Chapter 9

Conclusions and future directions

This chapter summarises the findings within this thesis, examines its main contributions, and provides suggestions for further directions.

9.1 Thesis summary

This thesis has presented background information about bone anatomy, fractures and methods of visualisation along with an overview of the previous research that has been performed in the field of computer aided diagnosis and more specifically fracture detection. It examined methods of detecting edges within long-bone x-ray images, and how the long-bone shaft could be approximated using a set of parameters. These parameters were then utilised for diaphysis segmentation and fracture detection within the segmented region. Finally, to decrease the diagnosis time, parallelisation of two of the slower parts of the algorithm was examined. Results showing the effectiveness of each of the algorithms were presented in each section.

After examining the relevant bone and joint anatomy and physiology, it was noted that despite their great strength and flexibility, bones do occasionally break. The incidence of bone fractures is relatively high, accounting for roughly 20% of the occupancy of orthopaedic wards [30], at any point in time. In addition, the treatment costs for fractures are increasing, as are the number of fractures associated with age-related bone loss [94]. The first step in providing adequate treatment for bone fractures is accurate fracture diagnosis. Unfortunately, missed diagnoses are relatively common, and account for a significant portion of the malpractice suits against radiologists in the USA [20]. Satisfaction of search—where the identification of an obvious abnormality distracts the observer from detecting other abnormalities, or causes the search to end prematurely—is a possible cause of this error during x-ray reading, along with lack of compliance in unconscious patients and interpretations made without localisation cues or a clinical history. A computer aided fracture diagnosis scheme that does not suffer from the SOS effect was suggested as a solution to this problem, and midshaft long-bone fractures were identified as the abnormalities that should be detected. In order to achieve this aim, a digital x-ray image database was created using long-bone x-ray images collected from an emergency department in metropolitan Adelaide. From this database, an algorithm development set of six images, and an algorithm test set of 44 images were produced. Ideally the test set would have been larger, but given the time and difficulty of obtaining this current set, it was not practical to continue image collection for a longer period.

The overview of previous work revealed, most importantly, that very little had been achieved in the way of automated bone fracture detection. The only closely related work [105, 110, 104, 59] involved detecting neck of femur fractures based on the angle of the neck of femur, and on the analysis of the disruption to the trabecular patterns in the femoral neck. Limited work had also been performed on osteoporosis diagnosis and bone age estimation. Similarly, only a small amount of work had been published on automated long-bone segmentation, with only two relevant algorithms identified. The algorithm by Jia and Jiang [51] utilised a geodesic active contour model to identify bones within the arm, but the model required a very accurate prior shape for the match to occur, and was only likely to be accurate for a very small range of images. El-Kwae [38] created a model by analysing a diverse range of bone shapes, and the model was then applied to the image in an unspecified manner, to determine if any matches occurred. In both cases, no results from their algorithms were presented. This lack of previous literature indicated that the creation of a CAD system for the detection of long-bone fractures should be a priority.

Within this thesis, the first stage of analysing the x-rays was performing low level analysis to detect edges within the image. To separate the true edges from the inherent noise, smoothing was required, leading to the notion of scale-space. In a scale-space approach to edge detection, the original image is replaced with a smoothed version of that image, at a particular scale. In this manner, local disturbing detail such as image noise or texture can be removed so that the features at that scale can be extracted. If the smoothing had to be linear, then the only suitable candidate was convolution with a Gaussian, which was equivalent to the solution of the heat equation [56]. However, because the heat equation was isotropic, smoothing occured across the boundaries in the image, resulting in blurring of the features that were to be detected. The solution was to limit the smoothing across the edge boundaries, but this required a non-linear smoothing function. One such function was the anisotropic diffusion by Perona and Malik [83]. Their equation was equivalent to the heat equation when the magnitude of the gradient was small, and the inverse heat equation when the magnitude of the gradient was large. The Perona and Malik equation was an improvement because smoothing was retained inside the boundaries, but it had several practical and theoretical difficulties that made it unsuitable for the low level analysis of x-ray images.

A non-linear partial differential equation called the mean curvature motion (MCM) was examined instead, and shown to diffuse only in the direction orthogonal to the gradient. A special case of this equation, called the affine and morphological scale space (AMSS) was chosen because it too only diffused in the direction orthogonal to the gradient, and was fully affine invariant, thereby satisfying all invariance requirements. A discrete numerical method to calculate the AMSS was implemented. Regardless of which smoothing method was chosen, selection of the most appropriate scale for analysis was still a significant and difficult problem. This limitation was tackled by examining the standard deviation of a sample of regions within the six development images, at a range of scales, to determine the effect of the smoothing. From this analysis two global scales were chosen for the various analyses. Testing of the AMSS algorithm on the development images showed that the edges it detected were much clearer than those created by the Canny edge detection algorithm.

After the low level analysis was performed, and a good representation of the important edges was obtained, it was necessary to interpret them. Rather than attempting to use traditional segmentation and extraction methods such as clustering or global segmentation algorithms—which can produce poor results on x-ray images—the long-bone shaft was instead approximated by a set of parameters. That is, only the information about the location and orientation of the bone was determined. This approximation was possible because long-bone diaphyses are generally straight, despite a large degree of natural variation in bone shape. To extract this global information from the image, the Hough Transform was utilised to convert spatially extended patterns into spatially compact features within the space of possible parameter values. As a result, the problem was altered from a difficult global feature analysis, into a more simple local peak detection. The standard Hough Transform was then modified slightly to reduce the likelihood of false accumulator peaks being created.

Once the transform had been performed, the detection of the appropriate peaks within the accumulator was still a significant problem that needed to be solved. Various peak detection algorithms already existed, but all of them assumed that the input image was relatively simple and that only one peak needed to be detected. As a result, they were not appropriate. It was shown by Van Veen and Groen [106] that the amount of peak spread in an ideal image can be quantified. This meant that a window of that size could be used to detect the peaks of interest, using the specially designed ranked sums method. The key feature of this new method was that due to the amount of peak spreading in a real x-ray image, the importance of a particular region should be based on the sum of the accumulator cells within that region, rather than the maximum value that it contained. Testing revealed that this method produced excellent results, with almost 90% of all peaks in the test set accurately detected. Manual intervention was used to correct the remaining peaks. In the future, a possible improvement to the algorithm could be to automatically increase the number of peaks to be detected from $e = 4$ to $e = 6$ in the case of radius and ulna fractures, since in many cases the radius was better approximated with four lines rather than two.

A unique long-bone shaft endpoint detection algorithm was also created in Chapter 5, to determine the extent of the parameterised lines. This algorithm was based on the magnitude of the gradient, which was modified to remove pixels where the direction of the gradient did not match the line angle parameters. A thick strip was extracted from underneath each line, and the sum of the pixels across the strip was calculated. Finally, a dual stage thresholding technique was used to identify the line endpoints within the strip. At this stage, the long-bone shaft should be accurately represented by a series of straight lines, whose parameters and endpoints were known. In the test image set, the line endpoints were identified with an accuracy of almost 98%, making this the most accurate part of the complete long-bone segmentation and fracture detection algorithm in this thesis. Analysis of the results also indicated that the method was more likely to perform well with images containing subtle fractures where there were no gross abnormalities, than in images where there were obvious, displaced fractures that confuse the long-bone shaft parameter approximation algorithm. This was significant because it is images containing subtle fractures that are more likely to require the assistance of a CAD system to improve the accuracy of the fracture detection.

After the long-bone shaft parameters were accurately identified, it was necessary to complete the segmentation to isolate the diaphysis. In Chapter 6 this was initially performed using an implementation of the AO segmentation, which involved locating the appropriate landmarks on and around the joints. The first segmentation stage involved locating all the bone centre-lines using the shaft approximation parameters, before determining which of the centre-line(s) should be retained. To determine all bone centre-lines, it was necessary to correctly choose pairs of shaft parameters that corresponded to the bone edges on each side of the desired centre-line, which was then located by averaging the angle and distance parameters. The difficult part of this stage was therefore to correctly determine how the parameters should be paired. Pairing by ranking the parameter pairs based on either ρ or θ was shown to work on two of the development images, but was unlikely to do so in more complicated cases. Instead, the lines were paired in the same way that a human would be likely to pair them in the absence of any image data. That is, based on how close the lines and the line endpoints were located to each other. To do this, a table of all the possible pairings was created, and the sum of the distances between each of the line endpoints in all the pairs was recorded. A second criterion consisting of the absolute distance between the ρ values of each line within the pair was also recorded. The sum of these two values was calculated and the combination of pairs with the smallest sum was chosen. For the image test set, 93% of the centre-lines were correctly detected using this method. From the complete set of centre-lines, the required centre-lines were selected based on the Hough inter-peak distances.

The next stage of the AO segmentation method involved locating the extreme articular surface and calculating the epiphyseal width. Both of these tasks were performed using projections obtained from the modified Hough Transform. In both cases, correctly interpreting the projections was a non-trivial problem that could even be very difficult to perform manually. A set of criteria were chosen for both tasks, and were shown to be only moderately successful in locating the required features in the six development images. However, this was not the case for the images in the test set. Not only was the AO segmentation method not suitable for application to many of the test set images, but it also did a very poor job of segmenting those that were appropriate. A significant portion of this poor performance was due to the articular surface identification step, with no single or multiple criteria able to accurately determine its location. Due in part to the poor results it produced, and also because it could not be applied to mediolateral images or images that did not completely contain the articular surface and epiphyses, the AO segmentation method was abandoned.

The AO segmentation algorithm was replaced with one that used the long-bone shaft approximation lines to mark the diaphysis end points. This was possible because the line endpoints correspond to the locations where the bone curved away from the approximation lines, and these points were by definition the diametaphyses. Not only could this algorithm be applied to all the images in the test set—unlike the AO method—but it also correctly segmented 83% of the images in the test set. The only significant problem with this method was that, in a small number of cases, the presence of a fracture caused the bone centre-line to end prematurely, resulting in the segmentation occurring in the incorrect location. Fortunately this only occurred with very obvious fractures, which are less likely to be the type requiring the use of a CAD system. In the case of subtle fractures, the segmentation was 100% accurate. Despite not using the AO segmentation method, the first of the two major aims of this thesis outlined in Section 3.3.4 on page 51 was satisfied, with the algorithm able to discern between the diaphyseal and epiphyseal segments.

Once the segmentation had been performed, a fracture detector could be applied. In Chapter 7, the fracture detection algorithm was constructed to locate abnormal gradients within the segmented image, using a tool called the gradient composite measure. The GCM was used to remove all the normal parts of the image, such as the bone shaft, leaving behind only the abnormal regions such as fractures. The GCM was a combination of the magnitude and direction of the gradient of the smoothed image that was calculated using the product of the magnitude of the gradient and two scaling factors called the importance rank and the distance rank. These two scaling factors measured how well the direction of the gradient at a chosen pixel matched the angle parameter θ of the long-bone shaft approximation lines, and how close that chosen pixel was to all of the approximation lines, respectively. Thus, the regions corresponding to the bone shaft received a very low importance rank because the angle of the gradient at those points matched the angle parameter. The resulting GCM was characterised by regions of high intensity where the fracture was located, and very low intensity everywhere else in the image.

Two methods of presenting the fractures were examined. The first involved artificially colouring the image to show the magnitude of the GCM, but this had a number of practical limitations including being very difficult to see. Instead, using the second method, the regions were marked more clearly using a three step fracture identification process. Part of this process involved a dual stage thresholding to retain filtered GCM points of sufficient intensity and size. To determine the correct thresholds, a sampling based sensitivity analysis was performed using a range of threshold values. From this analysis, two new thresholds were chosen, and used to evaluate the final fracture detection algorithm using the complete image test set. The first significant result was that, while hard to measure objectively, subjectively the GCM did a very good job of highlighting abnormalities. However, the ability of the fracture identification algorithm to mark those GCM regions as either fracture or non-fracture was not quite so good. Although 83% of the fractures were detected, the false positive rate was relatively high, with many images containing at least one normal region that was identified as a fracture. Analysis of the causes of the false detections showed that most were due to biological phenomena that the algorithm confused as being fractures, while a smaller number were due to image artifacts or bones overlapping.

Although the algorithm did not detect all fractures within the test set, it did manage to detect one fracture that was missed during the initial reading by the radiologist, indicating that for some images the sensitivity of the algorithm is very high. This effectively illustrated the most significant problem with the current implementation of the fracture detection algorithm, which was that the two thresholds are probably not appropriate for all images. The sensitivity analysis showed that it was possible to select global thresholds that allowed five of the six images in the development set to be correct, but no choices would allow all six images to be correct. Interestingly this was almost the same proportion of images that were correct in the 44 image test set. If it were possible to accurately adaptively select the two thresholds in some way, then it is likely that the results would be greatly improved. The second problem with this implementation was that the sensitivity with which fractures were detected was related to the angle at which the fracture occured, due to the assumption that all regions parallel to θ_i were likely to be normal. As a result, fractures parallel to the bone centre-line were not necessarily as well detected as those parallel to it. Although the accuracy of the fracture detection scheme was not as high as desired, and not all types of fractures could be equally well identified, the second of the two major aims outlined in Section 3.3.4 was still achieved.

The final body of work within this thesis examined the amount of time required to calculate the AMSS and Hough Transform. Both of these algorithms were originally relatively slow to compute, due to the very large number of iterations that had to be performed. Some methods by which the AMSS calculation time could be reduced were suggested, but the only practical method was to implement a parallel algorithm that could run on multiple CPUs simultaneously. Unfortunately, the input for each AMSS iteration was the output of the previous iteration, so for parallelisation it was necessary to split the image into a number of stripes or tiles. In addition, the image could not be split and sent to multiple CPUs and then simply recombined, as the smoothing across the boundary resulted in an incorrect output. Instead, it was necessary to perform some boundary extension at the points at which the image was split, and the amount of extension required was determined by the final scale of smoothing. Several methods by which the image could be split and then smoothed were examined. Analysis of the chosen method—the incremental iteration smoothing algorithm—showed that the final scale of smoothing $t_2 = 20$ was reached in around one third of the time taken by the standard method. Similarly, the modified Hough Transform algorithm was also parallelised, and was found to be over one hundred times faster than the standard non-parallelised Hough Transform. Both of these methods allowed a diagnosis to be made much more rapidly.

9.2 Key contributions

In sum, the key contributions of this thesis, listed in order of significance, are:

- The production of the first semi-automatic computer aided long-bone segmentation and fracture detection algorithms.
- Fracture detection using gradient analysis, including the design of the importance rank, distance rank and gradient composite measure.
- Implementation of the AO segmentation method, and the design of a superior segmentation method based on the analysis of the bone curvature.
- The idea of approximating the long-bones using a set of straight lines, incorporating creation of the ranked sums method for peak detection in the Hough accumulator, and the long-bone shaft endpoint detector.
- Centre-line detection by minimising the distance between the line endpoints.
- The identification of the AMSS as the only appropriate method of smoothing x-ray images.

Figure 9.1: Scale-space analyses such as the AMSS can also be used to detect fractures based on the curvature of features within the image. This example demonstrates the results of this type of analysis on the same image used throughout Chapter 7, and shows that the intensity is higher in the region surrounding the fracture.

- The method of analysing the standard deviation over a range of scales to determine the appropriate scales for AMSS smoothing.
- Parallelisation of the AMSS algorithm.

9.3 Future directions

Chapter 4 introduced the theory and implementation of the affine morphological scale space, but did not provide any other uses apart from image smoothing. Although important for long-bone shaft parameter approximation and fracture detection, smoothing is not the only use for the scale-space techniques described. The scale-space curvature function described in Equation 4.18—where $curv\,(I) = \kappa$ is the curvature—can be used as a shape descriptor that is invariant to translation, scale and rotation [64]. This is possible because the curvature describes the rate of change in the direction of a curve, per unit length. As a result, the most curved sections of a line will have the greatest curvature. As suggested in Section 4.3.3, each point on a curve moves in a direction perpendicular to the boundary with a speed given by its curvature, so that more curved sections become flattened more quickly. In images containing fractures, the features of the fracture are often points of very high curvature, which therefore become smoothed quickly. In theory, a curvature analysis could be performed to detect these points and use them in the fracture detection decision process. An example of the type of image that could be used is shown in Figure 9.1.

In this example, the curvature is obtained by simply calculating the absolute difference between the images at two different scales, in this case $t_1 = 5$ and $t_2 = 20$, the scales chosen in Section 4.6.1. This effectively determines which regions have changed the most over the scale evolution, and this in turn is proportional to the curvature of those regions. Other methods of utilising the curvature were also investigated. In many respects the features that are detected are very similar to those in the GCM image in Figure 7.6b. However, the advantage of the curvature analysis is that the GCM is no longer required, and fractures are more likely to be detected well regardless of their orientation. The problem with the curvature method is that it is highly dependent on the scales chosen, and many fractures are poorly highlighted at the scales demonstrated here. Again, this reverts to the difficult problem of automatic scale selection identified in Section 4.6. Reliable results were not obtained with this method, in part because an appropriate set of global scales could not be determined. In addition, some false detections (similar to those described in Table 7.3) still occured, because the curvature analysis detected similar features to the GCM. Further investigation into this area is required to determine how the curvature information can be better utilised for fracture detection. Although likely to be difficult, a future improvement could be to adaptively choose the scales on a per image basis, using an automated scale selection method. As a result, the scales would be more likely to be appropriate for the features within the image.

As demonstrated in Chapter 7, the detection of fractures was performed in two separate parts. The first involved the use of the GCM (or in future, some type of curvature analysis) to process the image to extract a set of features, while the second involved examining those features to make a decision about whether or not they constitute a fracture. The decision making process utilised in Section 7.3.2 was not particularly advanced, and essentially just selected regions based on their size and intensity, with the use of the thresholds T_3 and T_4. Although this method performed well, it is likely that the fracture detection accuracy could be improved if a more advanced decision process was used. For example, an artificial neural network (ANN) that was trained using the features extracted from the development images, may produce good results on the test set. Alternatively, a Bayesian classifier or support vector machine approach could be used. Again, further development in this area is required to determine not only what type of classifier should be used and how it should be trained, but also how the features in the image should be applied.

In Section 2.3 is was suggested that in order to effectively treat a fracture it is first necessary to determine if, where and how the bone is broken. The fracture detection system described in this thesis performs the first and second of these two tasks, to determine if and where the bone is broken, but it does not examine how the the bone is broken. Section 2.3 also stated that this third task is normally achieved by performing fracture classification, to determine the nature and severity of the fracture. One of the problems with the described classification schemes is that they can be subjective, with some studies showing that there is little inter-observer agreement on the group and even less agreement on the subgroup of some AO fractures [57]. A method of automatically performing the fracture classification—after the fracture detection has been performed—could improve the reliability and accuracy of a classification. Although beyond the scope of this thesis, it is possible that computer aided fracture classification could be performed by analysing the shape of the features extracted using either the GCM or the AMSS curvature, and assigning the resulting shape to the appropriate AO group and subgroup.

As stated in the summary at the start of this chapter, ideally both the image development and test sets would have been larger, but it was not practical to continue image collection for a longer period of time. However, the investigations made using these data sets highlighted the need for more openly accessible high quality image data sets that are large enough for training, testing and validation of fracture detection algorithms. This will become even more important when ANNs are utilised, due to the quantity of training data that is likely to be required. A larger image set will also certainly allow more accurate development before testing. Additionally, it is possible that the small size of the development set—consisting of only six images—contributed to the results obtained. In future studies it would be better to use the entire image set for algorithm development, thereby increasing the exposure to the number and types of fractures, possibly creating a more robust method. To test the algorithms, a cross validation method could then be used to estimate their performance on unseen data.

A significant limitation of this thesis was that the criteria for success were weak. Unfortunately, the final analysis of the fracture detector described in Chapter 7 relied partly on the subjective comments of a single radiologist. For future studies, in addition to a larger image set, it would also be advantageous to have a larger group of radiologists interpret both the original x-ray images, and the results produced by the fracture detection algorithm. An experiment could be constructed, in which images of normal bones and bones with very subtle fractures were shown to a group of three or four radiologists, who would be asked to identify all fractures within the images. The false positive, false negative and true detection rates of the radiologists could then be compared to those of the algorithm. This would also allow the inter-rater reliability to be measured, and a better opinion of the algorithm to be gauged. This will become even more important as the fracture detection accuracy improves and approaches that of a trained human observer, and direct comparisons between the CAD system and radiologist are made. However, in the context of this study, it is unlikely that having a larger number of radiologists would have had any impact on the results obtained.

Finally, rather than reporting only the number of true and false positive and negative results obtained by the algorithms, in the future it may be better to perform a receiver operating characteristic (ROC) analysis to determine how the chosen thresholds affect the algorithm sensitivity and specificity. It was suggested in Section 7.4 that adjusting the sensitivity of the algorithm on a per image basis, using the two thresholds T_3 and T_4, resulted in a better detection. Construction of a ROC curve—the true positive rate against the false positive rate—would allow the tradeoff between the sensitivity and specificity, as well as the accuracy to be quantified. However, for a ROC analysis to be meaningful, it is necessary to know what constitutes normal and abnormal results. In this thesis, the gold standard—that is the radiologist's diagnosis of either fractured or unfractured—was not necessarily completely accurate. This is because only one expert was used to make the diagnosis, and extremely subtle fractures—those that the algorithm was aimed at detecting—may have been (and in one case certainly was) missed. Thus, performing a ROC analysis would have had very little effect on the outcome of this research.

Utilising the results presented in this thesis, including the suggestions for future work, should allow the production of faster and more sensitive algorithms for the detection of fractures in long-bones. However, the most important future goal should be the extension of these methods from long-bone diaphyses to long-bone epiphyses and then to other anatomical regions. The development of a CAD system that can detect all types of fractures in all anatomical locations will be extremely valuable for physicians and radiologists.

9.4 Conclusion

Prior to this thesis, the computer aided detection of midshaft long-bone fractures had not previously been examined. This thesis presented a method by which semi-automated long-bone shaft segmentation could be performed, along with fracture detection within the segmented region. In the test set, 83% of the diaphysis segmentation boundaries were correctly identified, and subsequently 83% of the fractures within those segmented regions were also detected correctly. Incorporating the methods and results formulated and utilised in this thesis, along with the future research outlined above, will further expand the capabilities of today's CAD systems, and result in more accurate diagnosis of fractures and a reduction of the fracture miss rate.