FingerNet: Deep Learning-Based Robust Finger Joint Detection from Radiographs

Sungmin Lee†, Minsuk Choi†, Hyun-soo Choi†, Moon Seok Park‡ and Sungroh Yoon†
†EECS, Seoul National University, Seoul, 151-744, Korea
‡Orthopedic Surgery, Seoul National University Bundang Hospital, Seongnam, 463-707, Korea
pmsmed@gmail.com, sryoon@snu.ac.kr

***Abstract*—Radiographic image assessment is the most common method used to measure physical maturity and diagnose growth disorders, hereditary diseases and rheumatoid arthritis, with hand radiography being one of the most frequently used techniques due to its simplicity and minimal exposure to radiation. Finger joints are considered as especially important factors in hand skeleton examination. Although several automation methods for finger joint detection have been proposed, low accuracy and reliability are hindering full-scale adoption into clinical fields. In this paper, we propose FingerNet, a novel approach for the detection of all finger joints from hand radiograph images based on convolutional neural networks, which requires little user intervention. The system achieved 98.02% average detection accuracy for 130 test data sets containing over 1,950 joints. Further analysis was performed to verify the system robustness against factors such as epiphysis and metaphysis in different age groups.**

I. Introduction

Finger joints play an important role in hand radiograph assessment. Due to their simplicity and limited exposure to radiation, hand radiographs are typically used to assess the finger bones and joints to carry out bone age evaluation for children and adolescents, and to diagnose growth disorders, inherited disease [1] and joint inflammation, such as rheumatoid arthritis. Because of the large number of joints in the hand and the frequent use of this type of examination, it is beneficial to automate joint detection in hand radiograph images. Fig. 1 illustrates an example of finger joint detection.

Since Pal and King proposed the first automated system for detecting joints in hand radiographs [2], several advanced methods, such as an automation system by Giordano [3] and computer-assisted bone age assessment technique by Pietka [4], have been developed. However, medical studies [5] indicate that these methods require improvements in accuracy, reliability and consistency to cultivate full-scale acceptance into clinical fields.

Deep learning, a recent breakthrough in machine learning, is gaining popularity for its excellent performance in various recognition problems [6], [7], [8], [9]. For example, deep-learning architectures have achieved the best performance in various competitions in natural language processing and image recognition. Specifically, convolutional neural networks (CNN) [10], a variant of deep-learning architecture, is widely used in computer vision, such as detection or recognition, where systems have produced state-of-the-art results when performing a variety of tasks. The excellent performance of CNN in other applications motivated us to adapt CNN architecture for finger joint detection. In this paper, we propose a user-interactive CNN-based joint detecting system called FingerNet.

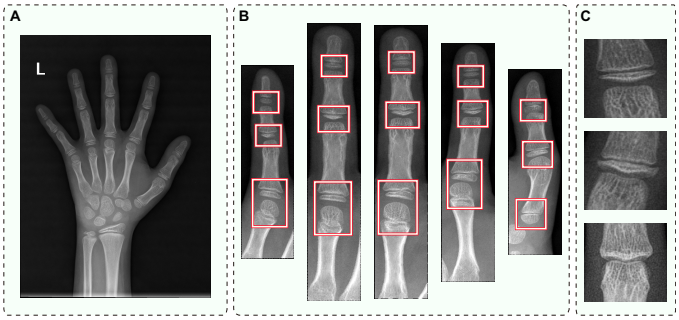

Fig. 1. Examples of finger joint detection. (a) Input hand radiograph. (b) Extracted fingers and detected finger joint regions. (c) Joint bone maturity.

The proposed method performs binary classification via CNN to each windowed image sub-block to determine whether it is a joint. The performance of the proposed method is compared with that of AdaBoost [11], which is another state-of-the-art approach to face detection. FingerNet outperformed AdaBoost, achieving 98% accuracy compared to AdaBoost's 92% in 130 test data sets.

II. Convolutional Neural Networks

CNN, a variant of artificial neural network architecture with multiple levels of learning layers, comprises convolutional layers, sub-sampling layers, and fully connected layers, taking advantage of 2D structure, such as image or video data [10]. Several processes underly the multi-layer information flow: The convolution layer performs convolution on the input image with an arbitrary filter before a sub-sampling layer sub-samples the results. Additional filtering is performed in the next layer, and so forth. These processes converge into a feature map that best reflects the features of input images [12].

CNN exploits a local connectivity pattern between units of adjacent layers, to limit errors in the spatial units that do not affect the others, with back-propagation. By back-propagating errors through each layer, weights can be optimized to form filters that achieve a minimum error rate. In addition, CNN adopts weight sharing, which assigns the same parametrization to each hidden unit in the same feature map. Applying weight sharing allows CNN to capture the relative rather than absolute distance between features. This property grants CNN robustness that prevents distortions or shifts [10].

III. Methods

Fig. 2 is a flow chart demonstrating the proposed algorithm. FingerNet consists of three stages: preprocessing (PP), finger extraction (FE) and joint detection (JD).

978-1-4799-7234-0/15/$31.00 ©2015 IEEE

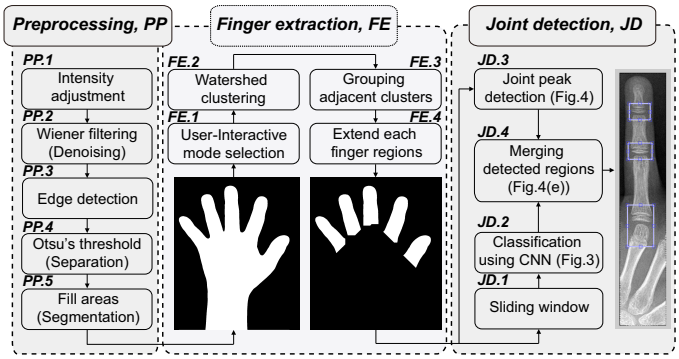

Fig. 2. Overview of the proposed methodology.

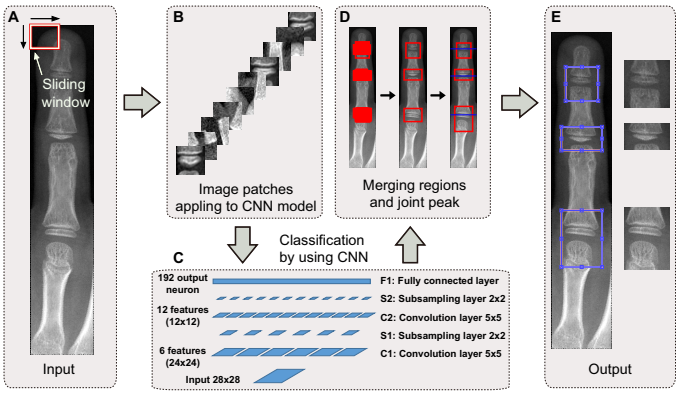

Fig. 3. Details of JD phase. (a) Applying sliding window. (b) Image patches extracted by (a). (c) CNN architecture. (d) Merging detected regions and joint peaks. (e) User interaction step. User can confirm and adjust the results.

A. Preprocessing phase (PP)

In the first phase (PP), hand mask segmentation is performed in the radiographic image. The extracted mask acts as a base for the latter FE phase.

In the PP.1 and PP.2 steps, hand radiograph image intensity adjustment and noise reduction are performed respectively. Prior to any further processing, all input images are resized to 1,200×1,400 pixels, which is a reasonable size for a radiograph. Results from statistical analysis are used for intensity adjustment, and a Wiener filter [13] is also applied for denoising (PP.2). The Wiener filter suppresses background noise without the loss of information in the original image due to the use of minimum mean squared error as its base.

Step PP.3 and PP.4 are carried out to detect edges in the image and separate the hand mask from the image background. FingerNet uses a gradient-based edge detector and returns an image that represents the intensity gradient of the input image. The resulting grayscale image consists of information regarding the presence of edges. Background is separated using Otsu's method [14], which divides the signal and background by maximizing the separability of the gray level distribution.

In step PP.5, dilation, erosion and fill methods are performed for postprocessing, and labels and noise in the background are removed to return a binary hand mask.

B. Finger extraction phase (FE)

The FE phase extracts five separate fingers from the hand mask image. This process is effective in enhancing detection performance because the search space of the JD phase can be limited and the fingers are always vertically aligned.

The FE.1 step involves manual user interaction to mark the end-point location of the thumb and fifth finger, as well as the forks adjacent to those two fingers. From this information, FingerNet calculates the rotation angles of the fingers, which are used in FE.3 and FE.4 (see Fig. 2).

The FE.2 step extracts each finger from a hand mask using the watershed algorithm [15], which is a region-based segmentation method that segments an image into a number of fragments based on a watershed where the pixel values surge.

In FE.3, the adjacent fragments are merged into large clusters based on distance. As a result, a hand is divided into fingers, palm, and others made up of carpal bones and the two long bones within the wrist. The determination of the position of fingers is relatively reliable due to their location in the extremity of the hand. If the hand is rotated excessively or the thumb is overly stretched and located under the palm, manual markings by the user, acquired in FE.1, are used to compensate.

The final step, FE.4, calculates the central axis and the rotation angle of each finger region and extends the regions of fingers to a predetermined distance using the direction of the rotation angle as a seed.

C. Joint detection phase (JD)

In the JD phase, three joints are detected for each finger image by combining two different approaches: machine learning-based CNN and signal processing-based joint peak detection.

1) CNN based joint detection: Fig. 3 shows the procedure for CNN-based joint detection. To extract sub-block images from the finger images, FingerNet applies a sliding window of 32×40 pixels to the images so that it reflects the ratio of the joint, which is measured empirically (Fig. 3(a)). The process is repeated for two additional iterations where window size is scaled 1.2 times. Extracted sub-block images are resized to 28×28 pixels. FingerNet detects joints by performing binary classification to evaluate whether each resized sub-block image is a joint, using a CNN model. Additional training samples were created by elastic distortion [12], which generated approximately 1.3 million joint images (see Section IV.A for further details).

The CNN architecture used in FingerNet is optimized and customized based on LeNet-5 [10]. The classification of grayscale joint images resembles the classification task of the MNIST handwritten digits [10]. The sliding window of this method is similar in size to that of the input data of LeNet-5.

The architecture of the CNN model consists of five layers: two convolution layers, two sub-sampling layers and a fully connected layer. As shown in Fig. 3(c), the convolution operation with a 5×5 square kernel is applied to the joint images of 28×28 and then connected to convolution layer C1 with six feature maps. Those six feature maps are connected to upper layer S1 after being sub-sampled to half size using the average

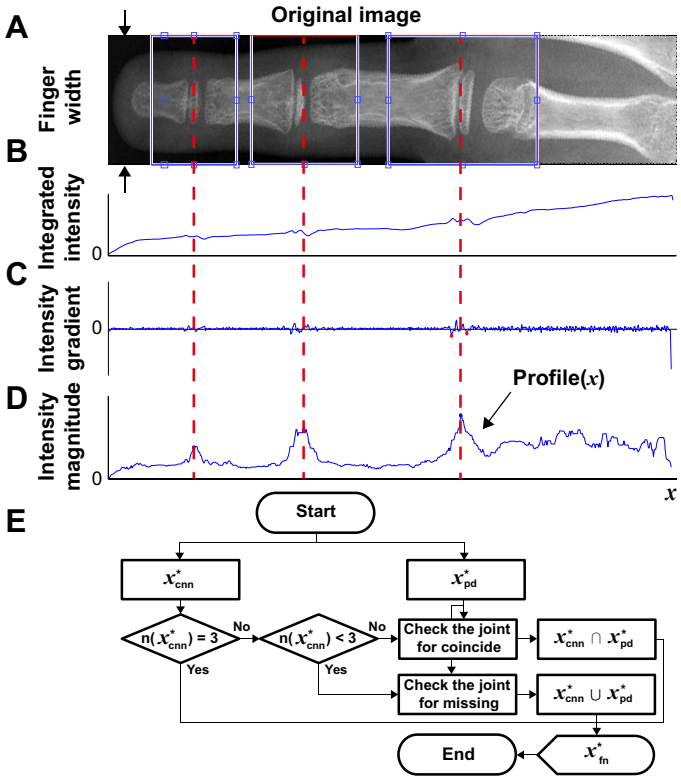

Fig. 4. Per phalange epiphyseal/metaphyseal ROI detection using peak finding method. (a) Original image. (b) Integrated intensity. (c) Intensity gradient. (d) Absolutized and median-filtered intensity gradient. (e) Flow chart of merging detected regions.

pooling method. Another 5×5 convolution filter is applied and the outputs are weighted summed to generate 12 feature maps for convolution layer C2. Through 2×2 average pooling, C2 is connected to sub-sampling layer S2. Finally, it is connected to fully connected layer F1, which is composed of 192 output neurons. The F1 layer linearly sums up output neurons to present an output that represents the probability of the input image being a joint. This procedure is the feedforward part of FingerNet. Furthermore, FingerNet transfers classification errors to lower layers using a back-propagation algorithm and iterates the above processes to update the weights of layers.

2) Joint peak detection: The joint peak detection method used in FingerNet is an image signal processing approach that scans each finger image from the end of the tip downward (in Fig. 4(a), left to right) to find peak locations in which intensity changed dramatically. Fig. 4(b) represents the intensity integrated over the finger width(vertical) direction of Fig. 4(a). Fig. 4(c) indicates the difference of Fig. 4(b) to clearly show the location where the integrated intensity changed significantly in the form of peaks. In Fig. 4(d), $\mathrm{Profile}(x)$ represents the amplified and denoised intensity curve resulting from absolutization and median filtering.

The positions marked with a red dotted line in Fig. 4(d) are the peak locations where joints exist. The location of a peak, x^*_{pd}, is given by

$$x^*_{\mathrm{pd}} = \arg\max_{x \in \mathrm{N}} \left\{ \mathrm{Profile}(x) \right\} \tag{1}$$

where x is the horizontal coordinate of Fig. 4(d) and N is the window size for peak location search. FingerNet gives local maxima for all examined x from Eq. 1.

By means of thresholding and hierachical clustering for postprocessing, it is possible to predict the joint locations x^*_{pd}. In this process, thresholding fulfills the role of denoising through exclusion of locations where the magnitude is less than a predetermined value and the clustering algorithm merges adjacent peaks. Lastly, the first three peaks from joint candidates are chosen to be x^*_{pd} because several peaks may exist at the fork where fingers are split. This method enhances the precision of an existing peak detection method [3], which simply estimates joint locations based on thresholding. The future revision of our method may employ more robust peak detection methods [16], [17].

3) Merging detected regions: Fig. 4(e) represents the flow chart of merging detected regions. Between the two algorithms, the CNN-based method leads the detection of finger joints. The joint regions detected by CNN, distributed in contiguous regions, are merged into a union, which is a rectangle covering all widths and heights of the elements (see Fig. 3(d)). Disconnected patches captured by false positive errors are removed. Then, the third joint region is expanded downward through the adaptation of domain knowledge to include the metacarpal.

The results from the joint peak detection method complement the results of the CNN model. This is especially useful, as the CNN model is prone to yield detection faults when the intensity of sub-regions in the radiographic image is too low. The joint peaks further contribute to the improvement of detection specificity by double checking the outputs of the CNN model. Ultimately, the final peak location x^*_{fn} was defined as

$$x^*_{\mathrm{fn}} = \begin{cases} x^*_{\mathrm{cnn}} & \text{if } n(x^*_{\mathrm{cnn}}) = 3 \\ x^*_{\mathrm{cnn}} \cup x^*_{\mathrm{pd}} & \text{if } n(x^*_{\mathrm{cnn}}) < 3 \\ x^*_{\mathrm{cnn}} \cap x^*_{\mathrm{pd}} & \text{if } n(x^*_{\mathrm{cnn}}) > 3 \end{cases} \tag{2}$$

where x^*_{cnn} denotes the joint regions detected by CNN.

The rule defined in Eq. 2 can be described as follows. If the element number of joint regions detected by the CNN model $n(x^*_{\mathrm{cnn}})$ is less than three, FingerNet deploys peak detection results to make up for the missing joints. The missing joint is determined by finding x^*_{pd} farthest away from elements of x^*_{cnn}. The distances between elements of x^*_{cnn} and elements of x^*_{pd} are evaluated based on the Euclidean distance. Incase the number of locations detected by CNN is more than three, FingerNet adopts the coincided regions of x^*_{cnn} and x^*_{pd}. The results (x^*_{fn}) as seen in Fig. 3(e), can also be adjusted through user interaction.

IV. Results and Discussion

In this paper, which aimed to evaluate the performance of CNN-based joint detection, the detection accuracy of CNN-based joint detection is compared with that of AdaBoost, which is well known for breakthroughs in face detection [11].

A. Experimental setup

The experiment was performed on an INTEL i5-2500 (3.3GHz quad-core), 16GB main memory and MATLAB version 2014A. Joint detection accuracy is known to be affected

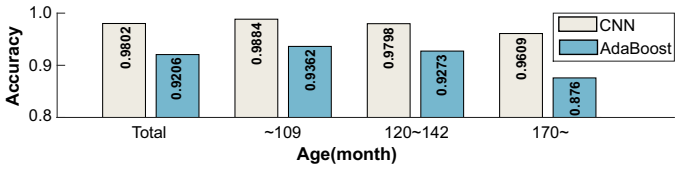

Fig. 5. Accuracy comparison: FingerNet based on CNN versus AdaBoost.

by subjects' ages due to the change in joint shapes as people age. To test the robustness of the method, the researchers thus deliberately selected three different age groups and prepared 130 left hand radiographs in DiCOM format of 1,200×1,400 pixels.

As golden standards, a professional clinician manually extracted 3,000 aligned middle phalange ROI images of second, third and fourth fingers from 1,000 hand radiographs distributed between subject ages of 80 to 180 months. Ten thousand non-joint sample images were also extracted and used as negative samples. All samples were expanded 100 times using elastic distortion [12], producing a total of 1.3 million data sets.

B. Accuracy of joint detection

Joint detection accuracy in random real-world radiographic images was evaluated under the supervision of an expert clinician. Detection results were evaluated for every joint in a single hand. Thus, all 15 joints successfully detected in an image would be calculated into 100% accuracy. Fig. 5 compares the performance of CNN-based FingerNet and AdaBoost-based FingerNet. The overall average accuracy of CNN (98.02%) is 6% point higher than that of AdaBoost (92.06%). The reliability of HOG features suffered with skewed or tilted images [18], whereas CNN is appropriate for the task because its robust features guard against such effects.

According to Fig. 5, both CNN and AdaBoost show the best average accuracy in the age group under 109 months, at 0.9884 and 0.9362 respectively. For 120 to 142 months and the over 170 months groups, the accuracy of the two models gradually decreased down to 0.9609 and 0.8759 each. However, the CNN model exhibited superior robustness against bone maturity.

V. Conclusion

In this paper, we proposed an advanced method for user-interactive finger joint detection in hand radiographs called FingerNet. Using 130 test data sets (over 1,950 joints) acquired from varying subject ages, average detection accuracy is 98.02%. FingerNet achieves this superb performance by combining two complementary strategies: CNN and peak detection. As a result, FingerNet presents robust, accurate results for joint detection from hand radiographs, and is expected to increase the efficiency of diagnoses.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT and Future Planning) [No. 2011-0009963, No. 2014M3C9A3063541, and No. 2015M3A9A7029735], in part by the SNUBH Research Fund [No. 12-2013-020], in part by the Brain Korea 21 Plus Project in 2015, and in part by Samsung Electronics Co., Ltd.

References

[1] P. Thangam, T. Mahendiran, and K. Thanushkodi, "Skeletal bone age assessment–research directions," *Journal of Engineering Science and Technology Review*, vol. 5, no. 1, pp. 90–96, 2012.

[2] S. K. Pal, R. King *et al.*, "On edge detection of x-ray images using fuzzy sets," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 1, pp. 69–77, 1983.

[3] 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," *Instrumentation and Measurement, IEEE Transactions on*, vol. 59, no. 10, pp. 2539–2553, 2010.

[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," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 8, pp. 715–729, 2001.

[5] M. Vincenzo De Sanctis, A. T. Soliman, M. Salvatore Di Maio, and S. Bedair, "Are the new automated methods for bone age estimation advantageous over the manual approaches?" *Pediatric Endocrinology Reviews (PER)*, vol. 12, no. 2, 2014.

[6] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] T. Lee and S. Yoon, "Boosted categorical restricted Boltzmann machine for computational prediction of splice junctions," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

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

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR.*, vol. 1. IEEE, 2001, pp. I–511.

[12] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *ICDAR*, vol. 2. IEEE, 2013, pp. 958–958.

[13] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

[14] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[15] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[16] J. Kim, S. Yu, B. Shim, H. Kim, H. Min, E.-Y. Chung, R. Das, and S. Yoon, "A robust peak detection method for rna structure inference by high-throughput contact mapping," *Bioinformatics*, vol. 25, no. 9, pp. 1137–1144, 2009.

[17] B. Shim, H. Min, and S. Yoon, "Nonlinear preprocessing method for detecting peaks from gas chromatograms," *BMC bioinformatics*, vol. 10, no. 1, p. 378, 2009.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR.*, vol. 1. IEEE, 2005, pp. 886–893.