Automatic detection of wrist fractures from posteroanterior and lateral radiographs

DOI:

[10.1007/978-3-030-11166-3_10](https://doi.org/10.1007/978-3-030-11166-3_10)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Ebsim, R., Naqvi, J., & Cootes, T. F. (2019). Automatic detection of wrist fractures from posteroanterior and lateral radiographs: A deep learning-based approach. In G. Zheng, T. Vrtovec, J. Yao, & J. M. Pozo (Eds.), *Computational Methods and Clinical Applications in Musculoskeletal Imaging - 6th International Workshop, MSKI 2018, Held in Conjunction with MICCAI 2018, Revised Selected Papers* (pp. 114-125). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11404 LNCS). Springer Nature. https://doi.org/10.1007/978-3-030-11166-3_10

Published in:

Computational Methods and Clinical Applications in Musculoskeletal Imaging - 6th International Workshop, MSKI 2018, Held in Conjunction with MICCAI 2018, Revised Selected Papers

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Automatic Detection Of Wrist Fractures From Posteroanterior and Lateral Radiographs: A Deep Learning-Based Approach

Raja Ebsim¹, Jawad Naqvi², Tim Cootes¹

¹ The University of Manchester.

{[raja.ebsim](mailto:raja.ebsim@manchester.ac.uk), [tim.cootes](mailto:tim.cootes@manchester.ac.uk)}@manchester.ac.uk

² Health Education North West School of Radiology, Manchester, UK.
naqvi.jawad@gmail.com

Abstract. *We present a system that uses Convolutional Neural Networks (CNNs) to detect wrist fractures (distal radius fractures) in posteroanterior (PA) and lateral (LAT) radiographs. The proposed system uses Random Forest Regression Voting Constrained Local Model (RFCLM) to automatically segment the radius. The resulting automatic annotation is used to register the object across the dataset and crop patches. A CNN is trained on the registered patches for each view separately. Our automatic system outperformed existing systems with a performance of 96% (Area under Receiver Operating Characteristic Curve AUC) on cross-validation experiments on a dataset of 1010 patients, half of them with fractures.*

Keywords: Medical Image Analysis with Deep Learning, X-ray Fracture Detection, Wrist Fracture Detection, Computer-aided diagnosis

1 Introduction

Wrist fractures are the commonest type of fractures seen in Emergency departments (EDs). They are estimated to be 18% of the fractures seen in adults [5, 9] and of 25% of fractures seen in children [9]. They are usually identified in EDs by doctors examining lateral (LAT) and posteroanterior (PA) radiographs. Yet wrist fractures are one of the most commonly-missed in ED-examined radiographs [10, 20]. Systems that can identify suspicious wrist areas and notify ED staff could reduce the number of misdiagnoses.

In this paper we describe a fully-automated system for detecting radius fractures in PA and LAT radiographs. For each view, a global search [17] is performed for finding the approximate position of the radius. The detailed outline of the bone is then located using a Random Forest Regression Voting Constrained Local Model (RFCLM) [4]. Convolutional neural networks are trained on cropped patches containing the region of interest on the task of detecting fractures. The decisions from both views are averaged for better performance. This paper is the first to show an automatic system for identifying fractures from PA and LAT view radiographs of the wrist by using convolutional neural networks, outperforming previously-published works.

2 Previous Work

Early work on fracture detection used non-visual techniques: analysing mechanical vibration [11], analysing acoustic waves traveling along the bone [24], or by measuring electrical conductivity [25]. The first published work on detecting fractures in radiographs was that in [27] where an algorithm is developed to measure the femur neckshaft angle and use it to determine whether the femur is fractured. There is a body of literature on radiographic fracture detection on a variety of anatomical regions, including arm fractures [28], femur fractures [27, 29, 15, 18, 1], and vertebral endplates [22]. Cao *et al.* [2] worked on fractures in a range of different anatomical regions using stacked random forests to fuse different feature representations (Schmid texture feature, Gabor texture feature, and forward Contextual-Intensity). They achieved a sensitivity of 81% and precision of 25%. Work on wrist fracture detection from radiographs is still limited. The earliest works [15, 18] used active shape models and active appearance models [3] to locate the approximate contour of the radius and trained Support Vector Machine SVM on extracted texture features (Gabor, Markov Random Field, and gradient intensity). They worked on a small dataset with only 23 fractured examples in their test set and achieved encouraging performance. In previous work [6, 7] we used RFCLMs [16] to segment the radius in PA and LAT views and trained random forest (RF) classifiers on statistical shape parameters and eigen-mode texture features[3]. The fully automated system achieved a performance of 91.4% (Area under Receiver Operating Characteristic Curve AUC) on a dataset of 787 radiographs (378 of which were fractured) in cross-validation experiments and was the first to combine the both views. Instead of hand-crafting features Kim *et al.*[13] re-trained the top layer (i.e. classifier) of inception v3 network [26] to detect fractures in wrist LAT views from features previously-learned from non-radiological images (ImageNet [23]). This was the first work to use deep learning in the task of detecting wrist fractures. The system was tested on 100 images (half of which fractured) and reported an AUC of 95.4%. However, they excluded images where lateral projection was inconclusive for the presence or absence of fracture which would bias the results favorably but contradict the goal of developing such systems (i.e. helping clinicians with difficult usually-missed fractures). Olczak *et al.*[19] re-trained five common deep networks from Caffe library [12] on dataset of 256,000 wrist, hand, and ankle radiographs, of which 56% of the images contained fractures. The dataset was divided into (70% training, 20% validation, and 10% testing) and used to train the networks for the tasks of detecting fractures, determining which exam view, body part, and laterality(left or right). Labels were extracted by automatically mining reports and DICOMs. The images were rescaled to 256 x 256 and then cropped into a subsection of the original image with the network’s input size. The pre-processing causes image distortion but they justified that as the nature of tasks does not need non-distorted images. The networks were pretrained on the ImageNet dataset [23] and then their top layers (i.e. classifier) were replaced with fully connected layers suitable for each task. The best performing network (VGG 16 [30]) achieved a fracture detection accuracy of 83% without report-

ing false positive rate. The model deals with various views independently but it does not combine them for a decision. Another related work [21] used a very deep CNN-based model (169 trainable layers) for abnormality detection from raw radiographs. Images are labeled as normal or abnormal, where abnormal does not always mean 'fractured'-it sometimes means there is metalwork present. Their dataset contains metal hardware in both categories (normal and abnormal) and also contains different age groups. This makes the definition of abnormality is rather unclear as what is considered abnormal for a certain group can be seen as normal for another age group and vice versa.

3 Background

3.1 Shape Modeling and Matching

Statistical Shape Models [3] are widely used for studying the contours of bones. Shape is the quality left after all differences due to location, orientation, and scale are omitted in a population of same-class objects. Statistical shape models (SSMs) assume that each shape instance is a deformed version of the mean shape describing the object class. The training data is used to identify the mean shape and its possible deformations. The contour of an object is described by a set of model points (x_i, y_i) packed in a $2n$ -D vector $\mathbf{x} = (x_1, \dots, x_n, y_1, \dots, y_n)^T$. An SSM is a linear model of shape variations of the object across the training dataset built by applying principal component analysis (PCA) to aligned shapes and fitting a Gaussian distribution in the reduced space. A shape instance $\mathbf{x}$ is represented as:

$$\mathbf{x} \approx T_\theta(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b} : \theta) \quad (1)$$

where $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}$ is the set of the orthogonal eigenvectors corresponding to the t highest eigenvalues of the covariance matrix of the training data, $\mathbf{b}$ is the vector of shape parameters and $T(\cdot : \theta)$ applies a similarity transformation with parameters θ between the common reference frame and the image frame. The number of the used eigenvectors t is chosen to represent most of the total variation (i.e. 95-98%).

One of the most effective algorithms for locating the outline of bones in radiographs is Random Forest Regression Voting (RFCLM) Constrained Local Model [4]. This uses a collection of Random Forests (RFs) to predict the most likely location of each point based on nearby image patches. A shape model is then used to constrain the points and encode the result.

3.2 Convolutional Neural Network (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep feed-forward artificial neural networks for processing data that has a known grid-like topology. They emerged from the study of the brain's visual cortex and benefited from the recent increase in the computational power and the amount of available training data.

A typical CNN (as in Figure 1) stacks few convolutional layers, then followed by a subsampling layer (*Pooling layer*), then another few convolutional layers, then another pooling layer, and so on. At the top of the stack fully-connected layers are added outputting a prediction (e.g. estimated class probabilities). This layer-wise fashion allows CNNs to combine low-level features to form higher-level features (see Figure 2), learning features and eliminating the need for hand crafted feature extractors. In addition, the learned features are translation invariant, incorporating the 2D spatial structure of images which contributed to CNNs achieving state-of-the-art results in image-related tasks.

Fig. 1: A CNN-based Classifier applied to a single-channel input image. Every convolutional layer (Conv) transforms its input to a 3D output volume of neuron activations. The pooling layer (Pool) downsamples the volume spatially, independently in each feature map of its input volume. At the end Fully-Connected layers (FC) output a prediction.

A convolutional layer has k filters (or kernels) of size $r \times r \times c$ (receptive field size) where r is smaller than the input width/height, and c is the same as the input depth. Every filter convolves with the input volume in sliding-window fashion to produce feature maps (see Figure 2). Each convolution operation is followed by a nonlinear activation, typically ReLU (Rectified Linear Unit) which sets any negative values to zero. A feature map can be subsampled by taking the mean or maximum value over $p \times p$ contiguous regions to produce translation invariant features (Pooling). The value of p usually ranges between 2-5 depending on how large the input is. This reduction in spatial size leads to fewer parameters, less computation, and controls overfitting.

The local connections, tied weights, and pooling result in CNNs have fewer trainable parameters than fully connected networks with the same number of hidden units. The parameters are learned by back propagation with gradient-based optimization to reduce a cost function.

Fig. 2: In CNN: k neurons receive input from only a restricted subarea (receptive field) of the previous layer output. Convolutioning the filters with the whole input volume produces k feature maps.

4 Methods

4.1 Patch Preparation

Because most parts of a radiograph are either background or irrelevant to the task, we chose to train CNNs on cropped patches rather than raw images. The steps of the automated system are shown in Figure 3. Following our previous work [7] we used a global search with a Random Forest Regression Voting (RFRV) technique to find the approximate radius location (red dots in Figure 3) followed by a local search performed by a sequence of RFCLM models with an increasing resolution to find its contour. The automatic point annotation gives information on the position, orientation and scale of the distal radius accurately. This is used to transfer the bone to a standardized coordinate frame before cropping a patch of size $(n_i \times n_i)$ pixels) containing the bone. We used the resulting patches to train and test a CNN. This process is completely automatic. Figure 4 shows examples of radiographs and extracted patches.

Fig. 3: Fully automated system for detecting wrist fractures.

Fig. 4: Example pairs of radiographs for four subjects with (a) a normal radius , (b-d) fracture radiuses. The first and third rows show the PA and LAT views respectively. The corresponding cropped patches appear below each view.

4.2 Network Architecture

We trained a CNN for each view. The two CNNs were classical stacks of CP layers (CP refers to one ReLU-activated Convolutional layer followed by a Pooling layer) with two consecutive fully-connected (FC) layers. No padding was used. Weights were initialised with the Xavier uniform kernel initializer [8] and biases initialised to zeros. The loss function was binary cross entropy optimised with Adam [14] (default parameter values used). An input patch size of 121x121, and of 151x151 were used for PA, and LAT networks respectively. Architecture details are summarised in Table 1. In our experiments we gradually increased the

number of CP layers and chose the network with the best performance. Figure 5 shows an example network with three CP layers followed by two fully-connected (FC) layers.

Fig. 5: An example network with an architecture of CP1-CP2-CP3-FC1-D-FC2. This network performed the best for the LAT view.

Layer	Type	Maps	Size	Kernel Size	Stride	Activation
In	Input	1	121x121	-	-	-
CP1	Convolution	32	119x119	3x3	1	ReLU
	Max Pooling 2D	32	59x59	2x2	2	-
CP2	Convolution	32	57x57	3x3	1	ReLU
	Max Pooling 2D	32	28x28	2x2	2	-
CP3	Convolution	64	26x26	3x3	1	ReLU
	Max Pooling 2D	64	13x13	2x2	2	-
CP4	Convolution	64	11x11	3x3	1	ReLU
	Max Pooling 2D	64	5x5	2x2	2	-
CP5	Convolution	64	3x3	3x3	1	ReLU
	Max Pooling 2D	64	1x1	2x2	2	-
FC1	Fully Connected	-	64	-	-	ReLU
D	Dropout (rate=0.5)	-	-	-	-	-
FC2	Fully Connected	-	1	-	-	Sigmoid

Table 1: The overall architecture detailed with maps' sizes corresponding to an input wrist patch of size 121x121. Same architecture also used with 151x151. In our experiments we gradually increased the number of CP layers and chose the network with the best performance.

5 Experiments and Results

5.1 Data

We collected a wrist dataset containing 1010 pairs of wrist radiographs (PA, and LAT) for 1010 adult patients (505 of whom had fractures) (see Fig. 4). Images for 787 patients (378 of whom had fractures) were gathered from two local emergency departments (EDs) while the rest were gathered from the MURA dataset[21] with fractures as abnormality. Fractured examples do not contain any plaster casts or metalware to make sure the network learns features for detecting fracture not hardware.

5.2 Fracture Detection FD

We carried out 5-fold cross validation experiments. During each fold 802 radiographs were used as training set, 102 as validation set, and 102 as testing set. The validation and testing sets were then swapped so that all the data were tested exactly once. Every time a network was trained from scratch for 20 epochs with batch size = 32 and the model with the lowest validation loss was selected. Training data was randomly shuffled at the start of each epoch to produce different batches each time. We found the architectures with three CP layers, and with four CP layers performed the best for the LAT view, and PA view respectively. Having trained the two CNNs, one for each view, their outputs are combined by averaging (see Figure 6). Figure 7 shows the average performance and learning curves. We achieved an average performance of AUC=95% for PA view , 93% for LAT view, and 96% from both views combined.

Fig. 6: During testing the outputs for both views are combined by averaging.

Fig. 7: Fracture Detection. (a) ROC for PA view. (b) ROC for LAT view. (c) ROC of from both views combined. (d) Example of learning curves for a model.

Kim *et al.* in [13] used features originally learned to classify non-radiological images [23] and used them to detect fractures in LAT views and reported an AUC of 95.4%. Unlike their work we have not excluded images where lateral projection was inconclusive for the presence or absence of fracture which would bias the results favorably. In our case, we performed 5-fold cross-validation and reported an overall AUC of 96%. For the sake of comparison with our previous RF-based technique in [7] we repeated all experiments in [7] on the current dataset with the same fold divisions and found an AUC of 92% from two views combined, 89% and 91% for PA view, and LAT view respectively (See Table 2 and Figure 8). The CNN-based technique clearly outperforms the RF-based one.

Table 2: Comparison between CNN-based and RF-based techniques on the same dataset in terms of $AUC \pm \text{stdev}$.

Method	PA view	LAT view	Both Views
CNNs	0.95 ± 0.01	0.93 ± 0.02	0.96 ± 0.01
RFs on shape and texture params [7]	0.89 ± 0.02	0.91 ± 0.02	0.92 ± 0.02

Fig. 8: Comparison between ROC Curves for the proposed CNN-based technique and the relevant RF-based work in [7] on: a) PA view, b) LAT view, and c) both views combined for the same dataset in terms of $AUC \pm \text{stdev}$

5.3 Conclusions

We presented a system for automatic wrist fracture detection from plain PA and LAT X-rays. The CNN is trained from scratch on radiographic patches cropped around the joint after automatic segmentation and registration. This directed preprocessing ensures meaningful learning from only the targeted region in scale which in turn reduces the noise a CNN is exposed to compared to when trained on full images containing parts that are not relevant to the task. Radiographs, unlike photos, have predictable contents that allow model-based techniques to

work well and therefore they can provide CNNs with an input that dispense with the need to: (1) perform any data augmentation and (2) unnecessarily complicate the deep architecture and its learning process. Our work was the first to train CNNs from scratch on the task of detecting wrist fractures and to combine the two views for a decision. The experiments showed that combining the results from both views leads to an improvement in overall classification performance, with an AUC of 96% compared to 95% for PA view and 93% for LAT view.

6 Acknowledgment

The research leading to these results has received funding from Libyan Ministry of Higher Education and Research. The authors would like to thank Dr Jonathan Harris, Dr Matthew Davenport, and Dr Martin Smith for their help setting up the project.

References

1. F. Bayram and M. Çakırolu. DIFFRACT: DIaphyseal Femur FRacture Classifier SysTem. *Biocybernetics and Biomedical Engineering*, 36(1):157–171, 2016.
2. Y. Cao, H. Wang, M. Moradi, P. Prasanna, and T. F. Syeda-Mahmood. Fracture detection in x-ray images through stacked random forests feature fusion. In *International Symposium on Biomedical Imaging (ISBI)*, pages 801–805, April 2015.
3. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
4. T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291. Springer, 2012.
5. C. M. Court-Brown and B. Caesar. Epidemiology of adult fractures: A review, 2006.
6. R. Ebsim, J. Naqvi, and T. Cootes. Detection of wrist fractures in x-ray images. In *Clinical Image-Based Procedures*, pages 1–8, Cham, 2016. Springer International Publishing.
7. R. Ebsim, J. Naqvi, and T. Cootes. Fully automatic detection of distal radius fractures from posteroanterior and lateral radiographs. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, pages 91–98, Cham, 2017. Springer International Publishing.
8. X. Glorot and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks*, pages 249–256. 2010.
9. C. A. Goldfarb, Y. Yin, L. A. Gilula, A. J. Fisher, and M. I. Boyer. Wrist fractures: what the clinician wants to know. *Radiology*, 2001.
10. H. R. Guly. Injuries initially misdiagnosed as sprained wrist (beware the sprained wrist). *Emergency medicine journal : EMJ*, 2002.
11. J. J., A. Chiabrera, M. Hatem, N. Z. Hakim, M. Figueiredo, P. Nasser, S. Lattuga, A. A. Pilla, and R. S. Siffert. A neural network approach for bone fracture healing assessment. *IEEE Engineering in Medicine and Biology Magazine*, 9(3):23–30, Sept 1990.

12. 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 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
13. D. Kim and T. MacKinnon. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, 2017.
14. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
15. S. E. Lim, Y. Xing, Y. Chen, W. K. Leow, T. S. Howe, and M. A. Png. Detection of femur and radius fractures in x-ray images. In: *Proc. 2nd Int. Conf. on Advances in Medical Signal and Information Processing*, 1:249–256, 2004.
16. C. Lindner, P.A.Bromiley, M.C.Ionita, and T. Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 37(9):1862–1874, 2015.
17. C. Lindner, S. Thiagarajah, J. M. Wilkinson, T. Consortium, G. A. Wallis, and T. F. Cootes. Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting. *Medical Image Analysis*, 32(8):1462–1472, 2013.
18. V. L. F. Lum, W. K. Leow, Y. Chen, T. S. Howe, and M. A. Png. Combining classifiers for bone fracture detection in X-ray images. 1:I–1149–52, 2005.
19. J. Olczak, N. Fahlberg, A. Maki, A. S. Razavian, A. Jilert, A. Stark, O. Sköldenberg, and M. Gordon. Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms are they on par with humans for diagnosing fractures? *Acta orthopaedica*, 88(6):581–586, 2017.
20. B. Petinaux, R. Bhat, K. Boniface, and J. Aristizabal. Accuracy of radiographic readings in the emergency department. *American Journal of Emergency Medicine*, 2011.
21. P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Ng. MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs. *ArXiv e-prints*, Dec. 2017.
22. M. G. Roberts, T. Oh, E. M. B. Pacheco, R. Mohankumar, T. F. Cootes, and J. E. Adams. Semi-automatic determination of detailed vertebral shape from lumbar radiographs using active appearance models. *Osteoporosis International*, 23(2):655–664, 2012.
23. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
24. D. M. Ryder, S. L. King, C. J. Oliff, and E. Davies. A possible method of monitoring bone fracture and bone characteristics using a noninvasive acoustic technique. In *Acoustic Sensing and Imaging, 1993., International Conference on*, pages 159–163, Mar 1993.
25. V. R. Singh and S. K. Chauhan. Early detection of fracture healing of a long bone for better mass health care. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, volume 6, pages 2911–2912 vol.6, Oct 1998.
26. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
27. T. P. Tian, Y. Chen, W. K. Leow, and W. Hsu. Computing Neck-Shaft Angle of Femur for X-Ray Fracture Detection. *Proc. Int. Conf. on Computer Analysis of Images and Patterns*, pages 82–89, 2003.

28. Y. J. Y Jia. Active Contour Model with Shape Constraints for Bone Fracture Detection. *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*, 3:90–95, 2006.
29. D. W. H. Yap, Y. Chen, W. K. Leow, T. S. Howe, and M. A. Png. Detecting femur fractures by texture analysis of trabeculae. *Proceedings - International Conference on Pattern Recognition*, 3:730–733, 2004.
30. S. Zhong, K. Li, and R. Feng. Deep convolutional hamming ranking network for large scale image retrieval. In *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*, pages 1018–1023. IEEE, 2014.