Hand Pose Estimation for Pediatric Bone Age Assessment

María Escobar[1]⋆(🖂), Cristina González[1]⋆, Felipe Torres[1], Laura Daza[1], Gustavo Triana[2], and Pablo Arbeláez[1]

[1] Universidad de los Andes, Bogotá, Colombia
{mc.escobar11, ci.gonzalez10, f.torres11, la.daza10, pa.arbelaez}@uniandes.edu.co
[2] Fundación Santa Fe de Bogotá, Colombia

Abstract. We present a new experimental framework for the task of Bone Age Assessment (BAA) based on a local analysis of anatomical Regions Of Interest (ROIs) of hand radiographs. For this purpose, we introduce the Radiological Hand Pose Estimation (RHPE) Dataset, composed of 6,288 hand radiographs from a population that is different from the currently available BAA datasets. We provide Bone Age groundtruths annotated by two expert radiologists as well as bounding boxes and keypoints denoting anatomical ROIs annotated by multiple trained subjects. In addition to RHPE, we provide bounding boxes and ROIs annotations for the publicly available BAA dataset by the Radiological Society of North America (RSNA) [9]. We propose a new experimental framework with hand detection and hand pose estimation as new tasks to extract local information for BAA methods. Thanks to its fine-grained and precisely localized annotations, our dataset will allow to exploit local information to push forward automated BAA algorithms. Additionally, we conduct experiments with state-of-the-art methods in each of the new tasks. Our proposed model, named BoNet, leverages local information and significantly outperforms state-of-the-art methods in BAA. We provide the RHPE dataset with the corresponding annotations, as well as the trained models, the source code for BoNet and the additional annotations created for the RSNA dataset.

Keywords: Bone Age Assessment · Computer Aided Diagnosis · Hand radiograph · Regions Of Interest.

1 Introduction

Bone age is a quantification of the skeletal development of children. This measurement differs from chronological age as it goes from 0 to 20 years, while varying according to gender, age and ethnicity. Bone Age Assessment (BAA) is performed by radiologists and pediatricians to diagnose growth disorders or to determine the final adult height of a child [17]. This evaluation is commonly

⋆ Both authors contributed equally to this work.

accomplished through visual inspection of ossification patterns in a radiograph of the non-dominant hand and wrist. The most commonly used manual methods are Greulich and Pyle (G&P) [8] and Tanner and Whitehouse (TW2) [17]. In G&P the entire hand is classified into a stage, while in TW2 it is divided into 20 Regions Of Interest (ROIs) that are analyzed individually to estimate the patient's bone age. As a result, TW2 is more precise because it makes a *local analysis* of the hand. However, both manual approaches are prone to intra and inter observer errors due to the level of expertise of the radiologist or possible variations in the radiograph.

To improve the accuracy of BAA, there has been a growing interest in the development of automated methods. The commercial software BoneXpert [18], created in 2009 by Hans Henrik Thodberg and Sven Kreiborg, is currently used in clinical settings. This tool performs BAA based on edge detection and active appearance models [2] to generate candidates and make comparisons according to G&P and TW2. However, the method was developed using patients from a Danish cohort, therefore the reliability is not guaranteed when assessing data from other countries.

Recently, the Radiological Society of North America (RSNA) created a BAA dataset for the 2017 Boneage Pediatric Challenge [9]. The RSNA Challenge encouraged the development of several deep learning and machine learning approaches to accurately perform BAA [9]. The winners of the challenge achieved a Mean Absolute Difference (MAD) of 4.26 months over the test set[1]. Similarly, most top-performing methods used shallower neural networks to extract features from the entire image. Other methods uniformly extracted overlapping patches or first segmented the bones to localize the analysis, obtaining 4.35 and 4.50 MAD, respectively [9]. Similarly, [12] developed an approach to focus on the carpal bones, the metacarpal and proximal phalanges, achieving 4.97 MAD.

Although some of the approaches above build explicitly on local information when assessing bone age, existing BAA datasets [3,6,7,9] provide only bone age annotations at the image level, and hence they are not designed to exploit the information of anatomical ROIs. A suitable approach for identifying ROIs in the hand is through hand pose estimation. This task has been studied in the context of 3D hand models directed towards human computer interaction, virtual reality and augmented reality applications [4,5]. In this work, we propose a 2D framework focused on radiological hand pose estimation as a new task, enabling various medical applications in this field.

We present the Radiological Hand Pose Estimation (RHPE) dataset containing 6,288 images of hand radiographs from a population with different characteristics than the currently available datasets, ensuring a high variability for a better model generalization. In addition to the new dataset, we introduce hand detection and hand pose estimation as new tasks to extract local information from images. To establish a robust framework, we collect manually annotated keypoints for anatomical ROIs and bounding boxes of each hand radiograph. An example of our annotations is presented in Figure 1. We also provide bounding boxes and keypoint annotations for the RSNA dataset. We evaluate the perfor-

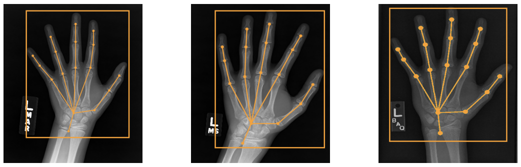

Fig. 1. Different examples of our keypoint and bounding box annotations. We provide groundtruth for hand detection and hand pose estimation for both RHPE and RSNA datasets.

mance of state-of-the-art methods on our proposed tasks on both RSNA and RHPE datasets, and we propose a new local approach to BAA called BoNet that significantly outperforms all existing approaches. Additionally, we prove that both datasets are complementary and can be combined to create a robust benchmark with a better model generalization, regardless of the population's characteristics.

Our main contributions can be summarized as follows:

1. We introduce RHPE, a new dataset from a diverse population, and create a new benchmark for the development of BAA methods.
2. We provide the first manually annotated bounding boxes and keypoints in the RSNA and RHPE datasets. These annotations enable a new experimental framework including hand detection and hand pose estimation as tasks for the extraction of local information from hand radiographs.
3. We present BoNet, a novel CNN architecture that leverages anatomical ROIs to perform BAA. BoNet significantly outperforms all state-of-the-art methods on the RSNA and RHPE datasets.

To ensure reproducibility of our results and to promote further research on BAA, we provide the RHPE dataset and the corresponding annotations for train and validation, as well as the trained models, the source code for BoNet and the additional annotations created for the RSNA dataset [9]. We also deploy a server for automated evaluation of the test set.[1]

2 Radiological Hand Pose Estimation Dataset

2.1 Dataset Description

We collect the RHPE data from a population that is different from the ones in the currently available datasets for BAA. The database comprises images of radiographs taken from left and right hands of both male and female patients between 0 and 240 months of age (0-20 years), with bone age annotations made by two expert radiologists for each patient. The dataset is composed of 6,288 images divided into 3 sets: 5,492 for training, 716 for validation and 80 for

[1] https://biomedicalcomputervision.uniandes.edu.co/index.php

testing, maintaining the proportion of images used in each split of the RSNA dataset. 54% of the dataset corresponds to female patients and 46% to male patients. This division has the same proportions as the RSNA dataset and the Gaussian distribution of bone age on our dataset and on the RSNA dataset is approximately the same, centered around 126 months of age. A similar bone age distribution between the datasets suggests that they are compatible and can be used to study the influence of ethnicity on bone age assessment algorithms. See supplementary material for a further analysis of the similarities and differences of both datasets.

We gather anatomical landmark annotations from 96 trained subjects. For each image, the subject is shown an example of where the keypoints should be located. We obtain multiple annotations per image and perform outlier rejection by identifying the annotations that are 2 standard deviations away from the mean. From this procedure, we obtain at least 4 annotations per image made by different trained subjects. With 17 keypoints per hand radiograph, this accounts for more than 1.3 million annotated keypoints. These annotations correspond to the proximal, middle and distal phalanges, the carpal bones, and the distal radius and ulna. For compatibility, we store all our annotations using the MS-COCO [13] format. For the detection groundtruth, we include the upper-left coordinates, width and height of the bounding box that encloses the hand.

2.2 Tasks

Hand Detection The goal of hand detection is to determine the location of a specific hand in the image. The importance of including detection as a task in our dataset lies in the fact that the images in RHPE include both hands and it is necessary to isolate the non-dominant hand for the assessment. For evaluation, we use the same standard metrics as in MS-COCO for object detection: mean Average Precision (mAP) and mean Average Recall (mAR) at Intersection over Union (IoU) thresholds @[.50 : .05 : .95].

Hand Pose Estimation In this task, the objective is to estimate the position of anatomical ROIs in the hand radiograph. For the evaluation of hand pose estimation algorithms, we use the mAP and mAR at Object Keypoint Similarity (OKS) [13] @[.50 : .05 : .95]. It is worth noting that the evaluation code used in MS-COCO only takes into account instances that were accurately detected. To obtain a more realistic assessment of performance, we modify this metric to consider the effect of every image regardless of the detection mAR@[.5:.05:.95]. Additionally, the standard deviation σ_i with respect to object scale s varies significantly for different keypoints. In full-sized images, our σ penalizes any keypoint estimation 10 pixels away or more from the mean location. See supplementary material for additional information.

Bone Age Assessment In the BAA task, we aim at estimating bone age in months for a given hand radiograph. To evaluate the performance, we use the same metric as in the RSNA 2017 Pediatric Bone Age Challenge : the Mean Absolute Distance

(MAD) between the groundtruth values and the model's predictions. We evaluate our performance on the RSNA dataset using the challenge evaluation server.

2.3 Baselines

Hand Detection As baseline for the hand detection task, we use Faster R-CNN [15] with an ImageNet pre-trained model and ResNet-50 [11] as backbone. This widely used object detector consists of a network that generates and scores object proposals. We train the Faster R-CNN [15] method for 5,000 iterations with a constant learning rate of 2.5×10^{-3} using the implementation released in [14].

Hand Pose Estimation To adress hand pose estimation, we build on the recent state-of-the-art architecture proposed by Xiao *et al.* in [19] for human pose estimation. This model consists of an encoder-decoder based on deconvolutional layers added on a backbone network. We train the model initialized on ImageNet with ResNet-50 as backbone [11] using the suggested training parameters for 20 epochs.

Bone Age Assessment As baseline for the BAA task, we re-implement the method proposed by the winners of the RSNA 2017 Pediatric Bone Age Challenge. However, our inference differs from that method because [1] used an ensemble of the best models at inference while we use only a single model. This model uses an Inception-V3 [16] architecture combined with a network to include gender information, and adds two 1000-neuron densely connected layers to produce the final prediction. For the baseline, we train the model for 150 epochs, using Adam optimizer with an initial learning rate of 3×10^{-3}.

3 BoNet

Inspired by the way physicians perform BAA, we introduce a method that leverages local information to accurately address this task. For this purpose, we first locate the hand and find the anatomical ROIs. Afterwards, we create attention maps by generating a Gaussian distribution around each anatomical landmark. Our network, which we call BoNet, takes as input both the X-ray image and the

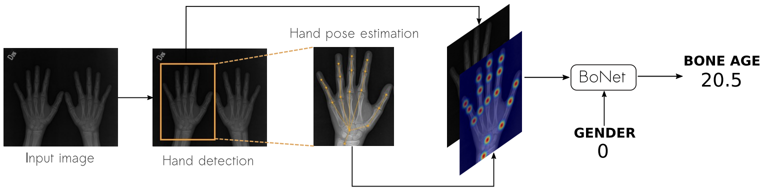

Fig. 2. Overview of the pipeline used in BoNet. The original image goes through hand detection and hand pose estimation to identify ROIs. Then we use as input for BoNet the cropped image, the heatmap and the gender for BAA.

attention map, and extracts high-level visual features from them using two independent pathways. We then combine these features from both pathways through a mixed Inception module and follow the suggestion in [1] to include gender information. Finally, after two fully-connected layers, BoNet regresses to the bone age using an L1 loss. See supplementary material for additional information. Figure 2 illustrates an overview of the complete method. To train BoNet, we start from our BAA baseline's weights. We use Adam optimizer with an initial learning rate of 1×10^{-4} over 50 epochs and reduce the learning rate by a factor of 0.1 once a loss plateau is reached.

4 Experiments

Hand Detection For this task, we evaluate the performance of our baseline method on the RSNA and RHPE datasets. Since the RHPE dataset contains both left and right hands on the image, we evaluate the detection of the left one, statistically assuming that it is the non-dominant hand [10]. Table 1 shows the results obtained in the validation split for both datasets. The performance of Faster R-CNN given by the mAP and the mAR at different IoU thresholds (@[.5:.05:.95]) is considerably high. Specifically, the mAP@[.75] indicates an excellent localization of the detections. This behaviour is appropriate considering that detecting the bounding box of the hand is the first step towards using local information for BAA. Consequently, precision and recall in hand detection will significantly affect the performance of BAA. The errors in finding bounding boxes can be associated to the detection of false positives contained inside the annotation bounding box and to the low IoU of most true positives. Since the RHPE dataset contains both left and right hands in the image, the hand detection task is more complex than for the RSNA dataset.

Hand Pose Estimation With the aim of determining the relevance of bounding box detection for hand pose estimation, we use the datasets separately following three modalities: full image of the radiograph, image cropped with groundtruth bounding boxes and image cropped with detected bounding boxes. In the RHPE dataset we consider *full image* as the left half of the radiograph to only include the non-dominant hand. The results obtained for the validation set are reported in Table 2. The results prove that, for both datasets, the performance of the hand pose estimation task is considerably affected by the input used. Thus, we establish the upper bound for this task by using the groundtruth bounding boxes. Consequently, the performance of our predicted bounding boxes

Table 1. Results of the hand detection task in the validation split for the RSNA and RHPE datasets using our baseline.

	mAP@[.5,.95] (%)	**mAP@[.5] (%)**	**mAP@[.75] (%)**	**mAR@[.5,.95] (%)**
RSNA	93.7	99.0	98.9	96.1
RHPE	90.1	96.9	96.9	93.1

Table 2. Comparison of results in the validation set of RSNA [9] and RHPE datasets using our baseline for the hand pose estimation task.

		mAP@[.5,.95] (%)	**mAP@[.5] (%)**	**mAP@[.75] (%)**	**mAR@[.5,.95] (%)**
RSNA	Full image	73.0	96.3	90.0	77.6
	Groundthruth bounding boxes	**81.4**	**97.8**	**96.8**	**84.1**
	Detected bounding boxes	80.8	94.1	93.2	83.0
RHPE	Full image	53.1	91.2	60.1	59.3
	Groundthruth bounding boxes	**81.4**	**97.8**	**96.8**	**84.1**
	Detected bounding boxes	80.8	94.1	93.2	83.0

is lower than using groundtruth information since it depends on the results from the hand detection task. In contrast, the full image includes noise associated with background, tags and other artifacts in the radiograph, hence it obtains the lowest precision. The low performance in the full image setup show that it is necessary to use as input for this task a bounding box of the hand radiograph.

Bone Age Assessment We design three sets of experiments to study the effect of training on different data. The first set uses only RSNA, the second one uses only RHPE, and the third one combines both datasets. For each set we assess the importance of local information by training on whole and cropped images. We use the training and the validation splits during the training stage and evaluate our results on the test split. The results shown on Table 3 demonstrate that hand detection is beneficial for accurate bone age assessment. Additionally, we observe that BoNet leverages effectively local information, achieving a significant improvement in performance over the re-implementation of state-of-the-art which is our baseline with full image. We also find that combining both datasets during training produces better results than training on a single dataset. These results indicate that increasing and diversifying the data is beneficial for model generalization. Regarding the time complexity of the algorithm, the final model using BoNet and cropped images takes 0.079 seconds per image on inference, making it a suitable choice for a future real time implementation.

5 Conclusions

We introduce the Radiological Hand Pose Estimation Dataset as a benchmark for the development of robust methods for BAA, hand detection and hand pose estimation in radiological images as a way of exploiting local information as done by physicians in current clinical practice. For each task, we propose an experimental framework and validate state-of-the-art methods as baselines. Our results prove that the use of local information is beneficial for BAA. We also develop BoNet, a new method based on exploiting local information that outperforms the state-of-the-art method that exploit only global information. The RHPE Dataset and its associated resources will push the envelope further in the development of robust BAA automated methods with better generalization regardless of the populations characteristics.

Table 3. BAA results on the RSNA and RHPE test sets.

Experiment		**MAD**	
Training on RSNA	Baseline (full image)	4.45	
	BoNet (full image)	4.37	
	Baseline + cropped image	4.20	
	BoNet + cropped image	**4.14**	
Training on RHPE	Baseline (full image)	8.57	
	BoNet (full image)	7.78	
	Baseline + cropped image	8.05	
	BoNet + cropped image	**7.60**	
Training on RSNA + RHPE		**RSNA**	**RHPE**
	Baseline (full image)	4.41	8.25
	Baseline + cropped image	4.09	7.99
	BoNet + cropped image	**3.85**	**6.86**

Acknowledgments This project was partially funded by Colciencias grant 841-2017. The authors thank Edgar Margffoy-Tuay for his support in developing the annotation server and the students of IBIO-3470 at Uniandes for their help as annotators.

References

1. Cicero, M., Bilbily, A.: Machine Learning and the Future of Radiology: How we won the 2017 RSNA ML Challenge (2017)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). pp. 484–498. Springer (1998)
3. Gaskin, C.M., Kahn, M.M.S.L., Bertozzi, J.C., Bunch, P.M.: Skeletal development of the hand and wrist: a radiographic atlas and digital bone age companion. Oxford University Press (2011)
4. Ge, L., Cai, Y., Weng, J., Yuan, J.: Hand PointNet: 3d hand pose estimation using point sets. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8417–8426 (2018)
5. Ge, L., Ren, Z., Yuan, J.: Point-to-point regression pointnet for 3d hand pose estimation. In: European Conference on Computer Vision (ECCV). vol. 1. Springer (2018)
6. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. Computerized Medical Imaging and Graphics **31**(4-5), 322–331 (2007)
7. Gilsanz, V., Ratib, O.: Hand bone age: a digital atlas of skeletal maturity. Springer Science & Business Media (2005)
8. Greulich, W.W., Pyle, S.I., Todd, T.W.: Radiographic atlas of skeletal development of the hand and wrist, vol. 2. Stanford University Press (1959)
9. Halabi, S.S., Prevedello, L.M., et al: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019).
10. Hardyck, C., Petrinovich, L.F.: Left-handedness. Psychological bulletin **84**(3), 385 (1977)

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Iglovikov, V.I., Rakhlin, A., Kalinin, A.A., Shvets, A.A.: Pediatric bone age assessment using deep convolutional neural networks. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 300–308. Springer (2018)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
14. Massa, F., Girshick, R.: maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. (2018)
15. 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 (NIPS). pp. 91–99 (2015)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308,
17. Tanner, J., Whitehouse, R., Marshall, W., Carter, B.: Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height. Archives of disease in childhood **50**(1), 14–26 (1975)
18. Thodberg, H., Kreiborg, S., Juul, A., Pedersen, K.: The BoneXpert Method for Automated Determination of Skeletal Maturity. IEEE Transactions on Medical Imaging **28**(1), 52–66 (2009). https://doi.org/10.1109/tmi.2008.926067
19. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV) (2018)