Extract Bone Parts Without Human Prior: End-to-end Convolutional Neural Network for Pediatric Bone Age Assessment

Chuanbin Liu[1], Hongtao Xie[1(✉)], Yizhi Liu[2], Zhengjun Zha[1], Fanchao Lin[1], and Yongdong Zhang[1]

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China
htxie@ustc.edu.cn

[2] Hunan University of Science and Technology, Xiangtan 411201, China

Abstract. Pediatric bone age assessment (BAA) is a common clinical practice to investigate endocrinology, genetic and growth disorders of children. The morphological characters of different specific bone parts, such as wrist and phalanx, have important reference significance in BAA. Previous deep learning approaches can be divided into two branches, (1) the single-stage structure ignores the attention on specific bone parts, thus it can be trained end-to-end but suffers from low accuracy, (2) the multi-stage structure extracts the bone parts with human prior, thus it exhibits high accuracy but suffers from model generalization and resource consumption problem. To enable an end-to-end training method extracting discriminative bone parts automatically without human prior, in this paper, we propose a novel single-stage Attention-Recognition Convolutional Neural Network (AR-CNN). The AR-CNN consists of one attention agent for discriminative bone parts proposing and one recognition agent for feature learning and age assessment. The attention agent can discover and extract bone parts automatically, meanwhile the recognition agent can learn the features from the proposing bone parts and assess the bone age. Furthermore, the assessment result will be fed back to attention agent for the optimization of bone parts extracting. Therefore, the two agents can reinforce each other mutually and the overall network can be trained end-to-end without human prior. To the best of our knowledge, this is the first end-to-end structure to extract bone parts for BAA without segmentation, detection and human prior. Experimental results show that our approach achieves state-of-the-art accuracy on the public RSNA datasets with mean absolute error(MAE) of 4.38 months.

Keywords: Bone age assessment · Deep learning · Object detection

1 Introduction

Pediatric bone age assessment (BAA) is a common clinical practice to investigate endocrinology, genetic and growth disorders of children [1,2]. Based on

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11769, pp. 667–675, 2019.
https://doi.org/10.1007/978-3-030-32226-7_74

the discrepancy between the reading of the bone age and the chronological age, physicians can make accurate diagnoses of abnormal development in children. Currently, the left-hand X-ray image is widely used for assessing the bone age, and the morphological characters of different specific bone parts, such as wrist and phalanx, have important reference significance in BAA. There are a series of popular standards of BAA, i.e., the Greulich and Pyle (G&P) standard and the Tanner-Whitehouse (TW) standard, which extract a different set of specific bone part as regions of interest (ROIs) for assessment. Taking different standard as the reference, conventional manual assessment methods mainly rely on personal experience and opinion of the clinicians, which show some intrinsic limitations with low efficiency, unstable accuracy, and expensive time-consuming. Recent years, benefitting from a huge amount of data, deep learning methods have achieved impressive success [3–5] and a series of deep learning approaches have been proposed for BAA.

Related Work: The deep learning methods for bone age assessment can be divided into two categories: The first single-stage structure adopts an end-to-end learning method [6–8], where the entire image is taken as input to a convolutional neural network (CNN) and predicts the bone age directly. Larson et al. [8] take ResNet50 as the backbone to output a probability score for each month. Spampinato et al. [7] design a customized six-layer network with one deformation layer for age regression. Their models can be trained end-to-end, but ignore the attention on the specific bone parts as regions of interest. Consequently, the precision is limited. Moreover, it is confusing to visualize and interpret the results to clinicians [9].

The second category uses a multi-stage structure with image preprocessing and human prior knowledge [9–12], segmenting the hands out from original radiography, detecting and extracting the bone parts with human prior knowledge, then generating the prediction result. Iglovikov et al. [11] segment the hands out from radiography by U-Net and then crop out the carpal bones, metacarpals and proximal phalanges for ensemble regression. Wang et al. [10] detect the distal radius and ulna areas from the hand by Faster-RCNN to estimate the bone age. The multi-stage structure with human prior brings improvement in accuracy, as well as a series of limitations [7]: (1) the visual features identified by domain experts may not suitable for automated methods, and the strict human prior limits the generalization of deep learning. (2) it requires extra labels and algorithms in detection and segmentation, which brings additional costs. (3) it cannot be trained end-to-end and has high complexity and time expenditure.

Contribution: To combine the advantages of single-stage structure and multi-stage structure, specifically, to enable an end-to-end training method extracting discriminative bone parts automatically without human prior, in this paper, we propose a novel single-stage Attention-Recognition Convolutional Neural Network (AR-CNN) for bone age assessment. As shown in Fig. 1, the AR-CNN consists of one attention agent for discriminative bone parts proposing and one

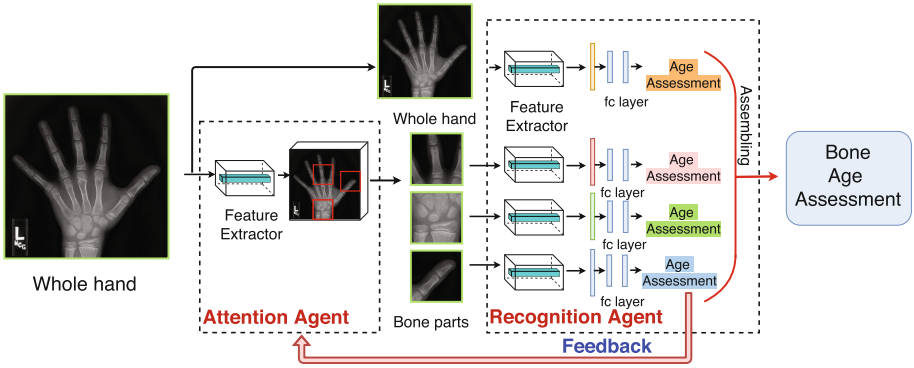

Fig. 1. The framework of AR-CNN.

recognition agent for feature learning and age assessment. The attention agent can discover and extract the discriminative bone parts automatically without human prior knowledge. Meanwhile the recognition agent can learn the features from the proposing ROIs and assess the bone age. Furthermore, the assessment result in recognition agent will be fed back to attention agent for the optimization of bone parts extracting. To the best of our knowledge, this is the first end-to-end structure to discover and extract bone parts for bone age assessment without any segmentation, detection and human prior.

2 Method

Overall Framework: Figure 1 illustrates the overall AR-CNN framework for bone age assessment. The attention agent takes the entire hand image as input and produces the proposal of discriminative bone parts as ROIs. These regions are cropped from the input radiography and fed into the recognition agent. The recognition agent learns the features from these bone parts as well as the input image, and produces the corresponding assessment results. Meanwhile, the assessment results are fed back to attention agent for optimization. To the end, it assembles all the assessment results and makes a final prediction.

Attention Agent: It has been proven by weakly-supervised object detection [13], that the higher-layer activation maps of a CNN for classification, can indicate the location of the discriminative parts. Therefore, the detection of an object can be realized even without bounding box annotations. Inspired by this idea, the attention agent of AR-CNN employs a Region Proposal Network (RPN) [14] to detect the discriminative bone parts without human prior.

Specifically, the RPN takes the radiography as input and produces a list of rectangle regions $\{R'_1, R'_2, ...R'_A\}$, each with an objectness score $S_{R'_i}$ of the region. Here we resize the input radiography X with the size of 448, and choose anchors with scales of{48, 96, 192} and ratios of {1:1, 3:2, 2:3}.

To reduce redundancy, we adopt non-maximum suppression (NMS) on the proposal regions based on their objectness scores. After NMS, the attention agent chooses top-M discriminative part regions $\{R_1, R_2, ...R_M\}$ according to S_{R_i}, then the regions are cropped from the input radiography and resized to predefined size. The attention agent feeds the resized regions into the recognition agent for feature learning and age assessment.

Recognition Agent: After being resized to the predefined size, the top-M regions are fed into feature extractor to generate feature vectors, each with length L. Then the feature vectors are fed into a fully-connected layer, which has L neurons to generate the age assessment $\{A_{R_i}\}$, here A_{R_i} denotes the age assessment according to region R_i. The input radiography X is also fed into the recognition agent and we generate its prediction as A_X. To further leverage the benefit of part feature ensemble, we concatenate the feature vectors of the input radiography and the top-K ($K \leq M$) regions. The concatenated feature vector, denoted as C, is fed into a fully-connected layer, which has $L(K+1)$ neurons, to generate the assessment A_C. Then we average A_C, A_X, $\{A_{R_1}, A_{R_2}, ..., A_{R_K}\}$ and get the assembling age assessment A_{asb} as Eq. 1.

$$A_{asb} = \frac{1}{K+2}\{A_C + A_X + \sum_{i=1}^{K} A_{R_i}\} \tag{1}$$

Feedback: In AR-CNN, a feedback flow is established from recognition agent to attention agent. The absolute deviation of the extracted regions $\{D_{R_1}, D_{R_2}, ...D_{R_M}\}$ between assessment and ground-truth A_{gt} will be fed back to the attention agent for optimization.

An accurate attention agent means that, if a region R_i is predicted with high objectness score S_{R_i} in the attention agent, it will get low absolute deviation D_{R_i} to ground-truth age A_{gt} in the recognition agent, as in Eq. 2.

$$D_{R_i} < D_{R_j} \quad if \quad S_{R_i} > S_{R_j} \tag{2}$$

Accordingly, an optimization strategy for attention agent can be set up, and the discriminative bone parts can be extracted without human prior.

Loss Function and Optimization: The strategy of optimizing attention agent is to make $\{S_{R_1}, S_{R_2}, ..., S_{R_M}\}$ and $\{-D_{R_1}, -D_{R_2}, ..., -D_{R_M}\}$ have the same order. Hence the attention agent loss function L_{att} is defined as a pairwise ranking loss in Eq. 3. The function ϕ is hinge loss function $\phi(x) = max\{1-x, 0\}$ in our experiment.

$$L_{att} = \sum_{s=i}^{M-1} \sum_{j=s+1, D_{R_i} > D_{R_j}}^{M} \phi\left(S_{R_i} - S_{R_j}\right) \tag{3}$$

Since the recognition agent can be viewed as an assemblage of multiple regressor. Its loss function L_{cls} can be defined as the sum of the total regression loss in Eq. 4. $\mathcal{R}$ denotes the regression loss, and we employ L1 loss as regression loss in our approach.

$$L_{cls} = \mathcal{R}(C) + \mathcal{R}(X) + \sum_{i=1}^{M} \mathcal{R}(R_i) \tag{4}$$

The total loss of our MAR-CNN is defined as below:

$$L_{total} = L_{cls} + L_{att} = \sum_{s=i}^{M-1} \sum_{j=s+1, D_{R_i} > D_{R_j}}^{M} \phi\left(S_{R_i} - S_{R_j}\right) + D_c + D_X + \sum_{i=1}^{M} D_{R_i} \tag{5}$$

3 Experiments and Results

Data: We run experiments on the RSNA Pediatric Bone Age Challenge dataset[1], which consists of 12611 images for training, 1425 images for validation and 200 images for testing. And the ground-truth age ranges from 0 to 18 years. The Mean Absolute Error(MAE) between predicted age and ground-truth age on test set is the final evaluation standard for models performance.

Experiment Setup: We use Pytorch to implement AR-CNN and the algorithm are trained separately for male and female. The input hand radiography and extracted bone parts are resized to 448×448 and 224×224 respectively. ResNet50 is chosen as the backbone. We fix $M = 6$ in attention agent and the comparative experiments are carried out where K ranges from 1 to 4. Our AR-CNN is trained on a Ubuntu workstation with eight NVIDIA GeForce 1080Ti GPU, and the code and pre-trained model will be released for public research[2].

Result: We first make ablation study on hyper-parameter K, which means the recognition agent takes K bone parts for age assessment. Table 1 compares the MAE of male with different K. As we can see, with the increasing of K, the MAE reduces at the beginning. We obtain the best accuracy of 4.32 months when $K = 3$, then the MAE gets increased. This phenomenon indicates that, we do not need undue bone parts for assessment, which will bring misleading by introducing the less discriminative bone parts. Hence, extracting three specific bone parts is enough in AR-CNN.

With $K = 3$, the recognition agent takes the input radiography X, three bone parts R_1, R_2, R_3 and their contacted feature C for assembling age assessment. As we can see in Table 2, the MAE according to X are 5.76 and 6.20 months for males and females, respectively. The contacted feature C can reduce the MAEs

[1] http://rsnachallenges.cloudapp.net/competitions/4.
[2] https://github.com/liuboss1992/AR-CNN.

Table 1. Ablation study on hyper-parameter K.

Hyper-parameter K	1	2	3	4
MAE (months)	5.87	5.45	**4.32**	7.43

Table 2. Ablation study on assembling age assessment.

Unit	X	C	R_1	R_2	R_3	Assembling
MAE (male)	5.76	4.99	5.55	5.59	6.03	**4.32**
MAE (female)	6.20	5.46	6.58	6.31	7.04	**4.44**

to 4.99 and 5.46 months. Furthermore, when we take the bone parts R_i into assembling, the MAEs can be reduced to 4.32 and 4.44 months. This indicates the important reference significance of specific bone parts in BAA.

Table 3 compares AR-CNN with other state-of-art methods. Typically, the multi-stage method [9,11] can achieve better accuracy than the single-stage ones [7,8], since it employs human prior to focus on specific bone parts. Iglovikov et al. [11] segment the hands out from radiography by U-Net and then crop out the carpal bones, metacarpals and proximal phalanges for assessment. Thus, they achieve impressive accuracy. Human prior brings improvement in accuracy, however, it can limit the generalization. AR-CNN can extract the discriminative bone parts without human prior knowledge, and is free of segmentation or detection. Experimental results show that AR-CNN achieves the best accuracy with MAE of 4.38 months over all the methods. This indicates that, the human prior is not a one-size-fits-all reference for deep learning in BAA task.

Table 3. Comparison results with the state-of-art methods. *RSNA.* represents the experiment which is performed on RSNA dataset.

Method	Human [8]	Spampinato [7]	Larson [8]	Iglovikov [11]	Ren [9]	Our AR-CNN
Stage		Single-stage	Single-stage	Multi-stage	Multi-stage	Single-stage
RSNA.	✓		✓	✓	✓	✓
MAE	7.32	9.12	6.24	4.97	5.20	**4.38**

Analysis and Visualization: Figure 2 shows the assessment result and deviation on test set. From Fig. 2(a) we can see, our assessment result keeps strong consistency with the actual age. Meanwhile, from Fig. 2(b) we can see, the absolute deviation of our model is controlled within 20 months.

Moreover, visualization experiments are carried out to observe the extraction result of discriminative bone parts. We highlight the extracted bone parts in attention agent with colored bounding box on the input images. As shown in Fig. 3, our attention agent mostly extracts the carpal bones and proximal phalanges as the discriminative parts, which keeps consistent with human prior

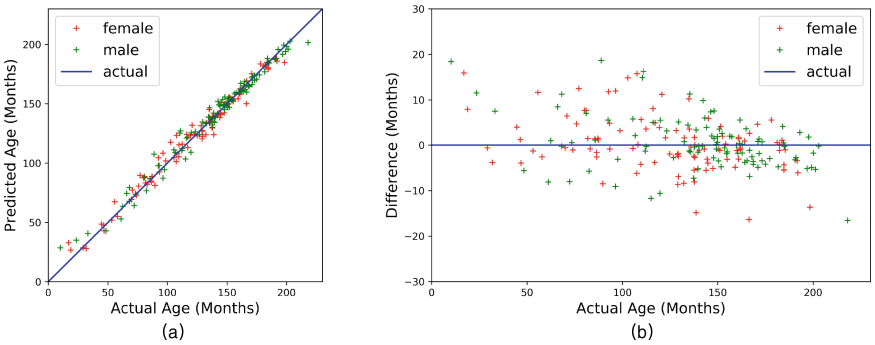

Fig. 2. Statistical results of our AR-CNN in bone age assessment. (a) shows the relationship between actual age and predicted age. (b) shows the relationship between actual age and deviation. [Best viewed in color] (Color figure online)

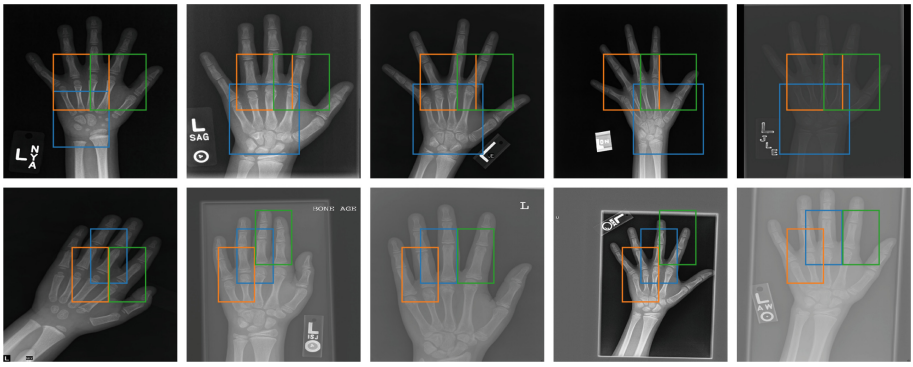

Fig. 3. Visualization experiments for bone parts extraction. The first row shows the results of female, and the second row shows the results of male. [Best viewed in color] (Color figure online)

knowledge. This indicates that AR-CNN can learn and establish correct knowledge without human prior. Meanwhile, AR-CNN embodies discrepant attention within gender, which differs from human prior. The carpal and the proximal phalanges are extracted for female, by contrast, the proximal phalanges of the index finger, middle finger and ring finger are usually separately extracted for male. This suggests that some of the human prior currently employed by clinicians might not be consummate. AR-CNN can slip the leash of human knowledge, and it could be a suggestion for a new clinical standard. In addition, AR-CNN presents stable performance with different rotation, scaling ratio and contrast ratio. This illustrates the strong reliability and expansibility of our approach.

4 Conclusion

This work presents AR-CNN, a novel single-stage approach for bone age assessment. It can discover and extract the discriminative bone parts without human prior knowledge, and can be trained end-to-end without segmentation and detection. The human prior is widely employed in BAA for bone parts extracting, however, the state-of-art performance of AR-CNN indicates the limitation of human prior. Moreover, this work can be an enlightening reference for other deep learning researches with human prior.

Acknowledgements. This work is supported by the Huawei-USTC Joint Innovation Project on Machine Vision Technology (FA2018111122). And we would like to thank Brain-inspired Technology Corporation (http://www.leinao.ai/) for its calculation support.

References

1. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.K.: Bone age assessment of children using a digital hand atlas. Comput. Med. Imaging Graph. **31**(4–5), 322–331 (2007)
2. Stern, D., Ebner, T., Bischof, H., Grassegger, S., Ehammer, T., Urschler, M.: Fully automatic bone age estimation from left hand MR images. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 220–227. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10470-6_28
3. Liu, A.A., Xu, N., Nie, W.Z., Su, Y.T., Wong, Y., Kankanhalli, M.: Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Trans. Cybern. **47**(7), 1781–1794 (2016)
4. Xie, H., Yang, D., Sun, N., Chen, Z., Zhang, Y.: Automated pulmonary nodule detection in ct images using deep convolutional neural networks. Pattern Recogn. **85**, 109–119 (2019)
5. Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y.: A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. In: AAAI 2019 (2018)
6. Mutasa, S., Chang, P.D., Ruzal-Shapiro, C., Ayyala, R.: MABAL: a novel deep-learning architecture for machine-assisted bone age labeling. J. Digit. Imaging **31**(4), 513–519 (2018)
7. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R.: Deep learning for automated skeletal bone age assessment in X-ray images. Med. Image Anal. **36**, 41–51 (2017). https://doi.org/10.1016/j.media.2016.10.010
8. Larson, D.B., Chen, M.C., Lungren, M.P., Halabi, S.S., Stence, N.V., Langlotz, C.P.: Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology **287**(1), 313–322 (2018)
9. Ren, X., Li, T., Wang, Q.: Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE J. Biomed. Health Inform. **pp**(c), 1 (2018)
10. Wang, S., Shen, Y., Zeng, D., Hu, Y.: Bone age assessment using convolutional neural networks. In: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 175–178. IEEE (2018)

11. Iglovikov, V.I., Rakhlin, A., Kalinin, A.A., Shvets, A.A.: Paediatric bone age assessment using deep convolutional neural networks. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 300–308. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_34
12. Hao, P., Chen, Y., Chokuwa, S., Wu, F., Bai, C.: Skeletal bone age assessment based on deep convolutional neural networks. In: Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W. (eds.) PCM 2018. LNCS, vol. 11165, pp. 408–417. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00767-6_38
13. Singh, K.K., Lee, Y.J.: Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV, pp. 3544–3553. IEEE (2017)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)