Bone Age Assessment with X-ray Images Based on Contourlet Motivated Deep Convolutional Networks

Xun Chen[†], Chao Zhang[‡], Yu Liu[‡]*
[†]Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230026, China
[‡]Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China
Email: xunchen@ustc.edu.cn, zhangc1994@mail.hfut.edu.cn, yuliu@hfut.edu.cn

***Abstract*—Bone age assessment (BAA) is a widely performed procedure for skeletal maturity evaluation in pediatric radiology. It has various clinical applications such as diagnosis of endocrine disorders, monitoring of growth hormone therapy and prediction of final adult height for adolescents. Recent studies indicate that deep learning techniques have great potential in developing automated BAA methods with significant improvements in terms of conventional computer-assisted approaches. In this paper, we propose a multi-scale feature fusion framework for bone age assessment based on deep convolutional neural networks. In our method, the non-subsampled contourlet transform (NSCT) is firstly performed on an input left-hand radiograph to obtain its multi-scale and multi-direction representations. Then, the decomposed bands at each scale are fed to a convolutional network that contains a series of convolutional and pooling layers for feature extraction, respectively. Finally, the feature maps from different branches are concatenated and put into a regression network consisting of several fully connected layers to obtain the bone age estimation. Experimental results on a public BAA dataset demonstrate that the proposed method can achieve state-of-the-art performance.**

I. Introduction

As a frequently performed procedure in pediatric radiology, bone age assessment (BAA) aims to determine the discrepancy between a child's skeletal age and chronological age, which may indicate abnormalities in skeletal development. BAA is of great significance for the diagnosis of endocrine disorders and monitoring of growth hormone therapy. Assessment of skeletal age is also typically used to predict a child's final adult height. The most popular approach for BAA is based on a radiological examination of left (non-dominant) hand and wrist to reveal the maturity of skeletal development, due to the advantages like simplicity and minimum radiation exposure. Fig. 1 shows an example hand radiograph for BAA.

Currently, there exist two widely-used clinical methods for radiologists and pediatricians to assess skeletal age based on the hand X-ray images: Greulich and Pyle (G&P) [1] and Tanner-Whitehouse (TW) [2]. The G&P method determines the skeletal age by comparing the whole target radiograph with a list of standard bone age maps known as an atlas. In contrast, the TW method gives a more detailed solution to this problem. It takes into account a set of local regions of interest (ROIs) in an X-ray image and provides an evaluation by a numerical score for each ROI that corresponds to a specific bone. The final bone age is estimated by combining all the scores using some pre-defined strategies. Although the G&P and TW methods have their own characteristics such as simplicity for G&P and precision for TW, these manual methods suffer from several common drawbacks. They both rely heavily on the domain knowledge and experience of radiologists. A considerable intra-rater and inter-rater variability always exists in clinical practice [3]. In addition, it is predictable that the time efficiency of assessing the bone age in a manual manner is relatively low. In the past few years, the computer-assisted methods for automatic bone age assessment have emerged an active topic in this area, and a variety of automatic BAA approaches have been proposed [4]–[12]. These methods usually handle BAA as a classification or regression problem which involves essential components like hand segmentation, ROI detection, feature extraction and classifier/regressor designing. Most recently, the methods based on deep learning techniques, and in particular convolutional neural networks (CNNs) [13], [14], have appeared in the literature [15], [16], leading to promising results and obvious advantages over conventional BAA methods.

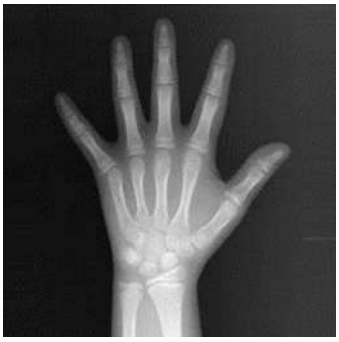

Fig. 1. An example hand radiograph for BAA. This image originates from the public BAA dataset *Digital Hand Atlas Database System* [17]. The image has been resized to 256*256 pixels, keeping aspect ratio by performing continuous extension on the short side of the image. The patient is a 10.49 years old (chronological age) girl while the estimated results of her skeletal age from two radiologists using this radiograph are 11.5 and 11.0 years old, respectively.

In this paper, we concentrate on the study of CNN-based skeletal age assessment. A multi-scale feature fusion framework for BAA based on deep neural networks is proposed. Specifically, we first perform non-subsampled contourlet trans-

* Corresponding author: Yu Liu

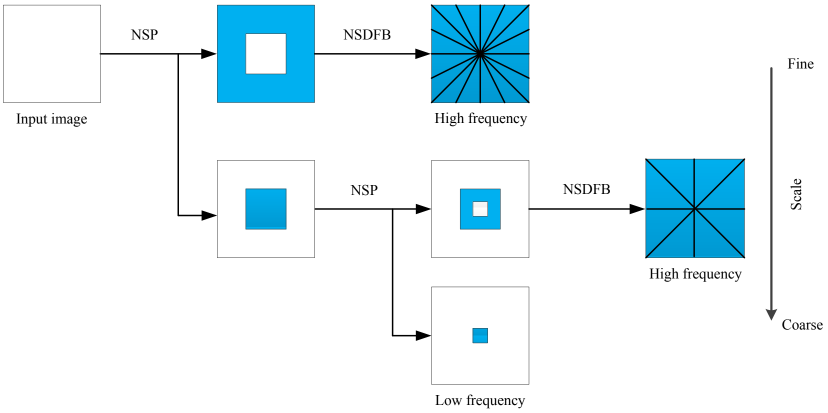

Fig. 2. The schematic diagram of a two-level NSCT decomposition.

form (NSCT) [18] on an input X-ray hand image to obtain its multi-scale and multi-direction representations. Then, the decomposed bands at each scale are fed to a convolutional network that consists of a series of convolutional and pooling layers for feature extraction, respectively. Finally, the feature maps from different branches are concatenated and put into a regression network containing several fully connected layers to output the estimated bone age. Experimental results on the public BAA dataset *Digital Hand Atlas Database System* [17] show that the proposed method can achieve more competitive results in comparison to some state-of-the-art methods.

The rest of this paper is organized as follows. In Section 2, some related work and the motivation of this study are provided. Section 3 presents the proposed BAA method in detail. Experimental results and discussions are given in Section 4. Finally, Section 5 concludes the paper and puts forward some future work.

II. Related work and motivation

A. Convolutional neural networks for bone age assessment

As a representative supervised deep learning (DL) architecture, convolutional neural networks (CNNs) have gained great success in a variety of communities including computer vision, remote sensing, medical image analysis, etc. In the field of bone age assessment (BAA), CNN-based studies appeared relatively late. In 2017, Lee et al. [15] proposed a fully-automated CNN-based BAA system involving data preparation, image preprocessing, CNN-based hand region detection and CNN-based age classification. The popular CNN model GoogLeNet [19] is employed for age classification using transfer learning in their method. They experimentally verified that the CNN-based method can outperform previous methods on a large-scale dataset (collected from the authors' hospital while not publicly available) containing more than 8000 radiographs from patients with chronological age of 5-18 years old. In the same year, Spampinato et al. [16] introduced a general DL-based BAA architecture consisting of two consecutive networks: a convolutional network for feature extraction and a regression network for bone age estimation. They tested several widely used convolutional networks such as GoogLeNet and VGGNet [20]. Two types of transfer learning techniques, namely, using off-the-shelf features and fine-tuning from networks trained on general imagery, are both studied in [16]. They also proposed a new architecture called BoNet by introducing a deformation layer in the convolutional network. The experiments in [16] are performed on a public BAA dataset [17], which contains nearly 1400 hand X-ray images from children with chronological age of 0-18 years old. Considerable improvements are achieved by the CNN-based method in comparison to many conventional BAA methods.

B. Non-subsampled Contourlet Transform

Non-subsampled Contourlet Transform (NSCT) [18] is a representative multi-scale geometric analysis approach that has been widely used in many image processing problems such as image denoising and image fusion. In comparison to some early multi-scale image decomposition approaches like pyramid, wavelet and curvelet, contourlet transform is verified to be a more optimal image representation method as it can capture the details of an image at diverse scales and directions more effectively. By applying the non-subsampling strategy to contourlet transform, NSCT also owns the shift-invariance property. The implementation of NSCT is mainly based on non-subsampled pyramid (NSP) decomposition and non-subsampled directional filter bank (NSDFB). The NSP decomposition is employed to obtain the multi-scale representations of an input image from fine to coarse. It is achieved by a two-channel non-subsampled 2-D filter bank. At each decomposition level, one high-frequency band and one low-frequency band of the same size as the input image are produced. As a result, by applying an L-level NSP decomposition, we can obtain $L+1$ sub-bands including L high-frequency bands and one low-frequency band. For each high-frequency band, the NSDFB is applied to obtain its multi-direction representations at the corresponding scale. The obtained band at each direction is also of the same size as the input image. The number of directions at each scale can be adjusted and it is typically set to a value that is the power of two. Fig. 2 shows the schematic diagram of a two-level NSCT decomposition.

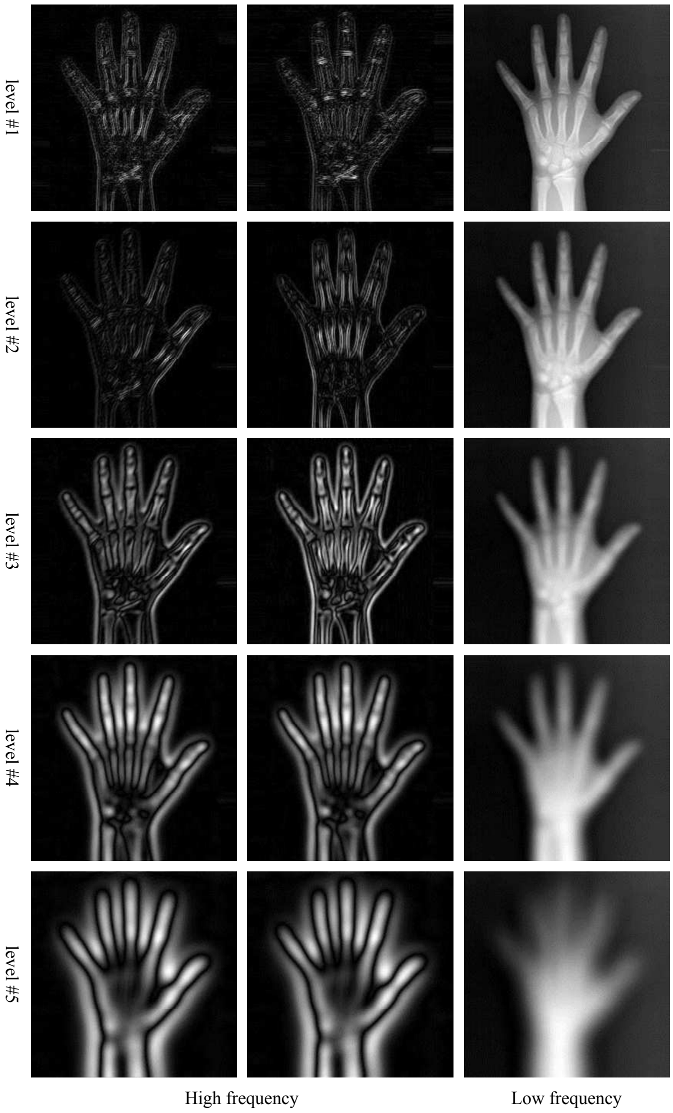

Fig. 3. A 5-level NSCT decomposition on the example image in Fig. 1. Only two of the high-frequency bands at each decomposition level are shown due to the limitation of space. The intermediate low-frequency band at each decomposition level is also shown.

C. Motivation of this work

Despite of the great potential exhibited by the CNN-based bone age assessment methods, the related study is just at the beginning stage and there still exists a lot of scope for further improvement in this area. In this work, we present a CNN-based multi-scale feature fusion framework for BAA. Instead of putting the original image into a deep convolutional network, the NSCT decomposition bands are employed as the network input. This is mainly based on the consideration that the quantity of training examples in BAA is usually very limited, which may cause difficulty in network training even through transfer learning could be applied. By pre-extracting some low-level features, the above problem is likely to be alleviated to some extent. To ensure the assessment performance, the pre-extracted features must be as rich as possible. Since NSCT can provide multi-scale and multi-direction representations for an image, it is selected as the feature extraction approach in this work. Moreover, the NSCT bands at different scales are put into independent convolutional layers for further feature extraction, so that a feature-level fusion mechanism can be naturally generated by combining the features coming from different scales. The improvement on algorithm performance is therefore expected.

III. The Proposed Method

A. NSCT for hand X-ray images

In consideration of computational efficiency and memory cost, all the X-ray images in the dataset are resized to 256×256 pixels, keeping aspect ratio by performing continuous extension on the short side of images.

Fig. 3 shows the results of performing a 5-level NSCT decomposition on the example image given in Fig. 1. The pyramid filter “pyrexc” and directional filter “vk” are employed because of their good performance in feature extraction for medical images [21]. The numbers of directional bands at different scales from fine to coarse are set to 16, 8, 8, 4 and 4 [21], [22] as a finer scale generally needs more directions for description. Due to limitation of space, we only show two of the high-frequency bands at each decomposition level in Fig. 3 (the absolute values of high-frequency bands are used for a better visualization). The intermediate low-frequency band at each decomposition level is also shown.

We can see from Fig. 3 that there is a clear trend with regard to the features captured in different decomposition levels. Features of different scales are captured at different decomposition levels. The features become coarser with the increase of the decomposition level. In particular, it can be seen that the features captured in the bands at level 5 are too coarse (only the main contours are extracted) that almost all the useful information for BAA has disappeared. For the other four levels, distinctive features always exist, even though they are at different scales. Therefore, a 4-level NSCT decomposition is selected in our BAA method, and the numbers of directional bands at scales from fine to coarse are set to 16, 8, 8 and 4.

B. Network architecture

Fig. 4 shows the proposed CNN-based multi-scale feature fusion framework for bone age assessment (BAA). The architecture consists of three main parts: NSCT decomposition, convolution and regression.

The NSCT decomposition part is used to generate multi-directional features at different scales. As discussed in the above subsection, a 4-level NSCT decomposition is employed to obtain four scales of high-frequency bands and one low-frequency band. At each scale, the obtained high-frequency bands are concatenated together to generate a multi-channel format as the network input. The low-frequency band is also employed as the input of an individual branch to make full use of the original information contained in the input image.

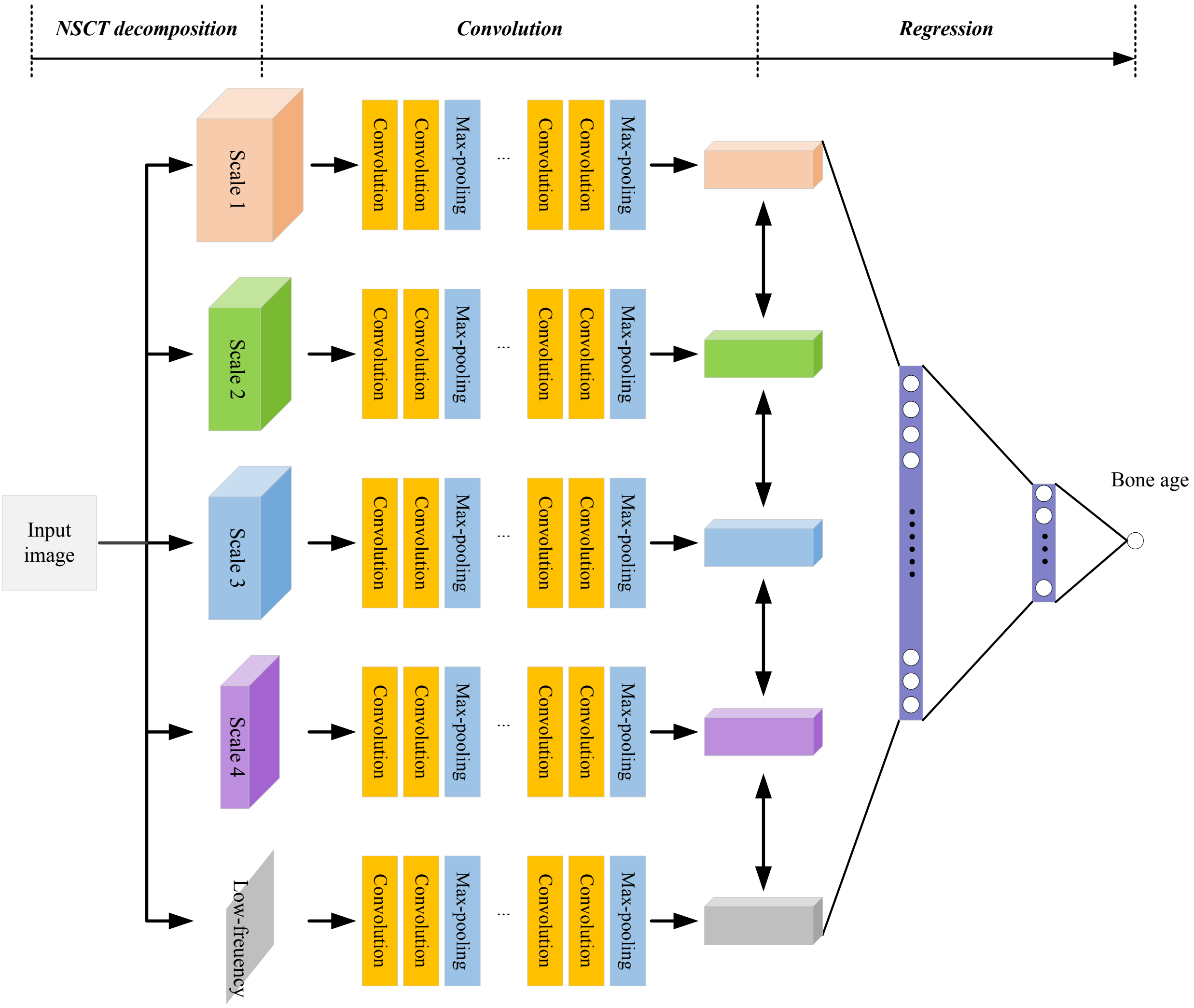

Fig. 4. The proposed CNN-based multi-scale feature fusion framework for bone age assessment.

The convolution part contains five branches to accept the above NSCT decomposition bands. Each branch is constructed by a series of convolutional layers and max-pooling layers. In this work, we employ the convolutional part of the 16-layer VGGNet [20] as the architecture of each branch as it has been verified to be a very effective network for BAA [14]. Transfer learning has shown its superiority on various deep learning (DL)-based medical image analysis problems including BAA [15], [16]. In this paper, the pre-trained VGGNet weights by general imagery (the ImageNet dataset), which is publicly available at Caffe Model Zoo [23], are used as the initial weights for fine-tuning. As the channel number of each branch input of our network is not equal to that of natural RGB images related to the pre-trained VGGNet model, the weights of the first convolutional layer in each branch are randomly initialized. The output feature maps of different branches are finally concatenated to achieve a feature-level fusion.

The regression part adopts the concatenated feature maps as the input. It consists of two fully-connected layers (1024 neurons and 128 neurons are used in our method, respectively) and one output layer with a single neuron estimating the bone age. We apply the mean squared error as our loss function:

$$L(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} |f(X_i, \Theta) - y_i|^2, \tag{1}$$

where Θ denotes the network parameters and n indicates the number of training examples in a mini-batch. X_i is the i-th training image in the batch and y_i is its ground truth bone age that is generally provided by radiologists or domain experts. The mini-batch stochastic gradient descent method is employed for network optimization.

IV. Experiments

In the field of bone age assessment (BAA), the public X-ray datasets are quiet limited and most previous methods were tested on non-public datasets, making a fair and comprehensive comparison impossible. In this paper, we verify the effectiveness of our method on the Digital Hand Atlas

TABLE I
THE PERFORMANCE OF DIFFERENT BAA METHODS ON MEAN SQUARED ERROR (MAE) OVER THE DIGITAL HAND ATLAS DATABASE.

Method	Reading 1	Reading 2	Average of two MAEs	Average of two readings
Ref. [6]	2.78	2.37	2.57	-
Ref. [7]	2.60	1.70	2.15	-
Ref. [9]	2.46	2.38	2.42	-
Ref. [11]	1.92	1.73	1.82	-
Fine-tuned VGGNet [16]	0.88	0.79	0.83	-
Fine-tuned GoogLeNet [16]	0.86	0.79	0.82	-
BoNet [16]	0.80	0.79	0.79	-
Proposed method	**0.75**	**0.75**	**0.75**	**0.72**

Database System [17], which is a publicly available BAA dataset released by Gertych et al. [7]. This dataset was also employed in a recently proposed CNN-based BAA method [16]. The dataset contains nearly 1400 left-hand radiographs of children of ages ranging from 0 to 18 years old, categorized by race and gender. Two bone age values evaluated by two expert radiologists are provided for each X-ray image.

As mentioned in Section 3, before performing NSCT decomposition, all the X-ray images in the dataset are resized to 256×256 pixels, keeping aspect ratio. For the decomposed NSCT bands, data augmentation was carried out by extracting nine uniformly spaced crops of size 224×224, which is in accord with the input size of VGGNet. In our training procedure, the batch size is set to 4. The momentum factor and weight decay are fixed as 0.9 and 0.0005, respectively. The training process is conducted for about 80 epochs over the augmented dataset. The learning rate is initially set to 0.0001 for the first 60 epochs and is dropped by a factor of 10 for the last 20 epochs. For each example, the average value of two expert readings is employed as the label (ground truth target) for training. The experiments are conducted on Caffe [24].

For the sake of fair comparison, we follow the evaluation approach that was performed in [16]. Specifically, a 5-fold cross validation is adopted to evaluate the algorithm performance over the whole dataset. The mean absolute error (MAE) between the estimated bone age and the ground truth is employed as the evaluation metric. We also calculate the MAE between the estimated value and each individual reading. Several existing bon age assessment methods [6], [7], [9], [11], [16] are used for performance comparison. In fact, most previous methods are either tested on non-public datasets or the related source code is not available, which increases the difficulty of objective comparison. Fortunately, the authors in [16] reported their experimental results about the performances of some existing methods on the Digital Hand Atlas Database. The comparison in this paper is based on the related results reported in [16].

Table I lists the performance of different BAA methods on MAE over the Digital Hand Atlas Database. The best results are shown in bold. The column "Average of two MAEs" contains the average MAE values over the two readings (its left two columns). The last column denotes the MAE between the estimated age and the average of two readings. Please note that this value can be smaller than each of individual MAE from a mathematical point of view. Actually, this is always the case in practice, which mainly because the average of two readings acts as the ground truth for training. However, these MAE values of the compared methods were not provided in [16]. It can be seen from Table I that the CNN-based BAA methods (last four) significantly outperform other methods. The proposed method achieves an improvement of about 0.04-0.08 years in terms of the state-of-the-art VGGNet, GoogLeNet and BoNet based methods presented in [16]. For the computational efficiency, it takes about 0.03 seconds to assess the bone age for an X-ray image on a computer with a NVIDIA GTX 1080 GPU and a 4.0 GHz CPU.

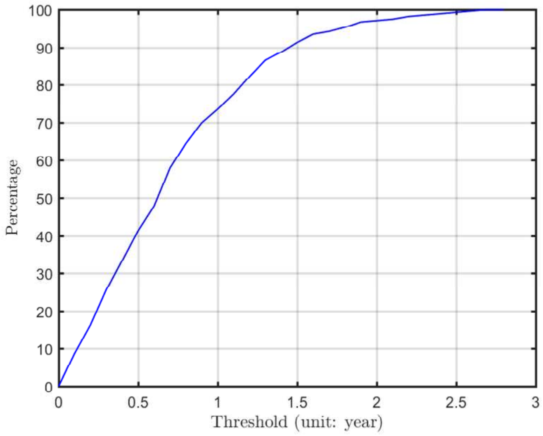

Fig. 5. The percentage of examples that have an absolute error less than a given threshold.

To have some further insights into the performance of our method, we calculate the percentage of examples that have an absolute error less than a certain threshold. By changing this threshold from 0 with an increasing step of 0.1 years, we can obtain a series of percentage values to construct a curve. The result is shown in Fig. 5. It can be seen that more than 40% examples have an absolute error less than 0.5 years. When the threshold changes to 1 year, the percentage increases to nearly 75%. This curve actually provides a new evaluation approach for bone age assessment, but unfortunately the detailed output results of other BAA methods are not available to generate this type of curves for comparison.

V. Conclusion

Contribution- In this paper, we propose a multi-scale feature fusion framework for bone age assessment based on deep convolutional neural networks. To overcome the training difficulty caused by limited quantity of training examples in bone age assessment, we pre-extract low-level features from the original X-ray image by performing non-subsampled contourlet transform (NSCT), which can effectively capture sufficient multi-scale and multi-direction features for the input image. In addition, a feature-level fusion mechanism is designed by applying a multi-branch network architecture to combine the features obtained from different scales individually. Performance comparison with some state-of-the-art methods verifies the effectiveness of the proposed method.

Future Work- It should be noted that this paper just provides a preliminary study on the presented CNN-based bone age assessment framework. Actually, this framework owns very high flexibility that a number of issues could have impacts on the final performance, such as the image preprocessing techniques, the parameters for multi-scale and multi-direction decomposition, the specific network architecture, etc. More experiments and comparisons are expected to fully study these issues. In the future, we will conduct a comprehensive and systematic study on this topic. We believe the performance of this method could be further improved.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grants 61701160, 81571760 and 61501164), the Provincial Natural Science Foundation of Anhui (Grant 1808085QF186) and the Fundamental Research Funds for the Central Universities (Grant JZ2018HGTB0228).

References

[1] W. Greulich and S. Pyle, “Radiographic atlas of skeletal development of the hand and wrist,” *The American Journal of the Medical Sciences*, vol. 238, no. 3, p. 393, 1959.

[2] J. Tanner, R. Whitehouse, N. Cameron, W. Marshall, M. Healy, and H. Goldstein, *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. Academic Press, London, 1983.

[3] M. Berst, L. Dolan, M. Bogdanowicz, M. Stevens, S. Chow, and E. Brandser, “Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the greulich and pyle standards,” *American Journal of Roentgenology*, vol. 176, no. 2, pp. 507–510, 2001.

[4] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H. Huang, and V. Gilsanz, “Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal roi extraction,” *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 715–729, 2001.

[5] E. Pietka, “Computer-assisted bone age assessment-database adjustment,” *International Congress Series*, vol. 1256, pp. 87–92, 2003.

[6] C. Hsieh, T. Jong, and C. Tiu, “Bone age estimation based on phalanx information with fuzzy constrain of carpals,” *Medical & Biological Engineering & Computing*, vol. 45, no. 3, pp. 283–295, 2007.

[7] A. Gertych, A. Zhang, J. Sayre, S. Kurkowska, and H. Huang, “Bone age assessment of children using a digital hand atlas,” *Computerized Medical Imaging and Graphics*, vol. 31, pp. 322–331, 2007.

[8] A. Vega and J. Arribas, “A radius and ulna tw3 bone age assessment system,” *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 5, pp. 1463–1476, 2008.

[9] D. Giordano, C. Spampinato, G. Scarciofalo, and R. Leonardi, “An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones,” *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 10, pp. 2539–2553, 2010.

[10] M. Harmsen, B. Fischer, H. Schramm, T. Seidl, and T. Deserno, “Support vector machine classification based on correlation prototypes applied to bone age assessment,” *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 190–197, 2013.

[11] D. Giordano, I. Kavasidis, and C. Spampinato, “Modeling skeletal bone development with hidden markov models,” *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 138–147, 2016.

[12] S. Simu and S. Lal, “A study about evolutionary and non-evolutionary segmentation techniques on hand radiographs for bone age assessment,” *Biomedical Signal Processing and Control*, vol. 33, pp. 220–235, 2017.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based leaning applied to document recognition,” *Proceedings of The IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[15] H. Lee, S. Tajmir, J. Lee, M. Zissen, B. Yeshiwas, T. Alkasab, G. Choy, and S. Do, “Fully automated deep learning system for bone age assessment,” *Journal of Digital Imaging*, vol. 30, pp. 427–441, 2017.

[16] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, “Deep learning for automated skeletal bone age assessment in x-ray images,” *Medical Image Analysis*, vol. 36, pp. 41–51, 2017.

[17] http://ipilab.usc.edu/BAAweb/.

[18] A. L. da Cunha, J. Zhou, and M. N. Jo, “The nonsubsampled contourlet transform: theory, design, and applications,” *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3089–3101, 2006.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *arXiv*, vol. 1409.4842, pp. 1–12, 2014.

[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv*, vol. 1409.1556v2, pp. 1–10, 2014.

[21] S. Li, B. Yang, and J. Hu, “Performance comparison of different multi-resolution transforms for image fusion,” *Informaton Fusion*, vol. 12, pp. 74–84, 2011.

[22] Y. Liu, S. Liu, and Z. Wang, “A general framework for image fusion based on multi-scale transform and sparse representation,” *Information Fusion*, vol. 24, pp. 147–164, 2015.

[23] https://github.com/BVLC/caffe/wiki/Model-Zoo.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 675–678.