Computers in Biology and Medicine 68 (2016) 67–75

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? A general question answered for bone age assessment

Muhammad Kashif*, Thomas M. Deserno, Daniel Haak, Stephan Jonas

Department of Medical Informatics, RWTH Aachen University, Pauwelsstr. 30, 52057 Aachen, Germany

ARTICLE INFO

Article history:
Received 2 July 2015
Accepted 10 November 2015

Keywords:
Feature extraction
Classification
Bone age assessment
Epiphyseal regions of interest (eROIs)
Computer-aided diagnosis

ABSTRACT

Solving problems in medical image processing is either generic (being applicable to many problems) or specific (optimized for a certain task). For example, bone age assessment (BAA) on hand radiographs is a frequent but cumbersome task for radiologists. For this problem, many specific solutions have been proposed. However, general-purpose feature descriptors are used in many computer vision applications. Hence, the aim of this study is (i) to compare the five leading keypoint descriptors on BAA, and, in doing so, (ii) presenting a generic approach for a specific task. Two methods for keypoint selection were applied: sparse and dense feature points. For each type, SIFT, SURF, BRIEF, BRISK, and FREAK feature descriptors were extracted within the epiphyseal regions of interest (eROI). Classification was performed using a support vector machine. Reference data (1101 radiographs) of the University of Southern California was used for 5-fold cross-validation. The data was grouped into 30 classes representing the bone age range of 0–18 years. With a mean error of 0.605 years, dense SIFT gave best results and outperforms all published methods. The accuracy was 98.36% within the range of 2 years. Dense SIFT represents a generic method for a specific question.

1. Introduction

Problems in computer science and especially computer vision can be tackled by two different approaches: (i) a specific solution that utilizes a lot of prior knowledge or (ii) a general solution that is applicable to other problems, too. Bone age assessment (BAA) is the process of determining the skeletal maturity of a person. For BAA, many specific solutions have been proposed. However, it might be efficient to solve such a problem using general algorithms.

In clinical practice, the bone age of a child (developmental age of the bones) is assessed based on a radiological examination of the left hand and wrist and compared to the chronological age. This allows anticipating the adult height as well as diagnosis and management of endocrine disorders and pediatric syndromes [1]. Moreover, BAA is used in forensic medicine [2]. Another relevant application is found in social fields. According to United Nations Children's Fund (UNICEF), only half of the children under 5 years in the developing world have their births registered. In sub-Saharan Africa and South Asia, about 65% of all births go unregistered [3]. Without documented proof of age, children are recruited to fighting forces, exposed to hazardous forms of work, forced to early marriages, and treated as adult in legal proceedings. In all of these cases, skeletal maturity can help to estimate the chronological age of a person.

However, BAA is a time-consuming and cumbersome task in radiology. In clinical routine, two methods are applied: Greulich and Pyle (GP) [4] and Tanner and Whitehouse (TW) [5]. Following the GP method, the radiologist compares all bones of the left hand to those in a radiograph of a standard atlas and assesses the bone age according to his visual perception. Following the TW method, certain subsets of bones are examined with respect to epiphyseal distances. Hence, the GP method is more subjective, while the TW method is more complex and time consuming. Based on the physicians expertize, the examination time varies. In [6], an average time of 40 s and 80 s was reported for GP and TW methods, respectively. Conversely in [7], the average reading time is 84 s and 474 s for GP and TW methods, respectively. Hence, automated BAA is desired.

Many approaches have already been adopted to automate BAA. In 1996, Al-Taani et al. presented an automatic BAA approach that is based on a point distribution model (PDM) of 130 feature points [8]. The distal and middle phalanxes of third finger were classified. A set of 120 images was used for classification. The evaluation rates for two experiments were 73.7% and 70.5%. In 2001, Pietka et al. comprehensively reviewed early approaches for BAA and presented a method for feature extraction from left hand radiograph by measuring the gap between metaphyses and diaphyses

* Corresponding author. Tel.: +49 241 80 88790; fax: +49 241 80 3388790.
E-mail address: Muhammad.kashif@rwth-aachen.de (M. Kashif).

http://dx.doi.org/10.1016/j.compbiomed.2015.11.006
0010-4825/

[9]. A solid view on fundamental principles in BAA was provided but computations were not performed.

Bocchi et al. proposed a system to implement the TW method using neural networks [10]. A set of 120 images for training and a set of 40 images for testing were used. A maximum error of 1.4 years with standard deviation 0.7 was reported. BAA based on phalangeal features was presented by Chang et al. [11]. In this method, the back propagation of neural networks was applied to train the features of phalanges. A quite large error of 1.5 years was reported.

In 2007, Kim and Kim used epiphyseal regions of interest (eROIs) [12]: the discrete cosine transform and a linear discriminant analysis were applied on nine relevant eROIs. A mean error of 0.6 years was reported. The data set in use was quite small and the error rate could not be confirmed by any others.

The use of private data restricts the comparability of BAA approaches. To improve this situation, a digital hand atlas of carefully selected radiographs was released at the University of Southern California (USC) and has been established as standard reference database. First experiment on that dataset was performed by Gertych et al., where a fuzzy classifier was applied on carpal bone and phalangeal ROIs [13].

In our previous work, a method based on content-based image retrieval (CBIR) was presented, where eROIs patches were extracted automatically and similar patches were retrieved using the Image Retrieval and Medical Application (IRMA) framework [14]. Classification was done with a k-nearest neighbor (kNN) approach. A mean error of 0.97 years for the age range of 0–18 years was reported on the USC data. The method was extended by Harmsen et al. introducing class prototypes. Applying the support vector machine (SVM) for classification, a mean error of 0.83 years was achieved [15]. Haak et al. have improved to 0.768 years by replacing the SVM with a support vector regression (SVR) [16]. To obtain the features, cross correlation between test and reference images was used [14–16].

The leading commercial product for BAA (BoneXpert) applies an active shape model [17]. Within the bone age ranges of 2.5–17 years and 2–15 years, BoneXpert obtains a root mean square error of 0.61 years for boys and girls, respectively [18].

However, all BAA methods published so far are specific rather than generic. In the last decade, many robust methods to extract distinct features from the image have been presented, which are being used in large variety of computer vision applications. The most prevailing methods are scale invariant feature transform (SIFT), speeded up robust feature (SURF), binary robust independent elementary features (BRIEF), binary robust invariant scalable keypoints (BRISK), and fast retina keypoint (FREAK). A lot of research has been published to compare these methods (Table 1), but a superior method has not yet been identified in general. Rather, the performance of the methods depends on the application domain.

The process of feature extraction is composed of feature detection and feature description. In feature detection, an algorithm determines the appropriate keypoints that represent the most informative part of the image. In feature description, a local image descriptor is computed for every keypoint. The descriptor possesses the neighborhood information of the keypoint to identify the same keypoint across various images.

The majority of previous research is concentrated on the comparison of feature detectors rather than feature descriptors. For instance, Juan and Gwun compared the feature detection performance of SIFT, PCA-SIFT and SURF methods for scale, rotation, and affine transforms as well as for blur and illumination changes [19]. SIFT performed superior in all experiments but showed the longest processing times. In other experiments, SURF was found fastest and stable. PCA-SIFT performed good in rotation and illumination changes.

Tuytelaars and Mikolajczyk presented a survey on local invariant feature detectors [20]. They compared corner, blob, and region detectors. Again, the study was focused on feature detectors.

Canclini et al. evaluated the performance of feature detectors and descriptors in terms of processing time, repeatability, and matching accuracy for image retrieval application [21].

Several feature extractors were compared for visual simultaneous localization and mapping (VSLAM). Klippenstein and Zhang have compared the Harris detector, the Lucas–Kanade–Tomasi tracker, and the detector part of SIFT for VSLAM [22]. They concluded that the choice of feature detector is irrelevant in terms of VSLAM performance. However, feature descriptors were not evaluated. More recently, Hartmann et al. have evaluated the feature descriptors for accuracy and speed in a typical graph-based VSLAM algorithm [23].

Nevertheless, the appropriate choice of SIFT, SURF, BRIEF, BRISK, or FREAK cannot be answered yet, since the proof of the pudding is in the eating. Moreover, a comparison of feature descriptors with respect to a certain application is not yet presented. From an evaluation-methodology point of view, a well-defined reference problem and a large, public available database of ground truth is required. Furthermore, the application domain shall be researched comprehensively on that database. Therefore, we selected the BAA problem to analyze SIFT, SURF, BRIEF, BRISK and FREAK methods for feature description rather than extraction.

2. Material and methods

The image processing chain in this work is composed of several steps: eROIs extraction, feature points specification, features description, and classification (Fig. 1).

2.1. EROI Extraction

From prior work, a semi-automatic approach is used to extract the eRIOs from radiographs [14]. For proper localization, the user simply clicks into the centers of relevant epiphyses. Fourteen eROIs were extracted from each radiograph and rotated into a normalized upright position (Fig. 2).

2.2. Feature point specification

Except BRIEF, all other methods (SIFT, SURF, BRISK, FREAK) can be used for keypoints detection, too. SIFT, SURF, BRISK and FREAK

Table 1
Feature extraction methods compared by previous authors.

	SIFT	PCA-SIFT	SURF	BRIEF	ORB	BRISK	FREAK	Others
[19]	x	x	x					
[20]	x		x					x
[21]	x		x	x	x	x	x	x
[22]	x							x
[23]	x		x	x	x	x	x	

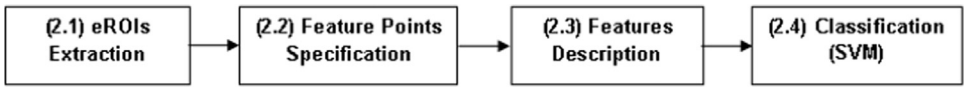

Fig. 1. Different steps in the proposed method.

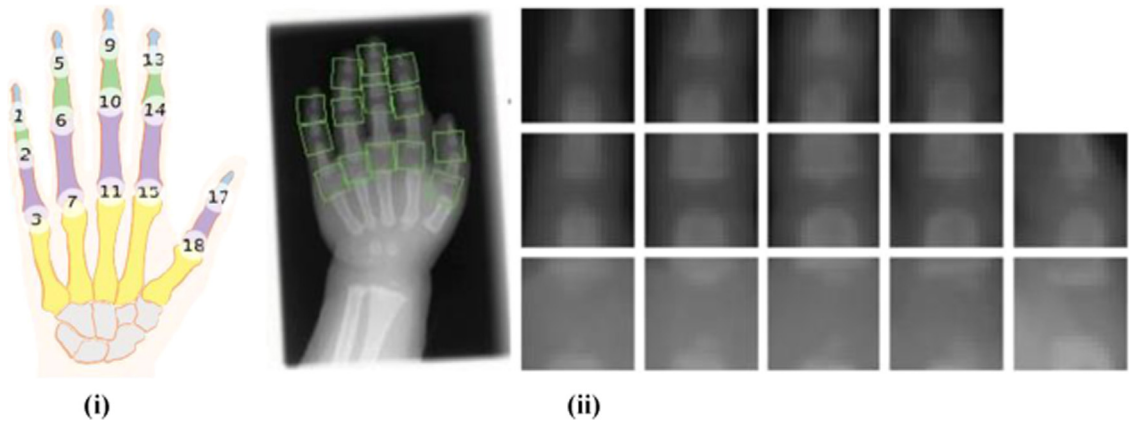

Fig. 2. (i) Corresponding eROI numbers, (ii) eROIs are extracted and rotated into upright position [15].

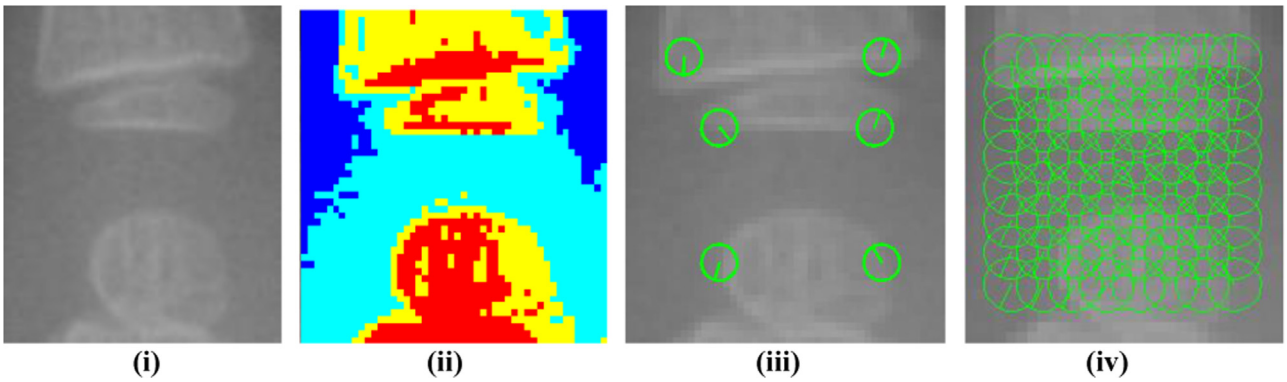

Fig. 3. (i) Original image of an eROI, (ii) RGB segmented image obtained using multi-thresholding where red and yellow segments show bone tissues, while cyan and blue colors represent the background (iii) sparse feature points, and (iv) dense feature points. (For interpretation of the reference to color in this figure legend, the reader is reffered to the web version of this article.)

may detect different number of keypoints on different locations. To assess the feature describing properties specifically, a fixed number of keypoints was chosen on identical locations and a fixed scale was used for all images. This discharged the scale invariance property of the methods. For a more detailed analysis, sparse and dense methods for keypoint selection were applied (Figs. 3 and 4).

2.2.1. *Sparse feature points*

Sparse feature points (or keypoints) are six points located at distinct positions of each eROI. Here, a simple but innovative and task-specific keypoint selection method was implemented [24]. A multi-level thresholding using Otsu's method [25,26] was applied: each eROI was quantized into four discrete levels using three thresholds. Two levels representing bony and osseous tissues are shown as red and yellow segments in Fig. 3. Scanning the image from left to right and right to left, the first pixel of eROI with a value equal or greater than the upper threshold value (which represents the bone) is considered as a keypoint. In this way, 6 keypoints were detected on upper, middle, and lower parts of the bone. The epiphysis may not be present for children under the age of 1.5 years. In this case, the two (left and right) feature points were defined in the middle of the eROI (Fig. 3).

2.2.2. *Dense feature points*

Dense features points were located closely together along a grid of specific step size. Starting from left-to-right and top-to-bottom, every third pixel was considered as a keypoint. Hence, a squared grid of keypoints was defined within each eROI. Fig. 3 (iv) exemplifies a grid of dense feature points for the SIFT method. For BRIEF, BRISK, and FREAK, a smaller grid was defined (Fig. 5), because the keypoints close to the image border are invalid. A feature descriptor was computed for every keypoint. In this way, a large number of feature descriptors was obtained.

2.3. *Features description*

SIFT, SURF, BRIEF, BRISK, and FREAK are used for feature description. In addition, SIFT and SURF were applied to extract both, sparse and dense features, while BRIEF, BRISK, and FREAK were used only on dense feature points (Table 2). This is because the sparse feature points were located close to the eROI border and, hence, they are invalid for BRIEF, BRISK, and FREAK.

2.3.1. *Scale invariant feature transform (SIFT)*

In 2004, Lowe has proposed the SIFT method [27]. It robustly extracts distinctive local features, which are used to match objects in different images. For each keypoint, a feature descriptor is

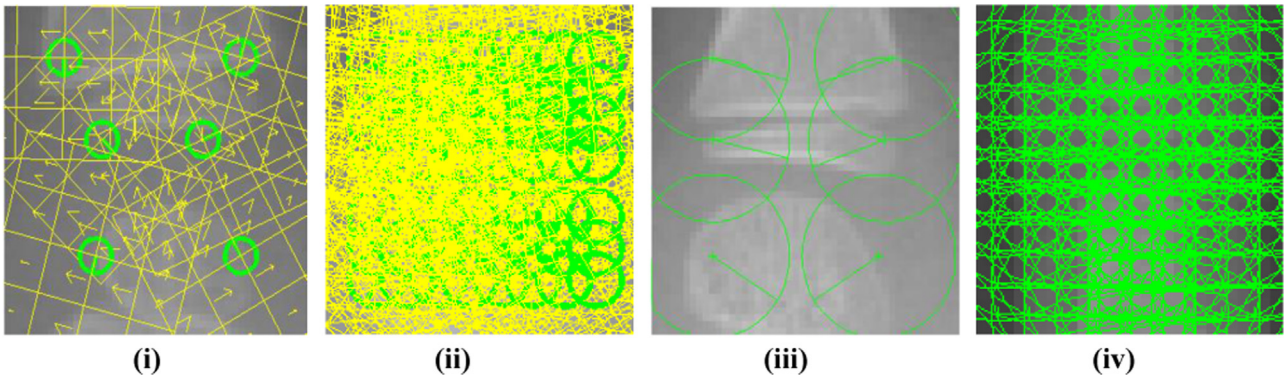

Fig. 4. (i) Six sparse SIFT feature descriptors: green circles and green lines in the circles represent key points and their orientations respectively. The yellow box and arrows show feature descriptors, (ii) dense SIFT feature descriptors, (iii) six sparse SURF feature points where lines show the orientations, and (iv) dense SURF feature points. (For interpretation of the reference to color in this figure legend, the reader is reffered to the web version of this article.)

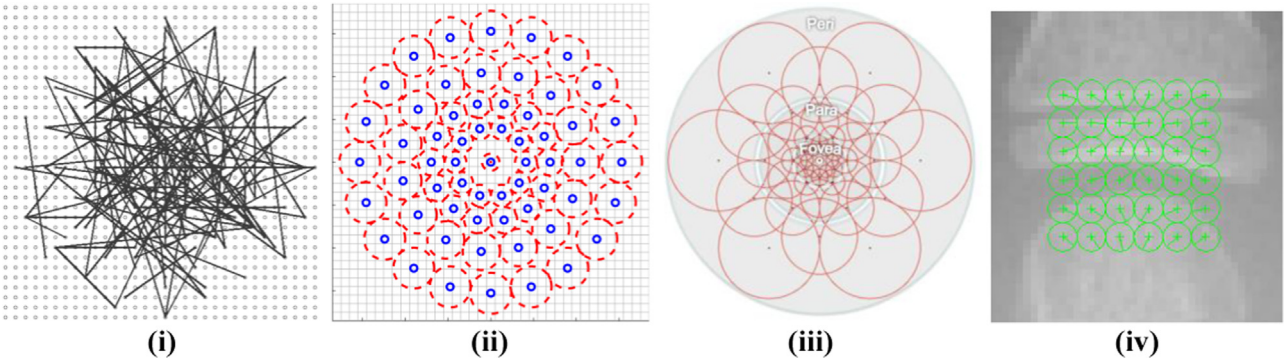

Fig. 5. (i) BRIEF sampling pattern [29], (ii) BRISK sampling pattern [30], (iii) FREAK sampling pattern [31], and (iv) dense feature points selected for BRIEF, BRISK and FREAK methods.

calculated from the gradient magnitude and the relative orientation in the local neighbourhood of pixels. A 16×16 region around each keypoint is sub-divided into 4×4 regions. Each sub-region contains a weighted directional histogram. This histogram is 8-dimensional (8D). It means that there are 8 bins, each turned by 45°. Since each of the 16 sub-regions contains 8D histograms, the resulting keypoint descriptor is $4 \times 4 \times 8 = 128$-dimensional (Fig. 4).

2.3.2. *Speeded up robust features (SURF)*

Similar to SIFT, SURF is a local invariant fast feature point detector and distinctive feature point descriptor [28]. After identifying the keypoint, the first step of processing assigns an orientation within a circular region around the keypoint. Then, a squared region is aligned to the selected orientation and the SURF descriptor is extracted computing Haar wavelet responses. A SURF descriptor is either 64- or 128-dimensional (Fig. 4).

2.3.3. *Binary robust independent elementary features (BRIEF)*

Presented by Calonder et al. [29], BRIEF is a binary descriptor based on pair-wise intensity comparisons (Fig. 5). A number of (128, 256, or 512) pairs are chosen randomly but on fixed locations in a patch around a keypoint. The BRIEF descriptor is computed by conducting the intensity difference test on these pairs. The test yields 1 if the intensity at one specific location is larger than at another. Otherwise, it gives 0. This descriptor is computationally faster than SIFT or SURF, as it is based on binary comparisons. BRIEF is invariant to illumination changes, but not to scaling or rotation. In the BAA application, there is neither scale nor rotation of the extracted eROIs and therefore, BRIEF is applicable. With standard configuration, a BRIEF descriptor is 32-dimensional.

Table 2
Comparison of descriptors (per eROI).

Feature	#Keypoints dense	Feature vector bytes dense	Feature vector bytes sparse
SIFT	64	8192	768
SURF	49	3136	768
BRIEF	16	512	–
BRISK	36	2304	–
FREAK	36	2304	–

2.3.4. *Binary robust invariant scalable keypoints (BRISK)*

Improves the concepts of BRIEF, BRISK is another binary descriptor [30]. Once keypoints are selected, a sampling pattern is applied in the keypoint's neighborhood (Fig. 5). Pairs of pixels around the keypoint are separated into two subsets: short-distance and long-distance pairs. Local intensity gradients from long distance pairs are computed and the orientation of the feature point is determined. Short distance pairs are rotated using this orientation. A BRISK descriptor is assembled by concatenating the results of short distance pair-wise brightness comparison tests. This descriptor is composed of a bit-string of length 512 (64 bytes) and represented as a 64-dimensional feature vector.

2.3.5. *Fast retina keypoint (FREAK)*

The binary descriptor FREAK is inspired by the human visual system [31]. A retinal sampling pattern is applied around the keypoint (Fig. 5), and a binary string is computed by comparing the pixel intensities over the sampling pattern. Similar to BRISK, the keypoint's orientation is adding from local gradients, but FREAK uses pairs with symmetric fields around the receptive

center instead of long pairs. A FREAK descriptor is constructed by thresholding differences of corresponding Gaussian kernels. It is a 512 bit long binary string formed by a sequence of 1 bit Difference of Gaussians (DoG).

2.4. Classification

The support vector machine (SVM) is a powerful classifier that is applied on a given set of training data in order to create a model [32]. That model is then used to classify new data. Harmsen et al. used the SVM for BAA on correlation prototypes [15]. A kernel function is used to map any non-linear input into high-dimensional feature space, where a hyperplane can be found that separates the classes. Usually, radial basis function (RBF) are used as kernel but RBF is not suitable for large numbers of feature [33]. Therefore, we applied the SVM with a 3rd degree polynomial kernel. To cope with 30 classes, a one-against-one approach was applied [34].

2.5. Age classes

The bone age corresponds to the time a skeleton matures from newborn infants to the bony structure of adults. It is described in progressive steps of predictable morphologic stages [35]. Based on these stages, Gilsanz and Ratib have defined an ontology (Fig. 6) [1], where reference bone ages are quantized in steps of 2, 4, 6, and 12 months for the intervals [8 month… 20 month), [20 month… 28 month), [2.5 year… 6 year), and [6 year… 18 year], respectively, where m and y denote month and year, respectively. This creates a set of 29 classes with four different bone age ranges. Harmsen et al. included a 30th class for bone ages larger than 18 years [15]. Accordingly, the clinical data of USC was grouped into 30 classes (Table 3).

2.6. Age computation

Since the classification results a predicted age class for a radiograph with unknown bone age, the estimated bone age is calculated by taking the arithmetic mean of upper and lower class bounds as follows

$$a = 1/2\ (U_B(c) + L_B(c))$$

where a, c, U_B and L_B are the estimated bone age, predicted age class, upper and lower bounds of the age class respectively.

2.7. Validation experiments

The USC hand atlas is composed of 1101 radiographs from different ethnics and age groups. All radiographs have been assigned with two reference readings of experienced radiologists. We have defined the average of the two readings as ground truth. Fourteen eROIs were extracted from each hand radiograph and rescaled to 32×32 and 48×48 pixel eROIs (Fig. 2). Dense SIFT and

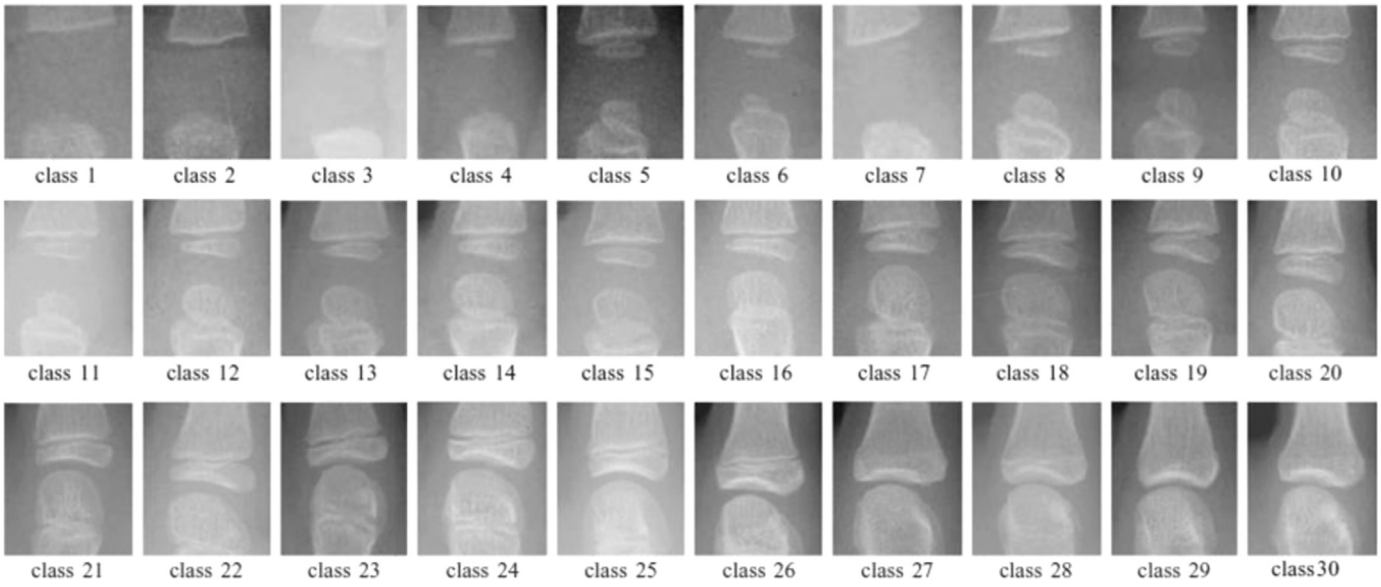

Fig. 6. Images showing different stages of the epiphyses development and corresponding class numbers.

Table 3
Class, corresponding BA range (years) and number of images.

Class	BA range	#Images	Class	BA range	#Images	Class	BA range	#Images
01	0.00–0.66	10	11	2.50–3.00	17	21	09.00–10.00	35
02	0.66–0.83	06	12	3.00–3.50	23	22	10.00–11.00	65
03	0.83–1.00	03	13	3.50–4.00	26	23	11.00–12.00	65
04	1.00–1.16	10	14	4.00–4.50	24	24	12.00–13.00	68
05	1.16–1.33	04	15	4.50–5.00	15	25	13.00–14.00	74
06	1.33–1.50	07	16	5.00–5.50	18	26	14.00–15.00	61
07	1.50–1.66	09	17	5.50–6.00	27	27	15.00–16.00	82
08	1.66–2.00	05	18	6.00–7.00	44	28	16.00–17.00	76
09	2.00–2.33	17	19	7.00–8.00	43	29	17.00–18.00	128
10	2.33–2.50	06	20	8.00–9.00	37	30	18.00–19.00	96

Table 4
Comparison of experiment outcomes for all methods.

Method	Regions	Correct classes	Accuracy (%)	Accuracy in 1 year range (%)	Accuracy in 2 years range (%)	Mean error (y)	Error variance
Dense SIFT	All	491	44.60	88.92	98.36	0.617	0.056
–	1,3,5,6,7,9,10,11,13,14,15,18	503	45.69	89.37	98.36	**0.605**	0.050
Dense SURF	All	459	41.69	86.65	97.64	0.655	0.054
–	3,6,7,9,10,11,14,15,17,18	454	41.24	86.92	98.18	0.647	0.053
Sparse SIFT	All	459	41.69	86.56	97.91	0.675	0.047
–	3,6,7,9,10,11,15,18	451	40.96	86.19	98.09	0.669	0.051
Sparse SURF	All	453	41.14	86.56	98.18	0.668	0.050
BRIEF	All	422	38.33	81.74	95.64	0.761	0.097
–	2,3,5,6,7,9,10,11,14,15,17,18	432	39.24	82.20	95.46	0.741	0.096
BRISK	All	458	41.60	85.20	96.73	0.679	0.067
–	3,5,6,7,9,10,11,13,14,15,17,18	455	41.32	85.20	97.00	0.662	0.070
FREAK	All	456	41.42	84.02	95.91	0.716	0.075

Table 5
Distance-wise classification results in the range of 0–18 years for dense SIFT method.

Distance	0	1	2	3	4	5	6	≥7
Class accuracy (%)	45.69	43.69	8.99	1.55	0	0.09	0	0
Accumulated accuracy (%)	45.69	89.38	98.36	99.91	99.91	100	100	100

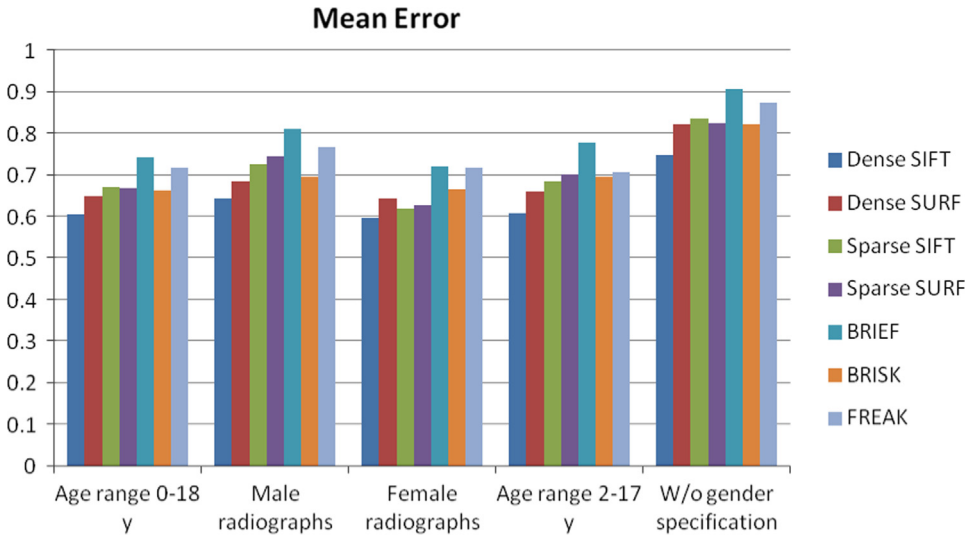

Fig. 7. Comparison of all methods for various experiments in terms of mean error (in years).

Table 6
Best method for various experiments.

Experiment no.	Name of experiment	Winner	Mean error (in years)
1	For BA range 0–18 years	Dense SIFT	0.605
2	For male radiographs	Dense SIFT	0.644
3	For female radiographs	Dense SIFT	0.596
4	For BA range 2–17 years	Dense SIFT	0.608
5	Without gender specification	Dense SIFT	0.746

dense SURF features were computed from 32×32 eROIs, while sparse SIFT, sparse SURF, BRISK, and FREAK features were extracted from 48×48 eROIs. After eROI extraction, a feature vector was computed for each eROI. Table 2 gives the corresponding numbers of features. Depending on experiment, the number of SVM input features differs. For instance, experiments performed on a single region using sparse SIFT result in 768 features. All experiments were performed on a complete set of 14 eROIs. For each radiograph, all features were combined into a vector of $14 \times 768 =$ 10,752 bins, which was then used as input to SVM.

Analysis was performed using 5-fold cross validation. Data was partitioned randomly with the condition that each fold or at least each training set contains minimally one instance of each class. We determine the number of correct classes, the age class accuracy, the accuracy in a 1-year range, the accuracy in a 2-year range, the mean error, and the error variance.

Out of the 1101 USC images, 552 were from male and 549 were from female children. The experiments were performed on combined and separated gender datasets. For the sake of comparison with previous works and BoneXpert, experiments were performed for the bone age ranges of 0–18 years and 2–17 years.

3. Implementation

The implementation was done in MATLAB. Using the mex-file functionality, C++ libraries were included. The computer vision toolbox of MATLAB provides built-in support for SURF, BRISK, and FREAK. SIFT features were computed using the VL-SIFT library version 0.9.17 [36]. BRIEF descriptors were obtained from the opencv2.4.6 library. Multiclass SVM was implemented using the LIBSVM library version 3.17, which has built-in support for multiclass problems and imbalanced datasets [37].

4. Results

Experiments were performed on single regions (eROI), on the subsets of regions, and on a complete set of regions, with and without gender specification. Some regions (e.g. regions of the middle finger) performed better than others. Lower error rates were achieved on subsets of regions. Using iterative adding and removal of regions, the best combinations were chosen. Consistent results were obtained on the complete set of eROIs (Table 4). A few subsets yielded similar or marginally better result.

For all regions and the bone age range of 0–18 years, mean errors of 0.617 ($\pm$0.056), 0.655 ($\pm$0.054), 0.675 ($\pm$0.047), 0.668 ($\pm$0.050), 0.761 ($\pm$0.097), 0.679 ($\pm$0.067), and 0.716 ($\pm$0.075) years were obtained by dense SIFT, dense SURF, sparse SIFT, sparse SURF, BRIEF, BRISK and FREAK, respectively. For the best subset of regions, mean errors of 0.605 ($\pm$0.050), 0.647 ($\pm$0.053), 0.669 ($\pm$0.051), 0.761 ($\pm$0.097), and 0.662 ($\pm$0.070) were achieved by dense SIFT, dense SURF, sparse SIFT, BRIEF, and BRISK. The set of all regions gave best result for sparse SURF and FREAK (Table 4).

For the distance-wise classification (Table 5), a distance 0 indicates the correctly labeled class, whereas a distance of 1 shows a wrongly labeled neighbored class in the range of 1-year and so on. Corresponding accumulated accuracies yielded 45.69%. 89.37%, and 98.36% for distances of 0, 1, and 2, respectively. Here, accuracy represents the correctly labeled class, while accuracy in 1 year shows the sum of distance 0 and 1 and accuracy in 2 years is the sum of distance 0, 1 and 2.

5. Discussion

Though SIFT performs better than SURF in case of individual regions, for all regions, SIFT and SURF perform equally well on sparse features. For dense feature points, SIFT yielded best results, followed by SURF, BRISK, FREAK, and BRIEF (Fig. 7). Overall, dense SIFT performed best in all experiments (Table 6). This is in line with the findings of Hartmann et al., who compared SIFT, SURF, BRIEF, ORB, BRISK, and FREAK descriptors for VSLAM and concluded SIFT as the most accurate method [23]. The average time to process one radiograph using the dense SIFT method is 102 ms, it takes 70 ms to extract the features and 32 ms for classification. In [23], the average time per keypoint is given for each descriptor.

Regions [3,7,11,15,18] belong to the epiphyseal center of proximal phalanges. They are more reliable and gave better results than the others. Region 11, the proximal eROI of the middle finger, performed best among all individual regions. This was also reported by Harmsen et al. [15]. Similarly, regions [9,10] of middle finger and regions [6,14] of index and ring fingers gave good results. A combination of these regions yielded best results, whereas regions [1,2,5,13,17] seemed to contribute minor. This order corresponds to the ossification of the epiphysis. The regions where ossification starts first gave better results than those where late ossification takes place. This association consolidates our method. In addition, better results were obtained in case of female radiographs as compared to male radiographs. These findings further strengthen the approach as female bones developed faster than male bones [1].

The mean error distributed similar for all methods (Fig. 8). The best classification was obtained in the range of 0–5 years.

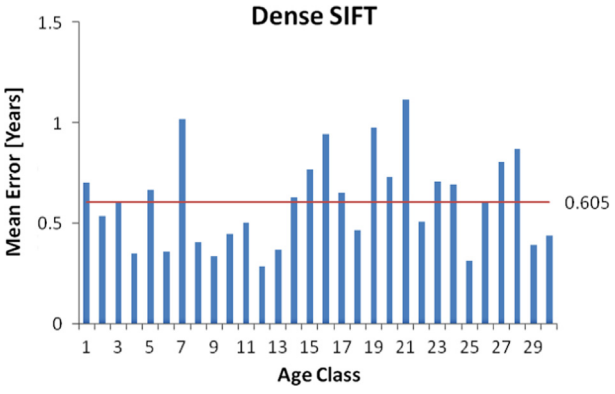

Fig. 8. Mean error itemized by class in [0, 18] years.

Table 7
Comparison to published results-mean error.

BA range	Used gender	All (14) regions	Best regions	Region numbers	Haak et al.	Harmsen et al.	Fischer et al.	BoneXpert
0–18	Yes	0.6172	0.6053	12	0.768	0.8320	–	–
0–18	No	0.7509	0.7461	12	–	0.9637	0.97	–
2–17	Yes	0.6082	0.6082	All	0.692	0.8265	–	–
2–17	No	0.8560	0.8447	12	–	0.9850	–	–
Female: 2.5–15 Male: 2-17	Yes	–	0.6195 (RMS)	12	–	—	–	0.61 (RMS)

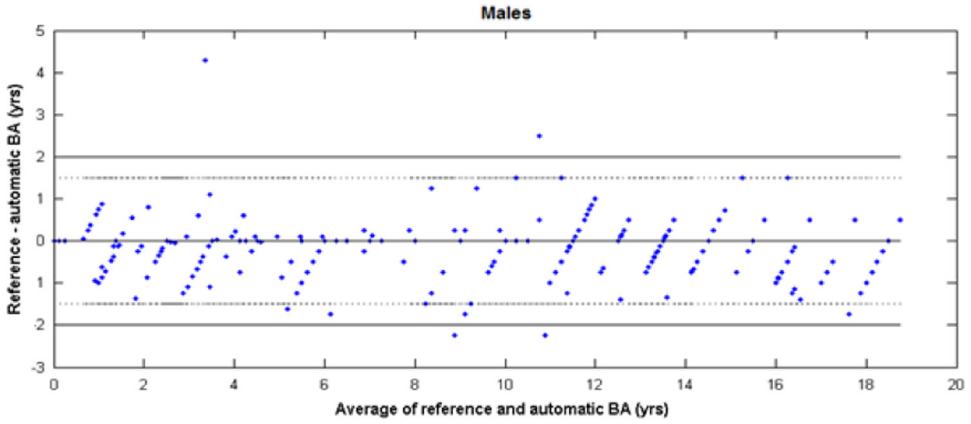

Fig. 9. Bland–Altman plot comparing manual and automated bone age rating for male radiographs.

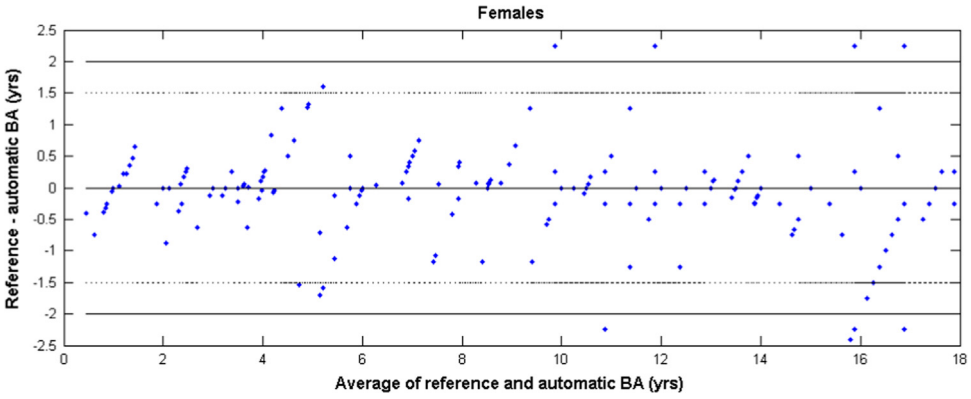

Fig. 10. Bland–Altman plot comparing manual and automated bone age rating for female radiographs.

Comparatively, a large error was obtained in the range of 6–12 years. Also, within the USC dataset, the two experts disagree more in the range of 6–12 years [18].

Overall, best result (mean error 0.605 ($\pm$0.050) years) was obtained when features were extracted using the dense SIFT method on the entire bone age range of 0–18 years, whereas Haak et al., Harmsen et al., and Fischer et al. reported mean errors 0.768, 0.832 and 0.97 years, respectively. Thereby, our approach outperforms all prior published methods except BoneXpert (Table 7).

The commercial product BoneXpert has been evaluated on a range of 2.5–17 years and 2–15 years for boys and girls, respectively. BoneXpert reached a root mean square (RMS) error 0.61 years [17]. Considering an age range of 2–17 years for both boys and girls, our method reached mean and RMS errors of 0.608 ($\pm$0.057) years and 0.653 years, respectively. An RMS error of 0.619 years was obtained for the exact range of BoneXpert. For the same BA range, the variation between the two expert readings yielded an RMS error of 0.63 years [18].

Our results improve prior publications and close-up to the performance of BoneXpert. Nevertheless, as discussed, BoneXpert is applied to the BA range of 2–17 years, whereas our approach is also applicable to the entire range of bone ages (0–18 years). However, our work uses a semi-automatic preprocessing step, where the user is required to indicate the center of relevant epiphyses.

Bland–Altman plots are constructed to analyze the agreement between reference and automatic bone age ratings. It compares reference and automatic bone age ratings of the cross validated results, separately for males (Fig. 9) and females (Fig. 10). Here, the *x*-axis represents the mean of the reference and automatic bone age ratings, while *y*-axis indicates differences between reference and automatic bone age ratings. Dashed and solid lines are drawn at $\pm$1.5 and $\pm$2 years, respectively. A few outliers can be seen in the plot. In [18], outliers were re-rated by the experts for further experiments. In this work, outliers were not re-rated and all experiments were performed on the original reference readings by the USC experts. Since the proposed method is evaluated using cross validation, the bias in the Bland–Altman plot is zero by construction.

The presented approach is robust and easy to implement, yet suitable for computer-aided diagnosis (CAD). In contrast to previous work, our approach does not require comparisons with all 1101 eROIs stored in database or with 30 prototypes as suggested by Fischer et al. and Harmsen et al., respectively. It does not need semantic features or atlases unlike method by Pietka et al., or the conventional methods by Greulich and Pyle or Tanner and Whitehouse.

As future work, carpal bones, distal radius, and ulna will be used in further experiments to further improve BAA quality. In addition, a new dataset of 987 radiographs of normal and abnormal children is composed at University Hospital Aachen and will be used for further evaluation of the algorithms.

6. Conclusion

In this paper, 5 feature descriptors were compared for a specific task of bone age assessment. SIFT is found to perform best in terms of accuracy, with SURF and BRISK as close competitive, while FREAK and BRIEF are slightly inferior. Two methods for keypoints selection, sparse and dense feature points, were implemented. This shows that task specific keypoints selection can improve the

outcome in other applications. We conclude that SIFT is most suitable feature descriptor for extracting features from radiographs subject to the proper selection/detection of keypoints.

We have also presented an effective and novel method for automatic bone age assessment, where SIFT (or other descriptive) features are extracted directly from the eROIs of left hand radiographs and classification is performed using SVM. The class accuracy of 46% seems low but most misclassification lies within the range of one or two classes (Table 5). A mean error of 0.605 years is achieved and the accuracy within the range of one or two years are 89.37% and 98.36%, respectively. In comparison, the difference between two expert readings in the USC data reaches up to 2.5 years [15], which is much higher.

Our solutions outperform most state-of-the-art BAA methods while utilizing a generic approach. A specialization of the feature detection, the more knowledge-driven sparse method, was inferior to the more general dense method. While this has also to be attributed to the higher amount of information covered in the dense key point extraction, it still indicates that generalization also might lead to better results.

Conflict of interest statement

None declared.

References

[1] V. Gilsanz, O. Ratib, Hand Bone Age: A Digital Atlas of Skeletal Maturity, Springer-Verlag, Berlin, Germany, 2005.

[2] S. Ritz-Timme, C. Cattaneo, M.J. Collins, E.R. Waite, H.W. Schütz, H.J. Kaatsch, H. I.M. Borrman, Age estimation: the state of the art in relation to the specific demands of forensic practise, Int. J. Leg. Med. 113 (3) (2000) 129–136.

[3] T. Smith, L. Brownlees, Age Assessment Practices: A Literature Review and Annotated Bibliography, United Nations Children's Fund (UNICEF), New York, 2011.

[4] W.W. Greulich, S.I. Pyle, Radiographic Atlas of Skeletal Development of Hand Wrist, Stanford Univ. Press, Stanford, CA, 1971.

[5] J. Tanner, M. Healy, H. Goldstein, N. Cameron, Assessment of Skeletal Maturity and Prediction of Adult Height (TW3), WB Saunders, London, 2001.

[6] M.J. Horter, S. Friesen, S. Wacker, B. Vogt, B. Leidiger, R. Roedl, F. Schiedel, Determination of skeletal age. Comparison of the methods of Greulich and Pyle and Tanner and Whitehouse, Orthopade 41 (2012) 966–976, http://dx.doi.org/10.1007/s00132-012-1983-y.

[7] V.D. Sanctis, S.D. Maio, A.T. Soliman, G. Raiola, R. Elalaily, G. Millimaggi, Hand X-ray in pediatric endocrinology: Skeletal age assessment and beyond, Indian J. Endocrinol. Metab. 18 (7) (2014) 63–71, http://dx.doi.org/10.4103/2230-8210.145076.

[8] A.T. Al-Taani, I.W. Ricketts, A.Y. Cairns, Classification of hand bones for bone age assessment, Proc. IEEE ICECS 2 (1996) 1088–1091.

[9] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H.K. Huang, Computer assisted bone age assessment: image preprocessing and epiphyseal/ metaphyseal ROI extraction, IEEE Trans. Med. Imaging. 20 (8) (2001) 715–729.

[10] L. Bocchi, F. Ferrara, I. Nicoletti, G. Valli, An artificial neural network architecture for skeletal age assessment, in: Proceedings of International Conference on Image Processing, 1, 2003, pp. 1077–1080.

[11] C.H. Chang, C.W. Hsieh, T.L. Jong, C.M. Tiu, A fully automatic computerized bone age assessment procedure based on phalange ossification analysis, Proc. IPPR 16 (2003) 463–468.

[12] H.J. Kim, Y.W. Kim, Computerized bone age assessment using DCT and LDA, Proc. ICC Vis./CGCT 4418 (2007) 440–448.

[13] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, H. Huang, Bone age assessment of children using a digital hand atlas, Comput. Med. Imaging Graph. 31 (4–5) (2007) 322–331.

[14] B. Fischer, P. Welter, R.W. Günther, T.M. Deserno, Web-based bone age assessment by content-based image retrieval for case-based reasoning, Int. J. Comput. Assist. Radiol. Surg. 7 (2012) 389–399.

[15] M. Harmsen, B. Fischer, H. Schramm, T. Seidl, T.M. Deserno, Support vector machine classification based on correlation prototypes applied to bone age assessment, IEEE J. Biomed. Health Inf. 17 (1) (2013) 190–197.

[16] D. Haak, H. Simon, J. Yu, M. Harmsen, T.M. Deserno, Bone age assessment using support vector machine regression, in: H.P. Meinzer, T.M. Deserno, H. Handels, T. Tolxdorff (Eds.), Bildverarbeitung für die Medizin 2013, Springer-Verlag, Berlin, 2013, pp. 164–169.

[17] H.H. Thodberg, S. Kreiborg, A. Juul, K.D. Pedersen, The bonexpert method for automated determination of skeletal maturity, IEEE Trans. Med. Imaging 28 (1) (2009) 52–66, http://dx.doi.org/10.1109/TMI.2008.926067.

[18] H.H. Thodberg, L. Sävendahl, Validation and reference values of automated bone age determination for four ethnicities, Acad. Radiol. 17 (11) (2010) 1425–1432.

[19] L. Juan, O. Gwun, A comparison of SIFT, PCA-SIFT and SURF, Int. J. Image Process. 3 (4) (2009) 143–152.

[20] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends Comp. Graph. Vis. 3 (3) (2007) 177–280, http://dx.doi.org/10.1561/0600000017.

[21] A. Canclini, M. Cesana, A. Redondi, M. Tagliasacchi, J. Ascenso, R. Cilla, Evaluation of low-complexity visual feature detectors and descriptors, in: International Conference on Digital Signal Processing, 2013, pp. 1–7. http://dx.doi.org/10.1109/ICDSP.2013.6622757.

[22] J. Klippenstein, H. Zhang, Performance evaluation of visual SLAM using several feature extractors, in: IEEE International Conference on Inteligent Robots System, 2009, pp.1574–1581.

[23] J. Hartmann, J.H. Kluessendorff, E. Maehle, A comparison of feature descriptors for visual SLAM, in: European Conference on Mobile Robots, 2013, pp. 56–61 http://dx.doi.org/10.1109/ECMR.2013.6698820.

[24] M. Kashif, S. Jonas, D. Haak, T.M. Deserno, Bone age assessment meets SIFT, Proc. SPIE 941439 (2015) 1–7, http://dx.doi.org/10.1117/12.2074572.

[25] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Sys. Man. Cyber 9 (1) (1979) 62–66.

[26] P.S. Liao, T.S. Chen, P.C. Chung, A fast algorithm for multilevel thresholding, J. Inf. Sci. Eng. 17 (5) (2001) 713–727.

[27] D.G. Lowe, Distinctive image features from scale invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[28] H. Bay, A. Ess, T. Tuytelaars, L.V. Gool, SURF: speeded up robust features, Comput. Vis. Image Underst. 110 (3) (2008) 346–359.

[29] M. Calonder, V. Lepetit, C. Strecha, P. Fua, BRIEF: binary robust independent elementary features, Proc. ECCV 2010 (2010) 778–792.

[30] S. Leutenegger, M. Chli, C.R. Siegwart, BRISK: binary robust invariant scalable keypoints, IEEE Int. Conf. Comput. Vis. 2011 (2011) 2548–2555.

[31] A. Alahi, R. Ortiz, P. Vandergheynst, FREAK: fast retina keypoint, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 510–517 http://dx.doi.org/10.1109/CVPR.2012.6247715.

[32] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[33] C. Hsu, C. Chang, C. Lin, A Practical Guide to Support Vector Classification, 2003.

[34] C.W. Hsu, C.L. Lin, A comparison of methods for multi-class support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425.

[35] A. Tomei, S. Battisti, M. Martino, D. Nissman, R.C. Semelka, Text-atlas of skeletal age determination, Wiley Blackwell, 2013.

[36] A. Vedaldi, B. Fulkerson, Vlfeat: an open and portable library of computer vision algorithms, MM'10 Proc. ICM (2010) 1469–1472.

[37] C.C. Chang, C.J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.