AUTOMATIC DETECTION OF WRIST FRACTURES IN RADIOGRAPHS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE & ENGINEERING

2018

Raja Ebsim

School Of Computer Science

Contents

Word Count: 36628

List of Tables

List of Figures

Abstract

Fractures of the wrist are usually identified in Emergency Departments (ED) by doctors examining lateral (LAT) and posterioanterior (PA) radiographs. Unfortunately missing such fractures is one of the most common diagnostic errors in EDs, leading to delayed treatment and more suffering for the patient. This is mainly because the majority of patients attending EDs are seen by less experienced junior doctors. This problem is widely acknowledged, so in many hospitals X-rays are reviewed by an expert radiologist at a later date - however this can lead to significant delays on missed fractures which can have an impact on the eventual outcome. There is an urgent need for automated methods to analyse radiographs of the wrist in order to identify abnormalities and thus prompt clinicians, hopefully reducing the number of errors. This project developed the first fully automated system to analyse the wrist in the two standard views (i.e. PA and LAT). The system achieves an encouraging fracture detection rate, with an AUC of 0.93 from LAT view, of 0.95 from PA view, and of 0.96 from both views combined. The project also worked on improving the state-of-art technique Random Forest Regression Voting Constrained Local Model (RFCLM) in order to perform better on overlapping structures in radiographs and showed significant performance improvements when segmenting the radius and ulna in wrist radiographs, and femoral condyles in lateral knee radiographs.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses

Acknowledgements

All praises are due to God, the Most Gracious and the Most Merciful. I would like to thank Him for giving me opportunity, determination and strength to do my research. I would also like to take this opportunity to express my deep gratitude to everyone that contributed, directly or indirectly, to me taking on this PhD project and to successfully completing it. Specifically, I would like to thank:

- My supervisor Prof. Tim Cootes for his great patience and dedication to his students. This thesis would not have been possible without his extraordinary support.
- My examiners Prof. Alejandro Frangi and Dr. Paul Bromiley.
- The University of Manchester for giving me the opportunity to do this PhD and the Libyan Ministry of Higher Education and Research for funding it.
- Dr. Jonathan Harris, Dr. Matthew Davenport, and Dr. Martin Smith for their collaboration to set up the project.
- Dr. Jawad Naqvi for generously dedicating his time collecting clinical datasets and providing help and guidance on the medical and clinical aspects of the project.
- Dr. Carole Twining, Dr. Claudia Lindner, and Dr. Henry Reeve for their insightful comments and helpful advice.
- Everyone in the DIIDS for their assistance, especial thanks to Shelagh and Catherine.
- My fellow colleagues : Jessie, Jon, Nora, Luca, Luke, Geo, Areej, Anna-Maria, Ethan, Sultan for their friendship, support, and great baking.

Finally I would like to thank from the bottom of my heart all my family for their continued prayers and unending support in all my life choices and struggles. I would not reach this stage without their love and care. This thesis is dedicated to them.

Abbreviations

AAM Active Appearance Model

APM Appearance Model

ASM Active Shape Model

AUC Area Under ROC Curve

CDF Cumulative Distribution Function

CLM Constrained Local Model

CNN Convolutional Neural Network

CV Cross Validation

ED Emergency Department

LAT Lateral

MSE Mean Squared Error

OA Osteoarthritis

PA PosteroAnterior

PCA Principal Component Analysis

RF Random Forest

RFCLM Random Forest Regression Voting Constrained Local Model

RFRV Random Forest Regression Voting

ROC Receiver Operating Characteristic

SOA The State Of the Art

SSM Statistical Shape Model

StdErr. standard error

stdev. standard deviation

SVM Support Vector Machine

List of Publications

- R. Ebsim, J. Naqvi and T. F. Cootes, “Detection of Wrist Fractures in X-Ray Images.”, In: Lecture Notes in Computer Science, vol 9958. Springer.
- R. Ebsim, J. Naqvi and T. F. Cootes, “Fully Automatic Detection of Distal Radius Fractures From Posteroanterior and Lateral Radiographs.” In: Lecture Notes in Computer Science, vol 10550. Springer.
- R. Ebsim, J. Naqvi and T. F. Cootes, “Automatic Detection Of Wrist Fractures From Posteroanterior and Lateral Radiographs: A Deep Learning-Based Approach”, In: Lecture Notes in Computer Science, vol 11404. Springer.
- In preparation: R. Ebsim, J. Thomson and T. F. Cootes, “Automatic Osteoarthritis Diagnosis From Posteroanterior Knee Radiographs”
- In preparation: R. Ebsim and T. F. Cootes, “Shape-Specific Local Models To Segment Overlapping Structures In Radiographs”.

Chapter 1

Introduction

1.1 Motivation

Radiographs (X-rays) are one of the most widely used forms of medical images due to their availability, low cost and modest radiation dose. They are used in Emergency Departments (EDs) to diagnose bone fractures and by clinicians to look for signs of joint diseases such as osteoarthritis and monitoring their progression. However, one of the commonest diagnostic errors when people visit an ED unit, is that a fracture which is visible on an X-ray is missed by the clinician on duty. That is mainly because extracting all available information from radiographs requires years of training and there is always the question of inter-observer reliability [2, 5, 99, 106]. This problem is widely acknowledged, so in many hospitals X-rays are reviewed by an expert radiologist at a later date - however this can lead to significant delays in catching missed fractures which can have an impact on the eventual outcome. For these reasons developing computer-aided image interpretation techniques is highly interesting from a clinical perspective. A system which prompts clinicians to look more carefully at certain regions that might be fractured holds the promise of reducing the number of fractures missed in the ED.

1.2 Aims and Objectives

The main aim of the project was to create a system which can detect wrist fractures automatically from radiographs. However, this end goal requires several developmental goals to be achieved throughout the project. The detailed objectives were to:

- Collect a dataset of wrist radiographs and associated clinical data about fracture status. These should be verified by a clinical expert.
- Identify a suitable point annotation model to capture the shape of the distal radius in plain PosteroAnterior (PA) and Lateral (LAT) wrist radiographs.
- Build a fully automatic system that accurately and robustly annotates the distal radius in PA and LAT wrist radiographs of varying quality and resolution.
- Explore various feature extraction, selection, and learning methods to analyse their suitability to capture the shape and texture variations due to fractures compared to normals.
- Evaluate the performance of the developed system.

1.3 Contributions

The contributions of the thesis include:

- A novel fully automated wrist fracture detection system was developed using combined information (shape and texture) from both views (plain PA and LAT views) with random forest classifiers. We showed for the first time that fractures can be better identified in the lateral view, and that combining information from both views leads to an overall improvement in performance.

- A novel deep-learning-based approach for automatically detecting wrist fractures from plain PA and LAT X-rays. Previous work used transfer learning and fine-tuning of off-the-shelf deep models, we train convolutional neural networks from scratch on registered patches containing the target bone. The same approach was also applied to the problem of diagnosing knee osteoarthritis from PA radiographs and achieved results comparable to the state of the art.

- Our work highlighted the problem behind segmenting overlapping structures in lateral radiographs. Limited constraints on how a patient is positioned result in the local appearance around a point on one bone changing dramatically as other bones move over it.

- We adapted a technique that has been used previously in [93] for facial point tracking over a wide range of head angles on the task of tracking driver faces to solve the problem of segmenting overlapping structures in LAT X-rays.
  The technique improves Random Forest Regression Voting Constrained Local Model (RFCLM) [74] by switching between different local models depending on the current global shape.

- We introduced a multi-start initialisation scheme for RFCLM and a new variant (*Switched RFCLM*).

1.4 Outline of the Thesis

The next chapter (**Chapter 2**) presents an overview of the relevant literature. The review describes the problem of misdiagnosing fractures at EDs and its current solutions, wrist fractures as one of the most commonly missed fractures, current image analysis literature on detecting and classifying fractures, and an introduction to

the state-of-the-art algorithm used to detect the wrist and localise the feature extraction/learning methods.

Chapter 3 describes the data used throughout the experiments (Chapters 4-6), the methods for wrist detection and segmentation which were based on the RFCLM, the feature extraction/learning methods and classifiers used, and the techniques used to evaluate the classification performance.

Chapters 4 presents the experiments and results for automatic wrist segmentation in the posterioanterior (PA) and lateral (LAT) wrist views, and the use of random forest classifiers trained on the extracted features of shape, texture and appearance from Chapter 3 on the task of fracture detection.

Chapters 5 presents and evaluates a novel deep-learning-based approach for automatically detecting wrist fractures from plain PA and LAT X-rays. The same technique was further evaluated on another problem: Automatic diagnosis of knee Osteoarthritis (OA) in plain PA X-rays. We performed experiments on detecting OA (OA vs Non-OA) and on classifying its severity according to Kellgren-Lawrence Grading (KL) [64].

Chapter 6 presents work on improving Random Forest Regression Voting Constrained Local Model (RFCLM) in order to perform better on overlapping structures in lateral radiographs, and shows how it improves the accuracy of segmenting (i) the radius and ulna in wrist radiographs, and (ii) femoral condyles in lateral knee radiographs.

Chapter 7 draws some conclusions from the work and outlines areas requiring future developments.

Chapter 2

Literature Review

This chapter describes the structure, function, and characteristics of human bones, and the literature relating to fractures missed in emergency departments. The chapter also describes wrist anatomy and its radiographic examination, reviews the literature related to the main image analysis techniques used for fracture detection and classification, and concludes with the best reported object detection/segmentation methods as they are essential building blocks of any computer-aided fracture detection system.

2.1 Bones and Fractures

Bones are the elements of the human skeleton system shaping the body, providing support and protection to the various organs of the body, and enabling mobility. They also produce blood cells and store minerals. Bones have different shapes and sizes with a complex internal and external structure and are mainly composed of approximately 70% minerals, 22% protein, and 8% water. There are 270 bones at birth but some bones are fused together to become 206 by adulthood. There are two main types of bone tissues, trabecular and cortical. Cortical tissues constitute 80% of the

bone mass of the human body because they are dense and compact. Inside the cortical envelope lies trabecular bone tissue which is a cancellous (mesh-like) structure whose surface area-to-weight ratio is higher than that of the cortical tissue in order to provide support for the skeleton, without the added weight that would be present with a denser structure.

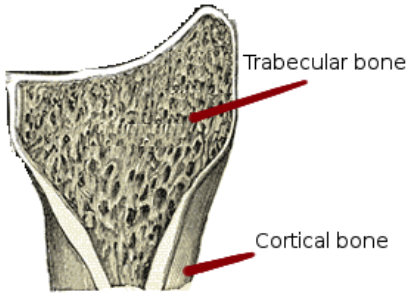


Figure 2.1: Bone tissues [61].

Despite appearances, bones are living tissues and are continuously being renewed in a natural process called *bone remodelling*. It involves bone breakdown (resorption) followed by formation of new bone tissue (ossification). Bone remodelling maintains bone integrity and strength and reshapes bone architecture in response to the mechanical forces placed on it. Imbalanced bone remodelling can result in bone disorders. Although the hardest organs in the body, bones are vulnerable to being cracked and fractured. In general the fractures can be caused by accidents of high force or pressure or by pathological reasons (e.g osteoporosis). In practice, clinicians in Emergency Departments (EDs) rely mainly on radiographs to detect fractures and to determine their nature. However, the literature shows that the vast majority of the ED diagnostic errors are missed fractures and the majority of them are in the peripheral bones [52, 91]. In the next section, we will overview the literature related to missing abnormalities in EDs: causes, types, and current practice to reduce missing rates.

This literature review along with meetings we held with ED consultants and radiologists from a local hospital was crucial for deciding on working on detection of wrist fracture as a research problem for this PhD project.

2.2 Radiographic Examination in EDs

2.2.1 The current state of ED workload

There is an increasing demand on emergency departments' services. In England alone there were 22.3 million attendances recorded in 2014-2015, with an increase of 2.7% from the previous year and an increase of about 25% over the last decade [85]. 29-50% of ED attendances are referred for radiographic examination [35, 86, 95]. The ED radiographs are initially interpreted by ED medical staff, however the accuracy of these interpretations has always been a source of concern.
The discrepancy in interpretation between emergency departments and radiology departments ranges from 1.5 to 7.8% [12, 91, 104]. 0.3-2.8% of the Radiology's interpretations dictates a significant change in patient treatments [12, 50, 122]. In order to reduce patient suffering and to avoid the potential of litigation all ED radiographs should be reported by a senior radiologist [12] as a study showed that the discrepancy rate between junior and senior radiologists' reports is 6.3% (i.e similar to the discrepancy rate between emergency departments and radiology departments) and two thirds of these differences are clinically significant [101].

2.2.2 Causes of ED Diagnostic Errors and Current Solutions

Many of the discrepancies between emergency departments and radiology departments are caused by senior house officers (SHO). SHOs are "the most junior doctors

working in EDs and often work in a 6-month stand-alone posts immediately following their pre-registration house jobs, they are inexperienced and will not return following completion of the post" [126]. The fact that the majority of patients attending EDs are seen by junior doctors [71, 81] makes ED a high-risk area of modern clinical practice [71]. One study [118], in which SHOs were given only abnormal radiographs, showed that ED junior doctors misread 35% of radiographic abnormalities, 39% of which have clinically significant consequences. Another study [81] showed that SHOs failed to diagnose two thirds of significant trauma abnormalities on X-ray. Among the strategies that have been suggested to reduce ED radiograph-related diagnostic errors are immediate radiology reporting, and involving radiographers in initial assessment of radiographs before review by ED clinician [12, 13, 77].

It is clear that there are insufficient resources to apply immediate radiology reporting. In many of UK radiology departments, all ED radiographs are routinely reported in retrospect. They endeavor to report within one working day but it might take up to 3 days [55]. In general, there is an increasing workload on radiologists in the UK who are basically in chronic shortage. Per million population there are 48 trained radiologists in the UK whereas there are 92 in Germany, 112 in Spain and 130 in France [87]. In a survey amongst NHS trusts in England, it is found that almost 330,000 patients are waiting for more than a month for results of their X-rays [88]. One practical solution was to involve radiographers (i.e. medical technologists who take X-rays) in detecting abnormalities. A radiographer marks a radiograph as abnormal if she believes it shows an acute abnormality. This solution was introduced in 1980s for its simplicity and the improved accuracy [13] and it is now known as Radiographer Abnormality Detection scheme (RAD) and implemented in 85%-90% of UK hospitals [109]. However, the low compatibility of the current practice (i.e. flagging and reporting) with digital imaging technologies [23] and the lack of communication standardization could increase the risk of errors [109] instead of reducing it.

Apart from patient suffering and the risk of loss of patients' confidence in the health provider, misinterpreting the radiograph at the first visit would lead to more visits and therefore more costs, missed time at work, and the potential of litigation. This urgent need of help in radiograph interpretation at EDs could be fulfilled by developing algorithms and tools for more efficient interpretation of the imaging examination focusing on the commonly-missed abnormalities and therefore reducing ED diagnostic errors.

2.2.3 Wrist Fractures in EDs

The literature suggests that a great deal of these diagnostic errors are missed fractures and the majority of them are in the peripheral bones. A retrospective study [52] of diagnostic errors over four years, in a busy district general hospital emergency department, reported that:

- 77.8% of the diagnostic errors were missing abnormalities seen on radiographs,
- 79.7% of the missed abnormalities were fractures,
- 17.5% of the missed fractures were wrist fractures,
- 85.3% of the errors were made by SHOs.

In another retrospective review [91] of all ED radiographs over nine-year period, it was found that almost 56% of the missed bone abnormalities were fractures and dislocations. Taking into consideration that the annual fracture incidence in England is 3.6% [36] these error rates are worrying.

Another study [123] about missed extremity fractures at ED showed that: wrist fractures are the most common among all extremity fractures (19.7% of total fractures) with miss rate of 4.1%. Other studies estimated wrist fractures to be about 18% of the fractures seen in ED in adults [27, 46] and of 25% of the fractures seen in children[46].

There has been an increase in the incidence of wrist fractures in all age groups with no clear reasons. Some put this increase down to lifestyle influence, osteoporosis, child-obesity and sports-related activities [92].
Given the high incidence of wrist fractures and the rates at which they are mis-reported, there is a clear need for the research in this project which aims to help clinicians identify fractures.

2.3 The Wrist Joint

2.3.1 Wrist Structure

The wrist joint (shown in Figure 2.2) is one of the most complex joints in the body, and comprises an articulation between: (1) the proximal row of the carpal bones (except the pisiform) at the distal side of the wrist joint, and (2) both the distal radius and the articular disk (fibrocartilaginous ligament) at the proximal side of the wrist joint. The distal ulna is not part of the wrist joint as it does not articulate with the carpal bones whereas the articular disk which lies over the distal ulna articulates instead. The distal radius articulates with the distal ulna at the distal radioulnar joint (DRUJ) and articulates with the proximal row of carpals with the radiocarpal joint (RCJ).
Like all other synovial joints, the wrist joint is covered by a layer of synovium responsible for secretion of viscous fluid to provide strength and lubrication and allows for movement along two axes, which means flexion, extension, adduction and abduction can all occur at the wrist joint.
Fractures of the carpal bones are usually referred to as *carpal fractures*, while fractures of distal radius are referred to as *wrist fractures*. A wrist fracture is described as

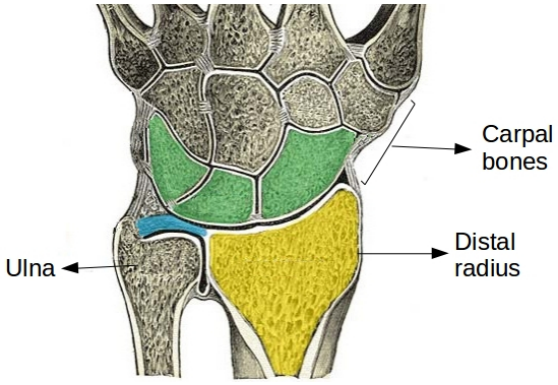


Figure 2.2: The wrist joint lies between (1) the proximal row of the carpal bones (in green), and (2) both the distal radius (in yellow) and the articular disk (in blue) [61].

intraarticular if it involves the radiocarpal joint, distal radioulnar joint, or both, otherwise it is *extraarticular*. It can be also described in terms of the distal component displacement in relation to the proximal component (i.e. volarly-displaced, dorsally-displaced, or non-displaced).

Fractures, in general, can be described as *simple* or *comminuted* (i.e. consists of more than two fragments). Simple fractures can be either *transverse* (i.e. bone breaks at a right angle to the bone's axis) or *oblique* (i.e. bone breaks diagonally).

2.3.2 Wrist Radiographic Views and Measurements

There are two standard wrist radiographic views: Posterioanterior (PA) view and Lateral (LAT) view (see Figure 2.3).

On these standard views there are three main measurements (shown in Figure 2.4) are usually taken to quantify wrist deformities [46]. These measurements are defined as follows [46]:

1. Radial length (height) It is measured on the posteroanterior (PA) radiograph as the vertical distance between a line perpendicular to the long axis of the radius and passing through the distal tip of the sigmoid notch (point C in Figure 2.4)

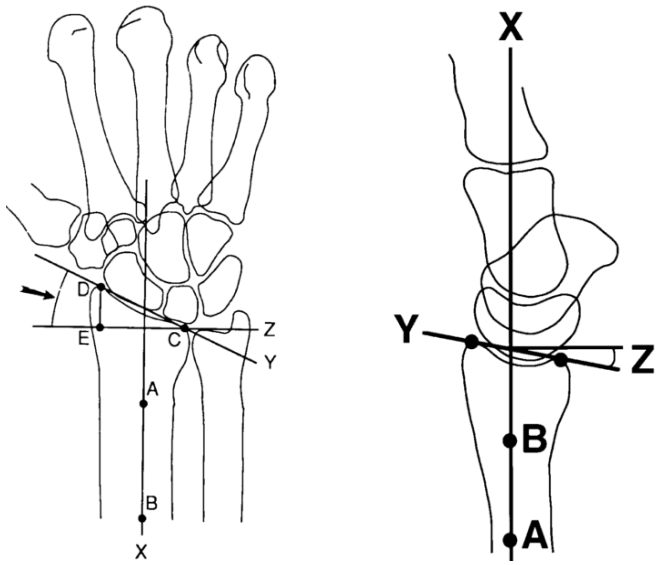
(a) Posterioanterior (PA) (b) Lateral (LAT)

Figure 2.3: The Two Standard X-ray Wrist Views.

and the distal tip of the radial styloid (point D in Figure 2.4). Its normal measure is between 10 - 13 mm.

2. Radial inclination (angle) It is measured on the PA radiograph as the angle between two lines: one passing through both the distal tip of the sigmoid notch and the distal tip of the radial styloid, the other is perpendicular to the long axis of the radius and passing through the distal tip of the sigmoid notch. Normally it measures between $21 - 25^\circ$.

3. Volar (palmar) tilt It is measured on the lateral (LAT) radiograph as the angle between two lines: the first is passing through the most distal points of the posterior and anterior rims of the distal articular surface of the radius, the second is



Schematic PA View [46] Schematic Lateral View [46]

Figure 2.4: Wrist Radiographic Measurements. In PA View: Radial length is the shortest distance between points D and E, Radial inclination in the angle DCE. In Lateral View: Volar tilt is the angle Z

perpendicular to the long axis of the radius. Both lines crossing the perpendicular at the same point. Its normal measure averages 11° and with range between $2-20^\circ$.

In clinical practice, eponyms are commonly used to describe common patterns wrist fractures take [46] such as Colles fractures, Smith fractures, Barton fractures, and Hutchinson fractures.

2.4 Fracture Detection

Early work on fracture detection used non-visual techniques: analysing mechanical vibration by a neural network model [57], analysing acoustic waves traveling along the bone [97], or by measuring electrical conductivity [107]. There are few papers published related on computer-aided radiographic fracture detection. This section summarises the literature for detection of fractures in different bone regions.

2.4.1 Hip and Wrist Fractures

The first published work on detecting fractures in X-ray images was that by Tian *et al.* [116] for femur fractures. The method consists of three steps. The first involves extraction of the femur contour, the second involves the measurement of the neck-shaft angle NSA (see Figure 2.5), and the third is a classification of femur fracture based on measured angle. Extraction of the femur contour is not described in detail, but is said to be performed using Canny edge detection, Hough Transform, and active snake contours [62] with the Gradient Vector Flow (GVF) method [127]. To measure

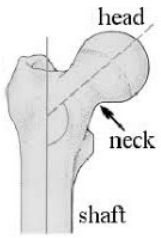


Figure 2.5: Femoral neck-shaft angle (NSA)[116]

the angle the axes of neck and shaft were required. The shaft axis was recovered by drawing normals (called level lines) to the almost-parallel shaft contour lines from one side of the shaft to the opposite side. The shaft axis passes through the midpoints of the shaft level lines. Shaft level lines were differentiated from other level lines by the fact that they were shorter and located in the lower part of the image (see Figure

2.6). To estimate the neck axis, they first clustered level lines within the head and neck areas into bundles according to their lengths and proximity of their midpoints. The mean direction in the cluster containing the largest number of long level lines gave an initial estimate of the neck axis, which was fed to an optimisation algorithm computing the neck axis as the best-fitting axis of symmetry of the head-neck contour. Finally, a threshold on the angle value was learnt from the training set and used for classification. The method was able only to detect severe fractures that cause significant change of neck-shaft angle but not those that cause local disruptions of the texture without displacing or rotating the head.

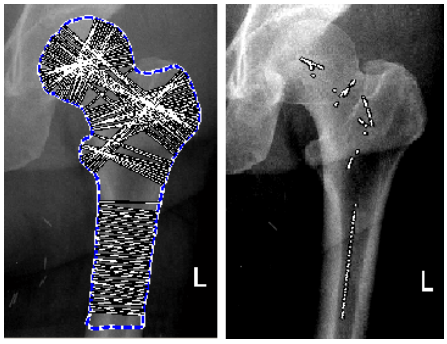


Figure 2.6: Level lines found in the femur contour (left). Mid-points of the level lines at the shaft are oriented along the shaft axis (right). [116]

To overcome this shortcoming, a complementary method performing texture analysis of the femoral trabecular pattern was proposed by Yap *et al.* [128]. The method also consisted of three stages, the first of which extracted the femur contour using active shape model [25] and active appearance models [26], the second analysed trabecular texture by extracting an orientation grid for the femur's upper extremity, and the third performed classification using both a Bayesian classifier and a Support Vector Machine (SVM). The extracted orientation grid (examples shown in Figure 2.7) had a fixed number of sampling locations. At each of them the orientation of the texture was calculated by a set of eight Gabor filters and was set to be the orientation of the Gabor filter whose response was the largest. The resulting vector map was converted

to a scalar one, for sake of classification convenience, by calculating the difference between it and the mean orientations of healthy training examples. The femoral neck was classified as fractured if either of the two classifiers or the NSA [116] method classified it as fractured. This combination of methods yielded a fracture detection rate of 84.6% compared to 61.5% for the NSA method [116] alone.

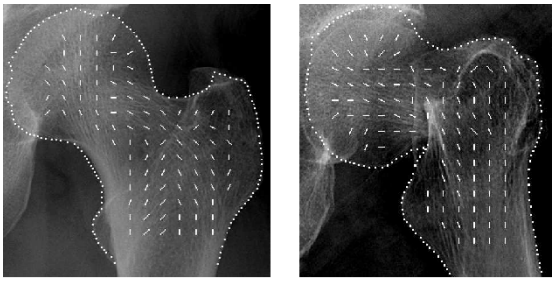


Figure 2.7: Orientation maps of healthy femur (left) and fractured femur (right). The short lines indicate trabecular orientations. [128]

Lim *et al.* [73] further improved the work in [128] by extending the feature extraction stage to include extracting texture feature maps of Markov Random Field (MRF) [30] and intensity gradients (IG) in addition to Gabor maps and the neck-shaft angle from [116]. The segmentation stage remained the same as in [128]. Six different classifiers were trained: neck-shaft angle with thresholding, Gabor maps with Bayesian classifier and SVM, intensity gradient maps with Bayesian classifier and SVM, and Markov Random Field texture with SVM. They showed that individual classifiers have low fracture detection rate but each of them can detect some fractures that are missed by the other classifiers so they can be used to complement each other. The bone was classified as fractured if any two of the six classification methods were positive. This produced an improved femur fracture detection rate of 92.2%, with a false positive rate of 1% on testing set containing 13 fractured examples out of 108 examples. The method was preliminarily tested on PA radiographs of the wrist (only 23 fractured examples in test set of total 74 examples). The method produced a fracture detection rate of 82.6% with a false positive rate of 17.6%. A further refinement of

these methods was undertaken by Lum *et al.* [79], who used same features with different probabilistic rules of combining classifiers [68] such as max, min, sum, product, majority vote and the simple *m*-of-*n*. They concluded that the OR rule (1-of-n) had the highest sensitivity and comparable accuracy.

Motivated by the success of learning feature representations using CNNs in many image processing and medical imaging analysis, instead of hand-crafting features, Kim *et al.* [66] showed that transfer learning from a deep convolutional neural network pre-trained on non-medical images can be applied to analyse X-rays. They re-trained the top layer (i.e. classifier layer) of Inception v3 network [112] to detect fractures in wrist lateral views from features previously-learned from non-radiological images (ImageNet [96]). 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 area under Receiver Operating Characteristic curve (AUC) of 95.4%. However, they excluded images where lateral projection was inconclusive for the presence or absence of fracture from both training and testing sets. This might be seen as a contradiction to the purpose of developing such systems (i.e. helping clinicians with difficult usually-missed fractures).

The current state-of-the-art for detecting hip fracture from frontal pelvis X-rays is also deep-learning-based. Gale *et al.* [43] used 172 layer-deep DenseNet [56] architecture optimising two loss functions: the first related to fracture detection (healthy vs fractured), and the second was more specific (intra-capsular fractured, extra-capsular fractured, or healthy). They collected a dataset of total 53,278 images from a teaching hospital over a decade, randomly split it into (training: 45,492 images, validation: 4,432 images, and test: 3,354 images containing 348 fractured), with no patient overlapping. The test set was labelled manually by a consultant radiologist using all of the available sources of information (i.e. the orthopaedic surgical unit records, and findings from the radiology report archive). They also trained other CNNs for tiding

up the dataset: a small CNN to filter out any non frontal pelvis X-rays, a regression-based CNN to localise the hip, and another CNN to filter out the hips with metal-works. They reported area under ROC curve of 0.994 which is claimed to be the highest level ever reported for automated diagnosis in any large-scale medical task, not just in radiology [43].

2.4.2 Diaphyseal Fractures

Injuries in bones like humerus, radius, ulna, femur, tibia, and fibula are usually referred as long-bone fractures. Each long-bone has three regions: proximal, distal (two extremities), and diaphyseal (shaft). This section summarises the work on detecting fractures on the middle part (diaphyseal) of long bones.

Jia and Jiang [59] worked on fractures of the arm shaft by developing a geodesic active contour segmentation model with a shape prior as a global constraint. The model evolved toward the desired shape by deforming the curve until the mutual information between the curve and the shape prior was maximized. This segmentation step was followed by a bone alignment calculation step for which no details were provided within the text. They reported that they had tested their algorithm on "more than 10 cases", and that their results showed that their algorithm is "robust and accurate". However, they have not reported any quantitative results to support their claim.

Donnelley *et al.* [37] proposed a computer aided diagnosis system for detecting the mid-shaft fractures of long bones. First the middle part (diaphysis) of the long bone was semi-automatically segmented by scale-space approach with the Hough transform to detect edges (straight lines). Fracture detection was done by gradient analysis by assuming large gradients occurring at angles not orthogonal to the bone edges are indicative of fractures. This assumption caused a high false positive rate (98%) as the algorithm detected bone overlap and biological phenomena not related to fractures as

fractures. They also reported that 83% of the diaphysis segmentation boundaries in the test set were correctly identified, and 83% of the fractures within those segmented regions were also detected correctly.

Chai *et al.* [21] also worked on the shaft of long bones. They first performed different image preprocessing steps: binary conversion by thresholding, Laplacian edge detection, and suppressing isolated noise by the median filter before using K-means clustering algorithm to separate (segment) the femur shaft area from non-shaft area ($K = 2$). For fracture detection four grey level co-occurrence matrices (GLCMs) [54], each sampled in a different direction, were calculated. A GLCM, in short, is a matrix whose element $[i, j]$ contains the frequency of a pixel with intensity value i adjacent to a pixel with intensity value j and normalised so that each element represents a probability of intensity i is found adjacent to intensity j. A GLCM is widely used as a statistical way to express texture structure. They [21] calculated four statistics per each GLCM, (energy, contrast, correlation, and homogeneity), and then averaged them to provide a total number of four statistics. A threshold was set on the values of these statistics to classify the area as fractured or not. They tested on 30 images, half of which fractured and reported accuracy of 86.67%.

Fuadah *et al.* [41] proposed a system to detect diaphyseal fractures by first applying image enhancement, edge detection, and filtering to remove noise and extract clean edges, then finding the maximum value of the difference distance from the right and left border margin of object for each scan line. The maximum value of scan line is then used as a threshold to classify normal cases or fracture cases. They tested on 70 images (40 of which were fractured) and reported specificity of 100% and sensitivity of 90%.

Bandyopadhyay *et al.* [6–8] developed an entropy-based segmentation technique [10] and integrated it with their adaptive thresholding approach [9] to improve the

segmentation quality before calculating geometrical indexes (concavity [8] and relative concavity [6]) along the contour line and monitor their changes to detect/classify fractures in the shaft of the long bone.

2.4.3 Pelvis Fractures

The only work in the literature we could find on detecting pelvis fractures from radiographs is that of Smith *et al.* [108]. They used an Active Shape model [25] to segment the pelvic ring and pubis. The resulting segmentation was used to measure the horizontal and vertical displacements between the left and right pubis as quantitative measurements mimicking the approach used by radiologists. Furthermore, the segmented pelvic ring is divided into overlapping windows based on the shape model landmarks. For a window 2-D Discrete Wavelet Transform (DWT) was performed for line detection. The choice of the right DWT coefficient that best detected the bone boundary was determined by the location of the window around the pelvis ring. The location was determined by landmark number. The last step was edge tracing so in case of fracture a window will contain multiple boundaries depending on the types and number of fractures. Their test set was very limited (10 radiographs, 3 of which contain fractured rings). All fractures were detected and the overall accuracy was 86.7%.

2.4.4 Fractures From Multiple Anatomical Regions

All the previous work on detecting fractures in X-ray images addressed one anatomical region at a time. Cao *et al.* [18] developed the first learning method with the aim to identify different kinds of fractures over different anatomical parts at the same time. The motivation behind the work was to develop a technique to extract a feature representation, for a patch, that can capture different types of fractures and then

fuse the different feature types in a way that prevents the classifier from being biased to the feature type with the highest dimensionality, which is the case when different feature-type vectors of different lengths are simply concatenated. The method uses a multi-layer classifier. Each layer contains a number of random forests. The classifier in the first layer calculates different feature representations for the training patches. Each representation is a vector of one feature type. They used three feature types: Schmid texture features [100] to capture orientation-invariant textures representing comminuted fractures, Gabor texture features to mimic the functionality of the mammal visual cortex, and forward Contextual-Intensity (CI) as an additional descriptor of edge and texture. Each random forest in the first layer was trained on one feature type. Each decision tree would provide a probability distribution of the two classes: fracture, healthy. For a training patch, concatenating such distributions provides the feature fusion representation for the sample, which was used to repeatedly train new layers of random forests in the same fashion. The random forest of the last layer gives the probability distribution for a patch to be in either of two classes. A score map for an image is composed, by scanning patches along it and feeding them to the stacked random forests, and passed to Subwindow Search algorithm [69] to obtain fracture bounding boxes. They achieved sensitivity $\approx$81%, and precision $\approx$25% from the top seven detected bounding boxes.

Olczak *et al.* [89] re-trained five common deep networks from Caffe library [60] on dataset of 256,000 wrist, hand, and ankle radiographs, of which 56% of the images contained fractures. The dataset was split (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 networks' inputs varied between squares of 224 pixels width to 227 pixels. 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 [96] and then their top layers (i.e. classifier layer) were replaced with fully connected layers suitable for each task. The best performing network (VGG 16 [132]) achieved a fracture detection accuracy of 83% without reporting a false positive rate. The model deals with various views independently but it does not combine them for a decision. When comparing the network with two senior orthopedic surgeons on 400 images at the same resolution as the network, they found that the network performed on par with the humans (kappa = 0.76) with accuracy of 69%. The two human observers agreed with each other with kappa = 0.8. Another related work [94] used a DenseNet model (169 trainable layers) for abnormality detection from raw radiographs. Images were labeled as normal or abnormal, where abnormal did not always mean 'fractured'-it sometimes meant there was 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 rather unclear as what is considered abnormal for a certain group can be seen as normal for another group and vice versa.

All the literature on computer-aided radiographic fracture detection reviewed in this section is summarised in Table 5.11.

Table 2.1: Summary of literature on computer-aided radiographic fracture detection.

Bone	Reference	Dataset in number of radiographs: Total number (radiographs containing fractures F)	Segmentation Technique	Detection Technique	Performance
Hip	Tian *et al.* [116]	Training: 126 (19 F) Testing: 320 (33 F)	Canny edge detection Hough Transform Active Snake [62] Gradient Vector Flow [127]	Thresholding	Accuracy: 92.5%
Hip	Yap *et al.* [128]	Training: 324 (39 F) Testing: 108 (13 F)	Active Shape Model [25] Active Appearance Models [26]	Thresholding [116], SVM & Bayesian classifier with Gabor maps.	Accuracy: 84.6%
Hip & wrist	Lim *et al.* [73]	For hip: Training: 324 (39 F) Testing: 108 (13 F) For wrist: Training: 71 (21 F) Testing: 74 (23 F)	Active Shape Model [25] Active Appearance Models [26]	Thresholding [116]. SVM & Bayesian classifier with Gabor maps, intensity gradient maps, and Markov Random Field texture.	Hip Accuracy: 92.2% Wrist Accuracy: 82.6%
Wrist	Kim *et al.* [66]	Training: 1,389 (695 F) Validation: 10% of training set Augmentation used. Testing: 100 (50 F)	-	Re-training the top layer of Inception v3 network [112]	AUC: 0.954
Hip	Gale *et al.* [43]	Training: 45,492 Validation:4,432 Testing: 3,354 (348 F)	CNN-based localisation	DenseNet [56]	AUC: 0.994
Arm shaft	Jia and Jiang [59]	-	Geodesic Active Contour	Bone alignment calculation	-
Long bone shaft	Donnelley *et al.* [37]	Training: 6 (6 F) Testing: 44 (38 F)	Semi-auto scale-space approach with Hough Transform	Gradient analysis	Accuracy: 83%
Long bone shaft	Chai *et al.* [21]	Training: - Testing: 30 (15 F)	Laplacian edge detection K-means	Thresholding on a statistic calculated from GLCMs [54]	Accuracy:86.67%
Long bone shaft	Fuadah *et al.* [41]	Training: - Testing: 70 (40 F)	Edge detection Noise filtering	Thresholding on the maximum distance between the right and left border lines	Specificity: 100% Sensitivity: 90%
Long bone shaft	Bandyopadhyay *et al.* [6–8]	Training: - Testing: 30 (21 F)	Entropy-based [10] with adaptive thresholding [9]	Monitoring changes in geometrical indexes (i.e. concavity [8] and relative concavity [6])	Accuracy: 100%
Pelvis	Smith *et al.* [108]	Training: - Testing: 10 (3 F)	Active Shape Model [25]	Line detection with Discrete Wavelet Transform followed by edge tracing	Accuracy: 86.7%
Multiple bones	Cao *et al.* [18]	145 (145 F) split into: 80% training 20% testing	-	Patch-wise fracture detection using Stacked Random Forests	Sensitivity: 81% Precision: 25%
Multiple bones	Olczak *et al.* [89]	256,000 (143,360 F) split into: 70% training 20% validation 10% testing	-	Re-training the top layer of VGG 16 network[132]	Accuracy: 83%
Multiple bones	Rajpurkar *et al.* [94]	Training :37,111 Validation: 3,225 Testing: 559 39% of the dataset contain fractures	-	DenseNet	Cohens kappa statistic: 0.749

2.5 Fracture Classification

Fracture Classification is the first step in fracture treatment and it involves determining the fracture location, its characteristics, and level. Medical literature has different methods to classify fractures depending on either the morphology of the fracture, degree of fracturing, or the severity of the damage to the soft tissue. The most widely used method is Muller AO classification based on the shape of fracture. Developing tools for fracture classification is useful because it would serve as a "second opinion" providing analytic justifications for diagnosis taking into consideration it is a rather difficult process for a physician to remember all 117 fracture types [11]. Automated image analysis literature has very limited work related to fracture classification. The first work [42] was an initial model considering the classification process as a tree traversal problem in which a series of questions are answered regarding the fracture to finally land at a leaf node. Implementing the whole system means implementing appropriate algorithms answering the questions at each node. The model implementation was left as an open research problem.

Wei *et al.* [124, 125] proposed methods to automatically classify the fracture type [125] and interpret the fracture site in the femur [124] (i.e. whether proximal, middle, or distal) by converting the problem to a shape detection problem since the three regions have different shapes. Proximal and distal regions have bumps in different directions while the middle region has a more uniform width. The algorithm segmented the bone to obtain the edges of different objects and then filled the areas inside each segmented objects. To identify each object a thinning operation was performed on the image into a single pixel-width. The Hough Transform was used to find the number of potential lines in the thinned image with their angles. Angles were used to classify fractures and their sites. It is hard to comment on their techniques as there were only two examples in their results.

A more comprehensive classification work by Bayram *et al.* [11] proposed an integrated system for classifying diaphyseal femur fractures automatically in nine different classes according to the Muller AO classification system [84]. For segmenting bone fragments Niblack thresholding was used for its ability to keep the information related to the fracture region compared to other techniques [11]. However, this was found to produce a lot of noise. For this reason the study proposed a method called *SVM-based sensitive noise remover* in which 11 different features are presented to SVM classifier in order to differentiate noise and bone segments. The extracted features ranged from basic (e.g. particle area) to more complex ones (e.g. fullness ratio, roughness). They reported a differentiation success of 93.7% between bone segments and noise generated by segmentation process. Finally the fracture classification was done by a SVM classifier with 8 designed features (e.g. number of fragments, angle of fracture ends). They reported an accuracy of 90% for 196 radiographs of fractured bones in ten-fold cross validation experiments. To avoid the need to segment the bone and extract high-level features such as number of fragments which might be prone-to-errors Kazi *et al.* [63] proposed a CNN-based method to classify hip (i.e. proximal femur) fractures according to the Muller AO 6-class classification standard [84]. For the tasks of femur localization and classification they adapted the spatial transformer (ST) [58], which implies unsupervised learning of region of interest and a classifier, both trained end-to-end. Their original 1221-image clinical dataset was class-unbalanced (some classes contained as few as 15 images while others as many as 195 images) so they used image augmentation leading to 195 images per class. They trained on 900 images and tested on a separate set of 270 images. They reported average accuracy of 89% and 68% for precision and recall.

Fracture image retrieval is a closely-related topic to fracture classification as it looks at the similarity between cases and allows access to visually-similar previous cases

for consulting. Zhou *et al.* [133] presented a case-based retrieval algorithm for images with fractures. The algorithm combines multi-image queries to search an image dataset of 2690 cases. The cases in the dataset were represented by a bag of visual keywords and a local scale-invariant feature transform SIFT [78] descriptor. Retrieval was achieved by calculating the similarity of every image in the query case with every image in the dataset to find the set of most similar images and therefore cases.

2.6 Bone Segmentation

In order to build an automated system for assessing radiographs, one needs to find a way to detect and identify regions of interest. This task, in general, is challenging because of (1) the high variability in image quality due to different types of X-ray systems, (2) the variation in imaging positioning, and (3) the presence of non-anatomical objects in the radiographs such as tags, bracelets, and implants. Any segmentation algorithm should be able to handle the differences in resolution, sharpness, contrast, different orientations, and noise in order to be robust to all above mentioned variabilities.

Another challenge is the the natural shape and appearance variability of anatomical structures across the population as a result of: anatomical variations between individuals, differences in clinical variables such as age and gender, or most importantly variations caused by abnormalities such as diseases or fractures.

There is a wide range of literature on automating the segmentation of structures in radiographs ranging from thresholding [98], edge detection [17], template matching, atlas-based techniques, deformable models and recently deep learning. These methods can be seen as two different paradigms: Model-based methods, and Non-model-based methods. Model-based methods, as the name suggests, try to match the evidence in the image to what it is expected to be a legitimate instance of the

object according to a prior model. In contrary, non-model-based methods look for local structures,(i.e. edges, regions), to assembly in an object without the use of any prior model of shape. The non-model-based methods are more prone to failure for the lack of shape information whereas the model-based methods may be more prone to bias (e.g. a method with a strictly enforced shape constraint cannot properly fit any examples that fall outside the learned shape model, which is a particular problem in medical applications where the pathological examples are of more interest.)
Model-based methods can use rigid or deformable models of the object. Rigid models are fixed templates of the object shape and are matched by measuring the correlation and similarity between the template and the object in the image. Although this approach might work well with geometrical objects, it does not work for anatomical structures due to the natural variability mentioned earlier. For a model to work well, it has to capture the key characteristics of the object including expected in shape and appearance.
Different segmentation techniques were used in the literature to segment healthy/fractured bones (see Section 2.4) as a first step before detecting fractures. As a segmentation algorithm for our fully automated system we decided to use Statistical Shape Models (SSMs) [25] with their matching algorithm Constrained Local Model with Random Forest Regression Voting (RFCLM) [24]. Human anatomy has strong shape priors which led to the wide use of statistical shape models in medical image analysis. Statistical Shape Models are deformable models providing a way to incorporate object geometric information which were shown to be affected by fractures in Section 2.3.2. Fractures might be seen as random and irregular so that they can not be represented with statistical shape models. However the medical literature shows that there are patterns according to which a bone fractures (see Section 2.3.2). We adopted these patterns as variants of normal shape. Such statistical shape models will not only be useful for detecting obvious fractures but also for detecting more subtle

fractures. Fractures cause deformities that are quantified in radiographic assessments in terms of measurements of bone geometry (i.e. angles, lengths). Slight deformities might not be noticeable by eye. For this reason we do not only use shape models to segment the targeted bones, as Lim *et al.* [73] did, but also for capturing these deformities in shape parameters and use them as features for classifiers. Statistical shape models and their matching algorithms will be discussed in the next sections.

2.6.1 Statistical Shape Models (SSMs)

Shape is the geometric information invariant to translation, rotation, and scaling of an object [65]. Statistical Shape Models [25] are deformable models assuming all legitimate shapes of an object to be deformations from an average shape. These deformations are learned from the training set and used as some linear combinations of modes of variations to the mean shape.

SSMs can be built for objects where a correspondence across examples can be defined, which is the case for many anatomical structures. The shape of an object is captured by a series of model points along the object's contour. A model point (also known as feature point, landmark, or contour point) is chosen to adequately cover the morphology of the object and to be found consistently. The mean shape of the object class and its modes of variations are learned from the training dataset. The process of finding the object in a new image became an optimisation process for finding the best combinations of these modes that best describe the object in the image.

Principle Component Analysis (PCA) [1] provides the functionality needed to learn modes of variations and help formulate the process of model matching mathematically as follows:

Building Statistical Shape Models

Given a set of training images, each training image is manually annotated with n points which are object landmarks (such as high curvature points, T junction, or anatomical landmarks) and evenly-spaced points along the contour. A point i in an image is represented by (x_i, y_i) which results in a vector x of length $2n$ representing all points as an object.

$$\mathbf{x} = (x_1, ..., x_n, y_1, ..., y_n)^T \tag{2.1}$$

Shapes from all training images are aligned first with Generalised Procrustes Analysis (GPA) [49] to remove the variations that come from different scaling, rotation, and translation. GPA applies similarity transformations to all shape vectors so that the distance between each shape and the mean shape ($|T_\theta(\bar{\mathbf{x}}) - \mathbf{x})|^2$) is minimised.
A shape instance $\mathbf{x}$ is represented as:

$$\mathbf{x} \approx T_\theta(\bar{\mathbf{x}} + \mathbf{Pb} : \theta) \tag{2.2}$$

where $\bar{\mathbf{x}}$ is the mean shape in the reference frame, $\mathbf{P}$ is the set of the orthogonal eigenvectors corresponding to the t highest eigenvalues λ_j where $j = 1, .., t$ of the covariance matrix of the training data, $\mathbf{b}$ is the vector of shape parameters and $T(. : \theta)$ applies a similarity transformation (i.e. scaling, rotation, and translation) with parameters θ between the reference frame and the image frame. The number of the used eigenvectors t is chosen to represent some proportion p (e.g. p= 0.95) of the total variance and calculated as:

$$\left(\frac{\sum_{j=1}^{t} \lambda_j}{\sum_{j=1}^{2*n} \lambda_j} \right) \geq p \tag{2.3}$$

The shape vector $\mathbf{b}$ can be calculated from $\mathbf{x}$ by applying:

$$\mathbf{b} = \mathbf{P}^T (T_\theta^{-1}(\mathbf{x}) - \bar{\mathbf{x}}) \tag{2.4}$$

The variance of each shape parameter, b_j, is given by the eigenvalue λ_j. This model makes it possible to generate legitimate instances of the object and also to check whether an object instance is a legitimate one. A shape $\mathbf{x}$ may be called plausible if the shape parameters, $\mathbf{b}$, are within some squared Mahalanobis distance M_t chosen from the χ^2 distribution.

$$\left(\sum_{j=1}^{t} \frac{b_j^2}{\lambda_j} \right) \leq M_t \tag{2.5}$$

Statistical shape models (SSMs) describe the shape of an object with a limited number of parameters providing a way to study a class of object and eliminating the need to carry geometric measurements on its instances. However, manual annotation is time-consuming and prone to inconsistency and subjectivity of annotators. Some techniques try to predict the parameters of the shape model without the need for annotating new images [34, 131]. Other techniques studied automating the annotation process and they will be addressed briefly in the next section.

Automating Annotations

A body of research has studied methods to automate the process of annotating objects in images, and methods can be divided into two categories: point-detection-based methods and registration-based methods. Point-detection-based methods estimate the position of each feature point and regularize the estimates with some shape constraint. This means a point might be shifted from its best texture-wise position and placed in a less good position in order to make an acceptable overall shape in the image. Registration-based methods [111] align the images into same reference frame before annotating them using a single annotated reference image. The alignment process can be guided by few manually-annotated points (semi-automated). There are other methods [53, 130] combined the two approaches of point detection and image registration.

Different point-detection-based algorithms have been developed to match statistical shape models to new images such as the Active Shape Model (ASM) [25], the Active Appearance Model (AAM) [26], the Constrained Local Models (CLM) [29], and its robust variant Random Forest Regression-Voting Constrained Local Model (RFCLM) [24]. All these algorithms are semi-automatic and in need of a good landmark initialisation to start the search. Some initiate the search by predicting the positions of landmarks directly from the whole image while others use object detection methods. In Section 2.7 we will provide details on the object detection methods used to fully automate statistical shape matching algorithms. Before that, we will introduce CLM as a statistical shape model matching framework on top of which the robust RFCLM is built.

2.6.2 Constrained Local Model (CLM)

The Constrained Local Model (CLM) [29] is an algorithm for fitting the points of a statistical shape model to a new image. The CLM requires a local texture model built for every feature point independently. A local texture model should be able to generate a response image over a region indicating the cost of having the feature point at each position in the region. During search, the CLM finds the shape and pose parameters $\rho = \{\mathbf{b}, \theta, \mathbf{r}\}$ that lead to the lowest cost subject to the constraints of the shape model by minimizing:

$$\begin{aligned} Q(\rho) &= \sum_{j=1}^{n} C_j(T_\theta(\bar{\mathbf{x}}_j + \mathbf{P}_j\mathbf{b} + \mathbf{r}_j)) \\ s.t.\ \mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} &\leqslant M_t \text{ and } \left|\mathbf{r}_j\right| < r_t \end{aligned} \tag{2.6}$$

where: C_j is a cost image for the feature point j, $\mathbf{S}_b$ is the covariance matrix of model parameters $\mathbf{b}$, M_t is a threshold on the Mahalanobis distance which is set using the

cumulative distribution function (CDF) of the χ^2 distribution so that 99% of the training examples lay within it, and r_t is a threshold on the residuals to allow some small deviations from the model.

Early CLM frameworks [29] used normalised correlation as a texture model. Later works in [24] incorporated regression-based voting into the CLM framework resulting in Random Forest Regression-Voting Constrained Local Model (RFCLM) and achieved excellent performance on a range of facial and medical datasets. Lindner *et al.* [75] showed that RFCLM outperformed all other matching algorithms with a mean point-to-curve error of 0.9mm for 99% of the images on the task of detecting the outlines of proximal femur from 839 radiographs. The accuracy and robustness of the RFCLM algorithm has lead to its use in many similar problems, such as segmenting knees [82, 114], vertebrae [15], and skull [121]. RFCLM was used throughout the project as it is the state of the art algorithm for matching SSMs, which can deal with the geometric information in the problem. RFCLM [24] and its local texture models (Random Forest Regression-Voting RFRV) will be explained in detail in Chapter 3.

2.7 Object Detection

For fully automated annotation statistical shape models need to be proceeded with some kind of object detection mechanism in order to provide them with initial estimates of position, orientation and scale of the object's bounding box. This can be done with building a template of the object and scanning the whole search image at a range of scales and orientations searching for the most similar patch to the template. This sliding window approach has been adapted to use machine learning techniques resulting in classification-based [32, 40, 44, 119], regression-based [75], and hybrid [28] object detection methods. In general, machine-learning-based object detection

methods outperform template matching as they are more able to tackle high intra-class variations, occlusions, and presence of noise. Compared to classification-based methods, regression-based methods (which predict the pose of the target given nearby image information) have the advantage of: (1) not requiring pre-determined selection of positive and negative examples for the training of classifiers which is not always easy, (2) integrating evidence from various image regions not necessarily only the patches centered at the target point, and (3) having the ability to perform significant subsampling without compromising accuracy which means less computations [74]. Moreover due to consistency of skeletal anatomy across individuals, any part of the image can predict the required area removing the need to label patches and perform classification. For these reasons we chose to use a regression-based method called Random Forest Regression Voting (RFRV) first used by Cootes *et al.* [24] for both object detection and CLM-based contour extraction (RFCLM). RFRV trains class-independent regression forests to cast votes from all image structures for finding the position, orientation, and scale of the object's bounding box. The use of random forest regression voting for detecting the object and for extracting its contour yielded a robust and fully automatic segmentation system for many anatomical structures in radiographs [15, 74–76, 82, 114]. We will describe RFRV in detail in Chapter 3 as the chosen object detection method throughout the project.

Recently, deep learning approaches claim state-of-the-art landmark localisation performance. Payer *et al.* [90] detected multiple landmarks by regressing a heatmap for all landmarks simultaneously. A heatmap is a new image with a Gaussian blob around each predicted landmark position. They trained a novel CNN (named Spatial Configuration Net) end-to-end and showed it was able to learn local features and imposed constraints on the spatial configuration of landmarks on experiments to detect 37 landmarks in hand X-rays and 28 in hand MRI images. Their technique achieved the same accuracy of RFCLM localised by RFRV on the 2D dataset. Sofka

et al. [110] used a fully convolutional network (FCN) to regress point locations. The regressed locations are mapped at the last convolutional layer into a location using a new center-of-mass (CoM) layer, which computes the mean position of the prediction. Spatial context is modeled with Convolutional Long-Short Term Memory (CLSTM) cells. Unlike direct heatmap regression, this approach could predict subpixel values and its objective function could penalise measurement length differences from the ground truth. Zhang *et al.* [129] proposed a two-stage deep CNN model. They first used millions of image patches to train a patch-based CNN regression model to predict 3D displacements to the target landmarks. The same architecture and network weights were used after adding extra layers to predict the coordinates of multiple landmarks jointly, with an entire image as input and the landmark coordinates as output. They predicted 1200 landmarks in 3D MRI brain scans and 7 landmarks from 3D tomography images of prostates.

2.8 Summary

Fractures of the wrist have high incidence rate and they are usually identified in Emergency Departments (EDs) by doctors examining lateral (LAT) and posterioanterior (PA) radiographs. Unfortunately missing such fractures is one of the most common diagnostic errors in EDs which constitute a clear need for this research. Few automated methods have been developed to detect and classify fractures in different bone regions. One of the first issues to tackle when designing these methods is to accurately extract bone contours in radiographs; the work of Lindner *et al.* [75] on segmenting proximal hips from radiographs has shown accuracy and robustness of the RFCLM algorithm and led to its use in many similar problems. In the remainder of the thesis we use: (1) RFCLM as our segmentation algorithm of choice, (2) shape and texture features found in the literature to detect wrist fractures. Data and methods are introduced in the next chapter.

Chapter 3

Data and Methods

The chapter is split into four parts describing (i) the data used throughout the project; (ii) object segmentation, to find the wrist in the image; (iii) extracting features to describe the object's shape and texture; (iv) and classification methods. The choice of the methods was following the key analyses from the literature in Chapter 2.

3.1 Data

To analyse fracture features and evaluate the accuracy of our methods, we collected a clinical wrist dataset containing 1010 pairs of wrist radiographs (i.e. PA, and LAT) for 1010 adult patients (half of whom had fractures). No other matching of fracture/non-fracture (e.g. sex, age) was done. The clinical images vary in resolution and in aspect ratios. None of the images contain any plaster casts or metalware in order to ensure the detection is targeting signs of fractures not signs of hardware. Radiographs for 787 patients (378 of whom had fractures) were gathered from two local emergency departments (EDs), revised and anonymised by a clinician while the rest were gathered from the MURA dataset [94] with fractures as abnormality. MURA is a dataset of clinical musculoskeletal radiographs containing 40,561 images for different body

parts (elbow, fingers, forearm, hand, humerus, shoulder, and wrist) labeled as either normal or abnormal. Abnormalities in MURA are of various types ranging from fractures, hardware, degenerative joint diseases, and other miscellaneous abnormalities, including lesions and subluxations. Table 6.1 shows the distribution of the collected dataset.

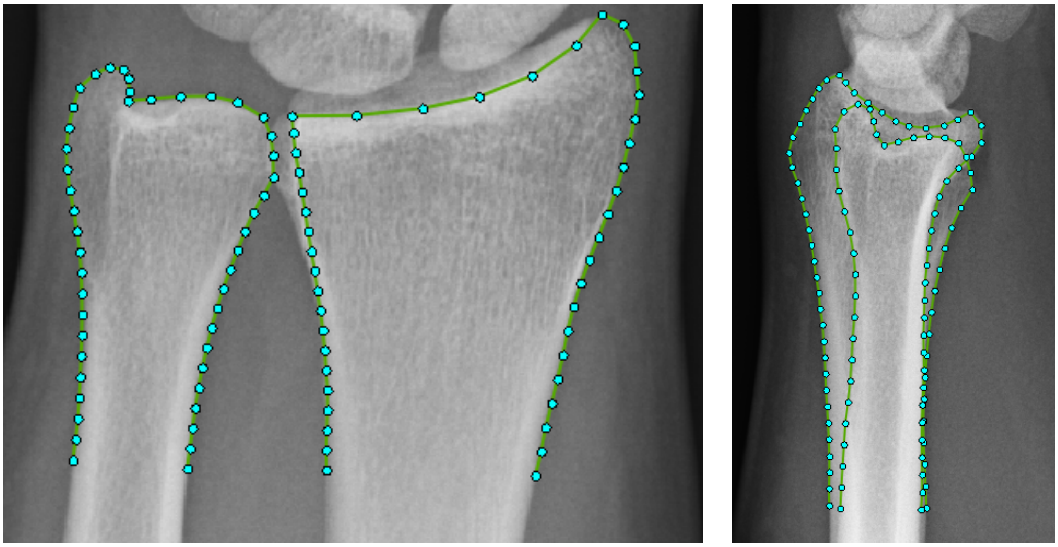
Table 3.1: Different sources of the dataset used for fracture detection with their sizes in number of adult patients.

Source	Normal	Fractured	Total
ED 1	211	193	404
ED 2	198	185	383
MURA	96	127	223
Σ	**505**	**505**	**1010**

3.2 Image Annotation

A sample of radiographs were studied to understand the geometry of the bones (i.e. radius, ulna) and the variations in shape and orientation of the wrist. As a result of this investigation a set of points were placed to describe the key shape characteristics of the radius and ulna such as the corner of the two bones, the sigmoid notch, the ulnar styloid process, and the radial styloid process. The sample dataset (containing 50 adult patients, 15 of which with fractures) was used to test different annotations. The annotations were optimised several times by trying different landmarks and different numbers of points. Building models from one half of the sample dataset and testing on the other half and vice versa. The evaluation metric was the point-to-curve error. Annotating the dislocated fractured bones was done by trying to contain the bone as a whole and not exactly following the contour. The annotation process was

done in two rounds. In the first round the set of points describing the key characteristics, such as corners and minimal points on curves, was placed. In the second round curves between these points were connected with a set of equally-spaced points. The final set of points (see Figure 3.1), on average, took 7 minutes to annotate by hand on each image and contains 93 points for PA view and 112 points for LAT view. All annotations were performed on the left wrist. All images containing right wrists were reflected. The manually annotated points were used for training segmentation models and also as ground-truth for testing.



(a) PA view (b) LAT view

Figure 3.1: Wrist Annotation with curves.

3.3 Wrist Detection and Segmentation

In this project the methods used for wrist detection and segmentation were based on an RFCLM [24]. This was chosen because of the high accuracy in analysing similar 2D bone shapes [15] [76] [75] [74] [114] [82]. The fully-automated system comprised a *global search* detecting the bones and a *local search* segmenting the bone

contours. The global search is performed by Random Forest Regression-Voting (RFRV) while the local search is performed by Random Forest Regression-Voting Constrained Local Model (RFCLM). RFRV and RFCLM are described in detail below before demonstrating how the two techniques were combined to build a fully-automated segmentation system.

3.3.1 Random Forest Regression-Voting RFRV

For a model point $\mathbf{x}_j$ a random forest [14] regressor F_j is trained to estimate the relative position of the point to an image patch. Training patches (see Fig. 3.2) are sampled at many random displacements between [-d_{max}, +d_{max}] in x and y from the true position of the point. Random perturbations in scale and orientation are also considered to compensate for inaccuracies in initial point estimate during matching. Haar features [119] $\mathbf{f}_i$ extracted from each training patch i with its corresponding displacement vector $\mathbf{d}_i$ are used to build a random forest regressor whose decision trees are trained on different bootstrap samples of training patches. When training a tree the aim is to increase the compactness of samples reaching the branches. So the set of pairs $\{(\mathbf{f}_i,\mathbf{d}_i)\}$ are split at each tree node by selecting a threshold value t on a feature f so that the split entropy G_T is minimised. The split entropy is defined as:

$$\begin{aligned} G_T(t) &= G(\{\mathbf{d}_i : f_i < t\}) + G(\{\mathbf{d}_i : f_i >= t\}) \\ and\ G(\{\mathbf{d}_i\}) &= Nlog|\mathbf{\Sigma}| \end{aligned} \tag{3.1}$$

where N is number of displacements $\{\mathbf{d}_i\}$ ended in the branch, and $\mathbf{\Sigma}$ is their covariance matrix. The splitting continues until reaching maximum tree depth or a minimum number of examples per node. Each tree leaf stores the mean displacement and the standard deviation of all the training examples which landed at that leaf node. Algorithm 1 demonstrates the training process.

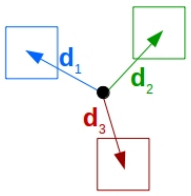


Figure 3.2: Patches sampled at random displacements $\{\mathbf{d}_i\}$.

Data: Training pairs T = $\{(\mathbf{f}_i,\mathbf{d}_i)\}$ of features and displacements representing patches sampled around the feature point $\mathbf{x}_j$ from different training images, minimum number of examples n_{min} to split a node.
Result: Random Forest regressor F_j. Each node contains a feature f, a threshold t, and two child nodes: LeftChild, RightChild.

```
for Each tree in Forest F_j do
    Sample a bootstrap T' from T
    BuildNode(T', rootNode)
end
Function BuildNode(Training pairs S, node):
    if |S| < n_min then
        Store mean displacement (dx,dy) and standard deviations: σ_x,σ_y of the
         examples reached this leaf.
        return
    end
    Choose a random subset of the features
    Choose feature f and threshold t for which the split entropy is minimum
     and store in node
    BuildNode ((S | f > t), LeftChild)
    BuildNode((S | f ≤ t), RightChild)
End Function
```

Algorithm 1: Training RF Regressor F_j to predict the displacement to point $\mathbf{x}_j$.

To search a new image for point $\mathbf{x}_j$ the regressor F_j scans the region around the current estimate of the point position and each tree in the forest votes for the point position. Different voting styles were explored in [74] and the most accurate and cost efficient style was found to be a single vote per tree ($w = 1$) at the mean displacement although weighted voting ($w = \frac{1}{\sqrt{\sigma_x \sigma_y}}$) was reported to perform equally well.

The votes from different trees are accumulated in a response image $V_j()$ where the most likely position of the feature point $\mathbf{x}_j$ has the highest number of votes (or lowest cost in a cost image $C_j(x,y) = -V_j(x,y)$). Algorithm 2 demonstrates the process of constructing the response image V_j for point $\mathbf{x}_j$.

Data: Region of interest ROI around the current estimate of $\mathbf{x}_j$;
RF Regressor F_j
Result: Response Image V_j
Set Response Image V_j to zeros;
for *Each x in ROI* **do**
  for *Each y in ROI* **do**
    Extract Haar features **f** for the patch centered on (x,y);
    for *Each Tree in Regressor* F_j **do**
      (dx,dy, w) = $Tree$(**f**) ;
      $V_j(x+dx,y+dy) = V_j(x+dx,y+dy) + w$.
    end
  end
end

Algorithm 2: RF Regressor F_j constructing response image V_j for point $\mathbf{x}_j$.

In the RFCLM framework, the response images $\{V_j()\}$ of all model points, resulting from their associated RF regressors $\{F_j\}$, are regularised by the statistical shape model learned from training data iteratively. This optimization process finds the most likely (highly-voted) point positions to form a likely shape (see below).

3.3.2 Random Forest Regression-Voting Constrained Local Model RFCLM

RFCLM applies Random Forest Regression-Voting (RFRV) in the Constrained Local Model (CLM) framework. The local search for a point position is done by RFRV for each feature point separately and all approximate locations are regulated by the statistical shape model to ensure forming a plausible shape.

Building an RFCLM

Given a set of training images, each annotated with points, we first build a shape model as described in Section 2.6.1. A reference frame of width f_w is defined and the mean shape is scaled to fit within it. All training images are re-sampled into the reference frame by applying the inverse of pose parameter θ found from minimizing the distance ($|(T_\theta(\bar{\mathbf{x}}) - \mathbf{x})|^2$). Training an RFCLM implies building a Random Forest Regressor for each model point in the reference frame as previously described in Section 3.3.1.

Shape Model Matching With RFCLM

Starting from the initial estimates of shape parameters $\mathbf{b}$ and object pose θ the region of interest in a new image is sampled to a reference frame of width f_w. In the reference frame, each local model F_j is used to search the area around its corresponding feature point $\mathbf{x}_j$ separately and to generate a response image V_j as explained in Algorithm 2. The search area is within the range of [-d_{search}, +d_{search}] in x and y. The problem of finding good candidates for feature points becomes an optimisation problem of shifting the current point estimates towards the ones having largest votes while

keeping the resulting shape plausible. The aim is to optimise the cost function:

$$\begin{gathered} Q(\mathbf{b},\theta) = \sum_{j=1}^{n} V_j(T_\theta(\bar{\mathbf{x}}_j + \mathbf{P}_j\mathbf{b})) \\ subject\ to\ shape\ constraints\ \mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} < M_t \end{gathered} \tag{3.2}$$

One search iteration is shown in Algorithm 3. The quality of fit QoF is defined as the mean displacement to the best individual point estimates from the current estimates:

$$QoF(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} |F_j(\mathbf{x}_j)| \tag{3.3}$$

Data: $\mathbf{I}$,$\mathbf{b}$,θ,$model$,$\{F_j\}$

Result: $\mathbf{x}$, $cost$

Function `SearchIteration`($\mathbf{I}$,$\mathbf{b}$,θ,$model$,$\{F_j\}$)**:**

  foreach *landmark* j **do**

    $\mathbf{S} \leftarrow$ sampleSearchAreaOfLandmark($\mathbf{I}$,j)

    $\mathbf{V}_j \leftarrow$getResponseImage($\mathbf{S}$,F_j)

  end

  $\mathbb{V} \leftarrow \bigcup_{j=1}^{n} \mathbf{V}_j$

  $\mathbf{x} \leftarrow$ fitModelToResponseImages($\mathbb{V}$,$model$)

  $cost \leftarrow$ QoF($\mathbf{x}$)

  update the points in the image frame as:

  $\mathbf{x} \leftarrow T_\theta^{-1}(\mathbf{x})$

  return $\mathbf{x}$, $cost$;

End Function

Function `fitModelToResponseImages`($\mathbb{V}$,$model$)**:**

  $radius \leftarrow d_{search}$

  while $radius > r_{min}$ **do**

    Pick the best points within $radius$

    Estimate the shape $\mathbf{b}$ and pose θ_{ref} for the selected points.

    if $\mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} > M_t$ **then**
      move $\mathbf{b}$ to the nearest valid point in the ellipsoid

    end

    update the points in the reference frame as:

    $\mathbf{x} \leftarrow T_{\theta_{ref}}(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b})$

    reduce $radius$

  end

  return $\mathbf{x}$

End Function

Algorithm 3: RFCLM algorithm matching Shape Model $model$ to a new image $\mathbf{I}$ starting from initial shape $\mathbf{b}$ and pose θ using local models $\{F_j\}$.

3.3.3 The Fully-Automated Wrist Segmentation System

After describing the two main techniques (i.e. RFRV and RFCLM) at the core of the fully-automated segmentation system used in our project, we describe the system and its two components: global search (detection) and local search (segmentation).

Global Search

The global search finds the approximate global pose parameters of the wrist (i.e. position, scale and orientation) using a Random Forest regression-voting technique (described in Section 3.3.1). *During training* two anatomical landmarks are used to define the horizontal axis of the reference frame (bounding box) so that their positions are fixed within the reference frame and will be used to approximate the frame pose (i.e. position, scale, orientation). For each training image, a number of training patches are cropped at a number of random displacements $\{d_i\}$ (in scale, position, and angle) from the center of bounding box. A displacement gives the difference in x and y in the reference frame coordinate system. A random forest regressor RF is trained on the pairs $\{(\mathbf{f}_i, d_i)\}$ where $\mathbf{f}_i$ are the Haar features of the patch with displacement d_i . *When searching* a new image, the image is scanned in a sliding window approach by the RF to cast votes $\{(dx,dy, w)\}$ at each position and to construct a response image. The scanning is repeated at a set of angles and scales. A response image is constructed for every angle-scale combination. The maxima of all response images (each response image associated with a different angle-scale combination) are ordered according to their votes and the highest is picked resulting in the most likely center, scale and orientation of the bounding box. From this the approximate positions of the two landmarks are calculated.

Local Search

The bounding box estimated by the global searcher is used to initialise the local searcher to segment the contours. A local searcher is a sequence of RFCLM models of increasing resolution. Bones are usually modeled together at first stages of a local search and then a separate RFCLM-based sequence of models for each bone can used to refine the search even further.
The sections below cover the methods used to extract shape and texture information guided by the annotation found by the fully-automated segmentation system. This information is used to train classifiers to detect fractures.

3.4 Feature Extraction

Feature extraction is the process of combining a set of features in order to come up with a new set of features in a lower dimensional space. Predictive models built using the high-dimensional feature space tend to overfit the data which results in poor model interpretability, and can be computationally expensive. For feature extraction, once the object annotation across the dataset is available (either manually or automatically), statistical shape models (SSM) [25] and statistical appearance models (SAM) [26] can be constructed and used to study the shape and texture of the object. These two techniques have been widely used to improve diagnosis and treatment of musculoskeletal disorders, such as osteoporosis [20, 47] and osteoarthritis [19, 51, 80, 120], and are thus suitable for the task of detecting wrist fractures as musculoskeletal diseases and fractures both involve disruption in intensity patterns and bone shape.
The parameters derived from these models are intended to be informative and non-redundant (i.e. as they are built using PCA which is a popular feature-extracting / dimensionality-reduction technique). This section describes three types of SSM- and

SAM-based features used in the project:

3.4.1 Shape Features

Statistical shape models describe the geometry of an object with limited number of parameters **b** (see Equation 2.2) and eliminate the need to carry out geometric measurements. Clinically, wrist deformities are quantified by predefined measurements such as lengths and angles (see Section 2.3.2). Shape features implicitly capture these measurements and therefore can be useful in detecting fractures.

3.4.2 Texture Features

Texture of an object is its intensity appearance. The variations of a bone's texture across a population may be due to the anatomical variation between individuals, different image acquisition protocols/machines, disease progression, or fractures. In order to compare the textures of two instances of the same object, sampling texture needs to be done at the same scale, orientation and location (i.e. in a shape-normalised frame). This problem is solved by warping the objects' textures to the mean shape first (using a triangulation algorithm). Similar to shape models, statistical texture models [26] are built by applying PCA to vectors of normalised intensity (**g**) sampled from the regions defined by the points of the mean shape in a reference frame of width f_w. The linear texture model has the form:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{3.4}$$

where $\bar{\mathbf{g}}$ is the mean normalised grey-level texture vector, $\mathbf{P}_g$ is a set of orthogonal modes of variation and $\mathbf{b}_g$ is a set of grey-level texture parameters.

3.4.3 Appearance Features

Because shape and texture variations are often correlated, learning this correlation leads to more compact models. Cootes *et al.* [26] applied PCA to the concatenation of the shape parameters **b** and texture parameters $\mathbf{b}_g$ to extract modes of variation of both shape and texture $\mathbf{P}_c$. The concatenation is performed in a weighted form to compensate for the difference in units:

$$\mathbf{b}_a = \begin{pmatrix} \mathbf{Wb} \\ \mathbf{b}_g \end{pmatrix} \tag{3.5}$$

where **W** is a diagonal matrix of weights for each shape parameters. **W** is chosen to balance the total variance in shape and texture:

$$\mathbf{W} = (\frac{TotalVar(texture)}{TotalVar(shape)})^{\frac{1}{2}}\mathbf{I} \tag{3.6}$$

where **I** is an identity matrix. Applying PCA on the concatenated vectors giving the model:

$$\mathbf{b}_a = \mathbf{P}_c\mathbf{c} \tag{3.7}$$

where $\mathbf{P}_c$ are the eignvectors and **c** is the resulting appearance vector.
Shape parameters **b** (in equation 2.2), texture parameters $\mathbf{b}_g$ (in equation 3.4), and appearance parameters **c** (in equation 3.7) were used in the project as features on which classifiers are trained to distinguish between normal and fractured bones.

3.5 Feature Learning

Engineering features is time-consuming, needs an expert knowledge, and might not generalize well. Instead of manually designing features such as those described in the previous section, features can be learned using Convolutional Neural Networks (CNNs) in a supervised manner. CNNs are trained at a specific task (using the features) while learning the features themselves. CNNs will be described in detail in Section 3.6.1.

3.6 Classification

A classifier is a supervised machine-learning algorithm that determines the class of an input element given a set of features based on a knowledge acquired from a training set whose elements have known classes. In this project we used Convolutional Neural Networks (CNNs) and Random Forest classifiers (RFs) to separate the data into discrete classes (Normal/Fractured) in cross-validation experiments with the area under Receiver Operating Characteristic as an evaluation metric.

3.6.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) [70] 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 3.3) stacks a 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 outputing 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 3.4), 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.

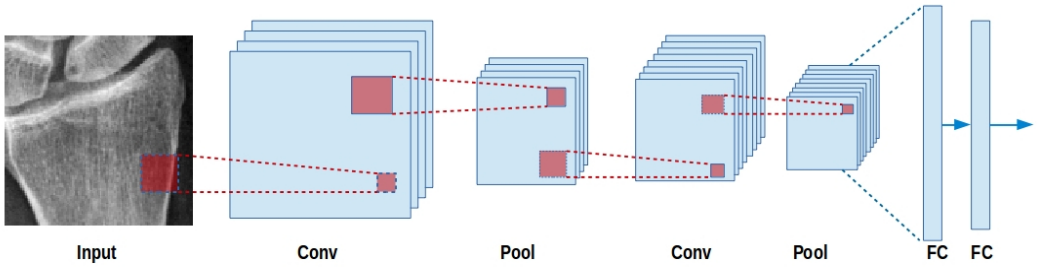


Figure 3.3: A CNN-based classifier applied to a single-channel input image. Every convolutional layer (Conv) transforms its input to a 3D output volume of neuron activations. The pooling layer (Pool) downsamples the volume spatially, independently in each feature map of its input volume. At the end, 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 3.4). 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 also leads to fewer parameters, less computation, and controls overfitting.

The local connections, tied weights, and pooling result in CNNs having 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.

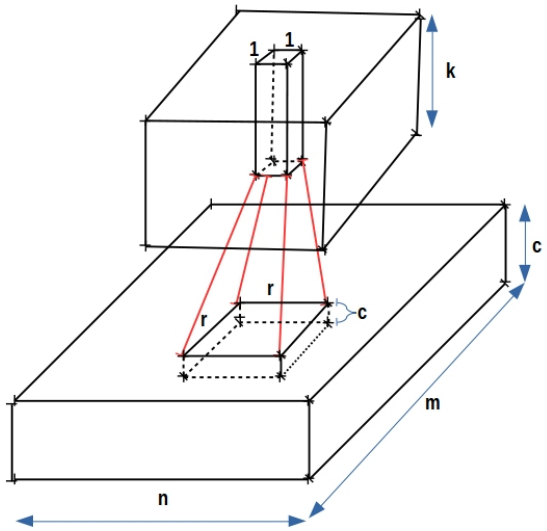


Figure 3.4: In CNN: k neurons receive input from only a restricted subarea (receptive field) of the previous layer output. Convolving the filters with the whole input volume produces k feature maps.

3.6.2 Random Forest (RF) Classifiers

Section 3.3.1 showed how random forests [14] are used as regressors to predict real-valued variables (displacements). We also used random forests as a classifier to predict a class label $x \in \{normal, fractured\}$. The RF classifier was trained on the features described in Section 3.4. The training algorithm is the same as in Algorithm 1 where random decision trees trained with bagging on bootstraps of training pairs of feature vectors and labels T= $\{(\mathbf{f}_i, label_i)\}$ and with split entropy defined as:

$$H(X) = -\sum_{x \in X} p(x) log p(x) \tag{3.8}$$

So the entropy before splitting is:

$$H(X) = -\frac{N_{fractured}}{N} log(\frac{N_{fractured}}{N}) - \frac{N_{normal}}{N} log(\frac{N_{normal}}{N}) \tag{3.9}$$

where N denotes the total number of examples at the node ($N = N_{normal} + N_{fractured}$), and the average entropy after splitting on feature f with threshold t is defined as:

$$H(X|f) = \frac{N_R}{N} \times H(X|f > t) + \frac{N_L}{N} \times H(X|f \leq t) \tag{3.10}$$

Where N_R and N_L denote the number of samples landed on the right and left branches of the split respectively. The information gained by branching on feature f is:

$$IG(X, f) = H(X) - H(X|f) \tag{3.11}$$

The feature that maximises IG (i.e. minimizing average entropy after split) is selected. The splitting continues until reaching maximum tree depth or a minimum number of examples per node. Each tree leaf stores the class distribution of all the training examples which landed at that leaf node. During testing, the decisions from different trees are combined by averaging their probabilistic prediction.

3.6.3 Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve [38] is a two-dimensional summary of classifier performance where the true positive rate TPR (Sensitivity) is plotted against the false positive rate FPR (1-Specificity) while a threshold on the classifier output is varied (see Figure 3.5) resulting in varied numbers of true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP). Sensitivity and Specificity are defined as follows:

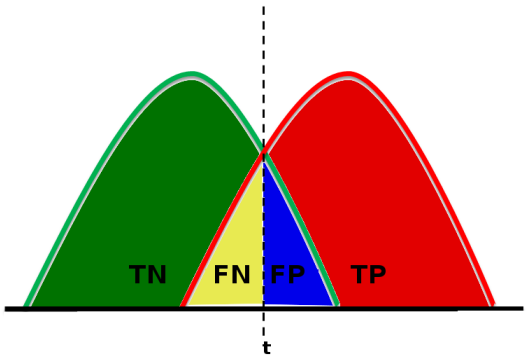


Figure 3.5: Red distribution curve is the pdf of of the positive class (fractured wrist) and the green distribution curve is that of the negative class(normal wrist) with respect to the classifier output. If threshold t decreases the sensitivity increases, the specificity decreases, and vice versa. ROC curve plots sensitivity/specificity pairs corresponding to different values of t. TN denotes number of true negatives. Similarly, FN (false negatives), FP (false positives), and TP (true positives).

$$TPR(Sensitivity) = \frac{TP}{TP+FN} \tag{3.12}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3.13}$$

$$FPR = (1 - Specificity) = \frac{FP}{TN+FP} \tag{3.14}$$

Figure 3.6 shows an example of a ROC curve where each point on the curve represents a sensitivity/specificity pair corresponding to a particular decision threshold t. The area under the curve (AUC) is a measure of how well a classifier can distinguish between two groups (e.g.fractured/normal) and is in the range [0,1]. Random guessing produces a diagonal line between (0,0) and (1,1) with area 0.5.

3.6.4 Cross Validation (CV)

Cross validation [103] was used to measure how well the classifiers (i.e. RFs and CNNs) classify unseen data. The method shuffles the dataset randomly and divides it into a K groups (refereed to as folds). Each fold will be held out once and treated as

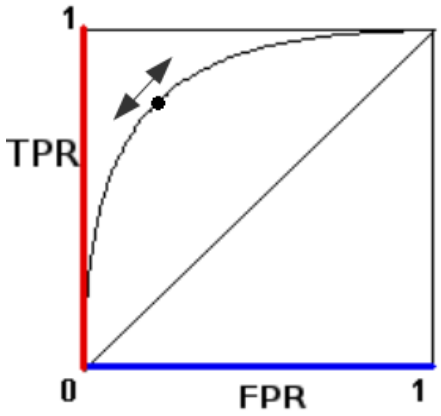


Figure 3.6: A Receiver Operating Characteristic (ROC) Curve.

a test set while a classifier is being trained on the remaining folds. K different classifiers will be trained and tested in K-Fold CV experiments. The project used 5-fold cross validation to give a mean AUC, standard deviations (stdev).

Chapter 4

Fracture Detection with Extracted Features

This chapter summarises experiments to evaluate the methods from Chapter 3 on the tasks: i) Bone segmentation in PA and LAT views for normal and fractured wrists; ii) Fracture detection from PA view, from LAT view, and from both views combined with random forest classifiers trained on the extracted features of shape, texture and appearance. Manual and automated annotations were used to guide the feature extraction process. Results from different feature types, different views, and different annotation methods (i.e. manual and automated) were compared.

4.1 Data

Experiments were run using images from two local EDs gathered and anonymised by a clinician (see Table 4.1). The dataset contains 787 pairs of wrist radiographs (i.e. PA, and LAT) for adult patients, 378 of which are fractured. None of the images contained any plaster casts or metalware in order to ensure the detection is targeting

signs of fractures and not signs of hardware. Although the dataset is not highly class-imbalanced (48% fractured vs 52% Normal), we decided to choose the area under ROC curve as a performance metric for its insensitivity to class distribution.

Table 4.1: Dataset used in this chapter with size (number of adult patients).

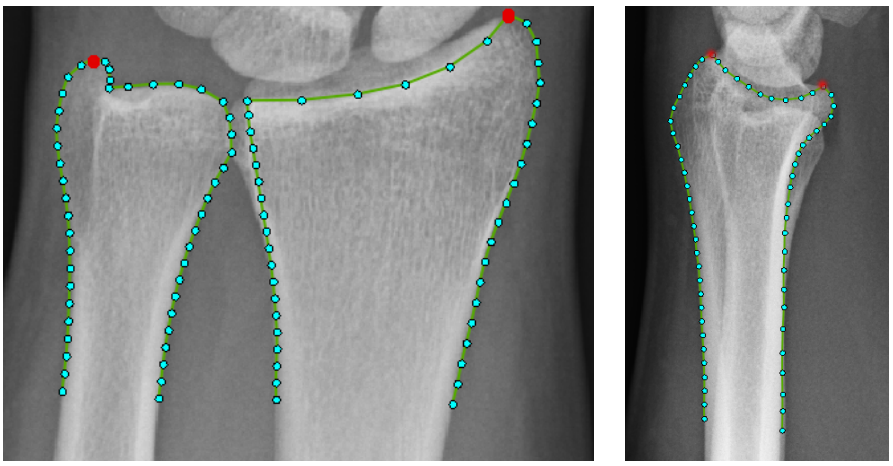
Source	Normal	Fractured	Total
ED 1	211	193	404
ED 2	198	185	383
Σ	409	378	**787**

Referring to the fracture classification described in Section 2.3.1, the dataset contains 291 extra-articular fractures and 87 intra-articular fractures. In terms of displacement, the dataset contains 110 dorsally displaced fractures, 18 volarly displaced fractures, and 250 non-displaced fractures.

4.2 Wrist Detection and Segmentation

As explained in Section 3.3.3 the automatic annotation was performed in two steps: (1) global search to detect the object and (2) local search to segment it. For segmenting the object the RFCLM algorithm (see Section 3.3.2) was trained on manually annotated images. A series of annotation models were tested during preliminary experiments as explained in Chapter 3. The models varied in the number of points and point locations until reaching a final set of points for each view. The radius was annotated with 45 points in the PA view, and 64 points in the LAT view while the ulna was annotated with 48 points in the PA view. This manual annotation gave the ground truth for evaluation. Figure 4.1 shows annotation examples and Figure 4.2 shows the shape modes of the resulting models. Two points per view were used as the reference points

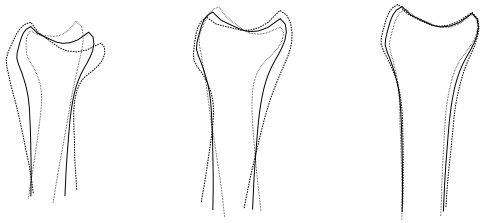
to train the object detector (i.e. the global search explained in Section 3.3.3). During testing these points were used to initialise the mean shape of the local searcher.



PA LAT

Figure 4.1: The annotation (local searcher output points) for each view. Global initialised points are highlighted red.

(PA.1) (PA.2) (PA.3)



(LAT.1) (LAT.2) (LAT.3)

Figure 4.2: The first three modes of the shape models of the radius. (Mean shape and $\pm$ 3 stdev. shapes).

The local searcher was split into stages as shown in Figure 4.3, with each stage containing an RFCLM model which was initialised using the point positions from the previous stage. The frame widths were scaled each stage. For the PA view we modeled the ulna and radius together in the first two stages for more stability. Extra parameters were optimised for each of these stages. These controlled: the search radius around the points, to set the search region for the Constrained Local Model (CLM) optimisation; and displacements of the point model in training, to displace the initial points during RFCLM training.

In order to generate the automatic annotation for the whole dataset, we trained PA models on radiographs from one ED and applied them to the radiographs from the other ED and vice versa. For LAT models, we divided the whole dataset into four subsets, trained models on three, and applied them to the fourth and so on. These four subsets were needed to successfully learn representative models for the LAT view because in addition to the changes in shape and texture due to fractures there is the overlap between the two bones, (i.e. radius and ulna) on the lateral view, which can take various orientations due to different positioning during acquisition [46] . This is not the case for PA view as the two bones appear side by side. Figure 4.4 shows some examples from our LAT dataset. For these reasons (i.e. various relative position of the ulna and the presence of fracture in the view) we were not able to produce a consistent manual annotation for the ulna in the lateral view when fractures were presented.

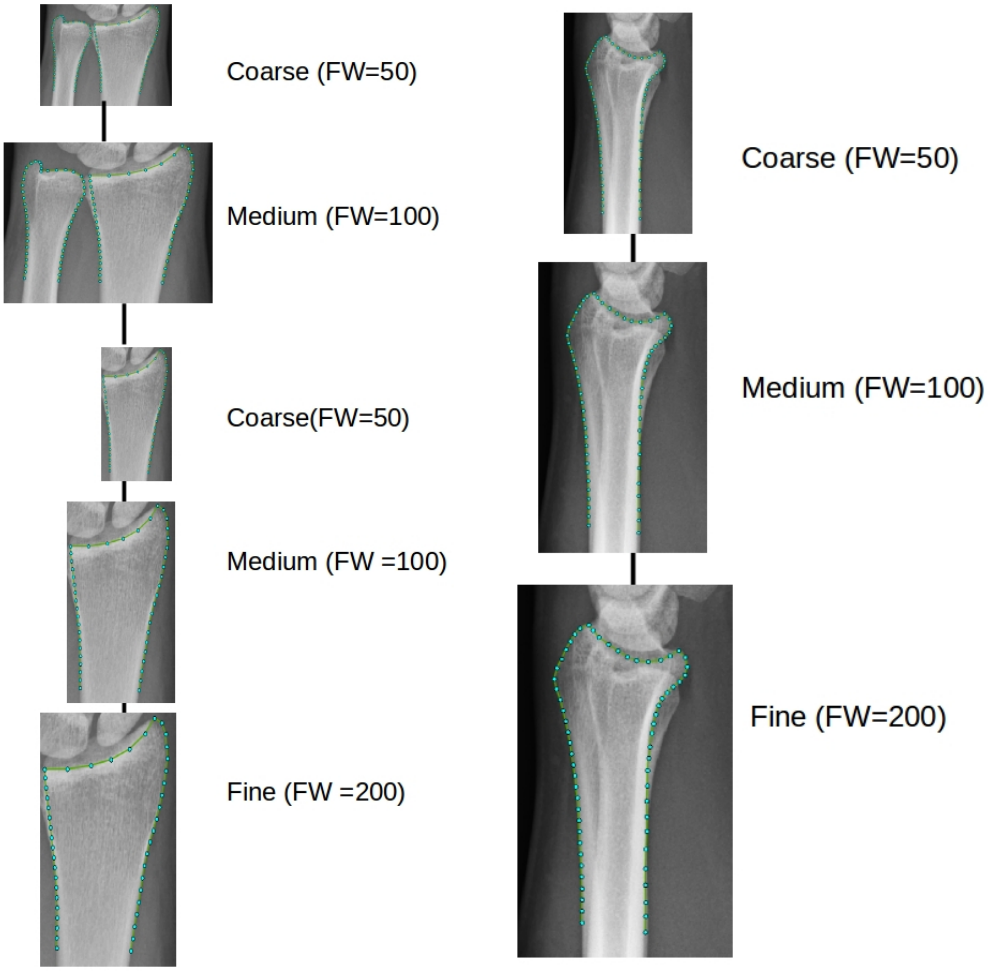


Figure 4.3: Illustration of local searchers with the models iterating over various frame widths (FW) for each view.

Figure 4.4: Different Relative Radius-Ulna Positions Appearing in Lateral Radiographs.

The error measure we chose is the point-to-curve Euclidean distance error (see Figure 4.5). This error was chosen over point-to-point error to deal with the aperture problem (i.e. to minimise the penalty of finding the shape outline but not the exact point positions), and was chosen over curve-to-curve error to reduce computation time over comparing points along each curve between the manual and automated points.

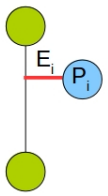
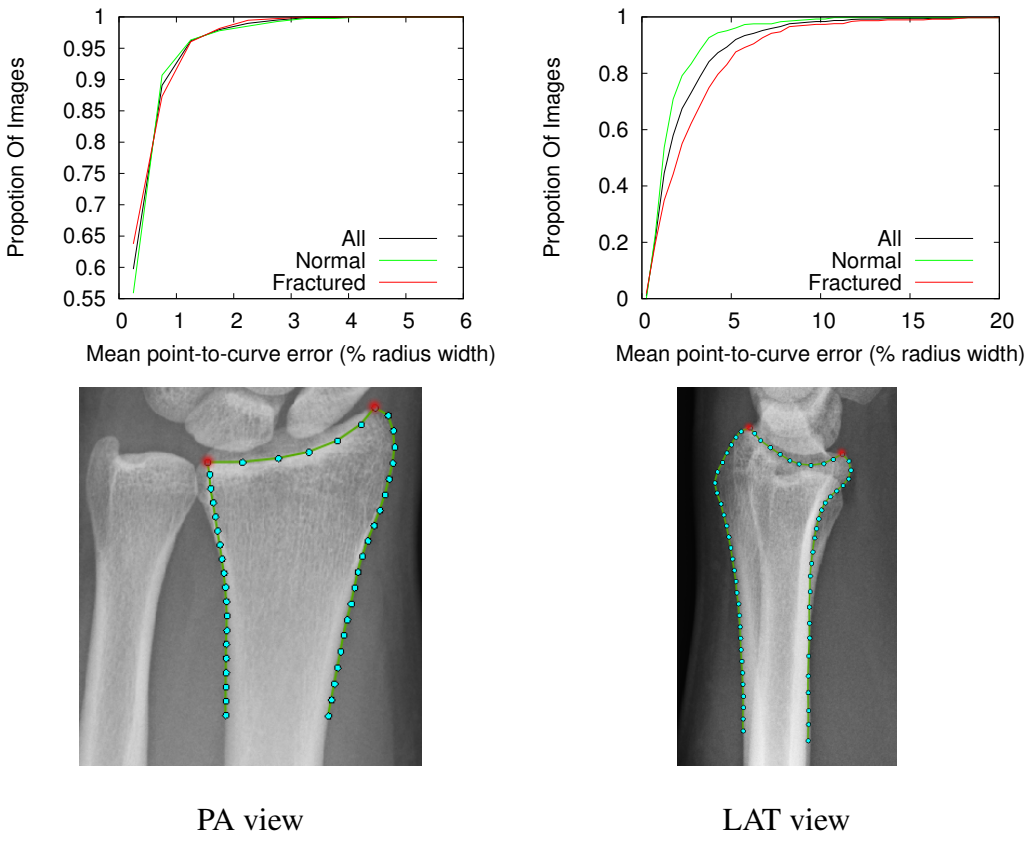


Figure 4.5: The point-to-curve error E_i is the distance highlighted red between the automated point P_i and the closest part of the curve between the manually-annotated points (drawn in green).

In order to provide invariance to image scaling the accuracy of the segmentation was calculated as the percentage of mean point-to-curve distance relative to a reference width. Results are presented in Table 4.2. The reference width of a view is the distance between the two reference points for that view (see Figure 4.6), the results are also presented as cumulative density functions (CDFs) for each class in Figure 4.6.

Table 4.2: The mean point-to-curve distance as a percentage of the reference length (radius width in the view).

View	Class	Mean	Median	90%	95%
PA	Normal	0.46	0.27	0.98	1.27
PA	Fractured	0.43	0.31	1.15	1.42
PA	Both	0.45	0.29	1.05	1.36
LAT	Normal	1.99	1.44	3.80	4.71
LAT	Fractured	3.11	2.24	6.33	8.08
LAT	Both	2.53	1.65	5.09	6.85



PA view LAT view

Figure 4.6: The CDF shows the relative distance error of all 787 images. The error is taken as a percentage of the reference distance (between the two red points).

The results show the ability of the models to successfully segment the targeted structure even when fractured. Supposing that the reference length is 25mm in PA view and 20mm for LAT view, the mean error would be less than 0.34 mm for more than 95% of the radiographs in the PA view which is in accordance with that reported in other similar studies (0.54mm for Knee joint in PA view [113], 0.6mm for proximal femur in PA view [75]), and less than 1.37mm for 95% of radiographs in the LAT view. The table also breaks down the results by class and shows that although the models performed equally well for fractured and normal cases in the PA view, in the LAT view errors are roughly 50% larger for fractured cases than normal cases. However overall the system was able to capture a good approximation to both the normal

and fractured shapes (see Figures 4.7 and 4.8 for examples). This is, to the best of our knowledge, the only work reporting results on the task of wrist segmentation. Other works in [73, 79] had segmented the radius before detecting fractures but they did not report any segmentation results.

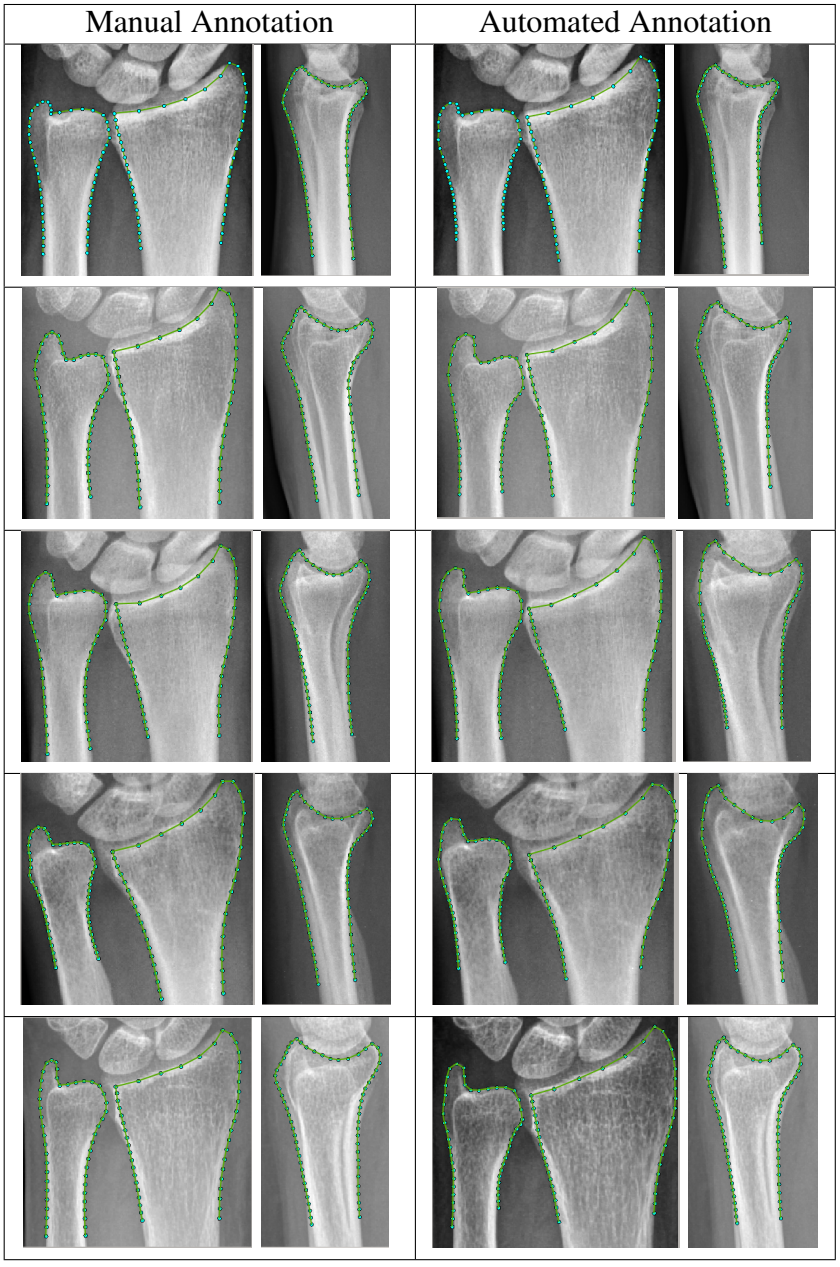


Figure 4.7: Annotation examples of normal wrists. Each row belongs to one patient.

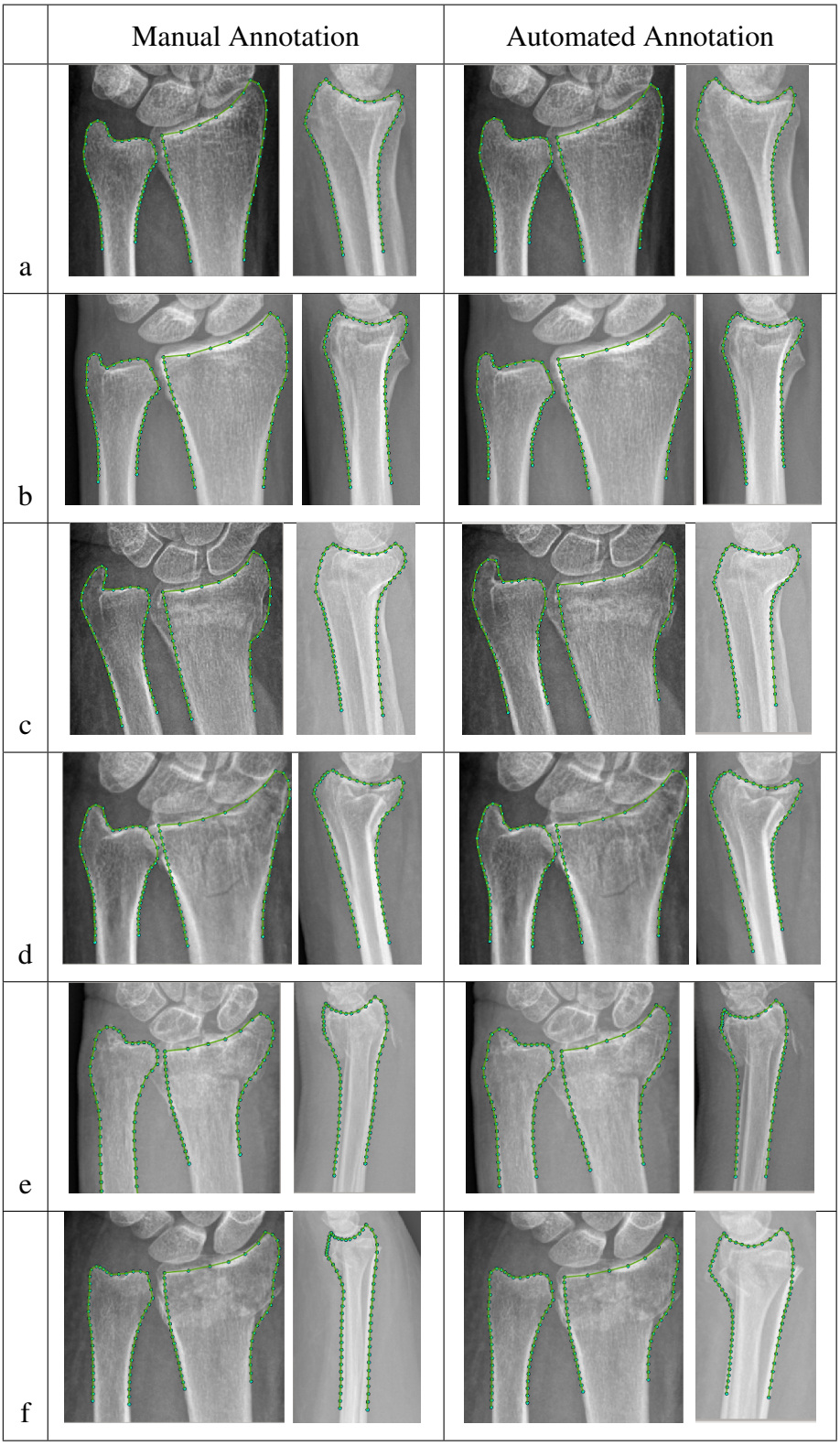


Figure 4.8: Annotation examples of fractured wrists. Each row belongs to one patient with a (a,b) non-displaced fracture, (c) extra-articular volarly displaced fracture, (d) intra-articular volarly displaced fracture, (e) intra-articular dorsally displaced fracture, (f) extra-articular dorsally displaced fracture.

The next section covers classification experiments given localization using this method. The relevant shape, texture, appearance features were extracted as described in Section 3.4.

4.3 Fracture Detection

For each view we performed 5-fold cross validation experiments with Random Forest classifiers on: (i) shape parameters only, (ii) texture parameters only, (iii) appearance parameters, and (iv) the combination of shape and texture parameters by concatenation and averaging. Figure 4.9 summarises the various algorithmic choices that were evaluated. During the training phase, shape, texture, and appearance models were built with the training images and their manual annotations. The number of modes of variation was constrained to model 99% of the variance of the training data (see equation 2.3). The resultant features were used to train random forest classifiers. The RF parameters (the number of trees n_{tree}, the maximum depth of each tree D_{max}, and the minimum number of training samples n_{min} allowed at a split node,) were optimised using the train/test data (i.e. no separate validation data) leading to n_{tree}= 100, D_{max}= 30, n_{min} =1. During testing, feature extracting models (i.e. shape, texture, and appearance) were fitted to the manual and automated annotations of the query image, and the resultant features were passed to the corresponding RF classifier. The results are presented in terms of the mean AUC and standard deviation (stdev) over the five folds. Below we present and discuss the classification results for each view.

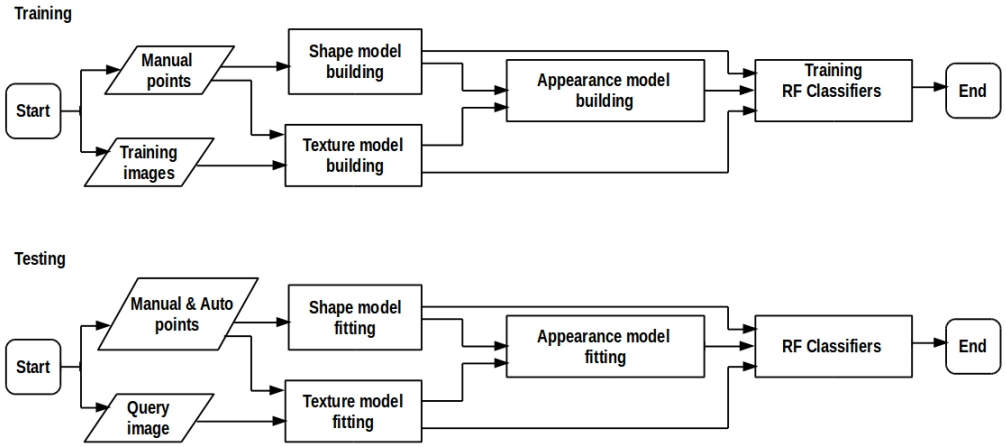


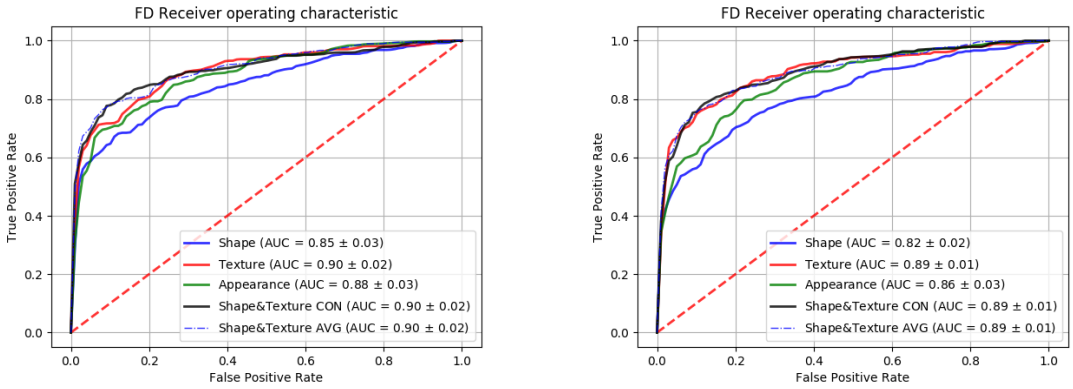
Figure 4.9: Flowchart summarising the process carried out in the cross validation expirements. All models were trained on the manual annotations of the training folds and tested with the manual and automated annotation of the test fold.

4.3.1 Classification From PA View

The results obtained from the PA view are shown in Table 4.3 and Figure 4.10. Texture features alone performed better than shape features and were more robust to contour inaccuracies of automated annotation, with no significant difference in performance compared to manual annotation mode. This was not the case for the shape features as the difference in performance between the two modes are significant despite the high segmentation accuracy reported in Section 4.2. Concatenating (CON) the two vectors of shape features and texture features and training a random forest on the resulting vector did not show any difference in the performance with averaging (AVG) the decisions from two random forests (one trained on shape features and another on texture features). The methods of combining shape and texture in the form of concatenation (CON) or averaging (AVG) did not add any new information to that provided by texture features alone although combining with a further PCA (i.e. appearance features) deteriorated the performance.

Feature Type	Manual	Automated
Shape	0.85±0.03	0.82± 0.02
Texture	**0.90±0.02**	**0.89±0.01**
Appearance	0.88± 0.03	0.86±0.03
Shape & Texture CON	**0.90±0.02**	**0.89±0.01**
Shape & Texture AVG	**0.90±0.02**	**0.89±0.01**

Table 4.3: The area under ROC curve (AUC±stdev) for classification based on PA view.



(a) Manual (b) Automated

Figure 4.10: ROC curves for fracture detection (FD) from PA view for images when paired with their: (a) manual annotation, (b) automated annotation.

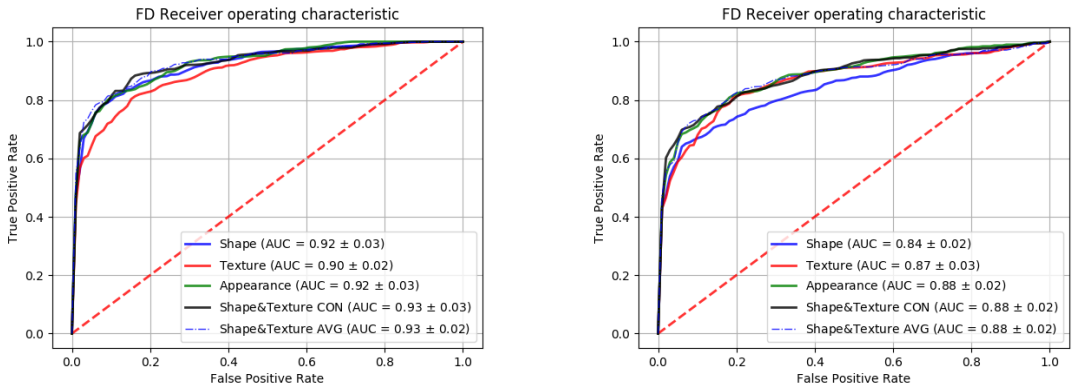
4.3.2 Classification From LAT View

Table 4.4 and Figure 4.11 show the classification results for the LAT view. Significant performance differences between the manual and automated modes are apparent which can be put down to the lower segmentation accuracy compared to that of PA view. Interestingly, classifying using the shape alone gives significantly better performance than that of the PA view shape. This might suggest that the shape in LAT view

is more informative than that of PA view although it was harder to segment as accurately (four folds for LAT view as opposed to 2 folds for PA view for segmentation experiments in Section 4.2). The difference in performance for shape between manual and automatic results suggests that classification performance can be improved by improving the accuracy of the point search (perhaps by increasing the size of the training set). Combining shape with texture, compared to shape alone, made a significant improvement in the automated mode but no significant improvement in manual mode. This could be because the two automated models (i.e. one for shape and the other for texture) make different mistakes while in the manual mode they tend to agree more. In general, there was no evidence for improved performance in the automated system when combining shape and texture either by PCA (i.e. appearance), concatenating, or averaging, as opposed to texture.

Feature Type	Manual	Automated
Shape	0.92±0.03	0.84±0.02
Texture	0.90 ± 0.02	0.87±0.03
Appearance	**0.93±0.03**	**0.88±0.02**
Shape & Texture concatenated	**0.93± 0.03**	**0.88±0.02**
Shape & Texture averaged	**0.93± 0.02**	**0.88±0.02**

Table 4.4: The area under ROC curve (AUC±stdev) for classification based on LAT view.



(a) Manual (b) Automated

Figure 4.11: ROC curves for fracture detection (FD) from LAT view for images when paired with their: (a) manual annotation, (b) automated annotation.

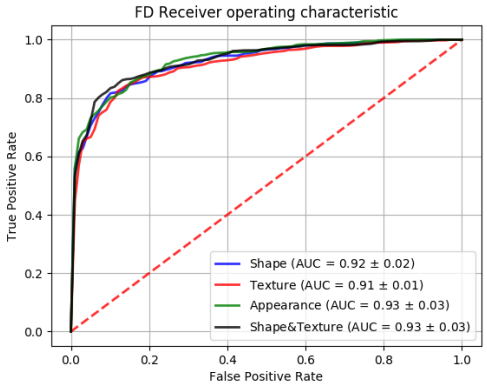
4.3.3 Classification From Both Views

Table 4.5 and Figure 4.12 show the classification results for combining two views by concatenating feature vectors (CON) and by averaging decisions from different classifiers (AVG). Combining information from both views resulted in the best classification performance in both manual and automated modes compared to results for each view separately. Averaging decisions from different RFs performed better than concatenating feature vectors and training one RF. This could be seen as an ensemble of random forests, each RF was trained on different set of variables leading to more robustness. Combining all shape and texture features from both views resulted in the best classification results (see Figure 4.13) achieving an AUC of 0.95 and of 0.92 for manual and automated modes, respectively. Interestingly, no improvements were achieved when combining shape and texture for PA view alone (see Table 4.3) or for LAT view alone (see Table 4.4) by averaging the decisions from two RFs compared to concatenating the two feature vectors. This could be because in both cases (PA and LAT) there was one classifier (either shape or texture) that performed significantly

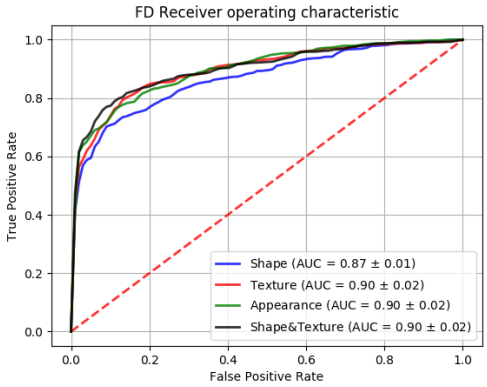
better than the other for the same view and the averaging would be still dominated by the stronger classifier. Averaging decisions across different views could be more able to lessen this tendency.

Feature Type	Manual	Auto
Shape CON	0.92±0.02	0.87±0.01
Shape AVG	0.93±0.02	0.88±0.01
Texture CON	0.91±0.01	0.90±0.02
Texture AVG	0.93±0.01	0.91±0.01
Appearance CON	0.93±0.03	0.90±0.02
Appearance AVG	0.94±0.02	0.91±0.02
Shape & Texture CON	0.93±0.03	0.90±0.02
Shape & Texture AVG	**0.95±0.02**	**0.92±0.01**

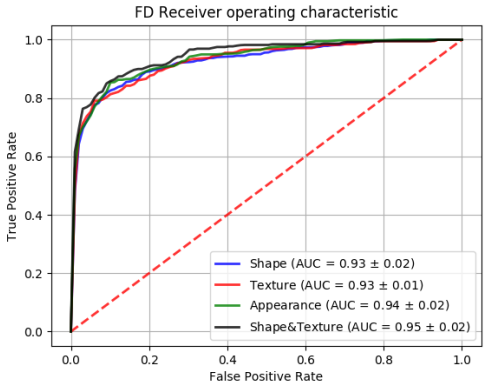
Table 4.5: AUC for classification based on features from both views. The view combining was performed either by concatenating feature vectors (CON) or by averaging decisions from different classifiers (AVG).



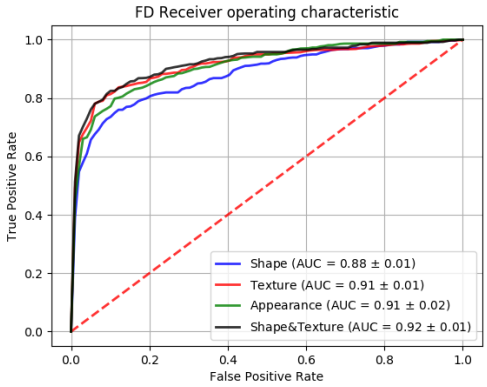
(a) Manual and CON



(b) Automated and CON

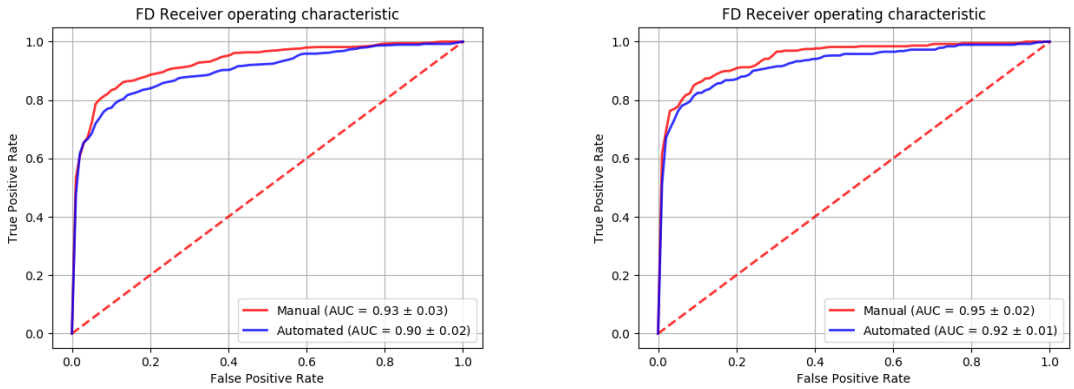


(c) Manual and AVG



(d) Automated and AVG

Figure 4.12: ROC curves for fracture detection from combining the two views for images when paired with their: (a) manual annotation, (b) automated annotation. The view combining was performed either by concatenating feature vectors (CON) or by averaging decisions from different classifiers (AVG).



(a) Features Concatenated (CON) (b) Decisions Averaged (AVG)

Figure 4.13: The ROC curves corresponding to classification based on combining shape and texture features from both PA and LAT views for manual annotation, and automatic annotation. In case of CON: four vectors of features concatenated (CON) and one RF trained while AVG means decisions averaged from four RFs, each trained on one vector of features.

4.4 Conclusion

This chapter presented a system that automatically locates the outline of the radius in both posteroanterior and lateral radiographs and extracts features by SSM and APM. These features capture statistically significant shape, and texture information from the images, which are not all related to fractures. Thus our choice of random forests as a classifier was driven by the need for further extraction of the informative features for fracture detection. RFs perform this selection in different feature subspaces at each node split. Our results showed good classification performances suggesting the suitability of both the extracted features and random forests to the task of fracture detection. A similar approach (combining RF classifiers with features extracted by SSM and APM) has been applied to classify osteoporotic vertebral fractures by Bromiley *et al.* [16] in X-rays and CT image volumes and showed significant improvements in

diagnostic accuracy.

We found that combining decisions from both views by averaging the outputs of all four RF classifiers achieved the best performance with AUC of 0.95 and of 0.92 for manual and automated modes respectively. All prior work on automated feature extraction for detecting wrist fractures that we are aware of [73, 79] used only the PA view. We showed for the first time that fractures can be better identified in the lateral view, and that combining information from both views leads to an overall improvement in performance. [73, 79] used active shape models [25] and active appearance models [26] to locate the approximate contour of the radius. They extracted various texture features (Gabor, Markov Random Field, and gradient intensity) and used Support Vector Machines (SVMs). They achieved encouraging performance (accuracy$\approx$sensitivity$\approx$96%) but were working on a rather small dataset with only 23 fractured examples in their test set.

In the next chapter, on the same task of fracture detection we explored the use of convolutional neural networks and learned features.

Chapter 5

A Deep Learning-Based Approach

This chapter presents a novel approach for automatically detecting wrist fractures from plain PA and LAT X-rays. A CNN is trained per view from scratch on radiographic patches cropped around the target bone after automatic segmentation and registration. The decisions from both views are combined by averaging.
Encouraged by the results we explored the use of the same technique on another problem: Automatic knee OA diagnosis from plain PA X-rays. The chapter is split into two main sections, one for each problem.

5.1 Automatic Wrist Fracture Detection

5.1.1 Data and Automatic Annotation

A wrist dataset containing 1010 pairs of wrist radiographs (PA, and LAT) for 1010 adult patients, 505 of whom had fractures (see Table 5.1). Images for 787 patients (378 of whom had fractures) were gathered from two local EDs while the rest were gathered from the MURA dataset [94] with fractures as abnormality. None of the

images contain any plaster casts or metalwork in order to ensure the detection is targeting signs of fractures not signs of hardware. In Chapter 4, the systems used for automatically annotating radiographs from EDs 1 and 2 were described and evaluated. Radiographs from the MURA dataset were automatically annotated by running the models previously built from ED datasets. All experiments were based on automatic annotations since no manual annotations were available for MURA radiographs. MURA automatic annotation was visually inspected to ensure quality. The goal was to use the point annotation to crop a patch containing the object.

Table 5.1: Different sources of wrist radiographs with their sizes (number of adult patients).

Source	Normal	Fractured	Total
ED 1	211	193	404
ED 2	198	185	383
MURA	96	127	223
Σ	505	505	**1010**

5.1.2 Methods

Because most parts of a radiograph are either background or irrelavent to the task, we chose to train CNNs on cropped patches rather than raw images. The automatic annotation of radiographs gives information on the position, orientation and scale of the distal radius accurately (see results in Chapter 4). 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. The steps of the automated system are shown in Figure 5.1. Chapter 4 covered the automated segmentation so below we only describe the system stages of cropping registered patches and training CNNs.

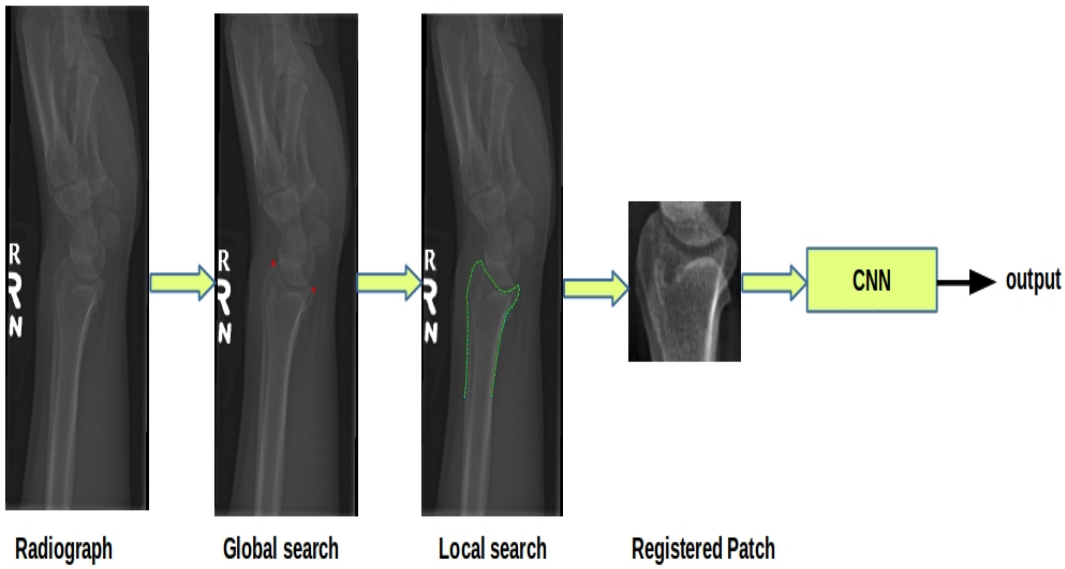


Figure 5.1: Fully automated system for detecting wrist fractures.

Patch Preparation

Given the automatic annotations Algorithm 4 produces registered patches. Figure 5.2 shows examples of radiographs and extracted patches.

Data: mean shape $\bar{\mathbf{x}}$, frame height f_h, border b, pairs of images and annotations = $\{(\mathbf{I},\mathbf{x})\}$
Result: Registered Patches $\{(\mathbf{P})\}$, each of size $(f_h + 2*b)$ x $(f_h + 2*b)$;
Calculate *box_height*: the height of the bounding box containing $\bar{\mathbf{x}}$;
Set up the mean shape in reference frame $\bar{\mathbf{x}}_{ref} = s\bar{\mathbf{x}}$ where $(s = \frac{f_h}{box\_height})$;
Compute the bounding box containing the mean shape $[x_{min}, x_{max}]$ $[y_{min}, y_{max}]$;
Translate points by $(b - x_{min}, b - y_{min})$ so that the expanded bounding box becomes $[0, 2*b + x_{max} - x_{min}]$ $[0, 2*b + y_{max} - y_{min}]$;
for *Each image* $\mathbf{I}$ *with its annotation* $\mathbf{x}$ **do**
    Compute θ so that $|T_\theta(\bar{\mathbf{x}}_{ref}) - \mathbf{x}|$ is minimal;
    for *i, j in [0,2*b + f_h)[0,2*b + f_h)* **do**
        $P(i,j) = \mathbf{I}(T_\theta(i,j))$
    end
    Save P;
end

Algorithm 4: Cropping registered patches.

(a) (b) (c) (d)

Figure 5.2: Example of 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.

Network Architecture

We trained a CNN for each view. The two CNNs were a classical stack 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 [45] and biases initialised to zeros. The loss function was binary cross entropy optimised with Adam [67] (default parameter values used). Architecture details are summarised in Table 5.2. In our experiments we gradually increased the number of CP layers and chose the network with best performance. Figure 5.3 shows an example network with three CP layers followed by two fully-connected (FC) layers. Models were developed using Tensorflow 1.0.

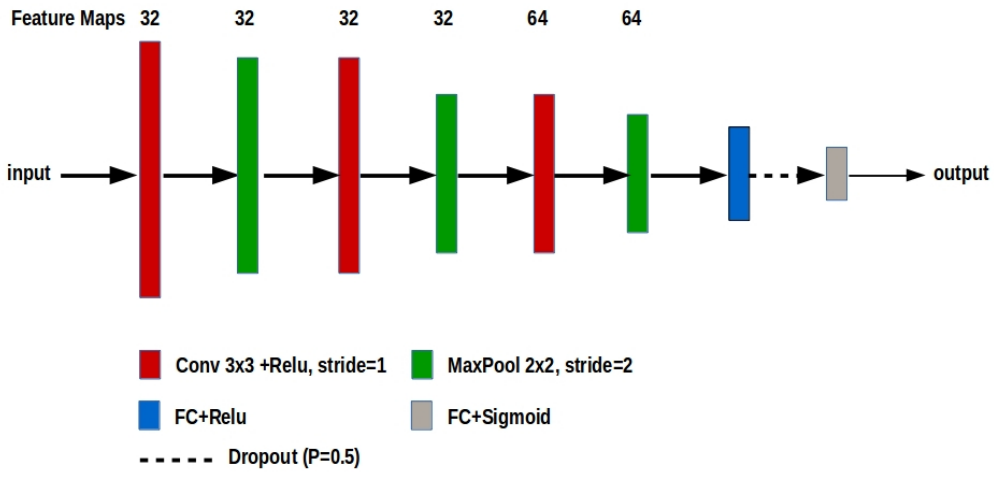


Figure 5.3: An example network with an architecture of CP1-CP2-CP3-FC1-D-FC2. This network performed the best with LAT view patches of size 151x151.

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 5.2: 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 one with best performance.

We aimed at an input patch size that is big enough to capture the trabeculae structure but not adding upsampling noise. The images in MURA have heights no longer than 512 pixels. Thus the area of interest around the radius is about 100 pixels high in most images. We tried patches of size: 121x121, and 151x151 corresponding to Algorithm 4's parameters: f_h=111, b= 5 and 20, respectively.

(a) 121x121 (b) 151x151

Figure 5.4: Wrist patches of different sizes.

5.1.3 Experiments

We carried out 5-fold cross validation experiments. During each fold about 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. Patches were transformed from [0; 255] range to [0; 1] range to avoid possible optimization issues. Figure 5.5 shows an example of learning curves. Training data is randomly shuffled at the start of each epoch to produce different batches each time. Having trained the two

CNNs, one for each view, their outputs are combined by averaging (see Figure 5.6).

Figure 5.5: Example of learning curves for a model (in the first fold).

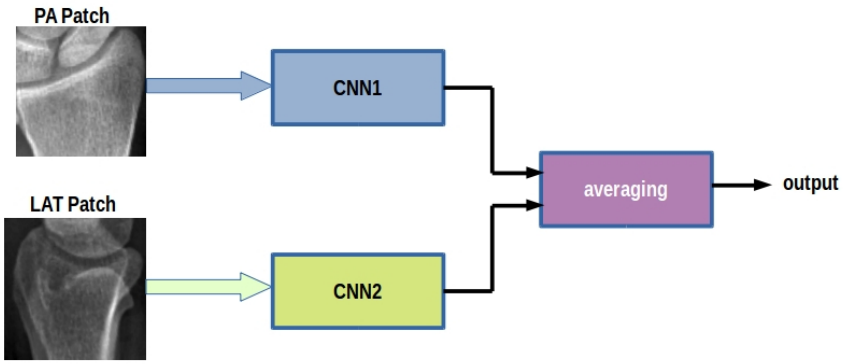


Figure 5.6: During testing the outputs for both views are combined by averaging.

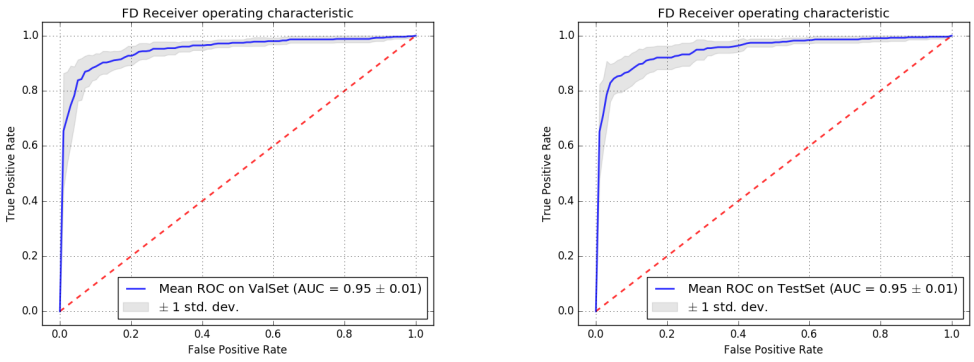
5.1.4 Results

PA View

The performance of different networks are summarised in Table 5.3. The results shows that the network NW3 performs significantly better on patch size 121x121 than on 151x151. This could be because in PA view bones appear side by side and therefore the smaller the border size the less noise the CNNs have to tackle. ROC curves on validation, and testing sets by network NW3 for patch size 121x121 are shown in Figure 5.7.

Table 5.3: The performance of different networks on PA view on different patch sizes in terms of average AUC $\pm$ stdev.

Network	Architecture	PA 121x121	PA 151x151
NW1	CP1-CP2-FC1-D-FC2	0.93±0.02	0.93±0.01
NW2	CP1-CP2-CP3-FC1-D-FC2	0.94±0.02	0.93±0.01
NW3	CP1-CP2-CP3-CP4-FC1-D-FC2	**0.95±0.01**	0.94±0.01
NW4	CP1-CP2-CP3-CP4-CP5-FC1-D-FC2	0.93±0.02	0.93±0.02



(a) Validation Set (b) Testing Set

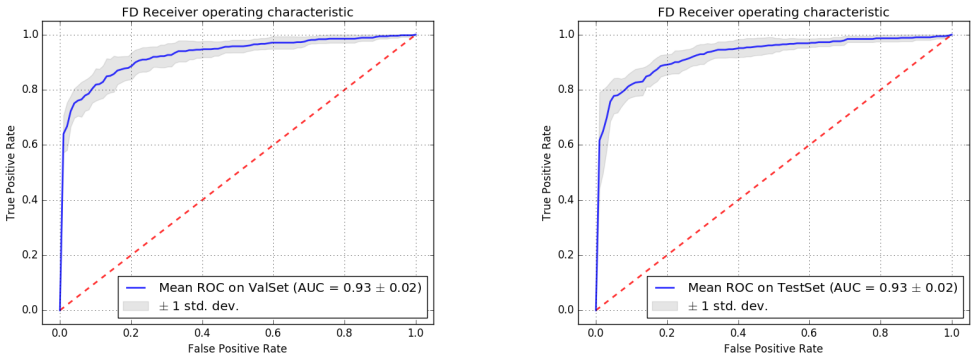
Figure 5.7: The best ROC Curves in PA View achieved with architecture NW3 and patch size 121x121.

LAT View

The performance of different networks are summarised in Table 5.4. Unlike the PA results, LAT results show that having larger borders improved the performance significantly. That could be because the relative position of ulna with respect to radius changes a lot due to different acceptable positioning in practice, and a wider border allows containing the whole wrist area. ROC curves on validation, and testing sets by network NW2 for patch size 151x151 are shown in Figure 5.8.

Table 5.4: The performance of different networks on LAT view on different patch sizes in terms of average AUC ± stdev.

Network	Architecture	LAT 121x121	LAT 151x151
NW1	CP1-CP2-FC1-D-FC2	0.91±0.02	0.92±0.02
NW2	CP1-CP2-CP3-FC1-D-FC2	0.91±0.02	**0.93±0.02**
NW3	CP1-CP2-CP3-CP4-FC1-D-FC2	0.91±0.02	0.92±0.02
NW4	CP1-CP2-CP3-CP4-CP5-FC1-D-FC2	0.90±0.03	0.90±0.02

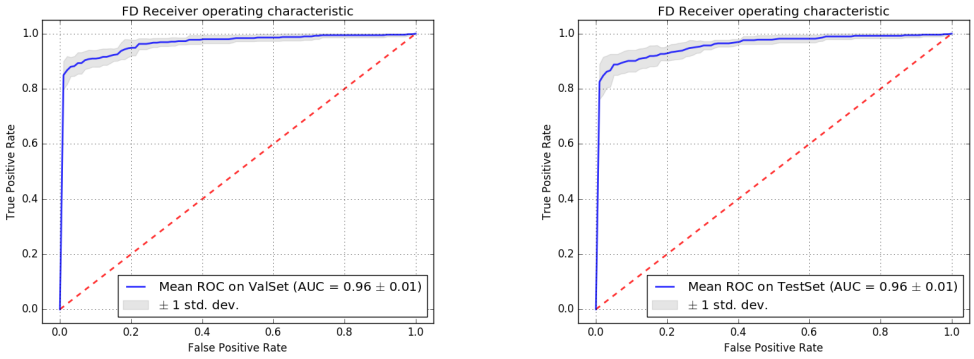


(a) Validation Set (b) Testing Set

Figure 5.8: The best ROC Curves in LAT View achieved with architecture NW2 and patch size 151x151.

Both Views

Combining the predictions from the best PA model (NW3 on 121x121 with AUC=95%) with the best LAT model (NW2 on 151x151 with AUC=93%) by averaging for each fold resulted in AUC=96%. ROC curves on validation, and testing sets are shown in Figure 5.9.



(a) Validation Set (b) Testing Set

Figure 5.9: ROC Curves in combined-view experiments.

5.1.5 Discussion

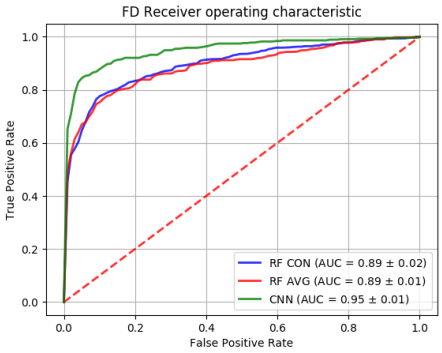
We presented a system for automatic wrist fracture detection from plain PA and LAT X-rays. The CNN is trained from scratch on radiographic patches cropped around the joint after automatic segmentation and registration. This directed preprocessing ensures meaningful learning from only the targeted region 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. The only prior work on the same problem was that of Kim *et al.* [66]. They used features originally learned to classify non-radiological images [96] by re-training the top layer of Inception v3 network on the task of detecting fractures in LAT views only. An initial set of 1,389 lateral wrist radiographs (695 fracture and 694 no fracture) were used for training after applying an eightfold data augmentation technique to get 11,112 images. They reported an AUC of 95.4% on previously unused data set of 100 lateral wrist radiographs (half of which contained fractures). However, their training and testing datasets did not contain images for which the lateral projection was inconclusive for the presence or absence of fracture. Unlike their work we have not excluded images on such criteria, which would bias the results favorably but contradict the goal of developing such systems (finding easily missed fractures). Compared to our network, which takes a registered patches of size 151x151 containing the targeted bone, they re-scaled the radiographs to fit the input size of Inception network, packing task-irrelevant details which resulted in an unnecessarily complex model.

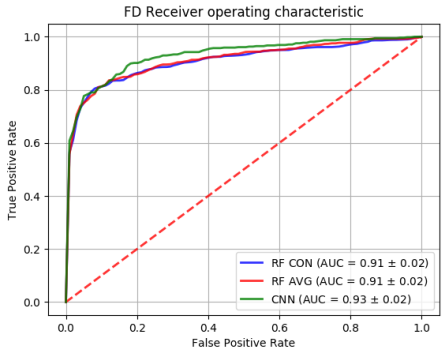
For the sake of comparison with our previous technique from Chapter 4, we repeated all experiments on the dataset used in this chapter (see Table 5.1) with the same fold divisions and found an AUC of 93% from two views combined, 89% and 91% for PA view, and LAT view respectively (See Table 5.5 and Figure 5.10). The CNN-based techniques clearly outperforms the RF-based. Unlike CNNs, training random forest does not need a validation set, so the number of folds were 5 for RF experiments compared to 10 folds (5 with swapping validation and testing sets) for CNNs. This difference is not expected to affect the comparison.

Table 5.5: Comparison between CNN-based and RF-based techniques on the same dataset in terms of AUC ± stdev.

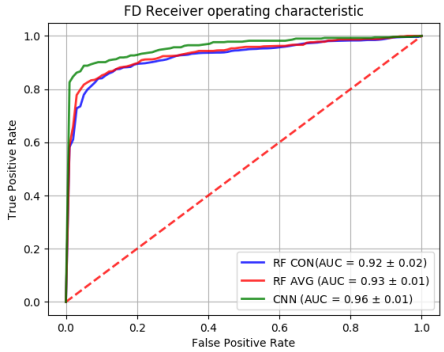
Method	PA view	LAT view	Both Views
CNNs	**0.95±0.01**	**0.93±0.02**	**0.96±0.01**
RFs on shape and texture params CON (concatenated as per Chapter 4)	0.89±0.02	0.91±0.02	0.92±0.02
RFs on shape and texture params AVG (averaged as per Chapter 4)	0.89±0.01	0.91±0.02	0.93±0.01



a) PA view



b) LAT view



c) Both views

Figure 5.10: Comparison between ROC Curves for CNN-based and RF-based techniques on: a) PA view, b) LAT view, and c) both views combined for the same dataset in terms of AUC ± stdev.

5.2 Automatic Osteoarthritis Diagnosis From Knee PA Radiographs

Osteoarthritis (OA) is a degenerative disease in which bones and surrounding soft tissue of the affected joint deteriorate. The disease is associated with pain, disability and substantial care costs each year [22]. Experienced clinicians currently perform the clinical OA severity grading of X-ray images by looking for characteristic features of knee OA on plain radiographs. Classification criteria based on categorisation into distinctive grades is, however, subject to errors of measurement and poor observer agreement. There is an urgent need for automated methods to measure radiographic features and remove, as far as possible, the element of subjectivity in assessment. Although OA has no cure, the development of improved methods for detecting and analysing OA will improve understanding of disease development and evaluation of new treatments that may slow or prevent progression of the disease. Aiming at applying our deep-learning-based wrist fracture detection technique to similar problems, we describe a fully automated system to diagnose knee OA from PA radiographs according to Kellgren-Lawrence (KL) method [64]. We performed experiments on both binary OA classification (OA vs Non-OA) and Multi-class (Kellgren-Lawrence Grading KL [64]) classification.

5.2.1 Background

Kellgren-Lawrence Osteoarthritis Grading

The most widely used OA grading is the Kellgren-Lawrence (KL) method [64], which splits disease development into five classes: normal (KL0), doubtful (KL1), minimal (KL2), moderate (KL3) and severe (KL4). Onset of the disease is usually taken to be

KL2 and above. KL grading is performed through visual inspection of knee radiographs looking for OA characteristic features which include narrowing of the joint space, thickening of the joint line (bone sclerosis) and new bone formation at the joint margin (osteophytes) system. Because the grading is discrete while the OA progress is continuous, there will be many in-between cases resulting in weaker ground-truth. Moreover, the reliance on experience and training can make the grading susceptible to subjective views of the observer leading to high number of inter-rater disagreements (quadratic Kappa 0.56 [48], 0.66 [105], 0.67 [31]) especially when distinguishing between the central grades (KL 1-3).

Previous Work

There are few published methods in the area of automatic OA classification on radiographs. Early work in [102] used template matching to automatically locate and extract the knee joint. It used a weighted nearest neighbor rule on a hand-crafted feature vector to classify four KL classes (KL0-KL3). Other works [83, 113, 115] used RFCLMs to segment knee bones and trained random forest classifiers on shape and appearance features [26] with different hand-crafted texture features. Apart from all above-mentioned approaches which used engineered features, there is a body of work [3, 4, 117] on learning discriminate features by CNNs. Antony *et al.* [4] detect the knee joint region by a linear SVM classifier trained on image gradients for positive and negative examples. Off-the-shelf CNNs were then used in two forms: (1) to extract features and train a SVM classifier, (2) and to be fine-tuned using classification loss (cross entropy) and regression loss (mean squared error MSE=$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$). The resulting performance outperformed that of [102] and showed that CNNs fine-tuned on MSE loss outperform those using classification loss which makes the case for using continuous MSE as an appropriate classification metric. In a follow-up

work, Antony *et al.* [3] used a fully convolutional network-based method to automatically localize the knee joints and a multi-objective CNN trained from scratch to classify OA. They reported a multi-class accuracy of 61.9% and MSE of 0.66 on test set of 2200 radiographs.

Tiulpin *et al.* [117] used a histogram of gradients (HoG) feature descriptor to describe the knee joint shape and a linear SVM classifier to localize the joint then used a Siamese deep neural network for OA classification. They tested their approach on test set of 5960 radiographs and reported AUC=0.93, average multiclass accuracy is 66.71%, and corresponding quadratic Kappa coefficient and MSE value are 0.83 and 0.48 respectively, outperforming all previous works.

5.2.2 Data and Automatic Annotation

We used the PA view of MOST Osteoarthritis (OA) public dataset [39]. The MOST cohort contains data from 3,026 patients and their six follow-up examinations. The radiographs were graded according to the semi-quantitative KL scale 5. We treat KL0,1 as Non-OA (10,602 images, 55.2%), and KL2-4 as OA (8,606 images, 44.8%) for the purpose of OA detection. The data distribution is shown in Table 5.6.

Grade	KL0	KL1	KL2	KL3	KL4
No. of images	7691	2911	3429	3547	1630

Table 5.6: MOST dataset [39] with total of 19,208 PA images.

To automatically segment the knee bones we ran the model built by Thomson *et al.* and described in [113]. They used 500 images sampled from OAI dataset [72] to train a 74-point shape model, global searcher, and the set of increasing resolution RFCLMs to find the contour of tibia and femur (green points in Figure 5.11).

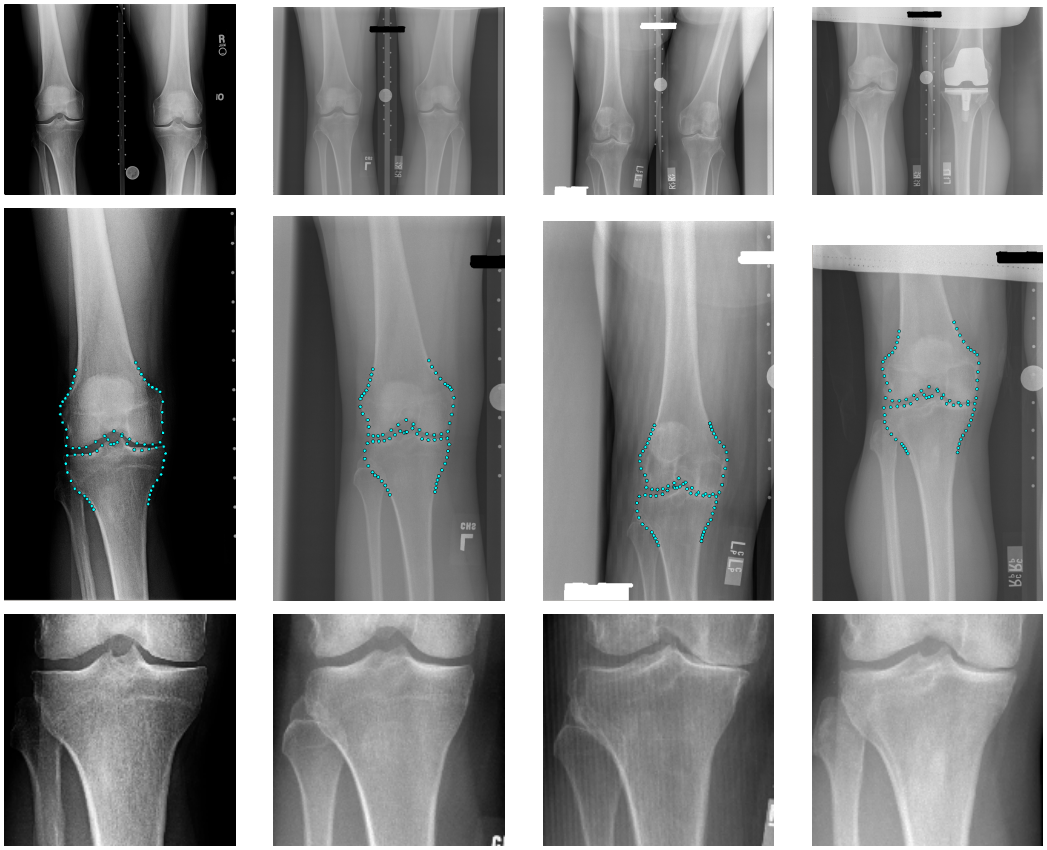


Figure 5.11: Example of PA bilateral Knee radiographs (upper row), and after RFRV localization and RFCLM segmentation (middle row), with their cropped patches after registration (lower row) containing left tibia of (a,b) a non-OA class , (c,d) OA class.

5.2.3 Methods

Patch Preparation

We chose to do registration using tibia model points only to ensure positioning the tibia in roughly same coordinates across all patches in the new coordinate frame. This way we help the CNN to incorporate joint mal-alignment and joint space narrowing associated with OA by fixing position of the tibia top surface. We tried patches of size: 121x121, and 151x151 (see Figure 5.12) corresponding to Algorithm 4's parameters: f_h=111, b= 5 and 20 respectively. The whole process is completely automatic. We then use the resulting patches to train and test a CNN.

(a) 121x121 (b) 151x151

Figure 5.12: PA knee patches.

Network Architecture

We trained two CNNs: one for binary classification and one for multi-class KL grade selection. The two networks are a classical stack of CP layers (CP: one ReLU-activated Convolutional layer followed by a Pooling layer) with two consecutive fully-connected layers. The output layer differs in the two networks. Architecture details are summarised in Table 5.7. No padding was used. Weights were initialised with Xavier

uniform kernel initializer [45] and biases initialised to zeros. The loss function was the binary cross-entropy for binary OA classification. For multi-class OA classification we used weighted categorical cross-entropy to overcome class imbalance. The weight assigned to a class in the weighted loss function is the ratio between the number of samples of the majority class to number of samples of that class in training set. Loss functions were optimized with Adam [67] with default parameter values. Models were developed using Tensorflow 1.0.

Layer	Type	Maps	Map Size	Kernel Size	Stride	Activation
In	Input	1	151x151	-	-	-
CP1	Convolution	32	149x149	3x3	1	ReLU
	Max Pooling 2D	32	74x74	2x2	2	-
CP2	Convolution	64	72x72	3x3	1	ReLU
	Max Pooling 2D	64	36x36	2x2	2	-
CP3	Convolution	128	34x34	3x3	1	ReLU
	Max Pooling 2D	128	17x17	2x2	2	-
CP4	Convolution	256	15x15	3x3	1	ReLU
	Max Pooling 2D	256	7x7	2x2	2	-
CP5	Convolution	512	5x5	3x3	1	ReLU
	Max Pooling 2D	512	2x2	2x2	2	-
FC1	Fully Connected	1	250	-	-	ReLU
D	Dropout (rate=0.5)	-	-	-	-	-
FC2 2-class Output	Fully Connected	-	1	-	-	Sigmoid
FC2 5-class Output	Fully Connected	-	5	-	-	Softmax

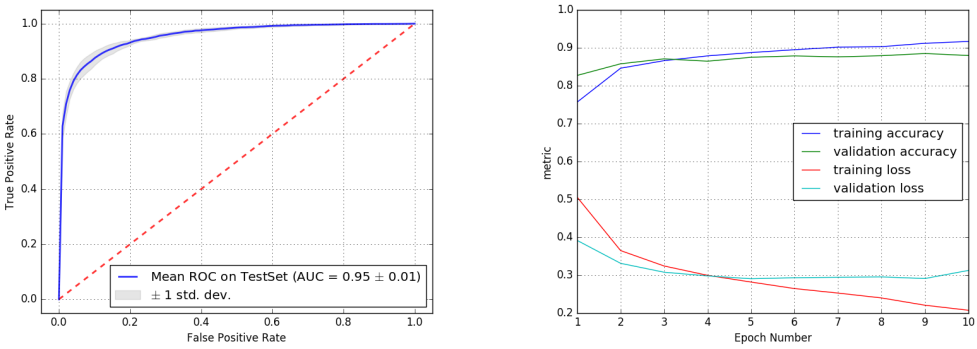
Table 5.7: The overall architecture detailed with maps sizes corresponding to an input knee patch of size 151x151. The same architecture were used with 121x121. Different output layers are used for different OA classification problems (i.e. binary or multi-class). In our experiments we gradually increased the number of CP layers and chose the one with the best performance.

5.2.4 Experiments and Results

We carried out 5-fold cross validation experiments, in which all images belonging to one patient were kept in the same fold. The fold-out subset was split into equal sized validation and testing subsets. The training process was repeated after swapping validation and testing subsets so that the whole dataset was tested. Every time the 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. The performances of different networks on the binary classification task are summerised in Table 5.9. Networks NW2 and NW3 performed the same. The performance, in general, was better on patch size 151x151 compared to 121x121. This might be because more joint area appears in 151x151 (see Figure 5.12). Figure 5.13 shows the average performance of AUC=0.95 (std 0.01) for NW3 on patches of size 151x151 and an example of learning curves during training a model.

Table 5.8: The performance of different networks on the task of OA vs Non-OA Classification in terms of average AUC $\pm$ stdev for different patch sizes.

Network	Architecture	121x121	151x151
NW1	CP1-CP2-FC1-D-FC2	0.90$\pm$0.13	0.90$\pm$0.13
NW2	CP1-CP2-CP3-FC1-D-FC2	0.94$\pm$0.01	**0.95$\pm$0.01**
NW3	CP1-CP2-CP3-CP4-FC1-D-FC2	0.94$\pm$0.01	**0.95$\pm$0.01**
NW4	CP1-CP2-CP3-CP4-CP5-FC1-D-FC2	0.92$\pm$0.01	0.94$\pm$0.01



(a) (b)

Figure 5.13: OA vs Non-OA Classification. (a) Mean ROC curve with AUC= 0.95 (std 0.01) (b) Example Learning curves during training a model.

For multi-class classification, we chose the networks NW2 and NW3, for their performance on the binary classification task, to carry on multi-class experiments with different patch sizes (see results in Table 5.10). The results suggest that there were no significant differences in the performance of the two networks and both perform better on 151x151.

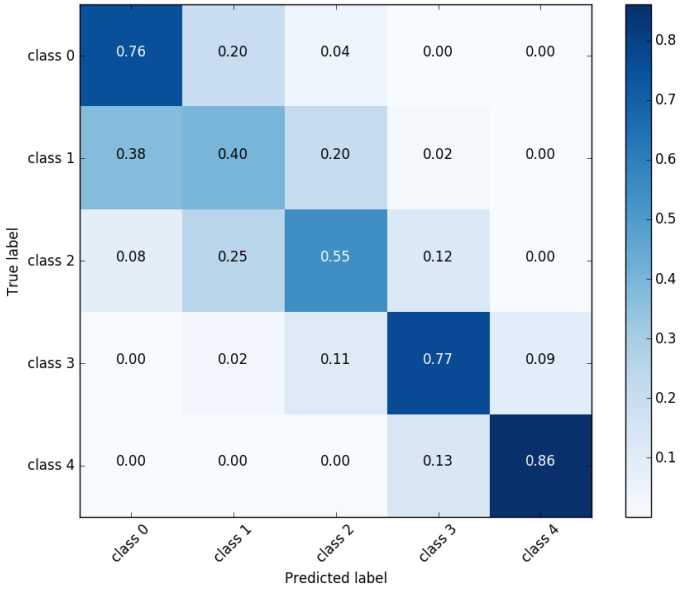
Table 5.9: The performance of models NW2 and NW3 on different patch sizes in terms of average AUC ± stdev and average multi-class accuracy%± stdev.

	NW2 121x121	NW2 151x151	NW3 121x121	NW3 151x151
AUC	0.94±0.01	0.95±0.01	0.94±0.01	**0.95±0.01**
Multi-class Accuracy%	64.8±5.4	66.0±5.0	63.4±6.0	**66.8**±7.0

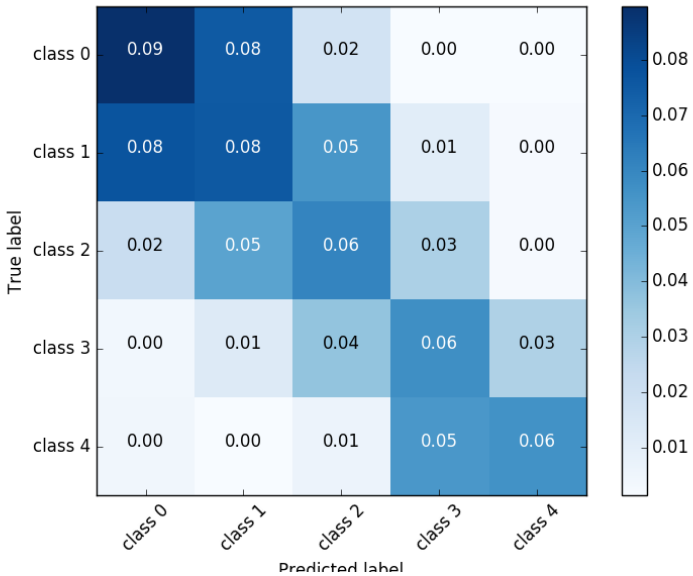
To report further on the multi-class task, we chose to select the network NW3 on patch size 151x151. These settings achieved an average multiclass accuracy of 66.8% with corresponding quadratic Kappa coefficient of 0.89 and mean squared error (MSE) of 0.46. Table 5.10 and Figure 5.14 show more performance metrics and average confusion matrix with standard deviations for these settings.

Class	precision	recall	F1 score
KL0	0.81	0.76	0.78
KL1	0.32	0.40	0.36
KL2	0.59	0.55	0.57
KL3	0.80	0.77	0.78
KL4	0.80	0.86	0.83
Mean	0.69	0.68	0.68

Table 5.10: Performance metrics for multi-class classification.



(a) Mean



(b) Standard Deviations

Figure 5.14: Average confusion matrix and standard deviations for multi-class classification. Average accuracy is 66.8% with quadratic Kappa coefficient= 0.89 and MSE= 0.46.

5.2.5 Discussion

We presented a system for automatic knee OA diagnosis from plain PA X-rays following the same methodology we developed initially for wrist fracture detection. Our system achieved an average AUC of 0.95±0.01 on the task of OA vs Non-OA Classification, and multi-class classification accuracy of 66.8% compared to the current state-of-art (SOA) [117] with AUC of 0.93 and accuracy of 66.7%. We also achieved a lower MSE (0.46) and higher quadratic Kappa (0.89) compared to MSE=0.48 and quadratic Kappa=0.83 for SOA. However, we used MOST dataset [39] in cross-validation experiments whereas the SOA [117] used MOST dataset [39] (18,376 images) for training and OAI dataset [72] for validation (2,957 images) and testing (5,960 images).

Comparing our results with methods by Antony *et al.* in [3] for they used a training set of 5,166 radiographs and testing set of 2,200 radiographs (for each fold we test on about 2,000 radiographs), both sets were collected from MOST dataset [39] and OAI dataset [72]. They reported lower performance than ours with a multi-class accuracy of 61.9% and MSE of 0.66. Other work in [83] extracted shape and appearance features from both PA and LAT knee views and trained random forest classifiers on. They reported a lower performance than ours with an average AUC of 0.85, 0.90, and 0.91 for PA, LAT, and both views combined respectively in 5-fold cross-validation experiments on a dataset collected from MOST dataset [39] with (4,628 OA images and 6,805 non-OA images). Table 5.11 summarises the results on the task of OA automated diagnosis.

Author	Dataset	AUC	Class accuracy%	MSE	Quadratic Kappa
Ours	5-fold CV on 19208 PA radiographs from [39]	**0.95**	**66.8**	**0.46**	**0.89**
SOA[117]	Trained on (18376 PA radiographs) from [39] validated on (2957 radiographs) from [72] Tested on (5960 radiographs) from [72]	0.93	66.7	0.48	0.83
[3]	Trained on (5166 PA radiographs) from [39] [72] Tested on (2200 radiographs) from [39] [72]	-	61.9	0.66	-
[83]	5-fold CV on 11433 pairs of PA and LAT radiographs from [39]	0.91	-	-	-

Table 5.11: Summary of the relevant results on the task of automated OA diagnosis from knee radiographs.

Chapter 6

Shape-Specific Local Models For Overlapping Structures

Accurately segmenting the outline of bones in radiographs can be challenging because structures often overlap, which causes significant variation in local appearance. The degree of overlap depends on the relative shape and position of the different bones which can be compactly encoded using a statistical shape model. Many shape matching techniques (e.g. Active Shape Models, Constrained Local Models) use a single shape model, together with one local model for each point which assumes that the appearance around each point is either independent of that around its neighbors or linearly related. This assumption is broken when two bones are superimposed. In this case there is a (non-linear) relationship between the local appearance and the overall shape, which can degrade overall performance of the above-mentioned approaches. In this chapter we show that using different local models depending on the global shape leads to significant improvements in accuracy and robustness when segmenting (i) the radius and ulna in radiographs of the wrist, and (ii) femoral condyles in lateral knee radiographs. Because of the inability to produce a consistent manual annotation for the ulna in LAT radiographs when fractures were presented (discussed in Section 4.2) the work in this chapter was only tested on non-fractured wrists.

6.1 Shape-specific local models in RFCLM framework

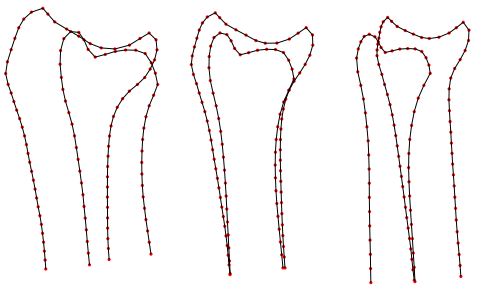
The patch around a model point can change significantly if there are overlapping structures. For instance, Figure 6.1 shows some examples of different relative positions of two bones (the radius and the ulna) in clinical lateral wrist radiographs (see also Figure 6.5). A local model trained on examples where the ulna is to the left will not work well on images for which it is to the right, and vice-versa. If all examples are included in the training set the local models (one per point) will not be able to deal well with any particular case. We overcome this limitation by building different sets of local models. Each set corresponds to a certain alignment of the overlapping structures and contains one local model per feature point.

(a) (b) (c) (d) (e) (f)

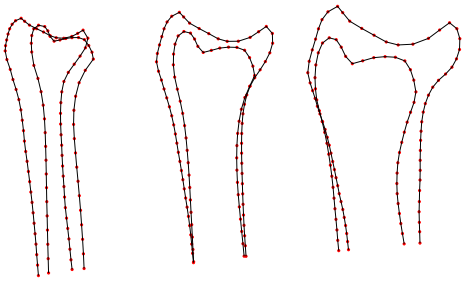
Figure 6.1: The variability of radius-ulna positions in lateral wrist radiographs.

6.1.1 Model Building

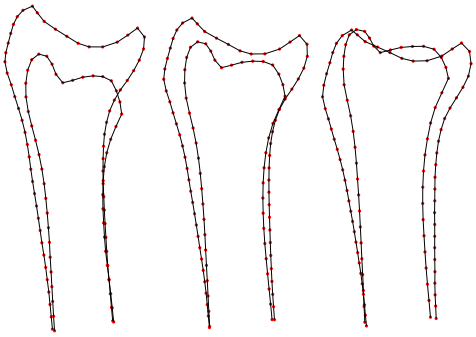
Figure 6.2 shows the effect of varying the first three parameters (b_1, b_2, and b_3) of a shape model built for the radius and ulna in lateral wrist view. The first shape mode b_1 explains most of the relative positioning as the ulna moves from right to left. The third mode b_3 shows a vertical displacement of the ulna compared to the radius. Given a training set of annotated images with n feature points, we first build a sin-



(a) Mode 1 ($\pm 3\sigma_1$)



(b) Mode 2 ($\pm 3\sigma_2$)



(c) Mode 3 ($\pm 3\sigma_3$)

Figure 6.2: Effect of varying each of the first three shape parameters of lateral wrist model.

gle global statistical shape model. We then divide the training images into m training subsets according to the values of shape parameters that most describe the target alignments (hence "shape-specific"). Within a training subset k we learn a local model $F_{j,k}$ for each feature point j. Thus each feature point j will have m different

RF regressors (i.e. $F_{j,k}$ where $j = 1,..,n$ and $k = 1,..,m$). The RFCLM fitting algorithm is then modified so that it selects the most appropriate set k of RF regressors depending on the current shape instance **b** during fitting. We refer to the new algortithm as Shape-Specific RFCLM (SSRFCLM) and it will be explained in Section 6.1.4. As an example of building the local models, if the partitioning of three training subsets was done on the value of the first shape parameter b_1 (i.e. m = 3) then we have three regions (i.e. R_k, k=1,..,3 as in Figure 6.3). During training, a dataset of N images, L= { I_1,...,I_N} with its corresponding annotations X= { X_1,...,X_N} will be split into three subsets L_1, L_2, and L_3 defined as:

$$\begin{aligned} L_1 &= \{I_i | I_i \in L, b_{i,1} \leq p\} \\ L_2 &= \{I_i | I_i \in L, p < b_{i,1} < q\} \\ L_3 &= \{I_i | I_i \in L, b_{i,1} \geq q\} \end{aligned} \tag{6.1}$$

where p and q are chosen so that the subsets are roughly equal in size. Each set of local models will be built from images laying in the same region, for example, the first set of local models (i.e. $\{F_{l,1}\}$, $l = 1,..,n$) are built from images $\{I_i | I_i \in L_1\}$. This way there will be three different sets of local models. Multi-dimensional shape space partitioning (i.e. considering more than one shape parameter) could be applied in case of large datasets containing overlapping structures with a wide range of different alignments.

6.1.2 Model Initialisation

As explained in Section 3.3.2, the standard RFCLM algorithm needs initial estimates of shape parameters **b** and object pose θ to start the search. The initial estimate **b** is set to the mean shape (i.e. **b = 0**) and used with feature point initialisations to estimate the initial pose. Initialisation are passed from an object detector or a previous

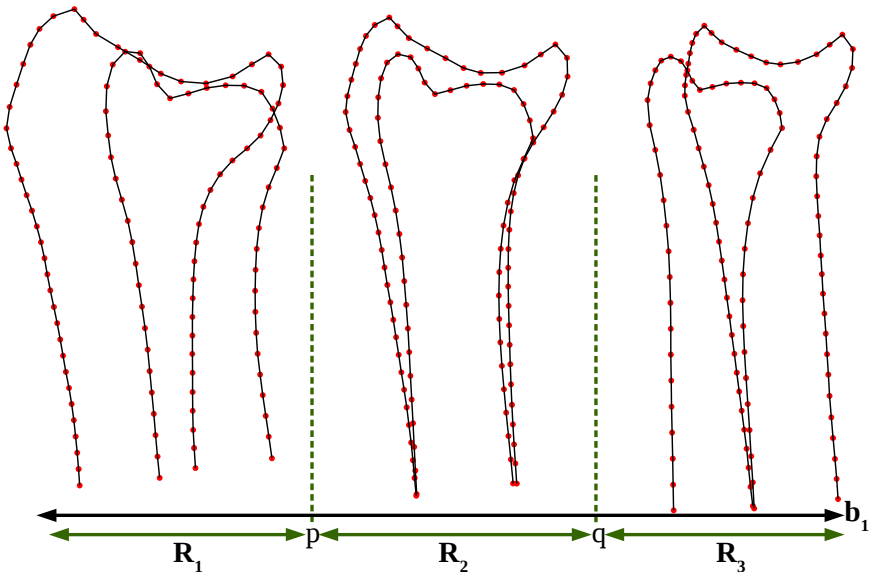


Figure 6.3: Lateral wrist shape space divided into three regions: R_1, R_2, and R_3 depending on the value of the first shape parameter b_1. Different sets of local models are trained from images lying in different regions. The values of p and q are set after inspecting the distribution of the training dataset.

model (in case of multi-stage model). However, initialisation from the mean does not always work well for SSRFCLM. Starting from the mean implies the use of regressors built only from examples very similar to the mean. Such regressors will not work well if the actual shape was away from the mean as that suggests a dramatic change in point appearance in case of overlapping objects. In other words, if the ulna was to the left or to right (See Figure 6.1) then local models built only from cases when the ulna is centered (i.e. mean shape) would not work well. We used a new initialisation scheme which uses multiple start shapes. We perform the search for each start shape separately and then select the result with the best quality of fit as explained in the coming sections. The start shapes are chosen to trigger a different set of local models at the start of each run. Each set of local models should be triggered by at least one initialisation. This can be best captured by the use of shape parameters' standard deviations calculated from the training set. So in the case of our example of partitioning

on the value of b_1, we use initialisations b_1 = {-2σ_1, -1σ_1, 0, 1σ_1, 2σ_1 } where σ_1 is the standard deviation of b_1.

6.1.3 Model Comparing

When comparing the results of different searchers/runs we define the Quality Of Fit (QoF) measure of a searcher as:

$$QoF(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} |F_j(\mathbf{x}_j)| \tag{6.2}$$

where $F_j(\mathbf{x}_j)$ is the result of applying the RF for point j to a patch centered at x_j and QoF is the mean displacement to the best individual point estimates from the current estimates. The lower the value of $QoF(\mathbf{x})$ the better the fit the searcher found. In the case of SSRFCLM where there are different sets of local models, the appropriate set is chosen according to the value of b_1 as

$$\begin{aligned} &SS\_QoF(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} |F_{j,k}(\mathbf{x}_j)| \\ &s.t.\ b_1 \in R_k \end{aligned} \tag{6.3}$$

6.1.4 Model Matching

The steps for one iteration of the SSRFCLM are given in Figure 6.4. The main developments over given in [24, 75] are: a) the use of multiple local models (as in [93]), and b) the use of multiple initialisations. So as in RFCLM, starting from the initial values of shape parameters **b** and object pose θ the region of interest is sampled to a reference frame of width fw. In the reference frame, the local models $\{F_{j,k}\}$ are chosen such that current value of $b_1 \in R_k$ and used to search an area around each feature point separately. The search area is within the range of [-d_{search}, +d_{search}] in x and y.

Data: $\mathbf{I}$,$\mathbf{b}$,θ,$model$,$\{F_{j,k}\}$

Result: $\mathbf{x}$, $cost$

Function `SearchIteration`($\mathbf{I}$,$\mathbf{b}$,θ,$model$,$\{F_{j,k}\}$)**:**

$k \leftarrow findRegionOf(\mathbf{b})$

foreach *landmark* j **do**

$\mathbf{S} \leftarrow$`sampleSearchAreaOfLandmark`($\mathbf{I}$, j)

$\mathbf{V}_j \leftarrow$ `getResponseImage`($\mathbf{S}$, $F_{j,k}$)

end

$\mathbb{V} \leftarrow \bigcup_{j=1}^{n} \mathbf{V}_j$

$\mathbf{x} \leftarrow$ fitModelToResponseImages($\mathbb{V}$,$model$)

$cost \leftarrow$ SS_QoF($\mathbf{x}$)

update the points in the image frame as:

$\mathbf{x} \leftarrow T_{\theta}^{-1}(\mathbf{x})$

return $\mathbf{x}$, $cost$;

End Function

Figure 6.4: One SSRFCLM search iteration for image $\mathbf{I}$ starting from shape and pose parameters ($\mathbf{b}$,θ) with aid of statistical shape model ($model$) and local models ($\{F_{j,k}\}$). SSRFCLM's search iterations are different from those of RFCLM (see Algorithm 3) although the main iterative nature and the function $fitModelToResponseImages()$ are the same.

At each iteration a shape model instance is fit to the best points from the resulting response images and therefore new values for $\mathbf{b}$ and θ are found. If new value of $\mathbf{b}$ failed the condition of $\mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} \leqslant M_t$, it will be moved to the closest point on the limiting ellipsoid. The new values of $\mathbf{b}$ and θ are used to calculate new estimate of $\mathbf{x}$ and its SS_QoF and to initiate a new search iteration. The values of $\mathbf{b}$ and θ corresponding to the iteration with the best SS_QoF (i.e. lowest cost) will be returned as the result of the search run. In case of more than one search run for the same model (i.e.

multiple initialisations discussed early in Section 6.1.2) the run with the best *SS_QoF* will be chosen.

6.2 Experiments

6.2.1 Data

To assess the performance of our method we carried out experiments on five different datasets: Two LAT Wrist datasets containing lateral view wrist radiographs, two PA Wrist datasets containing PosteroAnterior view wrist radiographs, and one LAT Knee dataset containing lateral view knee radiographs.

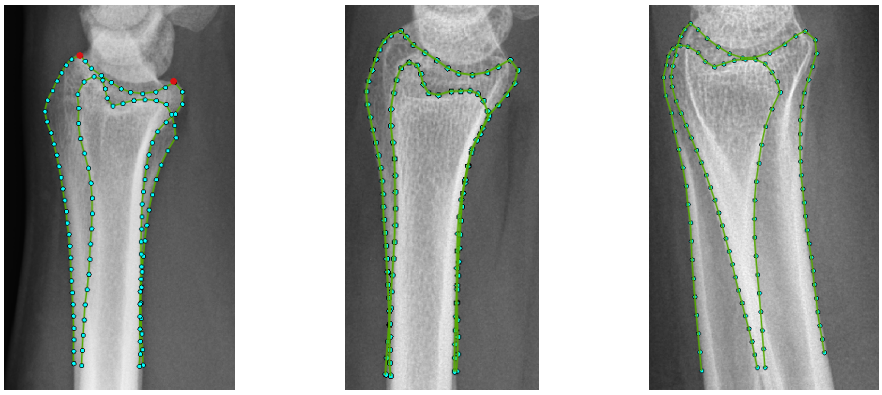
The four wrist datasets are from two local EDs gathered and anonymised by a clinician (i.e. one PA dataset and one LAT dataset per ED). All radiographs do not contain fractures. The LAT Knee dataset is sampled from the OsteoArthritis Initiative (OAI) [72] dataset. OAI is an observational prospective study of OA, taking participants across four sites across USA. The study began with 4796 participants, ranging between the ages of 45-79. Table 6.1 shows the size of each dataset.

Table 6.1: The used datasets' sizes (number of radiographs).

View	Set 1	Set 2	Number Of feature points
LAT Knee	434	-	49
PA Wrist	218	210	93
LAT Wrist	201	208	112

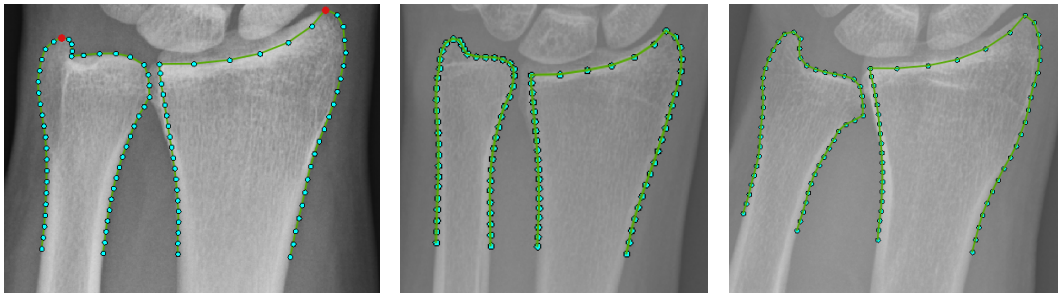
6.2.2 Manual Annotations

Anatomical landmarks and contour points of the targeted bones (see Table 6.1) are manually annotated in all radiographs. Figures 6.5, 6.6, and 6.7 show different annotation examples. Manual annotations are used to train models and also as ground-truth for testing.



(a) (b) (c)

Figure 6.5: The 112-point annotation of radius and ulna in wrist LAT view. The two red points define the reference length used to give the mean error as a percentage of the LAT wrist width and they are the two anatomical points contained in a box to be found by a global searcher.



(a) (b) (c)

Figure 6.6: The 93-point annotation of radius and ulna in wrist PA view. The two red points define the reference length used to give the mean error as a percentage of the PA wrist width and they are the two anatomical points contained in a box to be found by a global searcher.

6.2.3 Methodology

For each experiment we compare the accuracy of the automatic annotation found by our method with that of the original RFCLM in terms of the average point-to-curve error as a percentage of a reference length. In order to generate the automatic annotation for the LAT knee dataset, we divided it into two subsets, training models on the first subset and testing them on the second and vice versa. Parameters were optimised

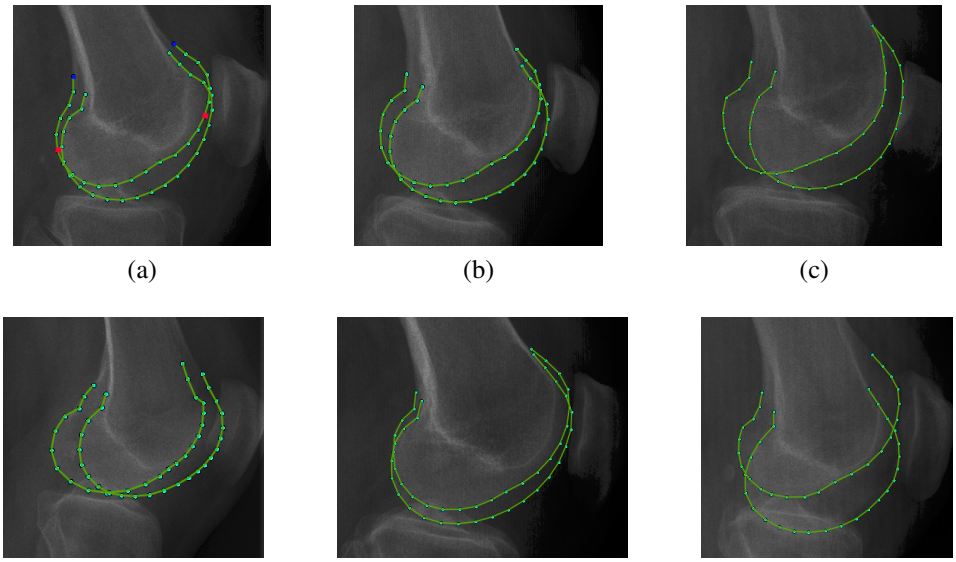


Figure 6.7: Examples of the two condyles in Knee LAT view annotated with 49 feature points. The two blue points define the reference length used to give the mean error as a percentage of the knee width. The two red points are the two anatomical points contained in a box that is found by a global searcher.

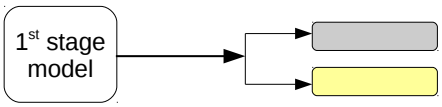
for RFCLM models following [75]. We used the same parameters for the corresponding SSRFCLM models. As for a wrist dataset, we do not divide it into subsets but rather we use the other dataset collected from the other ED and vice versa. To help the built models start from a reasonable initialisation we use global searchers. Global searchers provide the same initialisations to all models so that the differences in performance are a result of differences between local searches not between initialisations. Because of the relatively small size of the datasets we chose the number m of SSRFCLM subsets to be 3 and assigned the examples of a dataset, during training, to different subsets according to the value range of the first shape parameter b_1. Clearly the second and the third shape parameters (i.e. b_2 and b_3) contain some position information (see Figure 6.2) but because of the limited size of datasets we decided to only consider b_1 in splitting subsets. The initializations (i.e. each corresponding to a complete run) of SSRFCLM as discussed in Section 6.1.2 are chosen to be done in

three different schemes:

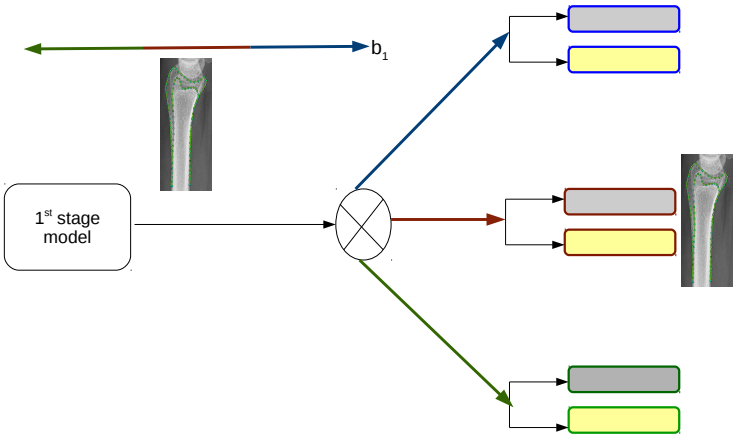
- 0σ-SSRFCLM: performs only one initialisation as the standard RFCLM does (i.e. **b** = **0**).
- 1σ-SSRFCLM : performs three different initialisations with $b_1 = \{-\sigma, 0, \sigma\}$ where σ is the standard deviation of b_1.
- 2σ-SSRFCLM: performs five different initialisations with $b_1 = \{-2\sigma, -\sigma, 0, \sigma, 2\sigma\}$ where σ is the standard deviation of b_1.

These different initialisations only consider different values for b_1 but not the other elements of **b**. This is in accordance with our previously-discussed design choice of assigning training examples to subsets based on the value of b_1 only.
Although the original RFCLM is only initialised at the mean (i.e. **b** = **0**) we applied the same above-mentioned initialisation scheme to result in: 0σ-RFCLM (i.e. original RFCLM), 1σ-RFCLM, and 2σ-RFCLM. This is to ensure that any performance differences would be only due to the algorithmic differences between RFCLM and SS-RFCLM not due to initialisation differences. We also investigated two-stage models where the first stage (modeling the two bones together) is either RFCLM, SSRFCLM or any of their variants followed by a second stage with separate refining models: one per bone. The aim of this multi-stage modeling is to check whether performance differences achieved by different single-stage models will persist when extended to multi-stage. Because SSRFCLM models two bones together it can not be used in the 2^{nd} stage where each bone is refined separately. For this reason, we either use RF-CLM or a new variant of it we call *Switched-RFCLM*. Switched-RFCLM is used in a second stage following a model of two bones together. It contains a switch and a set of RFCLM models. Depending on the region b_1 belongs to which is found by the 1^{st}-stage model the switch selects the refining model that was originally built from examples belonging to the same region. Figure 6.8(a) shows how separate RFCLM

models are used in the second stage preceded by either an RFCLM or SSRFCLM model while figure 6.8(b) shows the suggested Switched-RFCLM composing of three different RFCLMs per bone, each built from examples within the corresponding region of b_1. During matching the switch forwarded the results of the 1st-stage to the suitable 2nd-stage model depending on the value of b_1.



(a)



(b)

Figure 6.8: Architecture of two-stage models. The joint model (RFCLM, or SSRFCLM) in the 1st stage is followed by either (a) one separate RFCLM model per bone, or (b) three separate RFCLM models per bone, only one is chosen during matching depending on the value of b_1. Boxes filled with same colour model same bone.

6.3 Results and Discussion

6.3.1 Lateral Wrist View

To evaluate the performance of the system to automatically annotate radius and ulna in lateral wrist view we first ran different single-stage models and the results are shown in Table 6.2 and Figure 6.9. For the standard RFCLM we modeled the two bones together and separately (appearing as 0σ-RFCLM and Two-RFCLMs respectively in the results). The results confirm that modeling the bones separately perform poorly (see the performance of the model Two-RFCLMs). The single model encodes the relative positions of the bones, the additional constraints help improve the robustness of the matching. Moreover, the Two-RFCLMs model implies the use of two separate global searchers, one per bone, which is, at least in case of the ulna, is prone to error as it changes its relative position dramatically. 1σ-SSRFCLM and 2σ-SSRFCLM showed significant improvements over original RFCLM and its variants. 2σ-SSRFCLM performs better than 1σ-SSRFCLM as it has two more additional initialisations compared to 1σ-SSRFCLM. So 2σ-SSRFCLM has 1σ-SSRFCLM at its core. The poor performance of the model 0σ-SSRFCLM comes from the fact that it has one initialisation (i.e. mean shape instance) dictating the use of the information captured by only one training subset, as a start point, instead of the whole training dataset. For this reason, 0σ-SSRFCLM should not be used. The results also suggest that further gains can be achieved by RFCLM if it adopted this new initialisations scheme.

Table 6.2: The mean point-to-curve error of different one-stage models as a percentage of the LAT wrist width.

Model	Mean	StdErr.	Median	90%	95%
0σ-RFCLM	3.93	0.13	3.18	7.52	9.30
1σ-RFCLM	3.75	0.12	2.99	7.08	8.85
2σ-RFCLM	3.64	0.12	2.94	7.01	8.53
Two-RFCLMs	7.67	1.36	5.14	10.8	13.4
0σ-SSRFCLM	4.61	0.16	3.78	8.97	10.8
1σ-SSRFCLM	3.15	0.10	2.59	6.02	7.53
2σ-SSRFCLM	**3.02**	0.09	**2.53**	**5.48**	**6.73**

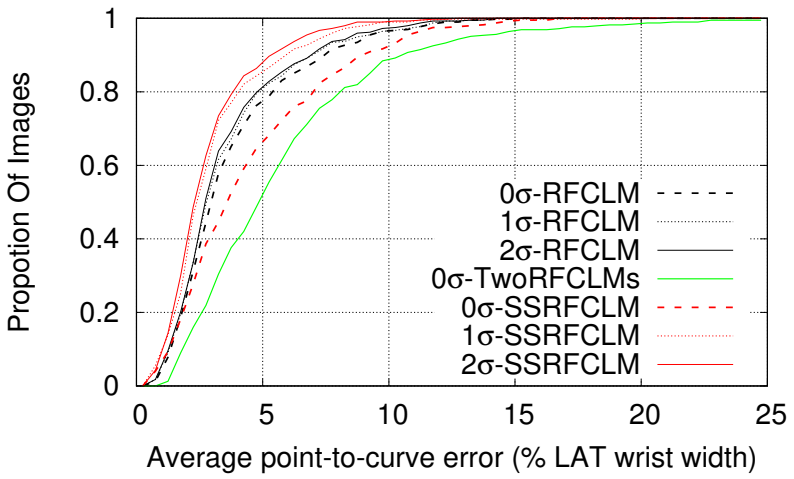


Figure 6.9: Fully automated single-stage search results of lateral wrist radiographs.

We also investigated whether these performance gains would continue to exist if we use two-stage models as described in section 6.2.3 and depicted in Figure 6.8. Table 6.3 and Figure 6.10 show that the refiner (i.e. 2^{nd}-stage model) proceeded by the original RFCLM has the highest error rates. RFCLM when adopted this new multi-initialisations scheme (i.e. 2σ-RFCLM followed by RFCLM refiner) achieved comparable results to that of 2σ-SSRFCLM with either refiner.

Table 6.3: The mean point-to-curve error (% LAT wrist width) obtained automatically by two-stage models.

1st-stage model	2nd-stage model	Mean	StdErr	Median	90%	95%
0σ-RFCLM	RFCLM	3.59	0.12	2.69	7.20	8.49
1σ-RFCLM	RFCLM	3.18	0.12	2.33	6.54	7.70
2σ-RFCLM	RFCLM	3.01	0.11	2.26	6.20	7.53
1σ-SSRFCLM	RFCLM	3.22	0.13	2.33	6.40	8.00
1σ-SSRFCLM	Switched RFCLM	2.92	0.10	2.27	5.59	**6.82**
2σ-SSRFCLM	RFCLM	2.95	0.10	2.29	**5.28**	**6.82**
2σ-SSRFCLM	Switched RFCLM	**2.87**	0.10	**2.16**	5.54	7.11

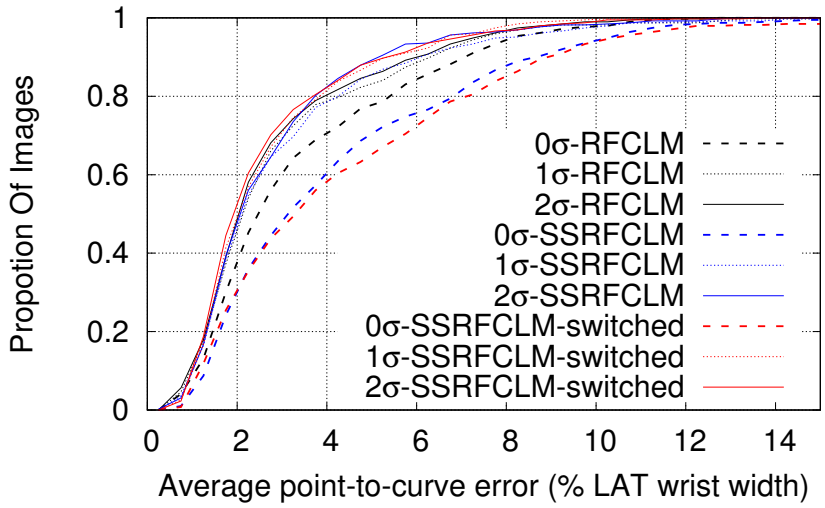


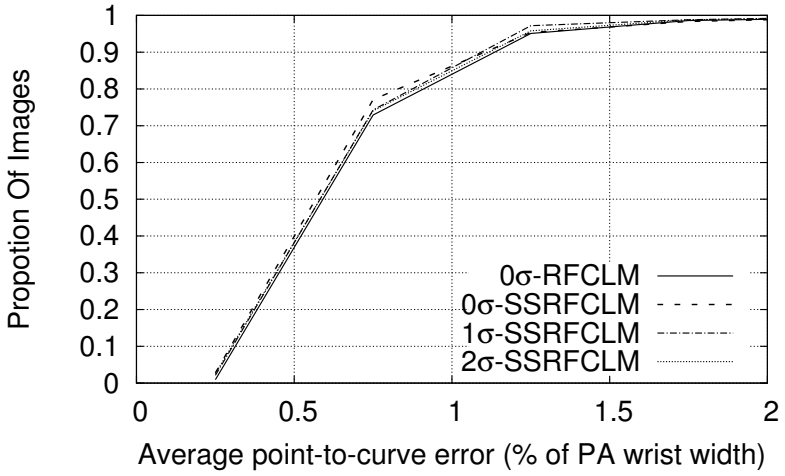
Figure 6.10: Fully automated two-stage search results of lateral wrist radiographs.

6.3.2 PA Wrist View

In this view radius and ulna are not overlapping. However, it was worth testing how different local models would perform in the absence of overlap. Table 6.4 and Figure 6.11 shows that the performance is slightly better than that of the standard RFCLM.

Table 6.4: The mean point-to-curve error (% PA wrist width) obtained automatically by different single-stage models.

Model	Mean	Median	90%	95%
0σ-RFCLM	0.959	0.861	**1.22**	1.48
0σ-SSRFCLM	**0.931**	0.820	1.23	1.48
1σ-SSRFCLM	0.933	**0.816**	1.23	**1.41**
2σ-SSRFCLM	0.943	0.822	1.27	1.47



(a)

Figure 6.11: Fully automated single-stage search results of PA wrist radiographs.

6.3.3 LAT Knee View

Femoral condyles look very similar and largely overlapping in lateral knee view. Table 6.5 and Figure 6.12 show that the model 2σ-SSRFCLM significantly outperforms all other models as half of radiographs with error less than 2.04% compared to 2.31% for original RFCLM (appears as 0σ-RFCLM) and 95% of radiographs with error less than 4.35% compared to 5.07% for original RFCLM.

Table 6.5: The mean point-to-curve error (% LAT Knee width) obtained automatically by different single-stage models.

Model	Mean	StdErr	Median	90%	95%
0σ-RFCLM	2.67	0.06	2.31	4.14	5.07
1σ-RFCLM	2.61	0.06	2.20	3.92	5.13
2σ-RFCLM	2.60	0.06	2.24	3.86	5.18
0σ-TwoRFCLMs	2.76	0.06	2.34	4.35	5.34
0σ-SSRFCLM	3.04	0.07	2.59	5.07	6.20
1σ-SSRFCLM	2.40	0.06	2.05	3.70	4.46
2σ-SSRFCLM	**2.35**	0.06	**2.04**	**3.56**	**4.35**

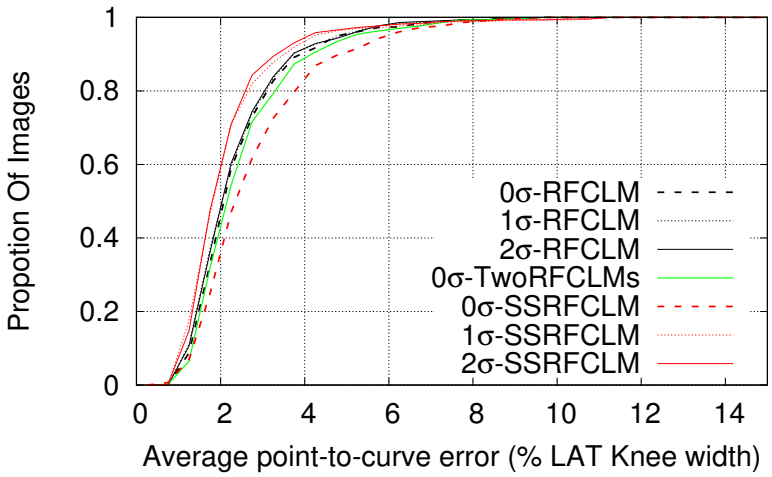


Figure 6.12: Fully automated single-stage search results of lateral knee radiographs.

Even when followed by a switched-RFCLM refiner, 2σ-SSRFCLM continued to outperform all other two-stage models as shown in Table 6.6 and Figure 6.13. It is worth noting that 2σ-SSRFCLM when followed by an RFCLM refiner showed almost the same performance as that of RFCLMs. This suggests that a single set of local models in later stages could deteriorate the gains achieved in previous stages where multi-sets are used.

Table 6.6: The mean point-to-curve error (% LAT knee width) obtained automatically by two-stage models.

1st-stage model	2nd-stage model	Mean	StdErr	Median	90%	95%
0σ-RFCLM	RFCLM	2.10	0.06	1.64	3.47	5.26
1σ-RFCLM	RFCLM	2.09	0.06	1.62	3.41	5.29
2σ-RFCLM	RFCLM	2.09	0.07	1.62	3.40	5.26
0σ-TwoRFCLMs	RFCLM	2.09	0.07	1.65	3.38	5.27
1σ-SSRFCLM	RFCLM	2.07	0.06	1.63	3.39	5.05
1σ–SSRFCLM	Switched RFCLM	1.64	0.06	**1.29**	2.60	3.20
2σ-SSRFCLM	RFCLM	2.08	0.06	1.62	3.36	5.29
2σ-SSRFCLM	Switched RFCLM	**1.61**	0.05	1.30	**2.51**	**3.09**

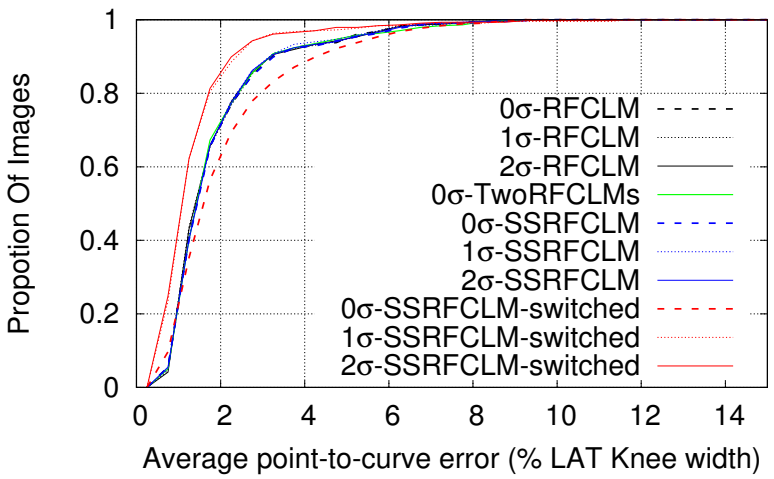


Figure 6.13: Fully automated two-stage search results of lateral knee radiographs. The model type of the 1st stage appears on the graph, the 2nd stage is RFCLM unless stated *switched*.

6.4 Discussion

We presented a system to segment overlapping bones in radiographs which is fully automatic and does not make any assumptions about the relative positions and degree of overlap. The algorithm SSRFCLM is an extended form of RFCLM. It has a set of local models for each point and switches between them depending on the global shape (i.e. current positioning). The idea of switching between different sets of feature point model has been used previously in [93] for facial point tracking over a wide range of head angles on the task of tracking driver faces accurately. They reported a marked speedup as a result of the reduced decision trees sizes and therefore the specificity of local models. We claim the novelty of the adaptation of their idea for solving a different problem. We investigated different initialisation schemes which showed to be effective even in improving original RFCLM performance. We carried experiments on single-stage models and two-stage models. We introduced an RFCLM-based refiner which has different RFCLMs, each built from examples belonging to certain region of shape space. One of them is selected depending on the result passed from the previous-stage model. This new type of refiner showed to work well with the SSRFCLM. SSRFCLM, as an adapted form of RFCLM, led to significant improvements in accuracy and robustness above the state-of-the-art technique RFCLM when segmenting the radius and ulna in wrist radiographs, and femoral condyles in lateral knee radiographs. There is a growing trend to use CNNs methods to locate points. Recent experiments on analysing bones in radiographs [33] show that a well-designed RFCLM performs better than CNN-based systems on datasets smaller than a thousand training examples.

Chapter 7

Conclusions and Future Work

The main contribution of this PhD project is the development of the first fully automated system to detect wrist fractures from the two standard wrist views (i.e. PA and LAT). For each view, a global search based on Random Forest Regression Voting (RFRV) was performed to find the approximate position of the radius. The detailed outline of the bone was then located using a Random Forest Regression Voting Constrained Local Model (RFCLM). Convolutional neural networks were trained from scratch on cropped patches containing the region of interest on the task of detecting fractures. The decisions from the two views were averaged for better performance. Based on our literature research, the system achieves the best yet published detection rate, with an Area Under the ROC Curve (AUC) of 0.93 from LAT view, of 0.95 from PA view, and of 0.96 from both views combined. Future work will be extending this approach to detect fractures in other skeletal structures or to generate automatic descriptions of the found fracture (i.e. fracture classification according the Muller AO classification). In addition, future methodological development will be aimed at modeling normality, for example by training auto-encoders and use them to extract discriminating features from trabacular patches for automated abnormality localisation.

Such systems could highlight the part which is considered as abnormal and bring it to the attention of a clinician.

We also have shown that the underlying methodology of the developed fully automatic system generalises well to solve the problem of knee osteoarthritis (OA) diagnosis from PA knee radiographs. It achieved a strong OA classification rate with an AUC of 0.95, multi-class accuracy of 66.8%, and weighted kappa of 0.89 on the MOST dataset [39]. The next step is to study how well the model performs using other datasets such as the OAI dataset [72]. Besides predicting current OA from PA radiographs, future work will extend the use of same methodology to cover the skyline and lateral knee X-rays for predicting the current pain, later onset OA, and later onset pain. Future work could also investigate training one CNN taking more than one view as an input at a time and minimise a joint loss function instead of averaging the outputs of many CNNs, each trained on a view.

This project proposed work on improving the state-of-art technique Random Forest Regression Voting Constrained Local Model (RFCLM) [24] in order to perform better on overlapping structures in lateral radiographs. The RFCLM uses a single local model (Random Forest regressor) for each point. Each local model votes for a point position, and the results of all the models are integrated using the statistical shape model. Since the radiograph is a 2D projection of a 3D object different structures may be superimposed, and the local appearance around a point on one bone may change dramatically because of the limited constraints on how a patient is positioned. Our results suggest that further gains can be achieved by RFCLM if it uses a multi-initialisation scheme or shape-specific local models (SSRFCLM). We showed that SSRFCLM, as an adapted form of RFCLM, led to significant improvements in accuracy and robustness above the state-of-the-art technique RFCLM when segmenting the radius and ulna in wrist radiographs, and femoral condyles in lateral knee radiographs. We also introduced an RFCLM-based refiner, which has different RFCLMs,

each built from examples belonging to certain region of shape space. During search only one of them is selected depending on the result passed from the previous-stage model. This new type of refiners showed to work well with SSRFCLM for some applications.

Bibliography

[1] H. Abdi and L. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[2] D. J. Andersen, W. F. Blair, C. M. Stevers, B. D. Adams, G. Y. El-Khouri, and E. A. Brandser. Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. *Journal of Hand Surgery*, 21(4):574–582, 1996.

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

[4] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *Proc. 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE, 2016.

[5] L. Audigé, M. Bhandari, B. Hanson, and J. Kellam. A concept for the validation of fracture classifications. *Journal of orthopaedic trauma*, 19(6):404–409, 2005.

[6] O. Bandyopadhyay, A. Biswas, and B. Bhattacharya. Classification of long-bone fractures based on digital-geometric analysis of X-ray images. *Pattern Recognition and Image Analysis*, 26(4):742–757, 2016.

[7] O. Bandyopadhyay, A. Biswas, and B. B. Bhattacharya. Long-bone fracture detection in digital X-ray images based on concavity index. In *Proc. International Workshop on Combinatorial Image Analysis*, pages 212–223. Springer, 2014.

[8] O. Bandyopadhyay, A. Biswas, and B. B. Bhattacharya. Long-bone fracture detection in digital X-ray images based on digital-geometric techniques. *Computer methods and programs in biomedicine*, 123:2–14, 2016.

[9] O. Bandyopadhyay, A. Biswas, B. Chanda, and B. B. Bhattacharya. Bone contour tracing in digital X-ray images based on adaptive thresholding. In *Proc. International Conference on Pattern Recognition and Machine Intelligence*, pages 465–473. Springer, 2013.

[10] O. Bandyopadhyay, B. Chanda, and B. B. Bhattacharya. Entropy-based automatic segmentation of bones in digital X-ray images. In *Proc. International Conference on Pattern Recognition and Machine Intelligence*, pages 122–129. Springer, 2011.

[11] F. Bayram and M. Çakirolu. DIFFRACT: DIaphyseal Femur FRActure Classifier SysTem. *Biocybernetics and Biomedical Engineering*, 36(1):157–171, 2016.

[12] J. Benger and I. Lyburn. What is the effect of reporting all emergency department radiographs? *Emergency Medicine Journal*, 20(1):40–43, 2003.

[13] L. Berman, G. de Lacey, E. Twomey, B. Twomey, T. Welch, and R. Eban. Reducing errors in the accident department: a simple method using radiographers. *British Medical Journal (Clinical Research Ed.)*, 290(6466):421–422, 1985.

[14] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[15] P. A. Bromiley, J. E. Adams, and T. F. Cootes. Localisation of Vertebrae on DXA Images using Constrained Local Models with Random Forest Regression Voting. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, volume 20 of *Lecture Notes in Computational Vision and Biomechanics*, pages 159–171. Springer, 2015.

[16] P. A. Bromiley, E. P. Kariki, J. E. Adams, and T. F. Cootes. Classification of osteoporotic vertebral fractures using shape and appearance modelling. In *Proc. International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, volume 10734 of *Lecture Notes in Computer Science*, pages 133–147. Springer, 2018.

[17] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[18] 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 *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 801–805, Apr 2015.

[19] M. Castano-Betancourt, J. B. Van Meurs, S. Bierma-Zeinstra, F. Rivadeneira, A. Hofman, H. Weinans, A. Uitterlinden, and J. Waarsing. The contribution of hip geometry to the prediction of hip osteoarthritis. *Osteoarthritis and cartilage*, 21(10):1530–1536, 2013.

[20] I. Castro-Mateos, J. M. Pozo, T. F. Cootes, J. M. Wilkinson, R. Eastell, and A. F. Frangi. Statistical shape and appearance models in osteoporosis. *Current Osteoporosis Reports*, 12(2):163–173, 2014.

[21] H. Y. Chai, L. K. Wee, T. T. Swee, S.-H. Salleh, A. Ariff, et al. Gray-level co-occurrence matrix bone fracture detection. *American Journal of Applied Sciences*, 8(1):26, 2011.

[22] A. Chen, C. Gupte, K. Akhtar, P. Smith, and J. Cobb. The global economic cost of osteoarthritis: how the UK compares. *Arthritis*, 2012(6), 2012.

[23] J. Coelho and P. Rodrigues. The red dot system: Emergency diagnosis impact and digital radiology implementation - a review. In *Proc. International Conference on Health Informatics (HEALTHINF 2011)*, pages 508–511. SciTePress, 2011.

[24] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *Proc. European Conference on Computer Vision (ECCV)*, volume 37 of *Lecture Notes in Computer Science*, pages 1–14. Springer, 2012.

[25] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995.

[26] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[27] C. M. Court-Brown and B. Caesar. Epidemiology of adult fractures: a review. *Injury*, 37(8):691–697, 2006.

[28] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression forests for efficient anatomy detection and localization in CT studies. In *Proc International MICCAI Workshop on Medical Computer Vision*, volume 6533 of *Lecture Notes in Computer Science*, pages 106–117. Springer, 2010.

[29] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[30] G. R. Cross and A. K. Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 5(1):25–39, Jan 1983.

[31] A. G. Culvenor, C. N. Engen, B. E. Øiestad, L. Engebretsen, and M. A. Risberg. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. *Knee Surgery, Sports Traumatology, Arthroscopy*, 23(12):3532–3539, 2015.

[32] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[33] A. Davison, C. Lindner, D. Perry, W. Luo, and T. F. Cootes. Landmark Localisation in Radiographs Using Weighted Heatmap Displacement Voting. In *the 6th MICCAI Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging (to appear)*.

[34] M. De Bruijne and M. Nielsen. Shape particle filtering for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 3216 of *Lecture Notes in Computer Science*, pages 168–175. Springer, 2004.

[35] G. De Lacey, A. Barker, J. Harper, and B. Wignall. An assessment of the clinical effects of reporting accident and emergency radiographs. *The British Journal of Radiology*, 53(628):304–309, 1980.

[36] L. Donaldson, I. Reckless, S. Scholes, J. S. Mindell, and N. Shelton. The epidemiology of fractures in England. *Journal of Epidemiology & Community Health*, 62(2):174–180, 2008.

[37] M. Donnelley, G. Knowles, and T. Hearn. A cad system for long-bone segmentation and fracture detection. In *Proc. International Conference on Image and Signal Processing (ICISP 2008)*, volume 5099 of *Lecture Notes in Computer Science*, pages 153–162. Springer, 2008.

[38] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[39] D. Felson, J. Niu, T. Neogi, J. Goggins, M. Nevitt, F. Roemer, J. Torner, C. Lewis, and A. Guermazi. Synovitis and the risk of knee osteoarthritis: the MOST study. *Osteoarthritis and Cartilage*, 24(3):458 – 464, 2016.

[40] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[41] Y. N. Fuadah, A. Rizal, and Y. S. Hariyani. Diaphysis fracture on tibia and fibula detection based on digital image processing and scan line algorithm. In *The 15th International Conference on Biomedical Engineering*, IFMBE Proceedings, pages 679–682. Springer, 2014.

[42] M. W. Funk, E. A. El-Kwae, and J. F. Kellam. Toward automated bone fracture

classification. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 755–766. International Society for Optics and Photonics, 2001.

[43] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv preprint arXiv:1711.06504*, 2017.

[44] J. Gall and V. Lempitsky. Class-specific Hough forests for object detection. In *Decision forests for computer vision and medical image analysis*, Advances in Computer Vision and Pattern Recognition, pages 143–157. Springer, 2013.

[45] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, pages 249–256, 2010.

[46] 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.

[47] S. Goodyear, R. Barr, E. McCloskey, S. Alesci, R. Aspden, D. Reid, and J. Gregory. Can we improve the prediction of hip fracture by assessing bone structure using shape and appearance modelling? *Bone*, 53(1):188–193, 2013.

[48] L. Gossec, J. Jordan, S. Mazzuca, M.-A. Lam, M. Suarez-Almazor, J. Renner, M. Lopez-Olivo, G. Hawker, M. Dougados, and J. Maillefert. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI–OMERACT task force. *Osteoarthritis and cartilage*, 16(7):742–748, 2008.

[49] J. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[50] M. C. Gratton, J. A. Salomone, and W. A. Watson. Clinically significant radiograph misinterpretations at an emergency medicine residency program. *Annals of Emergency Medicine*, 19(5):497–502, 1990.

[51] J. S. Gregory, J. H. Waarsing, J. Day, H. A. Pols, M. Reijman, H. Weinans, and R. M. Aspden. Early identification of radiographic osteoarthritis of the hip using an active shape model to quantify changes in bone morphometric features: can hip shape tell us anything about the progression of osteoarthritis? *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 56(11):3634–3643, 2007.

[52] H. Guly. Injuries initially misdiagnosed as sprained wrist (beware the sprained wrist). *Emergency Medicine Journal*, 19(1):41–42, 2002.

[53] J. Guo, X. Mei, and K. Tang. Automatic landmark annotation and dense correspondence registration for 3D human facial images. *BMC bioinformatics*, 14(1):232, 2013.

[54] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[55] M. Hardy and B. Snaith. The impact of radiographer immediate reporting on patient outcomes and service delivery within the emergency department: Designing a randomised controlled trial. *Radiography*, 17(4):275–279, 2011.

[56] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 2261–2269. IEEE, 2017.

[57] 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.

[58] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[59] Y. Jia and Y. Jiang. Active contour model with shape constraints for bone fracture detection. In *Proc. International Conference on Computer Graphics, Imaging and Visualisation*, pages 90–95. IEEE, 2006.

[60] 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 *Proc. the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[61] O. Jones. The wrist joint. Available at `http://teachmeanatomy.info/upper-limb/joints/wrist-joint/` (accessed 2018/09/03), 2017.

[62] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[63] A. Kazi, S. Albarqouni, A. J. Sanchez, S. Kirchhoff, P. Biberthaler, N. Navab, and D. Mateus. Automatic classification of proximal femur fractures based on attention models. In *Proc. International Workshop on Machine Learning in Medical Imaging*, Lecture Notes in Computer Science, pages 70–78. Springer, 2017.

[64] J. Kellgren and J. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494, 1957.

[65] D. G. Kendall. The diffusion of shape. *Advances in applied probability*, 9(3):428–430, 1977.

[66] D. Kim and T. MacKinnon. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, 2017.

[67] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[68] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, Mar 1998.

[69] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2129, 2009.

[70] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[71] C. Lee and A. Bleetman. Commonly missed injuries in the accident and emergency department. *Trauma*, 6(1):41–51, 2004.

[72] G. Lester. Clinical research in OA – the NIH Osteoarthritis initiative. *J Musculoskelet Neuronal Interact*, 8(4):313–314, 2008.

[73] 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. 2004.

[74] C. Lindner, P. Bromiley, M. Ionita, and T. Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1862–1874, 2015.

[75] 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.

[76] C. Lindner, J. Thomson, and T. F. Cootes. Learning-based shape model matching: Training accurate models with minimal manual input. In *The 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2015)*, volume 9351 of *Lecture Notes in Computer Science*, pages 580–587. Springer, 2015.

[77] C. Loughran. Reporting of fracture radiographs by radiographers: the impact of a training programme. *The British journal of radiology*, 67(802):945–950, 1994.

[78] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[79] 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. In *IEEE International Conference on Image Processing 2005*, volume 1, pages I–1149 – I–1152. IEEE, 2005.

[80] J. Lynch, N. Parimi, R. Chaganti, M. Nevitt, N. Lane, S. of Osteoporotic Fractures (SOF) Research Group, et al. The association of proximal femoral shape and incident radiographic hip OA in elderly women. *Osteoarthritis and Cartilage*, 17(10):1313–1318, 2009.

[81] C. McLauchlan, K. Jones, and H. Guly. Interpretation of trauma radiographs by junior doctors in accident and emergency departments: a cause for concern? *Emergency Medicine Journal*, 14(5):295–298, 1997.

[82] L. Minciullo and T. Cootes. Fully automated shape analysis for detection of osteoarthritis from lateral knee radiographs. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3787–3791. IEEE, 2016.

[83] L. Minciullo, J. Thomson, and T. F. Cootes. Combination of lateral and pa view radiographs to study development of knee oa and associated pain. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013411. International Society for Optics and Photonics, 2017.

[84] M. Müller. Müller OA Classification of Fractures - Long Bones. Available at `http://www.aofoundation.org/Documents/mueller_ao_class.pdf` (accessed 2017/09/03), 2004.

[85] NHS England. Quarterly Activity and Emergency Admissions statistics, NHS and independent sector organisations in England. Available at `http://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2014/04/Quarterly-time-series-2004-05-onwards-with-Annual3.xls` (accessed 2018/09/03), 2015.

[86] T. Nolan, F. Oberklaid, and D. Boldt. Radiological services in a hospital emergency departmentan evaluation of service delivery and radiograph interpretation. *Journal of Paediatrics and Child Health*, 20(2):109–112, 1984.

[87] T. R. C. of Radiologists. How the next government can improve diagnosis and outcomes for patients: Four proposals from the Royal College of Radiologists. Available at `https://www.rcr.ac.uk/sites/default/files/RCR(15)2_CR_govtbrief.pdf` (accessed 2017/09/03), 2015.

[88] T. R. C. of Radiologists. Unreported X-rays, computed tomography (CT) and magnetic resonance imaging (MRI) scans: Results of a snapshot survey of english national health service (NHS) trusts. Available at `https://www.rcr.ac.uk/sites/default/files/rcr_reporting_survey_sept15.pdf` (accessed 2017/09/03), 2015.

[89] 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.

[90] C. Payer, D. Štern, H. Bischof, and M. Urschler. Regressing heatmaps for multiple landmark localization using CNNs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Lecture Notes in Computer Science, pages 230–238. Springer, 2016.

[91] B. Petinaux, R. Bhat, K. Boniface, and J. Aristizabal. Accuracy of radiographic readings in the emergency department. *The American journal of emergency medicine*, 29(1):18–25, 2011.

[92] J. A. Porrino Jr, E. Maloney, K. Scherer, H. Mulcahy, A. S. Ha, and C. Allan. Fracture of the distal radius: epidemiology and premanagement radiographic characterization. *American Journal of Roentgenology*, 203(3):551–559, 2014.

[93] G. Rajamanoharan and T. F. Cootes. Multi-view constrained local models for large head angle facial tracking. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 18–25. IEEE, 2015.

[94] 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*, Jan 2018.

[95] P. Richards, B. Tins, R. Cherian, F. Rae, R. Dharmarajah, I. Phair, and I. McCall. The emergency department: an appropriate referral rate for radiography. *Clinical radiology*, 57(8):753–758, 2002.

[96] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[97] 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 *International Conference on Acoustic Sensing and Imaging.*, pages 159–163. IET, 1993.

[98] P. K. Sahoo, S. Soltani, and A. K. Wong. A survey of thresholding techniques. *Computer vision, graphics, and image processing*, 41(2):233–260, 1988.

[99] A. S. Sayed-Noor, P.-H. Ågren, and P. Wretenberg. Interobserver reliability and intraobserver reproducibility of three radiological classification systems for intra-articular calcaneal fractures. *Foot & ankle international*, 32(9):861–866, 2011.

[100] C. Schmid. Constructing models for content-based image retrieval. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 2. IEEE, 2001.

[101] S. Seltzer, S. Hessel, P. Herman, R. Swensson, and C. Sheriff. Resident

film interpretations and staff review. *American Journal of Roentgenology*, 137(1):129–133, 1981.

[102] L. Shamir, S. M. Ling, W. W. Scott Jr, A. Bos, N. Orlov, T. J. Macura, D. M. Eckley, L. Ferrucci, and I. G. Goldberg. Knee X-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415, 2009.

[103] J. Shao. Linear model selection by cross-validation. *Journal of the American statistical association*, 88(422):486–494, 1993.

[104] H. Sharma, S. Bhagat, and W. Gaine. Reducing diagnostic errors in musculoskeletal trauma by reviewing non-admission orthopaedic referrals in the next-day trauma meeting. *The Annals of The Royal College of Surgeons of England*, 89(7):692–695, 2007.

[105] L. Sheehy, E. Culham, L. McLean, J. Niu, J. Lynch, N. A. Segal, J. A. Singh, M. Nevitt, and T. D. V. Cooke. Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the multicenter osteoarthritis study (MOST). *Osteoarthritis and cartilage*, 23(9):1491–1498, 2015.

[106] A. Shehovych, O. Salar, C. Meyer, and D. Ford. Adult distal radius fractures classification systems: essential clinical knowledge or abstract memory testing? *The Annals of The Royal College of Surgeons of England*, 98(8):525–531, 2016.

[107] V. R. Singh and S. K. Chauhan. Early detection of fracture healing of a long bone for better mass health care. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 6, pages 2911–2912. IEEE, 1998.

[108] R. Smith, K. Ward, C. Cockrell, J. Ha, and K. Najarian. Detection of fracture and quantitative assessment of displacement measures in pelvic X-ray images. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 682–685. IEEE, 2010.

[109] B. Snaith and M. Hardy. Emergency department image interpretation accuracy: The influence of immediate reporting by radiology. *International emergency nursing*, 22(2):63–68, 2014.

[110] M. Sofka, F. Milletari, J. Jia, and A. Rothberg. Fully convolutional regression network for accurate detection of measurement points. In *Proc. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 10553 of *Lecture Notes in Computer Science*, pages 258–266. Springer, 2017.

[111] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.

[112] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, 2016.

[113] J. Thomson, T. O'Neill, D. Felson, and T. Cootes. Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, volume 9350 of *Lecture Notes in Computer Science*, pages 127–134. Springer, 2015.

[114] J. Thomson, T. ONeill, D. Felson, and T. Cootes. Automated shape and texture

analysis for detection of osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9350, pages 127–134, 2015.

[115] J. Thomson, T. ONeill, D. Felson, and T. Cootes. Detecting osteophytes in radiographs of the knee to diagnose osteoarthritis. In *Proc. International Workshop on Machine Learning in Medical Imaging*, volume 10019 of *Lecture Notes in Computer Science*, pages 45–52. Springer, 2016.

[116] T. P. Tian, Y. Chen, W. K. Leow, and W. Hsu. Computing Neck-Shaft Angle of Femur for X-ray Fracture Detection. In *Proc. International Conference on Computer Analysis of Images and Patterns*, volume 2756 of *Lecture Notes in Computer Science*, pages 82–89. Springer, 2003.

[117] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *ArXiv e-prints*, Oct 2017.

[118] C. Vincent, P. Driscoll, R. Audley, and D. Grant. Accuracy of detection of radiographic abnormalities by junior doctors. *Emergency Medicine Journal*, 5(2):101–109, 1988.

[119] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1. IEEE, 2001.

[120] J. Waarsing, R. Rozendaal, J. Verhaar, S. Bierma-Zeinstra, and H. Weinans. A statistical model of shape and density of the proximal femur in relation to radiological and clinical OA of the hip. *Osteoarthritis and cartilage*, 18(6):787–794, 2010.

[121] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis*, 31:63–76, 2016.

[122] J. Wardrope and P. Chennells. Should all casualty radiographs be reviewed? *British Medical Journal (Clinical Research Ed.)*, 290(6482):1638–1640, 1985.

[123] C.-J. Wei, W.-C. Tsai, C.-M. Tiu, H.-T. Wu, H.-J. Chiou, and C.-Y. Chang. Systematic analysis of missed extremity fractures in emergency radiology. *Acta Radiologica*, 47(7):710–717, 2006.

[124] Z. Wei and Z. Liming. Study on recognition of the fracture injure site based on X-ray images. In *Proc. International Congress on Image and Signal Processing (CISP)*, volume 4, pages 1947–1950. IEEE, 2010.

[125] Z. Wei, M. Na, S. Huisheng, and F. Hongqi. Feature extraction of X-ray fracture image and fracture classification. In *Proc. International Conference on Artificial Intelligence and Computational Intelligence (AICI 2009)*, volume 2, pages 408–412. IEEE, 2009.

[126] B. H. Willis and S. D. Sur. How good are emergency department Senior House Officers at interpreting X-rays following radiographers' triage? *European Journal of Emergency Medicine*, 14(1):6–13, 2007.

[127] C. Xu and J. L. Prince. Gradient vector flow: A new external force for snakes. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 66–71. IEEE, 1997.

[128] D. W. H. Yap, Y. Chen, W. K. Leow, T. S. Howe, and M. A. Png. Detecting

femur fractures by texture analysis of trabeculae. In *Proc. International Conference on Pattern Recognition*, volume 3, pages 730–733. IEEE, 2004.

[129] J. Zhang, M. Liu, and D. Shen. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Transactions on Image Processing*, 26(10):4753–4764, 2017.

[130] W. Zhang and M. Brady. Feature point detection for non-rigid registration of digital breast tomosynthesis images. In *Proc. International Workshop on Digital Mammography*, volume 6136 of *Lecture Notes in Computer Science*, pages 296–303. Springer, 2010.

[131] Y. Zheng, B. Georgescu, and D. Comaniciu. Marginal space learning for efficient detection of 2D/3D anatomical structures in medical images. In *Proc. International Conference on Information Processing in Medical Imaging*, volume 5636 of *Lecture Notes in Computer Science*, pages 411–422. Springer, 2009.

[132] S. Zhong, K. Li, and R. Feng. Deep convolutional hamming ranking network for large scale image retrieval. In *Proc. 11th World Congress on Intelligent Control and Automation (WCICA)*, pages 1018–1023, 2014.

[133] X. Zhou, R. Stern, and H. Müller. Case-based fracture image retrieval. *International journal of computer assisted radiology and surgery*, 7(3):401–411, 2012.