concept for the implementation of medical image retrieval systems," Methods of Information in Medicine, vol. 76, no. 2-3, pp. 252–9, 2007.

[16] H. Thodberg, S. Kreiborg, A. Juul, K. Pedersen, and H. V. Aps, "The bonexpert method for automated determination of skeletal maturity," IEEE Transactions on Medical Imaging, vol. 28, pp. 52 – 66, 2009.

[17] D. Martin, D. Deusch, R. Schweizer, G. Binder, H. Thodberg, and M. R. M, "Clinical application of automated greulich-pyle bone age determination in children with short stature," Pediatric Radiology, vol. 39, no. 6, pp. 598–607, 2009.

[18] de Oliveira JEE, A. Machado, G. Chavez, A. Lopes, T. M. Deserno, and de A Araujo, "Mammosys: a content-based image retrieval system using breast density patterns," Computerizes Methods and Programs in Biomedicine, vol. 99, no. 3, pp. 289–297, 2010.

[19] T. M. Deserno, D. Beier, C. Thies, and T. Seidl, “Segmentation of medical images combining local, regional, global, and hierarchical distances into a bottom-up region merging scheme,” Proc SPIE, vol. 5747, no. 1, pp. 546–555, 2005.
[20] C. Thies, M. Schmidt-Borreda, T. Seidl, and T. M. Deserno, “A classification framework for content-based extraction of biomedical objects from hierarchically decomposed images,” Proc SPIE, vol. 6144, pp. 559–568, 2006.
[21] B. Fischer, M. Sauren, M. O. Gueld, and T. M. Deserno, “Scene analysis with structural prototypes for content-based image retrieval in medicine,” Proc SPIE, vol. 6914; online first, DOI 10.1117/12.770541, 2008.
[22] B. Fischer, P. Welter, R. W. Gnther, and T. M. Deserno, “Web-based bone age assessment by content-based image retrieval for case-based reasoning,” International Journal of Computer Assisted radiology and Surgery, vol. 7, no. 3, pp. 389–399, 2012.
[23] V. Gilsanz and O. Ratib, Hand bone age. A digital atlas of skeletal maturity. Berlin: Springer-Verlag, 2005.
[24] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[25] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, “A practical guide to support vector classifcation,” last accessed on 22.03.2012 from http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf, 2010.
[26] C. W. Hsu and C. L. Lin, “A comparison of methods for multi-class support vector machines,” IEEE Transactions on Neural Networks, vol. 13, pp. 415–425, 2002.
[27] F. Cao, H. Huang, E. Pietka, and V. Gilsanz, “Digital hand atlas for web-based bone age assessment: system design and implementation,” Computerized Medical Imaging Graphics, vol. 24, pp. 297–307, 2000.
[28] P. QT and Nokia, “Qt 4.7.3: documentation,” http://doc.qt.nokia.com/4.7/index.html, online 22.03.2012.
[29] R. Hipp, “Sqlite: documentation,” http://www.sqlite.org/docs.html, online 22.03.2012.
[30] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 27:1–27:27, 2011.