Automatic Age Estimation and Majority Age Classification From Multi-Factorial MRI Data

Darko Štern, *Member, IEEE*, Christian Payer, *Student Member, IEEE*, Nicola Giuliani, and Martin Urschler, *Member, IEEE*

***Abstract*—Age estimation from radiologic data is an important topic both in clinical medicine as well as in forensic applications, where it is used to assess unknown chronological age or to discriminate minors from adults. In this paper, we propose an automatic multi-factorial age estimation method based on MRI data of hand, clavicle, and teeth to extend the maximal age range from up to 19 years, as commonly used for age assessment based on hand bones, to up to 25 years, when combined with clavicle bones and wisdom teeth. Fusing age-relevant information from all three anatomical sites, our method utilizes a deep convolutional neural network that is trained on a dataset of 322 subjects in the age range between 13 and 25 years, to achieve a mean absolute prediction error in regressing chronological age of 1.01 ± 0.74 years. Furthermore, when used for majority age classification, we show that a classifier derived from thresholding our regression-based predictor is better suited than a classifier directly trained with a classification loss, especially when taking into account that those cases of minors being wrongly classified as adults need to be minimized. In conclusion, we overcome the limitations of the multi-factorial methods currently used in forensic practice, i.e., dependence on ionizing radiation, subjectivity in quantifying age-relevant information, and lack of an established approach to fuse this information from individual anatomical sites.**

***Index Terms*—Information fusion, multi-factorial, convolutional neural network, age estimation, majority age classification, magnetic resonance imaging.**

I. Introduction

AGE estimation of living individuals or human remains is a very active research field in legal medicine and forensic anthropology [1] as well as in clinical medicine [2]. While clinical interest is largest in children close to puberty, e.g., to assess endocrinological diseases [3] or to plan orthopedic interventions [4], [5], interest from legal medicine focuses more on a broad age range around the majority age, i.e., between 13 and 25 years. Recently, majority age classification of children and adolescents without valid identification documents migrating to the European Union has seen a lot of attention, since it is a legally important question to distinguish adult asylum seekers from adolescents who have not yet reached majority age. To estimate unknown chronological age (CA) in children and adolescents, the gradual anatomical changes during physical maturation and growth [6] can be investigated by non-invasive, imaging based radiological methods, predominantly in skeletal [7], [8] and dental structures [9]. This allows experts in forensic radiology and forensic dentistry to examine biological development related to ossification of bones [10] and mineralization of third molars (wisdom teeth) [11]. However, CA estimation is prone to uncertainties [12]. Firstly, estimating CA based on the assessment of biological development is inherently limited due to biological variation among subjects of the same CA [13]. This biological variation defines the lowest error that any method for forensic age estimation can make. With no clear consensus in the literature, the biological variation is assumed to be up to one year in the forensically relevant age range studied in this manuscript. Secondly, due to visual examination, established radiological methods for assessing biological development involve intra- and inter-rater variability [14], which can be eliminated by utilizing software based automatic age estimation.

A. Multi-Factorial Age Estimation

The most extensively studied and widely accepted radiological CA estimation methods are the Greulich-Pyle (GP) atlas method [7] and the Tanner-Whitehouse RUS approach (TW2) [8]. In both methods, biological development of the hand is assessed from X-ray images. While GP is based on representative hand images of different age groups of a sample population, TW2 improves on intra- and inter-rater variability by proposing discrete stages of hand bones separately, according to textual and visual descriptions of their ossification process. These methods are well suitable to follow physical maturation in minors, since hand and wrist bones are finishing ossification at different times. While distal bones finish ossification earlier, proximal bones like radius and ulna close their epiphyseal gaps at an age of about 18 years. However, since the age range of interest for forensic age estimation is between 13 and 25 years,

Manuscript received March 30, 2018; revised July 23, 2018; accepted September 3, 2018. Date of publication August 29, 2018; date of current version July 1, 2019. This work was supported by the Austrian Science Fund (FWF): P28078-N33. The work of N. Giuliani was supported by the city of Graz (A16-021628/2013). *(Corresponding author: Martin Urschler.)*

D. Štern and M. Urschler are with the Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz 8010 Austria, and also with the Institute of Computer Graphics and Vision, Graz University of Technology, Graz 8010, Austria (e-mail: darko.stern@cfi.lbg.ac.at; martin.urschler@cfi.lbg.ac.at).

C. Payer is with the Institute of Computer Graphics and Vision, Graz University of Technology, Graz 8010, Austria (e-mail: christian.payer@icg.tugraz.at).

N. Giuliani is with the Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz 8010, Austria (e-mail: giuliani@student.tugraz.at).

Digital Object Identifier 10.1109/JBHI.2018.2869606

hand and wrist bones alone are not sufficient for accurate predictions. Thus, additional, complementary anatomical sites were investigated to allow an extension of the age estimation range up to 25 years. As recommended by the work group for forensic age diagnostics (AGFAD) [15], in the age range between 13 and 25 years, X-ray images of the hand and wrist bones should be combined with an orthopantomogram (OPG) of the wisdom teeth, and a computed tomography (CT) of the clavicle bones for assessment of biological development. In such a multi-factorial approach, most often the Demirjian method [9] is used for characterizing wisdom teeth development with different stages, and the method of Kellinghaus [16] for assessing the stages of clavicle bone maturation. Nevertheless, the problem of how to combine individual staging results from all three anatomical sites to a multi-factorial age estimate is still an open question in the literature [17], [18].

B. Magnetic Resonance Imaging Based Age Estimation

A major drawback of the above mentioned methods recommended for multi-factorial age estimation is their use of ionizing radiation, which is legally prohibited in healthy subjects for non-diagnostic reasons. However, due to the lack of an established forensic age estimation method without involving ionizing radiation, some European countries have made an explicit exemption to this law in the case of asylum seeking procedures. Recently, to overcome the drawback of ionizing radiation, a lot of research has focused on using magnetic resonance imaging (MRI) for forensic age estimation [19]–[21]. It is currently unclear if the same staging schemes developed for ionizing radiation based methods can also be used for MRI [22], [23]. Therefore, different MRI based methods have been developed for assessing biological development for each of the three anatomical sites [24]–[26], however, these methods still rely on the notion of discretizing biological development into a number of stages and on subjective visual examination.

C. Automatic Age Estimation

To enable objective age estimation without the drawback of intra- or inter-rater variability as introduced by radiologic visual examination, automatic age estimation from X-ray images of the hand has already been proposed in the literature with different methods. In the seminal work of [27], a statistical shape model was used for localization and age estimation. Overcoming the need for localization, very recently [28] showed a deep learning approach involving convolutional neural networks (CNNs) [29] for age estimation, which performed age regression on whole X-ray images of the hand. In 2017, Radiological Society of North America (RSNA) organized a Pediatric Bone Age Challenge intended to show the application of machine learning for estimating age from 14,036 clinical hand radiographs obtained from two children's hospitals [30]. Evaluated on 200 images, the winner of the competition used the deep Inception V3 CNN [31] with additional gender information. Differently to the large interest in automatic age estimation from hand X-ray images, up to our knowledge no machine learning based solutions have yet been proposed for estimating age from clavicle CTs, while for wisdom teeth OPGs a first approach has been shown in [32].

Our group has previously contributed to the development of automated age estimation methods from hand and wrist MRI. In [33] and [34], we have shown a method based on random forests [35], which performs nonlinear regression after dedicated anatomical landmark localization [36] of age-relevant bone structures. Later, we improved performance of the age regression component by training a deep CNN (DCNN) for age estimation in [37].

D. Contributions

In this work we propose a method for MRI based fully automatic multi-factorial age estimation from three anatomical sites (hand, clavicle and teeth) and we apply it for age regression and majority age classification in the forensically relevant age range between 13 and 25 years. Thus, with our novel method we make the following contributions:

- We overcome the problem of ionizing radiation of the recommended multi-factorial approach.
- We eliminate the need for defining discrete staging schemes for individual anatomical sites.
- We provide a solution how to fuse the age estimates from the three sites.
- We learn a nonlinear multi-factorial CA regression function directly from MRI data in an automatic manner.

We have already shown initial results on automatic multi-factorial age estimation in our conference paper at the MICCAI Machine Learning in Medical Imaging (MLMI) workshop [38]. This work extends our preliminary study [38] in the following aspects:

- We provide a more detailed explanation of our DCNN method for multi-factorial age estimation regarding architecture and parameters.
- We thoroughly investigate three DCNN architectures that differ in their strategies of fusing information from the three anatomical sites, compared to the single fusion strategy used in [38].
- We present a method that is robust to missing wisdom teeth, which was ignored in [38] where for each studied subject all wisdom teeth were available.
- We provide a visualization and thorough evaluation of the influence of each individual anatomical site to the estimated age.
- We now evaluate on a much larger dataset of 322 subjects compared to 103 subjects in [38].

II. Method

In our proposed method, we perform multi-factorial age estimation with a DCNN architecture predicting age (see Fig. 1). This nonlinear regression model is based on mapping appearance information from hand and clavicle bones as well as wisdom teeth to the continuous CA target variable. Thus, by extracting age-relevant information for different anatomical sites obtained through cropping from the input MRI data, our approach mimics the established radiological staging approaches

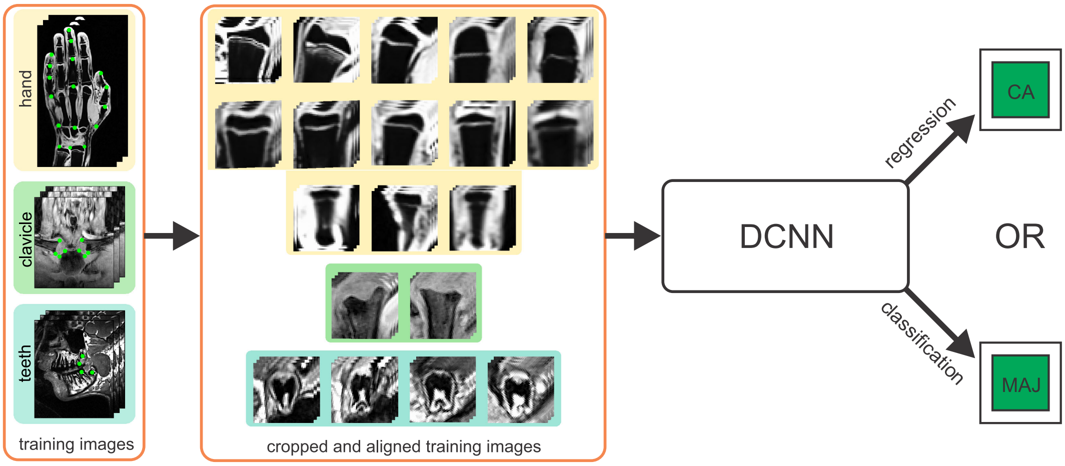


Fig. 1. Overview of our automatic multi-factorial age estimation framework. MRI volumes of the three anatomical sites hands, clavicles and wisdom teeth are cropped according to automatically located anatomical landmarks. A deep convolutional neural network (DCNN) performs the nonlinear mapping between appearance information and chronological age (CA) for both, regression and majority age classification (MAJ) tasks.

developed for each site separately, but without the need for defining discrete stages. In an end-to-end manner, we combine the information coming from different anatomical sites to automatically estimate the age of a subject from its MRI data depicting age-relevant anatomical structures.

A. Cropping of Age-Relevant Structures

Differently to [28], where a large dataset of whole X-ray images is available and used for age estimation, our motivation for cropping age-relevant structures in a separate preprocessing step is to simplify the problem of regressing age from appearance information, such that it is also applicable for datasets that are limited in size. Additionally, compared to the down-sampling of the original 3D images, which inevitably leads to the loss of valuable aging information from the epiphyseal gap regions, cropping of the age-relevant structures also reduces GPU memory requirements and allows us to work on a much higher image resolution. Different automated landmark localization methods as presented in [39], [40], or [36] could be used to accurately localize, align and volumetrically crop age-relevant anatomical structures from skeletal and dental 3D MRI data (see Fig. 1). By locating two anatomical landmarks per bone similar to [37], for the hand MRI data we crop the same 13 bones that are used in the Tanner-Whitehouse RUS method (TW2) [8]. In clavicle MRI data, the two clavicle bones are cropped separately based on two identified landmarks for each clavicle, respectively. The regions encapsulating wisdom teeth are extracted from the dental MRI data using the locations of the centers of second and third molars. In case of a missing wisdom tooth, we estimate its most likely location according to the second molars and extract the region containing the missing tooth as if it would be present.

B. DCNN Architecture

Motivated by how radiologists perform staging of different anatomical sites, we use DCNN blocks [29] to serve as an extractor of age-relevant features for each cropped input volume (see Fig. 2). Each DCNN block consists of three levels of two consecutive $3 \times 3 \times 3$ convolution layers without padding and a max-pooling layer that halves the resolution. Rectified Linear Units (ReLUs) are used as nonlinear activation functions [41]. A fully connected layer at the end of the feature extraction block (fc^{b}) leads to a dimensionality reduced feature representation for each cropped input volume individually, which serves as a feature extractor for that specific anatomical structure.

In this work, we explore three different strategies when to fuse information from anatomical sites within our CNN architecture. The first strategy is to fuse the three anatomical sites directly at the input by concatenating all cropped input volumes as channels before the single DCNN block, followed by two fully connected layers fc^{i} and fc^{o}. We refer to this DCNN as our *early fusion* architecture (see Fig. 3a). In our second *middle fusion* architecture, the sites are fused right after the DCNN blocks (one for each cropped volume) by concatenating the outputs of their fully connected layers fc^{b} before the two fully connected layers fc^{i} and fc^{o} (see Fig. 3b). Finally, in our *late fusion* architecture, the individual DCNN blocks are first combined with fully connected layers fc^{i} for each of the three anatomical sites separately, before fusing the sites with the last fully connected layer fc^{o} that generates the age prediction (see Fig. 3c).

For training, we associate each training sample $s_n, n \in \{1, .., N\}$, consisting of 13 cropped hand bone volumes $s^j_{n,h}, j \in \{1, .., 13\}$, two clavicle regions $s^l_{n,c}, l \in \{1, 2\}$ and four regions covering wisdom teeth $s^k_{n,w}, k \in \{1, .., 4\}$, either with CA as target variable y_n for a regression task, or with a binary variable y_n that is 1 for a minor (m), i.e., CA is smaller than 18 years, and 0 for an adult (a), i.e., CA is larger or equal than 18 years, in a classification task. Optimizing a regression DCNN architecture ϕ with parameters $\mathbf{w}$ is performed by stochastic gradient descent minimizing an L_2 loss on the regression target $\mathbf{y} = (y_1, ..., y_N)^T$:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \tfrac{1}{2} \sum_{n=1}^{N} \|\phi(s_n; \mathbf{w}) - y_n\|^2 \qquad (1)$$

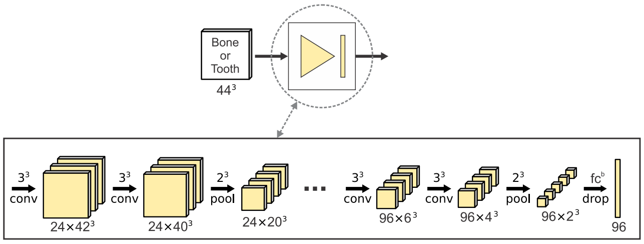
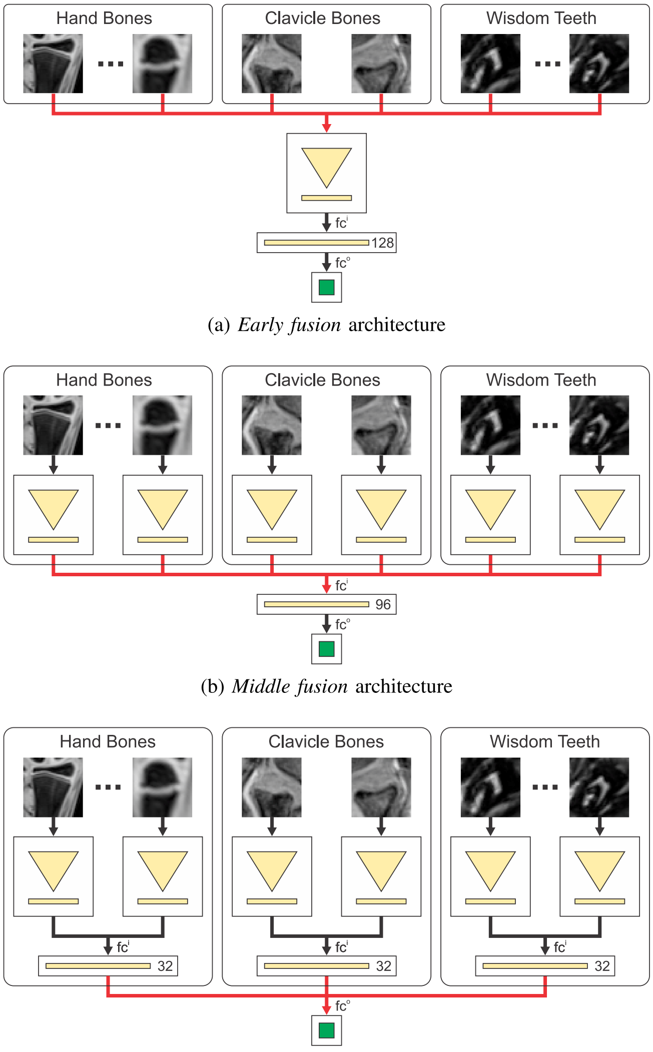


Fig. 2. Individual bone/tooth feature extraction block used in our DCNN architectures for multi-factorial age estimation.



(a) *Early fusion* architecture

(b) *Middle fusion* architecture

(c) *Late fusion* architecture

Fig. 3. Three DCNN architectures for multi-factorial age estimation with (a) early, (b) middle, and (c) late fusion strategies for combining information from the three anatomical sites. The red lines represent the depth level in the network architecture, where multi-factorial information is fused.

To regularize the regression problem, we use a standard weight decay regularization term as well as dropout [42]. For estimating whether a subject is a minor or an adult, the result of the regression DCNN architecture can be used for classification by thresholding the estimated age. In this work, we compare the classification results derived from the regression prediction with the classification results obtained by training the same DCNN architecture with a multinomial logistic classification loss computed as softmax:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \sum_{j\in\{m,a\}} -y_n^j \log \frac{e^{\phi_j(s_n;\mathbf{w})}}{\sum_{k\in\{m,a\}} e^{\phi_k(s_n;\mathbf{w})}} \tag{2}$$

Again weight decay and dropout are used for regularization.

C. Visualization of Influence Per Anatomical Site

To determine the importance of each bone or tooth as well as each anatomical site independently for different predicted ages in our multi-factorial age estimation method, we calculate the influence of the individual DCNN blocks on the network prediction. For each tested sample and its predicted age, we calculate the mean activation value after the fully connected layer fc^b at the end of each feature extraction block (see Fig. 2 and Fig. 3). To visualize the relative importance on the predicted age of each bone or tooth, the mean activation values are normalized to sum up to one. Additionally, we visualize the relative importance on the predicted age of each anatomical site independently, by first calculating the mean activation value of all feature extraction blocks contributing to hand, clavicle and teeth sites separately, followed by a normalization of the three calculated values to sum up to one.

III. Experimental Setup

A. Material

Our MRI dataset was collected at the Ludwig Boltzmann Institute for Clinical Forensic Imaging in Graz as part of a study investigating the role of MRI in forensic age estimation. This study involving male Caucasian volunteers was performed in accordance with the Declaration of Helsinki and approved by the ethical committee of the Medical University of Graz (EK 21399 ex 09/10). All eligible participants provided written informed consent and from underage participants written consent of the legal guardian was additionally obtained. Exclusion criteria were history of endocrinal, metabolic, genetic or developmental disease. We evaluate our proposed multi-factorial age estimation method on a dataset of 3D MRIs from $N = 322$ subjects with known CA ranging between 13.0 and 25.0 years (mean ± std: 19.1 ± 3.3 years, 134 subjects were minors below 18 years at the time of the MRI scan). For each subject, we use as our input for the DCNN architecture the three corresponding MRI volumes of left hand, upper thorax, and the jaw, which were acquired in a single MRI scan session. CA of subjects was calculated as difference between birthday and date of the MRI scan. T1-weighted 3D gradient echo sequences with fat saturation were used for acquiring the hand and clavicle data (physical voxel resolutions of $0.45 \times 0.45 \times 0.9$ and $0.9 \times 0.9 \times 0.9$ mm^3, respectively), while teeth were scanned using a proton density weighted turbo spin echo sequence ($0.59 \times 0.59 \times 1.0$ mm^3). Voxel sizes of the whole input volumes were $288 \times 512 \times 72$ for hand, $168 \times 192 \times 44$ for clavicle, and $208 \times 256 \times 56$ for wisdom teeth, respectively. Acquisition times of hand, clavicle, and wisdom teeth MR sequences were around 4, 6, and 10 min, respectively, but show potential for further acceleration through undersampling [43].

B. Implementation Details

In the architectures shown in Fig. 3, the two convolution layers from each of the three levels of the DCNN block generate 24, 48, and 96 intermediate outputs, while the fully connected layer fc^{b} generates 96 outputs, representing the extracted bone/tooth feature vectors. To address the increased number of inputs for the DCNN blocks of the *early fusion* architecture, this architecture uses four times as many intermediate outputs for both convolution layers and the fully connected layer. Further increasing the number of intermediate outputs was not feasible, due to its demands on GPU memory consumption. In the *early fusion* architecture, the fully connected layer fc^{i} generates 128, in the *middle fusion* architecture 96 outputs. Since the *late fusion* architecture does not combine all DCNN outputs directly, but only the outputs for the three sites separately, fully connected layers fc^{i} with 32 outputs generate the feature vectors of the three individual sites, i.e., hand, clavicles, and teeth. In all architectures, the fully connected layer fc^{o} generates either one output for regression, or two outputs for classification. Optimization was done with the TensorFlow framework [44] using the optimizer ADAM [45] with a maximum of 20,000 iterations, a mini-batch size of 8, and a learning rate of 10^{-4}. We perform L_2 weight decay with a factor of 0.0005, as well as Dropout [42] with a ratio of 0.5 before the fully connected layers to reduce overfitting.

The input volumes were cropped as described in Section II-A and trilinearly resampled to $44 \times 44 \times 44$ voxels for all individual bones/teeth. Due to the varying intensity ranges of the MR images and to be robust to intensity outliers, we scale and shift the intensity values of each cropped volume such that the median of 10% of the lowest intensity values is -1 and the median of 10% of the highest intensity values is 1. During training, images were additionally transformed on-the-fly in a data augmentation step using values sampled from a uniform distribution within the following intervals. The intensity values were shifted by $[-0.1,\ 0.1]$ and scaled by $[0.8,\ 1.2]$. Additionally, the cropped volumes were geometrically transformed using translation by $[-2\text{ mm}, 2\text{ mm}]$, scaling by $[0.85,\ 1.15]$, and rotation by $[-5^\circ, 5^\circ]$ in each dimension.

Training DCNNs for one fold of the cross-validation was around 7 hours, while testing a single subject takes around a second on our system with Intel Core i7 CPU and NVIDIA Geforce GTX 1080 GPU with 8 GB of RAM. These times include cropping of age relevant structures as well as model evaluation and they do not differ between DCNN architectures used for single or multiple anatomical sites.

C. Evaluation Setup

In our evaluation experiments, we used a four fold cross-validation such that each sample from our dataset is tested exactly once and in each cross-validation fold, the tested samples resemble the same age distribution as in our whole dataset. We trained networks for age regression, where we randomly sampled subjects from the training dataset such that the age distribution is uniform over the whole age range from 13 to 25 years. Further, with an age threshold of 18 years, we trained networks for majority age classification, where during training we randomly sampled subjects such that the two classes are equally represented. For all classification experiments, it has to be noted that we defined minors as positive samples and adults as negative samples.

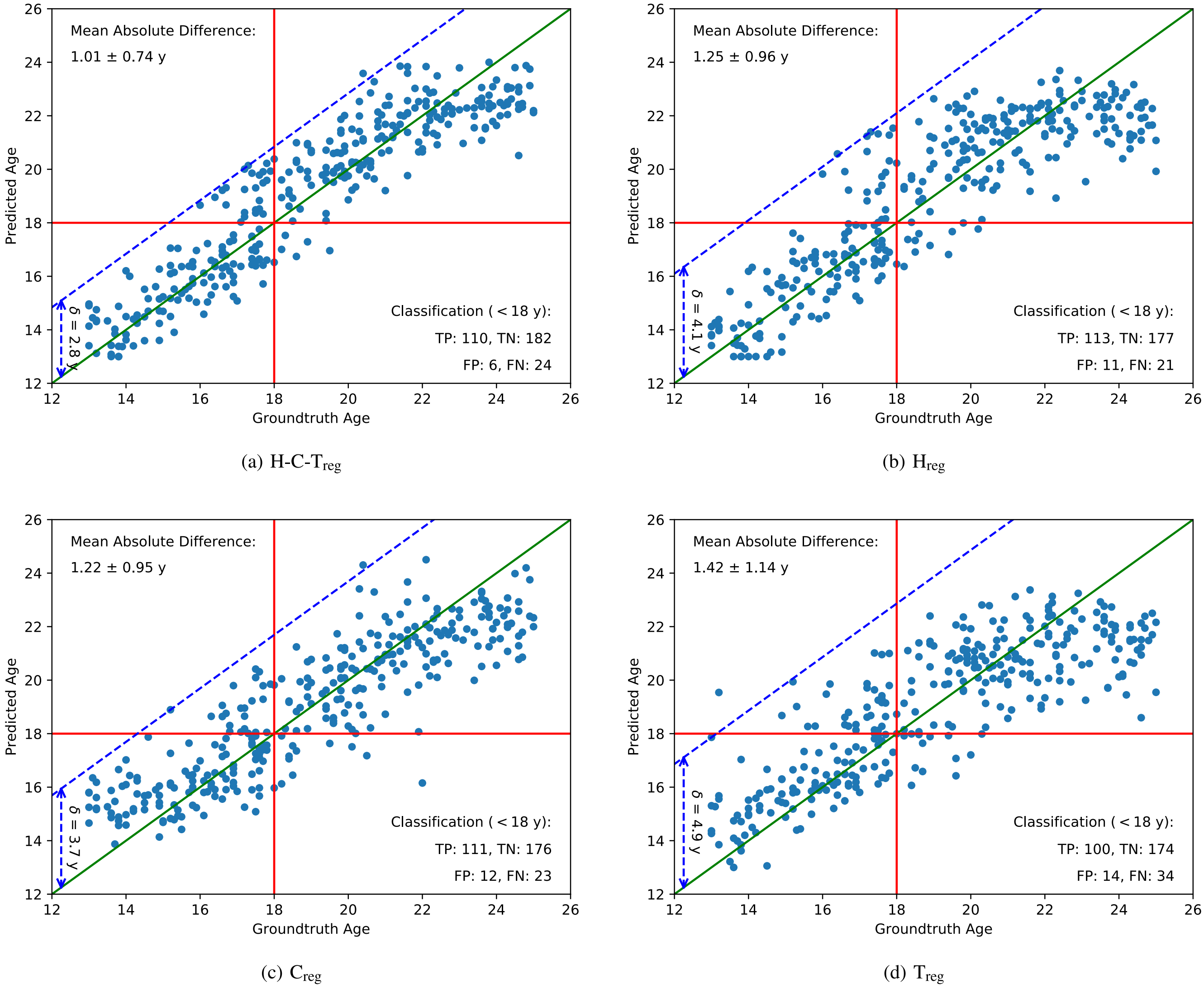
To evaluate the contribution of the three anatomical sites, we investigated networks trained on different combinations of sites. We designate the regression (reg) and classification (class) DCNNs with abbreviations of the anatomical sites on which they are trained, e.g., H_{reg}, H_{class} for hand bones, C-T_{reg}, $\text{C-T}_{\text{class}}$ for the combination of clavicle bones and wisdom teeth, or $\text{H-C-T}_{\text{reg}}$, $\text{H-C-T}_{\text{class}}$ for all three anatomical sites simultaneously. To confirm the radiologist's findings that age relevant information in hand images is only related to the closing of the epiphyseal gaps at around 18–19 years, we additionally evaluate the effect of randomly shuffling ground truth ages above 19 years for H_{reg}. In this experiment, we repeat the shuffling of ages above 19 years three times when training three different age estimation H_{reg} DCNN models.

For the age regression results, we compute mean and standard deviation of absolute differences between predicted and ground truth age as our error measure and provide the 95% confidence intervals of the mean absolute differences. Additionally we perform two-sided paired t-tests of our $\text{N} = 322$ samples, testing the null hypothesis that each method using a single or two anatomical sites shows the same mean absolute difference as our $\text{H-C-T}_{\text{reg}}$ method. Furthermore, we show graphs of the absolute differences over the age range. To generate these graphs, we calculate the mean and standard deviation at a certain age by considering all values that are within ± 1 years.

We evaluate classification experiments by inspecting true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). From the true positive rate (TPR, sensitivity) and the true negative rate (TNR, specificity) for different thresholds of the classifiers, we derive receiver operating characteristic (ROC) curves by plotting sensitivity over $1-$ specificity. To compare classifiers, we use the area under the ROC curve (AUC). Since we define correctly classified minors as true positives, sensitivity indicates the percentage of minors that are correctly classified as minors. To further evaluate our classifiers derived from thresholding the regression prediction of our DCNN architectures, we introduce δ, which quantifies the spread of the prediction error that is related to the forensically more crucial false negative predictions, i.e., the minors classified wrongly as adults. We define δ as the 99th quantile of all prediction errors, computed solely from subjects that were overestimated in age.

IV. Results

In our multi-factorial age estimation experiments we investigated three different DCNN architectures for fusing information from three different anatomical sites on our dataset of 322 subjects in the age range between 13 and 25 years. Cross-validation



(a) H-C-T_{reg}

(b) H_{reg}

(c) C_{reg}

(d) T_{reg}

Fig. 4. Scatter plots showing prediction results from the regression DCNNs for (a) multi-factorial age estimation H-C-T_{reg} as well as separately for (b) hand bones H_{reg}, (c) clavicle bones C_{reg} and (d) wisdom teeth T_{reg}. When thresholding predictions and groundtruth age with 18 years (red lines), the upper left quadrants contain false negatives (i.e., minors wrongly classified as adults), while the lower right quadrants contain false positives (i.e., adults wrongly classified as minors).

results show for our *early fusion* strategy a mean absolute error of 1.18 ± 0.95 years, for the *middle fusion* strategy 1.02 ± 0.76 years and for the *late fusion* strategy 1.01 ± 0.74 years. To allow a comparison to our previous work [38], we also evaluated our best performing *late fusion* architecture on the dataset of 103 subjects, which is a subset of our larger dataset used in this work, resulting in a mean absolute prediction error of 1.06 ± 0.79 years.

We used the *late fusion* architecture H-C-T_{reg} to compare the regression performance when individual anatomical sites alone are used for estimating age, as shown in the scatter plots of Fig. 4. In addition to the scatter plot of H_{reg} in Fig. 4(b), in Fig. 5 we show by shuffling three times the ground truth age of the subjects older than 19 years, that the hand alone cannot be used for predicting age in the forensically relevant age range between 13 and 25 years. Comprehensive regression results for all different combinations of anatomical sites can be found in Table I in terms of mean absolute errors, and in Fig. 6 by plotting the absolute error and standard deviations over our investigated age range. At a threshold of $\alpha = 0.05$ for rejecting our null hypothesis, statistically significant differences between H-C-T_{reg} and other methods can be seen for H_{reg}, C_{reg}, T_{reg}, and H-T_{reg} (see Table I). The influence of three anatomical sites, as well as each bone or tooth separately on the predicted age of the multi-factorial *late fusion* architecture H-C-T_{reg}, is visualized in Fig. 7.

For majority age classification, the ROC curves and their corresponding AUC for individual anatomical sites alone as well as all three sites combined are shown in Fig. 8. The ROC curves in Fig. 8(a) are obtained by varying the threshold of the regression predictions of our *late fusion* architecture. In addition to the best performing ROC curve from Fig. 8(a), Fig. 8(b) shows the ROC curves obtained by varying the threshold of the prediction output of the same *late fusion* architecture trained

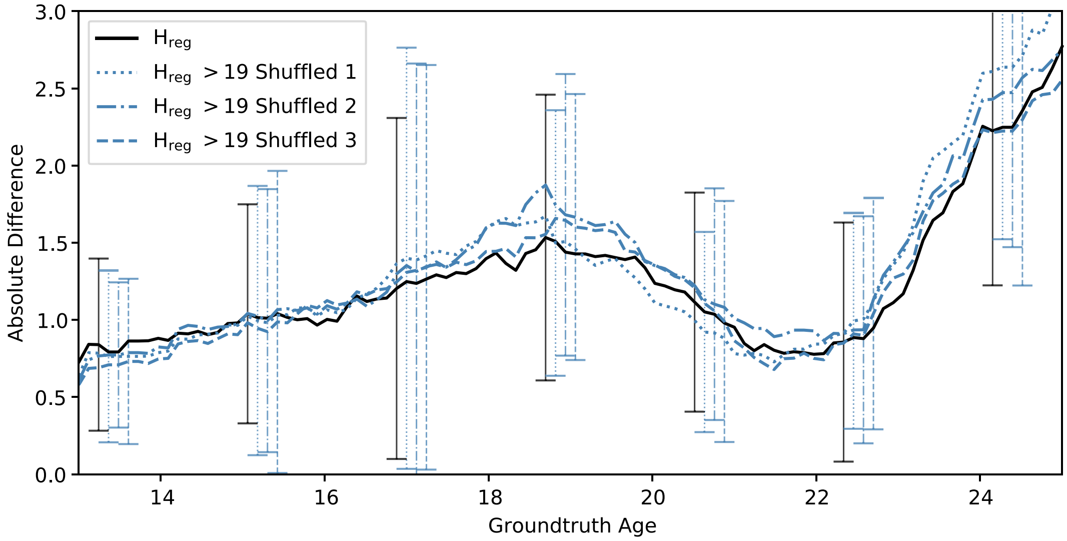


Fig. 5. Mean absolute difference of predicted and groundtruth age for DCNNs trained on hand bones only. The graphs show the results for networks trained on groundtruth age, and for DCNNs trained with ages randomly shuffled above 19 years.

TABLE I

Age Regression Results for DCNNs Trained on Different Combinations of Anatomical Sites, Shown as Mean Absolute Differences and Its Standard Deviation (SD), 95% Confidence Interval (CI) of the Mean, As Well As P Values for Evaluating If There Is a Statistically Significant Difference of Individual Methods Compared With H-C-T_{reg}

	Mean Absolute Difference ± SD	[95% CI]	p value
H-C-T_{reg}	**1.01 ± 0.74**	[0.93; 1.09]	-
H_{reg}	1.25 ± 0.96	[1.14; 1.36]	<0.001
C_{reg}	1.22 ± 0.95	[1.12; 1.32]	<0.001
T_{reg}	1.42 ± 1.14	[1.30; 1.54]	<0.001
H-C_{reg}	1.04 ± 0.78	[0.95; 1.13]	0.481
H-T_{reg}	1.11 ± 0.95	[1.01; 1.21]	0.029
C-T_{reg}	1.10 ± 0.87	[1.00; 1.20]	0.057

TABLE II

Majority Age Classification Performance of Regression and Classification DCNNs Evaluated As False Positive Rates (FPR) and the Corresponding Number of False Positives (FP) for Fixed False Negative Rates (FNR$\in \{0.5\%, 3\%, 6\%, 10\%\}$)

	FPR (FP)			
	FNR=0.5%	FNR=3%	FNR=6%	FNR=10%
H-C-T_{reg}	**25.0% (47)**	**18.1% (34)**	**11.7% (22)**	**7.4% (14)**
H_{reg}	54.3% (102)	43.1% (81)	25.5% (48)	10.6% (20)
C_{reg}	36.2% (68)	25.5% (48)	19.1% (36)	12.8% (24)
T_{reg}	55.3% (104)	24.5% (46)	21.8% (41)	17.0% (32)
H-C_{reg}	33.5% (63)	20.7% (39)	13.3% (25)	9.0% (17)
H-T_{reg}	57.4% (108)	19.7% (37)	16.5% (31)	8.5% (16)
C-T_{reg}	28.7% (54)	22.9% (43)	19.7% (37)	11.2% (21)
H-C-T_{class}	66.0% (124)	38.3% (72)	27.1% (51)	15.4% (29)
H_{class}	83.5% (157)	77.1% (145)	57.4% (108)	21.8% (41)
C_{class}	70.2% (132)	43.1% (81)	34.6% (65)	20.2% (38)
T_{class}	76.1% (143)	47.3% (89)	43.1% (81)	25.0% (47)

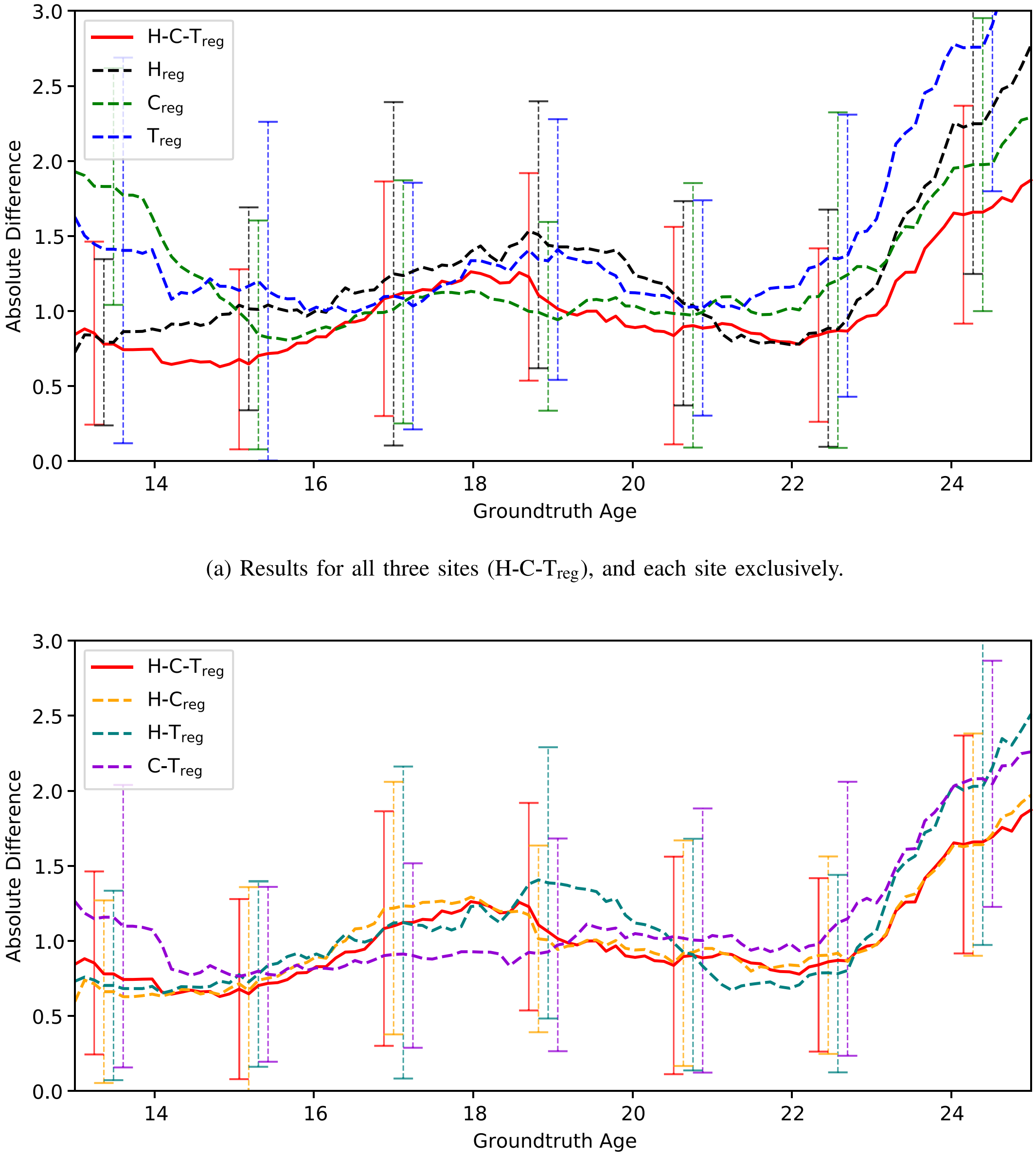
with a classification loss. For the specific threshold of 18 years, in Fig. 4 we show the classification result of the regression based classifiers in terms of TP, TN, FP and FN as well as δ, which robustly estimates the spread of the prediction error solely for the subjects whose age is overestimated. Finally, in Table II we compare the different regression and classification based majority age classifiers in terms of their specificity for several fixed sensitivities ($99.5\%, 97\%, 94\%$ and 90%).

V. Discussion

Motivated by the lack of a standardized way of fusing information from multiple complementary anatomical sites for age estimation, and with the aims to prevent the use of ionizing radiation as well as to reduce observer variability, in this work we proposed an automatic method for estimating age in the range between 13 and 25 years from MRI data. It is for the first time that a comprehensive evaluation of an automatic approach for information fusion from different anatomical sites was carried out on this large, forensically relevant age range.

A. Comparison of Different Fusion Strategies

In this work, we investigated with three different strategies when to fuse information from anatomical sites within our CNN architecture. Following the spirit of deep CNNs that the network is capable to extract all information relevant for an estimation task on its own, in our *early fusion* strategy input regions from all anatomical sites are fused by concatenating them before they are presented to the network. With a mean absolute error of 1.18 ± 0.95 years in regressing age, this strategy was outperformed by the other two, since the translation invariance, an important property of CNNs, cannot be fully exploited in our limited training dataset due to the large variations in the relative position of anatomical sites when being concatenated. In the other two strategies, we first extract age-relevant features in a CNN block and then combine features on two different levels. In the *middle fusion* strategy, information from all bones and teeth are fused immediately after features are extracted. This strategy corresponds to a forensic expert looking at the images of the individual anatomical structures simultaneously and mentally fusing all information when estimating age in a multi-factorial manner. The performance of this strategy was similar to our third *late fusion* network architecture, which first combines information for each anatomical site individually, followed by fusing the three anatomical sites with a fully connected layer. Also used



(a) Results for all three sites (H-C-T$_{reg}$), and each site exclusively.

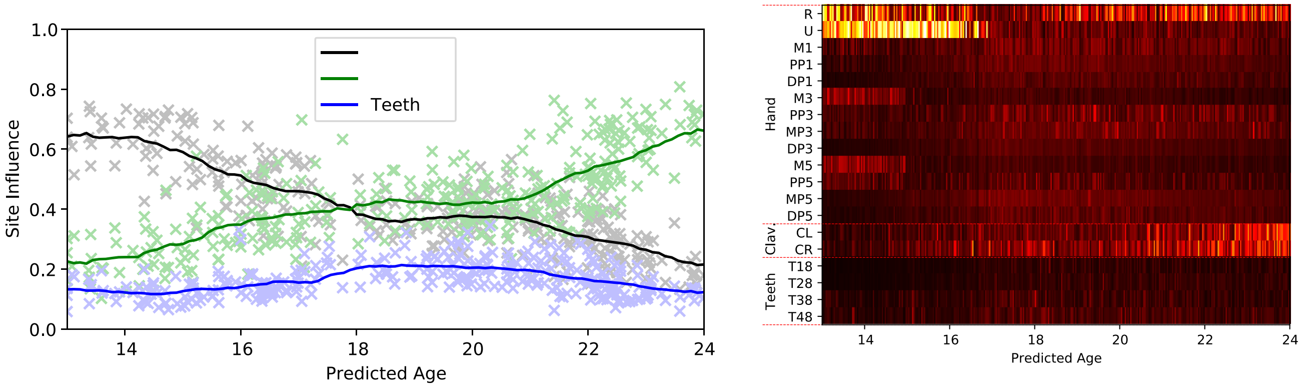
(b) Results for all three sites (H-C-T$_{reg}$), and combinations of two sites.

Fig. 6. Prediction performance of the DCNNs. The graphs show mean absolute differences and the standard deviation of the predicted age to the groundtruth for DCNNs trained on different sites and their combinations.

in our previous work [38], the *late fusion* strategy is inspired by how forensic experts are currently combining individual information from hand radiographs, wisdom teeth OPGs and clavicle CTs in practice when performing multi-factorial age estimation. We used the *late fusion* network architecture for our further evaluations due to its excellent age regression performance in terms of mean absolute regression error of 1.01 ± 0.74 years.

B. Comparison to Previous Work

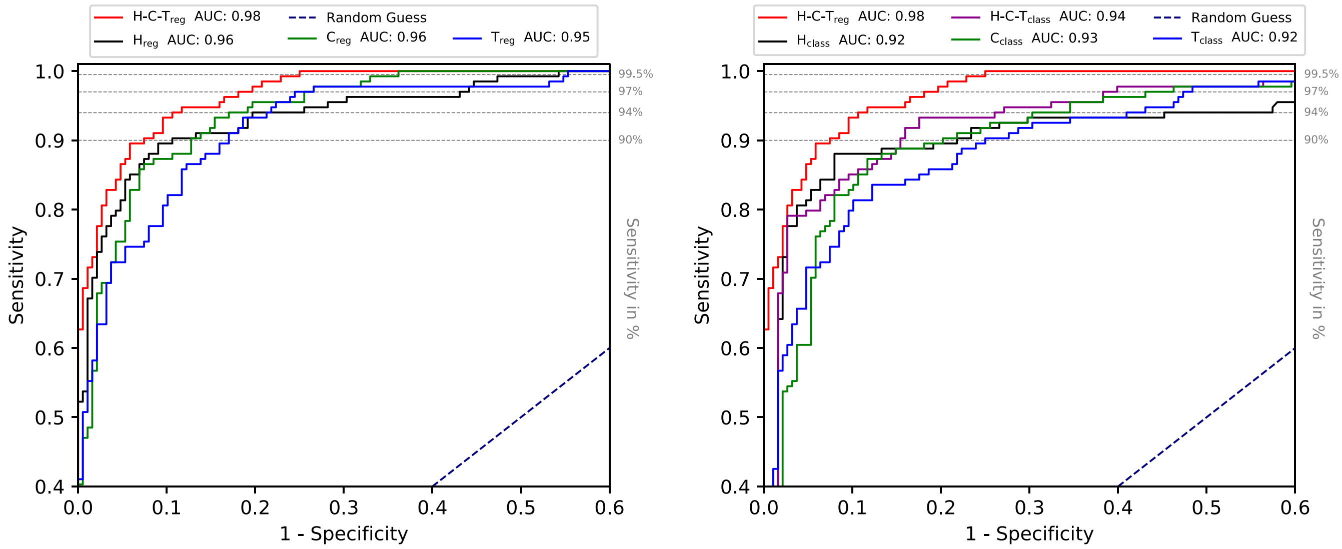
In our previous work [38], we solely studied the *late fusion* strategy and compared it to a random forest based approach on a much smaller MRI dataset of 103 subjects. This random forest used image intensity features generated by randomly selecting first an anatomical site, then an individual bone or tooth, and finally a location in the image to discriminate the age of a subject. Although this random forest showed state-of-the-art performance in [46] when only the hand was used for predicting age up to 19 years, in [38] when all three anatomical sites were fused, it resulted in a large mean absolute error of 1.93 ± 1.26 years on the dataset of 103 subjects. On the other hand, in [38] the *late fusion* DCNN architecture that we are using in this work achieved an error of 1.30 ± 1.13 years on the dataset of 103 subjects. For the same setup, but pretrained on the radiological estimation using the GP method for the hand, the Demirjian method for the teeth and the Kellinghaus method for the clavicle, the performance was further improved after finetuning to an error of 1.14 ± 0.96 years in [38]. In the present work we did not



(a) Influence from the fully connected layers, grouped by hand, clavicles, and teeth.

(b) Influence from the fully connected layers of all bones/teeth.

Fig. 7. Visualization of the contributions of (a) the three anatomical sites and (b) each bone and tooth to the age prediction of the H-C-T_{reg} DCNN architecture. The crosses in (a) show the individual site influence for each predicted age and curves the mean value of influence at each age. In (b) dark values indicate low, bright values high influence to the age prediction.



(a) Regression DCNNs.

(b) Classification DCNNs.

Fig. 8. ROC curves for the DCNNs performing majority age classification and corresponding AUC.

use pretraining, since it requires a costly and tedious procedure of annotating each anatomical site, which was not yet performed on our larger MRI dataset of 322 subjects. However, by changing the optimizer and better tuning our hyperparameters like mini-batch size and stopping criterion, we were able to outperform our previous results on the smaller dataset of 103 subjects (1.06 ± 0.79 years) without the need for pretraining. Furthermore, on our larger dataset of 322 subjects studied in this work, we achieved the best result presented so far for automatic age estimation from multi-factorial MRI data (1.01 ± 0.74 years).

C. Limitations of Age Estimation Solely From a Single Site

According to the recommendations of AGFAD [15], estimation of age in the forensically relevant age range between 13 and 25 years should not be performed by using information from the hand alone (H_{reg}), since epiphyseal gaps of hand bones close on average around the age of 18 years. The regression results from H_{reg} as well as from the multi-factorial H-C-T_{reg} DCNN for our dataset of 322 subjects are shown in Fig. 4(a) and Fig. 4(b). Comparing these two scatter plots, it can be seen that when only the hand is used for prediction, there is a saturation after 18 years, which leads to a prediction biased towards 22 years. The same behavior can be observed even better from the results of our shuffling experiments in Fig. 5, where the absolute error shows a local minimum around 22 years. This is also supported by theory, since the expected value of a model trained on data that carries no regression information and is uniformly distributed in a specific age range, is the average value of that age range. Thus, for subjects older than 19 years, our results confirm the forensic viewpoint that no age-relevant information can be extracted solely from hand images, but instead a DCNN trained solely on the hand learns the average age of the respective subjects with saturated age relevant information.

While H_{reg} can be used for predicting age up to 19 years, the use of T_{reg} or C_{reg} can extend this age range, since wisdom teeth mineralization can be followed until an age of 21 years [47] and ossification of clavicle bones finishes around the age of 24 years [16]. However, when using solely clavicle bones or wisdom teeth, the biological variation, i.e., the variation in chronological age of subjects with the same developmental progress, is much higher compared to using hand bones in subjects with an age below 19 years. This can be seen from the scatter plots in Fig. 4 and the plot of the absolute errors in Fig. 6(a), which further show that wisdom teeth on its own are the most unreliable anatomical site for age estimation. Thus, each site on its own is either not sufficient to predict age in the age range between 13 and 25 years, or shows high uncertainty in predictions due to biological variation.

D. Multi-Factorial Age Estimation

The main contribution of our work is the extension of the maximal age range for age regression from up to 19 years, which is possible when solely using hand images, to up to 25 years, which is enabled by fusing the multi-factorial data from three anatomical sites. The contribution of information from each site to multi-factorial age estimation for the whole age range is shown in Fig. 7(a). We can see that our *late fusion* DCNN mainly learned to estimate age up to 18 years by extracting age-relevant information from the hand bones, which is consistent with forensic findings that hand bones are the most informative source of information for estimating age in this age range. Confirming the recommendations of AGFAD [15], the relevance of the aging information coming from the clavicle bones is constantly increasing when subjects get older. By the age of 18 years, clavicle bones overtake hand bones regarding the importance as a source of information for predicting age. However, between 18 and 21 years our *late fusion* DCNN still relies on the information that the epiphyseal gaps of hand bones have closed to reduce the uncertainty coming from the biological variation of clavicle bone development in this age range. For the last two years of our forensically relevant age range, clavicle bones become the dominant source of information, since the late ossification stages of the clavicle bones are a strong indicator that a subject is older than 22 years. From Fig. 7(a) we can also see that information from wisdom teeth can be extracted by our method, although the teeth may be missing, and that they have some influence on predicting age in the age range between 17 and 22 years. However, overall the contribution of the teeth is much smaller than from the other two anatomical sites. The same behavior that is visible in Fig. 7(a) is shown in more detail for individual bones and teeth in Fig. 7(b), where it can be seen that radius and ulna carry most relative importance for the hand, while both clavicle bones are equally used by our *late fusion* DCNN.

E. Do We Need All Three Sites?

Inline with the finding that wisdom teeth are showing much less influence on age estimation in our *late fusion* DCNN, results in Table I show that the highest mean absolute regression error occurs when using solely wisdom teeth (T_{reg}), followed by H_{reg} and C_{reg}. This is further confirmed in the case of omitting one anatomical site from multi-factorial age estimation, since there we also see that the best regression performance can be achieved with omitting wisdom teeth (H-C_{reg}). Furthermore, there is no statistically significant difference between our overall best result when using all three sites (H-C-T_{reg}) and when combining only hand with clavicle (H-C_{reg}). However, in the age range from 16 to 19 years our detail results in Fig. 6(b) show that highest estimation accuracy is achieved when teeth is combined with clavicle information (C-T_{reg}). This supports the hypothesis that bone and teeth provide complementary information since teeth and bone cells originate from distinct embryonic germ layers, thus differently affecting the influence of genetic factors on growth. Currently it is not clear from our results, why our *late fusion* DCNN combining all three sites did not achieve the highest age estimation accuracy over the whole age range indicated by the lower bound of the graphs in Fig. 6(b). Especially for the age range between 16 and 19 years, this limitation has to be investigated in future work. Nevertheless, our findings confirm the recommendation of AGFAD [15] that a multi-factorial analysis should be done when predicting age in the forensically relevant age range between 13 and 25 years, while the use of teeth remains an open issue for further investigation.

F. Majority Age Classification

A specific challenge in forensic age estimation is majority age classification of asylum seekers lacking valid identification documents. In this ethically sensitive scenario, legal authorities have to take special attention to avoid mis-classifications of minors as adults. Thus, a low number of false negatives, i.e., minors misclassified as adults, has higher priority compared with a low number of false positives, i.e., adults misclassified as minors. We have performed two types of experiments for majority age classification. Firstly, we used the results of our *late fusion* DCNN for regression by thresholding its prediction output, and secondly, we trained a dedicated binary classifier with the same network architecture for majority age classification. From the results in the ROC curves of Fig. 8, we can see that in terms of AUC our multi-factorial majority age classifier derived from the regression DCNN (H-C-T_{reg}, AUC $= 0.98$) outperforms both regression and classification based types of classifiers that solely use information from a single anatomical site. This further strengthens the hypothesis that age estimation benefits from a multi-factorial approach, however, it can also be seen from Fig. 8(a) that solely using hand (H_{reg}) or clavicle (C_{reg}) for classification based on thresholding the regression prediction, the performance in terms of AUC is competitive (AUC $= 0.96$) within our dataset. In Fig. 8(b), similar behavior can be seen from the ROC curves obtained from the prediction probability of the binary classifiers. However, classification performance is inferior compared with the regression based classifiers, even for multi-factorial H-C-T_{class}. Overall, the results of the ROC curves in Fig. 8 indicate that a DCNN trained with the regression loss from Eq. (1) is better suited for classification than a DCNN trained with the binary classification loss from Eq. (2). Thus, we found that regression should be used for majority age classification, since the regression loss allows the network to explore the distance in age between different subjects, while in binary classification the inter-class distance is fixed and independent of the age difference of subjects belonging to different classes.

Motivated by the necessity from legal practice to reduce the number of false negatives, we evaluated in Table II how a decrease in sensitivity of the different classifiers affects their specificity. From this evaluation we see that for all classifiers, increasing the threshold of minors classified as adults from 0.5%, i.e., a single false negative, to 10%, i.e., 14 false negatives, decreases the number of adults being wrongly classified as minors, thus reducing the false positive rate. For our best performing H-C-T_{reg}, the false positive rate is reduced from 25% when allowing a single false negative to 7.4% when allowing 14 false negatives. To achieve the high sensitivity of 99.5%, i.e., allowing only a single false negative, the predicted age from H-C-T_{reg} has to be thresholded at an age of 20.22 years, while for 97%, 94%, and 90% sensitivity, the corresponding age thresholds are 19.91, 19.59, and 18.95 years, respectively. When thresholding predicted age exactly at 18 years, the sensitivity is low with 82.1%, but the specificity is very high with 96.8%, thus indicating that this threshold favors a low number of adults misclassified as minors. This low sensitivity and high specificity for the age prediction threshold of 18 years is also visible in the scatter plot of Fig. 4(a), where the upper left quadrant corresponds to the false negatives and the lower right quadrant corresponds to the false positives. For the case of regression being used for classification, we introduced the distance δ in this scatter plot. Additionally to the sensitivity, δ robustly quantifies the forensically relevant classification error for minors being wrongly classified as adults. Defined as the distance between the green line indicating the optimal prediction from regression and the dashed blue line capturing the spread of the prediction error solely for subjects whose age is overestimated, the distance $\delta = 2.8$ years for H-C-T_{reg} is much smaller than the distances for H_{reg}, C_{reg}, and T_{reg}, respectively. Thus, having the smallest spread in errors of the overestimated predictions, we expect H-C-T_{reg} to generalize best to a larger population of subjects when aiming for minimization of the number of minors wrongly classified as adults, a behavior that significantly impacts the involved subjects to their advantage.

G. Outlook

We see our work as a foundation for a novel multi-factorial age estimation method to be used in forensic practice, which would require that our results are reproduced on a different, potentially larger dataset and compared with multi-factorial age estimates combining individual estimates of forensic experts and dentists. As we have shown in our preliminary study [38], these individual estimates might also be beneficial for pretraining our prediction networks leading to an improvement of our automatic multi-factorial age estimation. Although in this work we have outperformed our previous results on a much larger dataset without the need for pretraining, we will investigate this interesting aspect in our future work.

VI. Conclusion

In this work, we extensively studied a deep learning based multi-factorial age estimation method from MRI data of 322 subjects from a large, forensically relevant age range between 13 and 25 years, which automatically fuses information from hand bones, clavicle bones, and wisdom teeth. We presented for the first time such an approach that overcomes several limitations of the method currently used in forensic practice, i.e., the use of ionizing radiation, the subjectivity due to assigning discrete staging schemes for the individual anatomical sites, and the lack of consensus in how information from individual sites should be fused into a final age estimate. After studying different network architectures, we showed that multi-factorial age estimation is possible by automatically fusing age-relevant information from all individual sites. With a prediction error of 1.01 ± 0.74 years, we outperformed age estimation results solely derived from hand, clavicle, or teeth data separately in the age range between 13 and 25 years. In this work, we also investigated the legally important question of majority age classification, by comparing thresholded predictions from the same regression method with results from a dedicated binary classifier that was trained with the same DCNN architecture. Our results showed that the regression based method is better suited for this task, however, due to the high biological variation of subjects with the same chronological age, special care has to be taken to select the compromise between minors being wrongly classified as adults and adults being wrongly classified as minors.

Acknowledgment

The authors would like to thank P. Kainz, N. Martinez, and T. Widek for their help in preparation of our multi-factorial age estimation dataset.

References

[1] K. Latham, E. Bartelink, and M. Finnegan, *New Perspectives in Forensic Human Skeletal Identification*. Amsterdam, The Netherlands: Elsevier, 2018.

[2] D. D. Martin *et al.*, "The use of bone age in clinical practice - part 1," *Hormone Res. Paediatrics*, vol. 76, no. 1, pp. 1–9, 2011.

[3] D. D. Martin *et al.*, "The use of bone age in clinical practice - part 2," *Hormone Res. Paediatrics*, vol. 76, no. 1, pp. 10–16, 2011.

[4] S. C. Lee, J. S. Shim, S. W. Seo, K. S. Lim, and K. R. Ko, "The accuracy of current methods in determining the timing of epiphysiodesis," *Bone Joint J.*, vol. 95-B, no. 7, pp. 993–1000, 2013.

[5] W. W. J. Wang *et al.*, Correlation of Risser sign, radiographs of hand and wrist with the histological grade of iliac crest apophysis in girls with adolescent idiopathic scoliosis. *Spine*, vol. 34, no. 17, pp. 1849–1854, 2009.

[6] J. M. Tanner, *Foetus Into Man: Physical Growth From Conception to Maturation*. London, U.K.: Open Books, 1978.

[7] W. W. Greulich and S. I. Pyle, *Radiographic Atlas of Skeletal Development of the Hand and Wrist*, 2nd ed. Stanford, CA, USA: Stanford Univ. Press, 1959.

[8] J. M. Tanner, M. J. R. Healy, N. Cameroon, and H. Goldstein, *Assessment of Skeletal Maturity and Predicion of Adult Height (TW2 Method)*, 2nd ed. New York, NY, USA: Academic, 1983.

[9] A. Demirjian, H. Goldstein, and J. M. Tanner, "A new system of dental age assessment," *Hum. Biol.*, vol. 45, no. 2, pp. 211–227, 1973.

[10] U. Baumann, R. Schulz, W. Reisinger, W. Heinecke, A. Schmeling, and S. Schmidt, "Reference study on the time frame for ossification of the distal radius and ulnar epiphyses on the hand radiograph," *Forensic Sci. Int.*, vol. 191, no. 1–3, pp. 15–18, 2009.

[11] H. M. Liversidge, "The assessment and interpretation of Demirjian, Goldstein and Tanner's dental maturity," *Ann. Hum. Biol.*, vol. 39, no. 5, pp. 412–431, 2012.

[12] N. Cameron, Can maturity indicators be used to estimate chronological age in children? *Ann. Hum. Biol.*, vol. 42, no. 4, pp. 300–305, 2015.

[13] T. J. Cole, "The evidential value of developmental age imaging for assessing age of majority," *Ann. Hum. Biol.*, vol. 42, no. 4, pp. 379–388, 2015.

[14] P. Kaplowitz, S. Srinivasan, J. He, R. McCarter, M. R. Hayeri, and R. Sze, "Comparison of bone age readings by pediatric endocrinologists and pediatric radiologists using two bone age atlases," *Pediatric Radiol.*, vol. 41, no. 6, pp. 690–693, 2011.

[15] A. Schmeling, P. M. Garamendi, J. L. Prieto, and M. I. Landa, "Forensic age estimation in unaccompanied minors and young living adults," in *Forensic Medicine - From Old Problems to New Challenges*, D. N. Vieira, Ed. London, U.K.: InTech, 2011, ch. 5, pp. 77–120.

[16] M. Kellinghaus, R. Schulz, V. Vieth, S. Schmidt, H. Pfeiffer, and A. Schmeling, "Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans," *Int. J. Legal Med.*, vol. 124, pp. 321–325, 2010.

[17] B. Gelbrich *et al.*, "Combining wrist age and third molars in forensic age estimation: how to calculate the joint age estimate and its error rate in age diagnostics," *Ann. Hum. Biol.*, vol. 42, no. 4, pp. 389–396, 2015.

[18] P. W. Thevissen, J. Kaur, and G. Willems, "Human age estimation combining third molar and skeletal development," *Int. J. Legal Med.*, vol. 126, no. 2, pp. 285–292, 2012.

[19] J. Dvorak, J. George, A. Junge, and J. Hodler, "Application of MRI of the wrist for age determination in international U-17 soccer competitions," *Brit. J. Sports Med.*, vol. 41, no. 8, pp. 497–500, 2007.

[20] E. Hillewig, J. De Tobel, O. Cuche, P. Vandemaele, M. Piette, and K. Verstraete, "Magnetic resonance imaging of the medial extremity of the clavicle in forensic bone age determination: A new four-minute approach," *Eur. Radiol.*, vol. 21, no. 4, pp. 757–767, 2011.

[21] Y. Terada *et al.*, "Skeletal age assessment in children using an open compact MRI system," *Magn. Reson. Med.*, vol. 69, no. 6, pp. 1697–1702, 2013.

[22] P. Baumann *et al.*, "Dental age estimation of living persons: Comparison of MRI with OPG," *Forensic Sci. Int.*, vol. 253, pp. 76–80, 2015.

[23] M. Urschler *et al.*, "Applicability of Greulich-Pyle and Tanner-Whitehouse grading methods to MRI when assessing hand bone age in forensic age estimation: A pilot study," *Forensic Sci. Int.*, vol. 266, pp. 281–288, 2016.

[24] J. De Tobel, E. Hillewig, and K. Verstraete, "Forensic age estimation based on magnetic resonance imaging of third molars: Converting 2D staging into 3D staging," *Ann. Hum. Biol.*, vol. 44, no. 2, pp. 121–129, 2017.

[25] E. Hillewig *et al.*, "Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: Towards more sound age estimates," *Int. J. Legal Med.*, vol. 127, no. 3, pp. 677–689, 2013.

[26] S. Serinelli *et al.*, "Accuracy of MRI skeletal age estimation for subjects 12–19. Potential use for subjects of unknown age," *Int. J. Legal Med.*, vol. 129, no. 3, pp. 609–617, 2015.

[27] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The BoneXpert method for automated determination of skeletal maturity," *IEEE Trans. Med. Imag.*, vol. 28, no. 1, pp. 52–66, Jan. 2009.

[28] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-ray images," *Med. Image Anal.*, vol. 36, pp. 41–51, 2017.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[30] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, 2017.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[32] J. De Tobel, P. Radesh, D. Vandermeulen, and P. W. Thevissen, "An automated technique to stage lower third molar development on panoramic radiographs for age estimation: A pilot study," *J. Forensic Odonto-Stomatol.*, vol. 35, no. 2, pp. 49–60, 2017.

[33] D. Štern, T. Ebner, H. Bischof, S. Grassegger, T. Ehammer, and M. Urschler, "Fully automatic bone age estimation from left hand MR images," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014* (Lecture Notes in Computer Science), vol. 8674, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Cham, Switzerland: Springer, 2014, pp. 220–227.

[34] M. Urschler, S. Grassegger, and D. Štern, "What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents," *Ann. Hum. Biol.*, vol. 42, no. 4, pp. 358–367, Jul. 2015.

[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[36] M. Urschler, T. Ebner, and D. Štern, "Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization," *Med. Image Anal.*, vol. 43, no. 1, pp. 23–36, 2018.

[37] D. Štern, C. Payer, V. Lepetit, and M. Urschler, "Automated age estimation from hand MRI volumes using deep learning," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016* (Lecture Notes in Computer Science), vol. 9901, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, Eds. Athens, Greece: Springer, Cham, 2016, pp. 194–202.

[38] D. Štern, P. Kainz, C. Payer, and M. Urschler, "Multi-factorial age estimation from skeletal and dental MRI volumes," in *Machine Learning in Medical Imaging. MLMI 2017* (Lecture Notes in Computer Science), vol. 10541, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds. Cham, Switzerland: Springer, pp. 61–69.

[39] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1862–1874, Sep. 2015.

[40] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using CNNs," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016* (Lecture Notes in Computer Science), vol. 9901, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016, pp. 230–238.

[41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[43] B. Neumayer *et al.*, "Reducing acquisition time for MRI-based forensic age estimation," *Sci. Rep.*, vol. 8, 2018, Art. no. 2063.

[44] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Oper. Syst. Des. Implementation*, Berkeley, CA, USA, 2016, pp. 265–283.

[45] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[46] D. Štern and M. Urschler, "From individual hand bone age estimates to fully automated age estimation via learning-based information fusion," in *Proc. IEEE 13th Int. Symp. Biomed. Imag.*, Prague, Czech Republic, 2016, pp. 150–154.

[47] A. Olze, A. Schmeling, K. Rieger, G. Kalb, and G. Geserick, "Untersuchungen zum zeitlichen Verlauf der Weisheitszahnmineralisation bei einer deutschen Population," *Rechtsmedizin*, vol. 13, pp. 5–10, 2003.