ARCHi: Automated Recognition and Classification of Hip Fractures

E.A. Murphy, C.L. Gregson, O.A. Von Arx, M.R. Whitehouse, C. Budd, H.S. Gill

April 4, 2018

1 Introduction

Hip fractures are a major cause of death and disability for older people and are one of the costliest treatments for the NHS. It is expected that by 2025, there will be more than 100,000 hip fractures a year in the UK and the current annual cost to the NHS is estimated at more than 1000 million (Leal et al., 2016). Given the high mortality rates for hip fractures (∼9% within 30 days; ∼30% within one year (Leal et al., 2016)), any improvement in classification and hence treatment will have significant benefits.

Hip fractures can be classified using the AO system[1], or by describing the location and displacement, as used by the National Hip Fracture Database (NHFD). However, there is no standardised system in the NHS to determine who performs this classification and there is wide variation between hospitals. As classification influences the chosen treatment, differences in the classification process can affect patient outcomes, potentially through misclassification. Misclassifications may undermine the validity of data captured by the NHFD. Our ongoing study proposes a machine learning based method to automatically classify hip fractures using X-rays, with the eventual aim of standardising classification across NHS hospitals.

There are two stages in our proposed classification process: automatically select the hip joints in the X-ray image; and then classify the hip fracture given this selected region. We train a fully convolutional network to automatically locate the hip joint,

[1] www2.aofoundation.org

using 636 manually annotated X-ray images split 60:20:20 into training, validation and test sets, respectively. The training set is increased artificially using standard augmentation techniques to improve learning.

Still to do: [*We then train a convolutional neural network to automatically classify the type of fracture. To ensure the validity of the classifier we have created a dataset of X-ray images from hip fracture databases in a large teaching hospital and a large DGH. From a period covering April 2007 to December 2016, 2000 hip fracture X-rays are extracted by stratified random sampling and anonymised. The X-ray images are classified by a panel of experts, including orthopaedic surgeons and radiologists.*

Using the classes given by the NHFD, the number of X-rays per class in our dataset are: intracapsular (displaced: 602), (undisplaced: 346); intertrochanteric: 602, (grade A1/A2: 184), (grade A3: 4); subtrochanteric: 240; other: 22. The unbalanced classes are due to natural differences in the frequencies of occurrence. This is counteracted in the training algorithm by a weighted sampling of the images. Results of our study are given, along with a discussion of the benefits of our system and its potential impact on hip fracture classification in the NHS.]

2 Related Work

Although convolutional neural networks (CNNs) have been used in applications since the 90's (LeCun et al., 1998), their recent explosion in popularity and use dates back to 2012 when Krizhevsky et al. won the ImageNet Large Scale Visual Recognition Challenge with AlexNet, beating algorithms that relied on traditional image analysis techniques (Krizhevsky et al., 2012). Since then, they have been improved to the extent that they outperform humans at certain tasks [ref] and are ubiquitous, used in many fields and applications.

A survey on deep learning in the field of medical imaging is given in Litjens et al. (2017). Notable success stories include using CNNs to classify skin cancer, achieving on-par performance with human experts (Esteva et al., 2017), and to detect diabetic retinopathy in retinal fundus photographs (Gulshan et al., 2016). Deep learning is becoming more common in musculoskeletal image analysis, e. g., Jamaludin et al. (2016) use a CNN-based system to achieve near-human performance to classify and qualitatively localise multiple abnormalities in sagittal lumbar MRIs while Spampinato et al. (2017) achieve state-of-the-art performance using CNNs to estimate skeletal bone age in X-ray images.

The method we use below for locating the hip joints uses a Fully Convolutional Network (FCN) (Long et al., 2015). These were developed to allow for pixel-wise classification, otherwise known as semantic segmentation. Our simplified version of an FCN is based on the work in Antony et al. (2017), who used a lightweight FCN to accurately locate knee joints in x-ray images.

3 Data and Classification System

3.1 Automated Detection Dataset

The dataset used to train the detection algorithm consists of a total of 638 antero-posterior (AP) pelvic x-rays, taken from [*EM: where did this data come from?*]. For each image, two regions of interest (ROIs) were marked using MATLAB's Training Image Labeler App. In order to have consistent ROIs, natural features of the hip joint were chosen as boundary markers of the ROIs. These boundary markers were chosen based on the advice of orthopaedic surgeons so that the ROIs provide sufficient coverage of the hip joint to allow for classification to be performed. The horizontal boundaries are the outermost part of the femoral head and the inner edge of Shenton's line. The upper boundary marker is the neck of [*EM: what is the anatomical term?*]; and the lower boundary is chosen so that the distance from the boundary to the lower trochanter is approximately the same as the distance from the lower trochanter to the upper trochanter. An example x-ray with marked ROIs is shown in Figure 1. Using the labelled ROIs, masks of the x-rays are created where pixels in the ROI have a value of 1 and pixels outside these regions have a value of 0. These are then used as the ground-truth label images.

3.2 Automated Classification Dataset

TO DO LATER

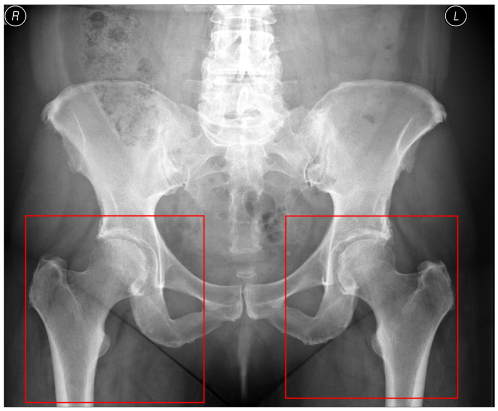


Figure 1: An example AP pelvic x-ray with the two ROIs surrounding the hip joints marked by red rectangles.

4 Methodology

4.1 Automated Detection of Hip Joints

4.1.1 Data Preparation and Augmentation

To prepare the data set for the FCN, the x-ray images and their corresponding labels were resized to 256 $\times$ 256. The images were then randomly assigned to either the training, validation or test sets, split 60%, 20%, 20% respectively. To artificially increase the size of the data set, each image in the training set was flipped horizontally, its inverse was taken, and random shifts were applied. This increased the size of the training set by a factor of 16.

4.1.2 Fully Convolutional Network

The success of in detecting knee joints inspired us to base our algorithm on the same method of using a lean fully convolutional network (FCN) to detect the ROIs, trained from scratch on our data set. The network, illustrated in Figure 2, takes 2-d greyscale images as input and applies four stages of 3×3 convolutions, each followed by a rectified linear unit activation layer (ReLU). The first three convolutional layers are each followed by a 2×2 max-pooling layer. To allow for dense pixel outputs, the output of the fourth convolutional layer is 8×8 upsampled to account for the three earlier stages of pooling. After upsampling, there is a 1×1 convolutional layer with softmax activation. The output of the network is the same size as the input, in our

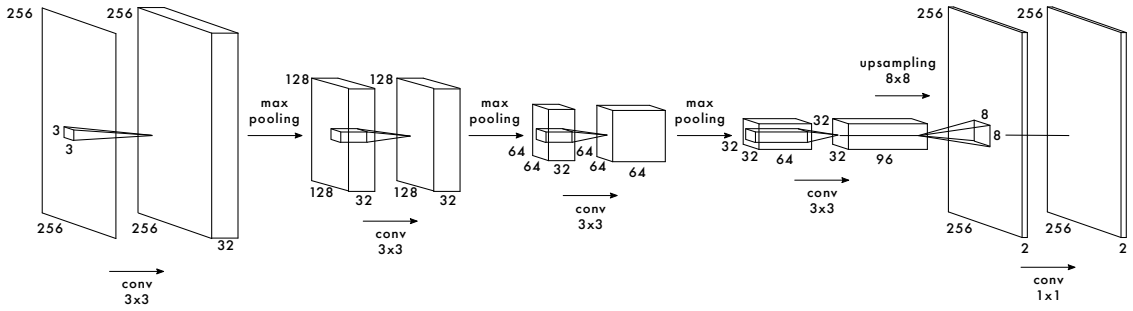
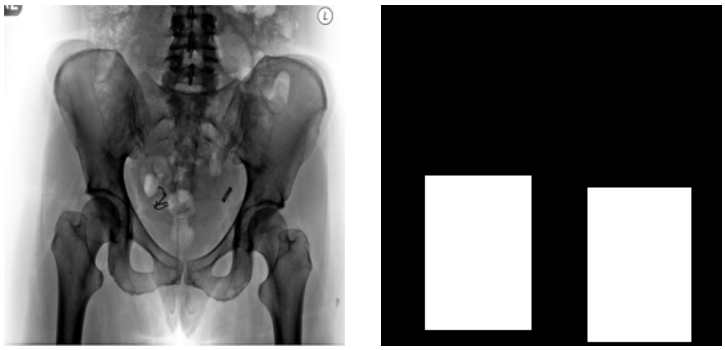


Figure 2: An illustration of the architecture of the FCN used for locating hips joints.

case 256×256. Figure 3 displays a sample input to the FCN, the x-ray image and the ground-truth label image.

The network was trained from scratch on the data set described in Section 4.1.1 to minimise the total binary cross entropy between the pixels in the ground truth labels and the corresponding outputs of the softmax layer.



(a) X-ray image. (b) Ground-truth label.

Figure 3: An example input to the FCN, taken from the test set, along with its ground-truth label.

4.1.3 Contour Detection

The ROIs outputed from the FCN are rough-edged must be converted to rectangular shape by post-processing. This was done in MATLAB, using standard image

processing techniques. First, small holes are filled and regions with less than XX pixels are removed so that only the two largest regions remain. For each region, R_O, a rectangle, R_F, is fitted by minimising $\sum \text{XOR}(R_O, R_F)$ i. e., the number of pixels that must change value when transforming R_O to R_F. The starting point for the optimisation algorithm is the midpoint between the outer and inner rectangular boundaries of the original region.

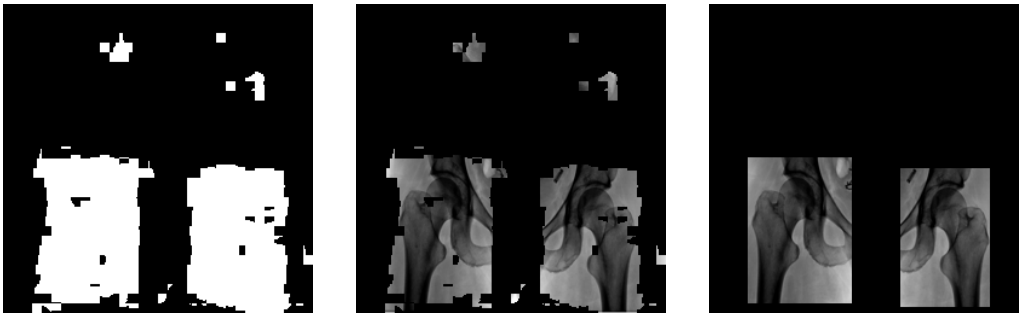
4.2 Automated Classification of Hip Joints

TO DO LATER

5 Results

5.1 Automated Detection of Hip Joints

The accuracy of the detection algorithm is measured using the standard Jaccard Index, J, also known as the Intersection-Over-Union (IoU). For two regions, in this case the ground truth ROI and the predicted ROI, J is computed as the area of intersection of the two regions divided by the area of the union of the two regions. By convention, the predicted ROI is considered correct for a value of $J > 0.5$.



(a) Output from the FCN before contour detection.

(b) Image masked with the output of the FCN.

(c) Image masked with contour detection output.

Figure 4: Output from the FCN and contour detection.

To examine the accuracy of the output, we can calculate the differences between the ground truth label and the output of the algorithm for the horizontal and vertical limits of the ROIs. There are four limits each for the (anatomically) right, R, and

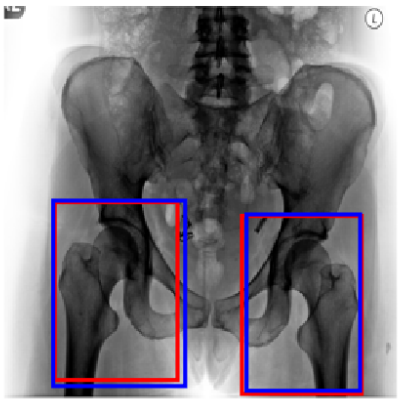


Figure 5: A comparison of the ground-truth ROI and the output from the detection algorithm shown by the red and blue rectangles, respectively. The overall J value between the red and blue rectangles is 0.9.

left, L, ROIs: *outer*, the outermost part of the femoral head; *inner*, the inner edge of Shenton's line; *top*, the neck of [*EM: what is the anatomical term?*]; and *bottom*, in the subtrochanteric region. For each of these limits for every example in the test set, the difference between the expected and actual value is calculated. The results are shown in Figure 6, using boxplots with whiskers. The central mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles, the whiskers extend to the furthest points that are not outliers. The outliers are marked with red crosses. From these results, we can see that the algorithm achieves better results for the horizontal limits, with average standard deviations of 4 pixels for both the outer and inner limits, than the vertical, with average standard deviations of 5 pixels and 8 pixels for the top and bottom. This is to be expected as there are definite physical features that limit the horizontal edges of the ROI, while defined features do not exist for the vertical limits.

An example output, taken from the test set, from the FCN is displayed in Figure 4, along with this output overlaid on the input x-ray image. Also in Figure 4 is the result of contour detection algorithm, also overlaid on the input x-ray image. A comparison of the ground-truth label and the output of the FCN, after contour detection has been applied, is shown in Figure 5. For this example the overall value of J is 0.9.

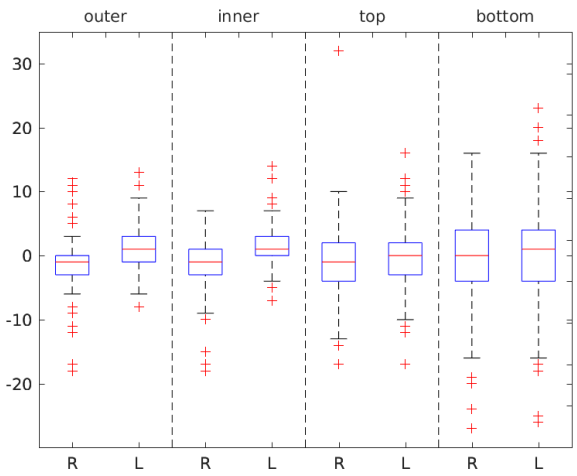
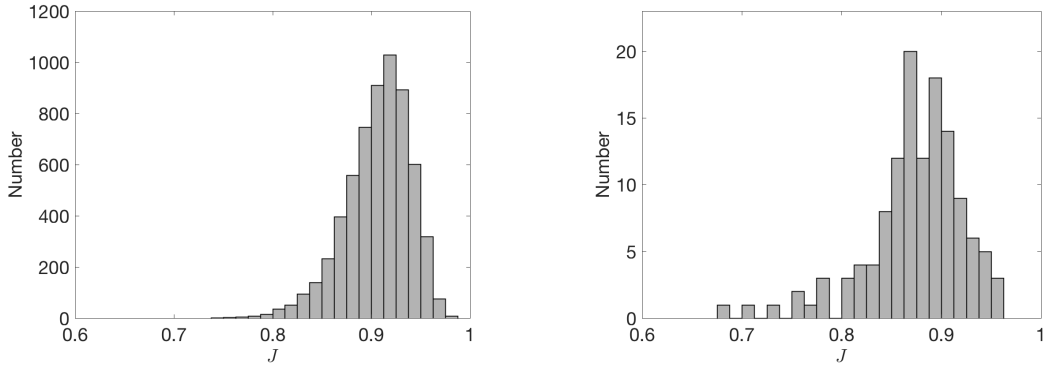


Figure 6: Box plots showing the differences between the limits of the ROIs for the ground truth label and the output of the algorithm. There are four limits for the (anatomically) right, R, and left, L, ROI: *outer*, the outermost part of the femoral head; *inner*, the inner edge of Shenton's line; *top*, the neck of [*EM: what is the anatomical term?*]; and *bottom*, in the subtrochanteric region. The central mark is the median, the edges of the box are the 25^{th} and 75^{th} percentiles, the whiskers extend to the furthest point that isn't an outlier. The outliers are marked with red crosses.

Data	J >0.5	J >0.6	J >0.7	J >0.8	Mean J	Std. Dev.
Training	100%	100%	100%	98.9%	0.89	0.03
Test	100%	99.2%	97.6%	82.7%	0.84	0.05

Table 1: Results of hip detection algorithm before applying contour detection.

The mean and standard deviation of J for the training and test sets is shown in Table 1, along with the proportion of samples scoring a value of J over 0.5, 0.6, 0.7 and 0.8. These are the results computed on the raw output of the FCN, before the contour detection algorithm has been applied. These results show that all samples score values of J higher than 0.5 and therefore, the hip joints are correctly located by the FCN in all samples in the training and test sets.



(a) Training set.

(b) Test set.

Figure 7: The distributions of J for the training and test sets. These values of J are computed after the contour detection algorithm has been applied.

Data	J >0.5	J >0.6	J >0.7	J >0.8	Mean J	Std. Dev.
Training	100%	100%	100%	99.4%	0.91	0.03
Test	100%	100%	99.2%	92.9%	0.87	0.05

Table 2: Results of hip detection algorithm after applying contour detection.

Table 2 displays the same information as Table 1 for the results after the contour detection algorithm has been applied. Again, all samples in the training and test sets score a value of J higher than 0.5. Applying the contour detection algorithm improves the average value of J, raising it by 0.02 for the training set and 0.03 for the test set. It also results in all samples in the test set scoring a value of J higher than 0.6 and increases the number of test samples scoring higher than 0.8 by 10 percentage points. Figure 7 displays the distributions of J, computed after the contour detection algorithm has been applied to the outputs of the FCN, for the training and test sets. For the test set, the average value of J is 0.87, showing that the combined algorithms are highly accurate at recognising the locations of the hip joints.

5.2 Automated Classification of Hip Joints

TO DO LATER

6 Discussion

References

Antony, J., K. McGuinness, K. Moran, and N. E. O'Connor (2017). Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 376–390. Springer.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542*(7639), 115.

Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama 316*(22), 2402–2410.

Jamaludin, A., T. Kadir, and A. Zisserman (2016). Spinenet: automatically pinpointing classification evidence in spinal mris. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 166–175. Springer.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

Leal, J., A. Gray, D. Prieto-Alhambra, N. K. Arden, C. Cooper, M. Javaid, A. Judge, R. study group, et al. (2016). Impact of hip fracture on hospital care costs: a population-based study. *Osteoporosis International 27*(2), 549–558.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*(11), 2278–2324.

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez (2017). A survey on deep learning in medical image analysis. *Medical image analysis 42*, 60–88.

Long, J., E. Shelhamer, and T. Darrell (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Spampinato, C., S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi (2017). Deep learning for automated skeletal bone age assessment in x-ray images. *Medical image analysis 36*, 41–51.