

Predictive Modelling of Bone Ageing

Luke Matthew Davis

A thesis submitted for the
Degree of Doctor of Philosophy

University of East Anglia
School of Computing Sciences



July 31, 2013

©This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

Bone age assessment (BAA) is a task performed daily by paediatricians in hospitals worldwide. The main reasons for BAA to be performed are: firstly, diagnosis of growth disorders through monitoring skeletal development; secondly, prediction of final adult height; and finally, verification of age claims. Manually predicting bone age from radiographs is a difficult and time consuming task. This thesis investigates bone age assessment and why automating the process will help. A review of previous automated bone age assessment systems is undertaken and we investigate why none of these systems have gained widespread acceptance. We propose a new automated method for bone age assessment, ASMA (Automated Skeletal Maturity Assessment).

The basic premise of the approach is to automatically extract descriptive shape features that capture the human expertise in forming bone age estimates. The algorithm consists of the following six modularised stages: hand segmentation; hand segmentation classification; bone segmentation; feature extraction; bone segmentation classification; bone age prediction.

We demonstrate that ASMA performs at least as well as other automated systems and that models constructed on just three bones are as accurate at predicting age as expert human assessors using the standard technique. We also investigate the importance of ethnicity and gender in skeletal development. Our conclusion is that the feature based system of separating the image processing from the age modelling is the best approach, since it offers flexibility and transparency, and produces accurate estimates.

Acknowledgements

I would like to acknowledge and thank my supervisory team of Dr Anthony Bagnall and Dr Barry-John Theobald for their help, support, guidance and encouragement throughout the project so far. I would also like to thank Dr Andoni Toms of the Radiology Academy at the Norfolk and Norwich University Hospital (NNUH). My fiancée, Anke Arkenberg for the many things she has helped me with including her support, manually labelling outlines and translating papers from German to English. As well as my family for their support and encouragement throughout my time in education. Special thanks go to Andrea De Marco for manually labelling outlines and Pia Unte for finding papers in Germany for the timeline of Bone Age Assessment. I would like to thank all of the people who allowed me to stay in a spare room or on a sofa whilst I was living in Swindon, as well as everyone in the Speech lab (past and present), Dr Bagnall's other PhD students Jason Lines and Jon Hills, and the other students and staff who have been here throughout my time in the School of Computing Sciences at UEA. I would also like to thank my examiners Dr Mark Fisher and Prof. Frans Coenen, University of Liverpool.

Contents

Acknowledgements	ii
List of Abbreviations	vii
List of Figures	viii
List of Tables	xi
Publications	xv
Awards	xvi
1 Introduction	1
1.1 Motivation	2
1.2 Contributions	3
1.3 Organisation of Thesis	6
2 Radiology and Bone Age Assessment	8
2.1 What is Bone Age?	9
2.2 Why is Bone Age Assessment Performed?	10
2.2.1 Diagnosis and Management of Health Issues	10
2.2.2 Prediction of Final Adult Height	10
2.2.3 Verifying Age Claims	11
2.3 X-ray Acquisition Process	12
2.4 A Brief History of Manual Bone Age Assessment	15
2.4.1 Atlas Methods	19
2.4.2 Oxford Method	24
2.4.3 Planimetry Method	26

2.4.4	Numerical Method	26
2.4.5	Age of Appearance Lists	27
2.5	How Bone Age Assessment is Currently Performed	28
2.5.1	The Greulich and Pyle Method	29
2.5.2	The Tanner and Whitehouse Method	29
2.6	The Need For Automation	32
2.7	Summary	32
3	Automated Bone Age Assessment	34
3.1	Image Processing Algorithms	34
3.1.1	Active Appearance Models (AAM)	35
3.1.2	Otsu Thresholding	37
3.1.3	Canny Edge Detector	38
3.2	Classification Algorithms	39
3.2.1	k -Nearest Neighbour (k -NN)	39
3.2.2	Naive Bayes	40
3.2.3	C4.5 Decision Tree	40
3.2.4	Support Vector Machines (SVM)	42
3.2.5	Random Forest	43
3.2.6	Rotation Forest	43
3.2.7	Multilayer Perceptron	44
3.3	A Brief History of Automated Bone Age Assessment	44
3.3.1	BoneXpert by Thodberg <i>et al.</i>	46
3.3.2	The Work by Pietka <i>et al.</i>	49
3.3.3	Other Proposed Methods for Automated Bone Age Assessment	53
3.4	Why is There a Lack of Widespread Acceptance of Automated Bone Age Assessment Systems?	62
3.5	The Pitfalls of Using AAMs for Automated Bone Age Assessment . .	63
3.6	Summary	64
4	ASMA Stages A & B: Hand Segmentation and Classification	66
4.1	Dataset Used	67
4.2	Outlining a Hand	69
4.2.1	Active Appearance Models	70

4.2.2	Otsu Thresholding	70
4.2.3	Canny Edge Detection	71
4.2.4	Contour Algorithm	71
4.3	Ensemble Algorithm	75
4.3.1	Dynamic Time Warping Outline Selection	76
4.3.2	Likelihood Ratio Outline Selection	79
4.4	Classification of Validity of an Outline	83
4.4.1	Transformation	83
4.4.2	Classifiers Used	88
4.5	Results	88
4.5.1	Classifying Outlines	88
4.5.2	Generating Test Outlines	90
4.5.3	Testing the Outlines	90
4.5.4	Manual Assessment of Outlining Algorithms	91
4.6	Conclusions	93
5	ASMA Stages C, D & E: Bone Segmentation, Feature Extraction and Bone Segmentation Classification	95
5.1	Bone Segmentation	97
5.1.1	Locating the ROIs	98
5.1.2	Segmenting Hard Tissue from ROI	100
5.2	Feature Extraction	102
5.3	Bone Segmentation Classification	110
5.3.1	Rejection Rules	110
5.3.2	Training a Classifier	111
5.4	Results	115
5.4.1	Classifying Segmentations	115
5.4.2	Performance on Test Outlines	118
5.4.3	Manual Assessment of Bone Segmentation Classification . . .	119
5.5	Conclusions	120
6	ASMA Stage F (Part One): Classification of Tanner-Whitehouse Stages	121
6.1	Exploratory Analysis of Features	122
6.1.1	Overall Information Gain of Features	123

6.1.1.1	Distal Phalange	124
6.1.1.2	Middle Phalange	125
6.1.1.3	Proximal Phalange	127
6.1.1.4	Overall	127
6.1.2	Information Gain based on Tanner-Whitehouse Stage	129
6.1.3	Feature Interaction	133
6.2	Classification of Tanner-Whitehouse Stages	139
6.2.1	Cross-Validation of Tanner-Whitehouse Stages	139
6.2.2	Performance on Testing Data	141
6.2.3	Comparison to Previously Published Work	142
6.3	Conclusions	143
7	ASMA Stage F (Part Two): Regression Onto Chronological Age	145
7.1	Linear Regression	146
7.2	Model Selection	148
7.2.1	Piecewise Regression	150
7.2.2	Multiple Bone Models	151
7.2.3	Outliers of Models	151
7.2.4	Heteroscedasticity of Models	156
7.3	Predicting Age	157
7.4	Effects of Gender and Ethnicity	163
7.5	Conclusions	167
8	Conclusions And Future Work	168
8.1	Conclusions	168
8.2	Future Work	171
8.2.1	Stage A: Hand Segmentation	171
8.2.2	Stage B: Hand Segmentation Classification	172
8.2.3	Stage C: Bone Segmentation	172
8.2.4	Stage D: Feature Extraction	172
8.2.5	Stage E: Bone Segmentation Classification	173
8.2.6	Stage F: Classification of Tanner-Whitehouse Stages	173
8.2.7	Stage F: Regression onto Chronological Age	173

List of Abbreviations

Abbreviation	Meaning
AAM	Active Appearance Model
ABAA	Automated Bone Age Assessment
AC	Active Contour
ASM	Active Shape Model
ASMA	Automated Skeletal Maturity Assessment System
BAA	Bone Age Assessment
CBA	Carpal Bone Age
CROI	Carpal Region of Interest
DTW	Dynamic Time Warping
EROI	Epiphyseal Region of Interest
EMROI	Epiphyseal-Metaphyseal Region of Interest
FFT	Fast Fourier Transform
GP	Greulich and Pyle Method for Bone Age Assessment
k -NN	k -Nearest Neighbour
LDC	Linear Discriminant Classifier
MAE	Mean Absolute Error
PBA	Phalangeal Bone Age
PCA	Principal Component Analysis
PDM	Point Distribution Model
PROI	Phalangeal Region of Interest
ROI	Region Of Interest
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
RUS	Radius, Ulna and Short Bones
SMS	Skeletal Maturity Score
SVM	Support Vector Machine
SVML	Support Vector Machine with linear kernel
SVMQ	Support Vector Machine with quadratic kernel
SVMR	Support Vector Machine with radial basis function kernel
TW	Tanner and Whitehouse Method for Bone Age Assessment

List of Figures

1.1	The six stages of the Automated Skeletal Maturity Assessment (ASMA) system.	4
2.1	An example radiograph of the distal phalange of the middle finger, with epiphysis, metaphysis and diaphysis labelled.	9
2.2	An X-ray of the hand of the wife of Wilhelm Conrad Röntgen [Spi00], taken on December 22, 1895.	13
2.3	A timeline of manual bone age assessment.	16
2.4	A radiograph of the hand with all bones numbered, the corresponding names can be seen in Table 2.1.	17
4.1	(a) and (b) Two examples of incorrectly located hand outlines, and (c) a hand outline correctly segmented. All outlines created using contouring algorithm.	68
4.2	An example of a contour \mathbf{c} being found at $l_i = 50$, on a simple input image \mathbf{I}	73
4.3	An example of the t value calculation.	74
4.4	An example radiograph of the hand where the heel effect is visible. Notice that the background pixels on the right hand side of the image are brighter.	76
4.5	γ corrected hand radiographs.	77
4.6	An example of a good hand outline from a radiograph (a) being converted into a one-dimensional series (b) and (c).	80
4.7	An example of a bad hand outline from a radiograph (a) being converted into a one-dimensional series (b) and (c).	81
4.8	An example of DTW between two one-dimensional series of hand outlines.	82
4.9	A graphical illustration of Algorithm 4.2.	87
4.10	Two examples of AAM finding an incorrect outline.	92

5.1	(a) and (b) two examples of incorrect bone segmentations, and (c) a correct bone segmentation.	97
5.2	An example radiograph labelled with (a) the correct hand outline extracted from stages A and B of ASMA, and (b) the finger tips and webs found using Algorithm 4.2.	98
5.3	An example radiograph labelled with (a) the axes of the middle finger, and (b) the ROI around each of the phalanges of the middle finger. . .	99
5.4	An example ROI of (a) distal, (b) middle and (c) proximal phalanges.	100
5.5	(a) An example of a distal phalanx ROI, (b) resulting binary mask, and (c) outline of segmentation imposed onto ROI.	102
5.6	(a) and (b) ROI before and after unwarp process, (c) and (d) bone segmentation of (a) before and after unwarp process.	106
5.7	Shows various features calculated from a bone segmentation (a) the best fitting ellipse for both the phalanx and epiphysis, along with the major (red) and minor (blue) axes, (b) the interpolated height (red) and width (blue) of the phalanx and epiphysis, and (c) the various widths calculated along the phalanx.	109
5.8	An example of a good bone segmentation from a ROI (a) being converted into a 1-D series (b) and (c).	113
5.9	An example of a bad bone segmentation from a ROI (a) being converted into a 1-D series (b) and (c).	114
6.1	A C4.5 tree showing the interactions of the extracted features for Tanner-Whitehouse stage classification on distal phalange III. . . .	136
6.2	A C4.5 tree showing the interactions of the extracted features for Tanner-Whitehouse stage classification on middle phalange III. . . .	137
6.3	A C4.5 tree showing the interactions of the extracted features for Tanner-Whitehouse stage classification on proximal phalange III. . . .	138
7.1	Epiphysis models. (a)(c)(e) Show the predicted age of each instance against the chronological age for D_e , M_e and P_e respectively. The dotted line is predicted = chronological. The solid line is the regression of predicted vs chronological. (b)(d)(f) Show the absolute standardised residuals against predicted age for D_e , M_e and P_e respectively. The solid line is the regression of absolute standardised residuals against predicted age.	153

7.2	Phalanx models. (a)(c)(e) Show the predicted age of each instance against the chronological age for D_p , M_p and P_p respectively. The dotted line is predicted = chronological. The solid line is the regression of predicted vs chronological. (b)(d)(f) Show the absolute standardised residuals against predicted age for D_p , M_p and P_p respectively. The solid line is the regression of absolute standardised residuals against predicted age.	154
7.3	Absolute Residuals plotted against predicted age for the Epiphysis models after Box-Cox transform where $\lambda = 0.67$ (a) D_e , (b) M_e , and (c) P_e	158
7.4	Absolute Residuals plotted against age based for the generalised linear models on the instances where the epiphysis is not present (a) D_p , (b) M_p , and (c) P_p	159
7.5	Absolute Residuals plotted against age after Box-Tidwell transformation of the regressors for the instances where the epiphysis is not present (a) D_p , and (b) M_p	160
7.6	Predicted ages vs actual age for the epiphysis model DMP with three bones present. The dotted line is for predicted = chronological. The solid line is the regression of predicted vs chronological.	161
7.7	Predicted ages vs actual age for the no epiphysis model DMP with three bones present. The dotted line is for predicted = chronological. The solid line is the regression of predicted vs chronological.	162

List of Tables

2.1	The corresponding names to the bones numbered in Figure 2.4. . . .	17
2.2	The Tanner-Whitehouse stages of distal phalange III [TWM ⁺ 75]. . .	31
3.1	Table showing research groups that have contributed to the field of ABAA.	45
3.2	Features extracted from segmented carpal bones in [PKKH93]. . . .	50
4.1	The amount of radiographs of each age in the complete dataset. . . .	69
4.2	The amount of radiographs of each gender in the complete dataset. .	69
4.3	The amount of radiographs of each ethnicity in the complete dataset.	69
4.4	Ten fold cross validation accuracy of hand segmentation classification (%)	89
4.5	Percentage of the 370 outlines classified as correct by the Random Forest classifier (100).	90
4.6	Confusion matrix for Random Forest on AAM outlines.	91
4.7	Confusion matrix for Random Forest on contour ensemble (DTW) outlines.	93
5.1	The Tanner-Whitehouse stages of middle phalange III [TWM ⁺ 75]. . .	103
5.2	The Tanner-Whitehouse stages of proximal phalange III [TWM ⁺ 75]. .	104
5.3	Features derived from Tanner-Whitehouse stages.	105
5.4	Number of instances for bone segmentation classification. In brackets is the number of good segmentations /number of bad segmentations. .	115
5.5	Overall cross-validation bone segmentation accuracy (%).	116
5.6	Distal phalanx cross-validation bone segmentation accuracy (%). . . .	117
5.7	Middle phalanx cross-validation bone segmentation accuracy (%). . .	117
5.8	Proximal phalanx cross-validation bone segmentation accuracy (%). .	118
5.9	Bone segmentation accuracy of SVMQ on unseen data (%).	119

5.10	Confusion matrices of the SVMQ classifier on the features dataset. . .	119
5.11	Confusion matrices of the SVMQ classifier on the one-dimensional series dataset.	120
6.1	Number of instances for exploratory analysis of features.	123
6.2	Features ranked by information gain for distal phalange III. The first column shows the ranks of the phalanx features on all images. The second column shows the top 15 ranks of all features on images where the epiphysis is present. The information gain is given in brackets. . .	125
6.3	Features ranked by information gain for middle phalange III. The first column shows the ranks of the phalanx features on all images. The second column shows the top 15 ranks of all features on images where the epiphysis is present. The information gain is given in brackets. . .	126
6.4	Features ranked by information gain for proximal phalange III. The first column shows the ranks of the phalanx features on all images. The second column shows the top 15 ranks of all features on images where the epiphysis is present. The information gain is given in brackets.	128
6.5	The information gain of every feature for each Tanner-Whitehouse stage of distal phalange III. The rank of each feature is given in brackets, where if two or more features are equally discriminatory, they are given the mean rank.	130
6.6	The information gain of every feature for each Tanner-Whitehouse stage of middle phalange III. The rank of each feature is given in brackets, where if two or more features are equally discriminatory, they are given the mean rank.	131
6.7	The information gain of every feature for each Tanner-Whitehouse stage of proximal phalange III. The rank of each feature is given in brackets, where if two or more features are equally discriminatory, they are given the mean rank.	132
6.8	Simple decision rules extracted for classifying Tanner-Whitehouse stages of distal phalange III. Rules extracted from Figure 6.1. . . .	134
6.9	Simple decision rules extracted for classifying Tanner-Whitehouse stages of middle phalange III. Rules extracted from Figure 6.2. . . .	135
6.10	Simple decision rules extracted for classifying Tanner-Whitehouse stages of proximal phalange III. Rules extracted from Figure 6.3. . . .	135
6.11	Number of instances for Tanner-Whitehouse classification.	139
6.12	Classification of overall Tanner-Whitehouse stage accuracy (%). . . .	140
6.13	Classification of distal phalange III Tanner-Whitehouse stage accuracy (%).	141

6.14	Classification of middle phalange III Tanner-Whitehouse stage accuracy (%).	141
6.15	Classification of proximal phalange III Tanner-Whitehouse stage accuracy (%).	142
6.16	Tanner-Whitehouse stage accuracy of SVMQ on unseen data (%).	142
6.17	Comparison of results with previously proposed method [TKJP09], with all percentages rounded to the nearest whole number.	143
7.1	Observations in individual models where absolute standardised residual > 2.5 . The threshold value for the model is the median value of the $f_{k+1,n-k-1}$ -distribution.	155
7.2	RMSE for regression models where the epiphysis is detected. GP1 and GP2 are the RMSE for the two clinical estimates.	161
7.3	RMSE for regression models where the epiphysis is not detected. GP1 and GP2 are the RMSE for the two clinical estimates.	162
7.4	MAE for alternative bone combinations.	163
7.5	RMSE for regression models based on gender. GP1 and GP2 are the RMSE for the two clinical estimates.	165
7.6	RMSE for regression models based on ethnicity. GP1 and GP2 are the RMSE for the two clinical estimates.	166

Publications

- On the Segmentation and Classification of Hand Outlines, International Journal of Neural Systems (IJNS), Volume 22, Number 5, L. M. Davis, B. J. Theobald, J. Lines, A. Toms, A. J. Bagnall
- Automated Bone Age Assessment using Feature Extraction, Intelligent Data Engineering and Automated Learning (IDEAL) 2012, L. M. Davis, B. J. Theobald, A. J. Bagnall
- On the Extraction and Classification of Hand Outlines, Intelligent Data Engineering and Automated Learning (IDEAL) 2011, L. M. Davis, B. J. Theobald, A. Toms, A. J. Bagnall

Awards

- Best Presentation Award - University of East Anglia, School of Computing Sciences Research Day 2011. Title: On the Extraction and Classification of Hand Outlines.

Chapter 1

Introduction

Bone age assessment (BAA) is the clinical estimate of the skeletal maturity of a patient in relation to a normal population and is performed in hospitals worldwide on a daily basis. The main reasons for a BAA is to compare it to chronological age in order to: 1) monitor skeletal development and therefore diagnose growth disorders; 2) predict final adult height; and 3) verify age claims made by asylum seekers who may have invalid age documents. This procedure is undertaken by obtaining a radiograph of the patient's non-dominant hand [CFMC06, HLW⁺11, Rot09, TWH⁺01]. There are three reasons why the hand is used for this task: firstly, it captures a large amount of development in a small area; secondly, it exposes the patient to a minimal amount of radiation when compared to other joints e.g. shoulder; and finally, it is an easy area to radiograph. BAA is most commonly undertaken using one of two methods: Greulich and Pyle (GP) [GP50, GP59] or Tanner and Whitehouse (TW) [TW62, TWM⁺75, TWH⁺01].

The most common way the GP method is performed in clinical use is for a clinician to compare the radiograph of a patient with a standard atlas of radiographs. They then decide which of the example radiographs is closest and assign the relevant age. The standard GP atlas is made up of radiographs from the mid-western United States from the 1930's and has been found not to be a good representation of modern populations [LEM⁺93, MBP⁺01, OIAB96].

The TW method involves a clinician categorising a set of bones. These individual bones are assigned a stage ranging from B-I (Immature to Mature). Once each bone has been rated, the ratings are converted to a numerical factor using a look-up table. The sum of the numerical factors is then calculated and this forms the basis for the bone age estimate. The TW method has been found to be more accurate than the GP method [BEK⁺99] and overcomes many of the major problems associated with using atlas based methods; however, it is used less frequently because it is more time consuming.

The bone age estimate obtained by one of these methods is compared with the chronological age to determine if the skeletal development is abnormal. If a significant difference between bone age and chronological age exists, the patient may be diagnosed with a disorder of growth or maturation [HJT07, HLW⁺11].

The task seemingly lends itself to being automated. The inventors of the TW system state *“From the beginning it seemed reasonable to suppose that bone age assessments were something a computer could do better than a human operator”* [TWH⁺01]. Age estimates are now often done by more than one assessor and this has highlighted the variability inherent in the estimation techniques. This can lead to the paediatrician diagnosing the patient having a low confidence in the result. This means there is an increasing need for software that is able to make quick, accurate assessments of a patient’s bone age directly from a radiograph.

1.1 Motivation

An automated bone age assessment (ABAA) system brings multiple advantages over the current manual methods used, in that:

- assessments are more objective and therefore more likely to give the paediatrician more confidence in the diagnosis and course of treatment prescribed;
- it gives paediatricians more effective use of their time;

- it can be built upon radiographs from the local population and thus incorporate sociological and environmental factors; and
- it will save money, as diagnosis are more accurate and efficient.

Systems aimed at automating BAA have been proposed previously [ACA09, Eff93, MSC⁺00, NvM⁺03, PPKGC03, Tho02, TKJP09]. These either attempt to recreate the TW or GP methods [Eff93, MSC⁺00, NvM⁺03, TKJP09], or construct regression models for chronological age [ACA09]. The majority of these systems use Active Shape Models or Active Appearance Models to segment the bones from the radiograph, and use the features of the model to calculate the bone age. However, none of these systems have gone on to gain widespread acceptance. We believe that this is due to two main factors: firstly, a lack of verification, and secondly, a lack of transparency.

This thesis proposes the Automated Skeletal Maturity Assessment (ASMA) algorithm, an ABAA system that can do both TW stage classification and regression onto chronological age. This system involves the clearly defined subtasks of: a) hand segmentation, b) hand segmentation classification, c) bone segmentation, d) feature extraction, e) bone segmentation classification, and f) bone age estimation. Figure 1.1 summarises these stages. The images require no manual landmarking, and by separating out the feature extraction from the segmentation and regression, ASMA retains the potential for quickly and simply constructing new models for regional populations. This offers the possibility of producing age estimates tailored to local demographics based on data stored locally in film free hospitals. By performing validation checks at each stage of ASMA, the problem of lack of validation is addressed. Along with the lack of transparency problem being addressed by extracting features derived from TW stages in stage D of ASMA.

1.2 Contributions

The contributions of this thesis are as follows:

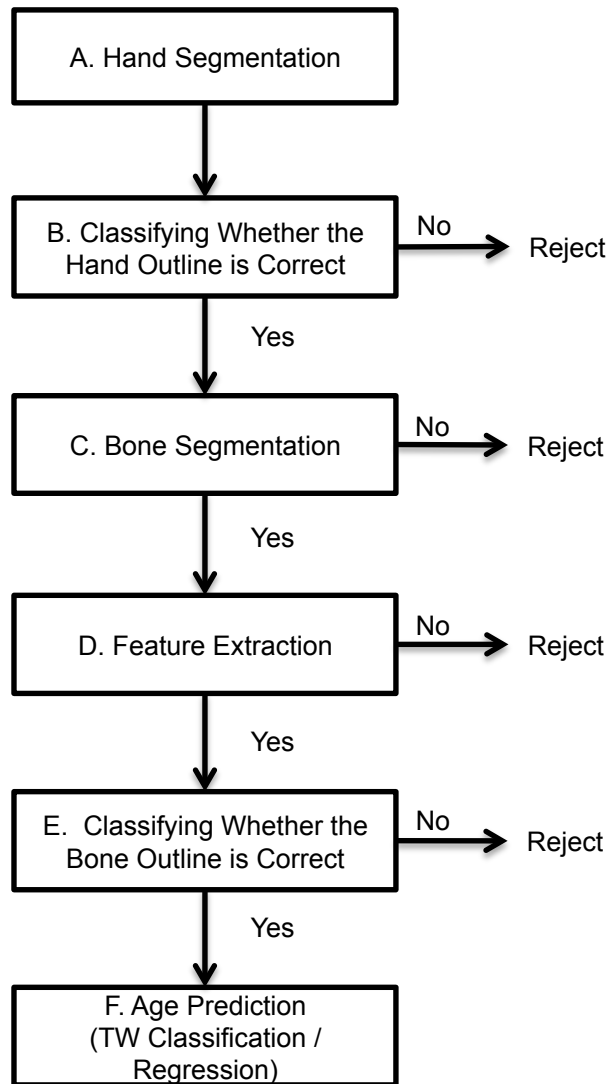


Figure 1.1: The six stages of the Automated Skeletal Maturity Assessment (ASMA) system.

- The ASMA algorithm. This is a stage based system and has the advantages that: an individual stage can be updated without affecting the other stages and that validation checks are performed after each segmentation. These are used to ensure that no bad segmentations get through to the latter stages of the system. Furthermore, the assessments are based on features derived from the TW method and this allows for transparency when explaining the system to clinicians.

- For the hand segmentation stage of ASMA we propose the use of a contouring algorithm (Section 4.2), that has previously not been used for this task. We show that this algorithm is capable of performing as well as three previously proposed methods [DTTB11, DTL⁺12].
- In order to overcome problems identified within the hand segmentation process, a novel ensemble algorithm for combining outlines using two voting schemes is introduced (Section 4.3). The first of the voting schemes is based upon shape, with the one-dimensional series of the outline being compared to a set of idealised outlines using the Dynamic Time Warping (DTW) distance metric. The second voting scheme is based upon the log-likelihood ratio derived from the pixel intensities both inside and outside the outline. We also demonstrate that, when used with the DTW voting scheme, the ensemble improves the performance of all image segmentation algorithms [DTL⁺12].
- For the classification of the segmentations (both hand and bone) we investigate (Sections 4.4 and 5.3) the use of a variety of representations/transformations and machine learning classifiers to make classification schemes that have not been used for this purpose before [DTTB11, DTL⁺12].
- A novel technique for bone segmentation from a region-of-interest (ROI) box is introduced (Section 5.1) [DTB12]. This uses Canny edge detection in conjunction with a Gaussian pyramid in order to find the bony tissue.
- In order to extract features relating to height and width, a novel technique that uses the elliptical Hough transform in conjunction with Gaussian pyramids is described (Section 5.2) [DTB12].
- We perform an exploratory analysis of the features extracted and examine how these affect the skeletal development process (Section 6.1) [DTB12]. Also, we investigate the use of a variety of classification techniques to classify bone age according to the TW standards, and prove that classifications done in this way

are as accurate as previously proposed methods [NvM⁺03, TKJP09], whose classifications are based upon AAM/ASM features (Section 6.2).

- Finally, we present methods for regressing onto chronological age using individual and multiple bone models. We demonstrate that ASMA performs at least as well as previously proposed methods [ACA09] and two manual GP raters (Section 7.3). We also demonstrate the use of gender and ethnicity independent models, verifying the claim that in the future ABAA systems should be tailored for local populations (Section 7.4).

1.3 Organisation of Thesis

The rest of this thesis is structured as follows. Chapter 2 introduces bone age and BAA. It also covers the history of BAA and radiology, describes how the GP and TW systems are used in a clinical setting in more detail, and finally, how automating the process will help. Chapter 3 describes relevant image processing and machine learning techniques and previously proposed ABAA systems. We investigate why none of these systems have gone on to gain widespread acceptance and how the proposed ASMA approach differs from previously published work. Chapter 4 introduces the first two stages of the ASMA system, hand segmentation and classification. It discusses the use of four different segmentation methods, proposes the ensemble outline detector, and finally, describes a variety of classification schemes in order to predict if a hand segmentation is correct. Chapter 5 discusses the methods used to locate the ROIs around the bones of the middle finger given the hand outline and how to segment the hard tissue from the ROI. After the bone has been segmented the features and the methods used to extract them are described. Finally, as with the hand segmentation, a variety of classification schemes are investigated. The TW classification part of ASMA is discussed in Chapter 6. In this chapter an exploratory analysis of the predictive power of the features is performed. We then investigate the use of a variety of machine learning classifiers for classifying TW stages. In

Chapter 7, we investigate using the extracted features in a linear regression onto chronological age, by building general overall models, as well as independent gender and ethnicity models. Finally we discuss the conclusions of this work and possible future directions in Chapter 8.

Chapter 2

Radiology and Bone Age Assessment

In this chapter we describe the background for the project in relation to clinical radiology. Bone age assessment (BAA) is a task regularly performed by paediatricians to monitor skeletal development and the effects of certain drugs in hospitals around the world [vRLR⁺01]. Manually predicting bone age from radiographs is a difficult and time consuming task. The aims of this chapter are as follows:

- introduce bone age and show how it differs from chronological age (Section 2.1);
- describe the uses of BAA (Section 2.2);
- describe the X-ray acquisition process (Section 2.3);
- discuss the history of radiology and BAA (Section 2.4);
- describe how BAA is performed currently in clinical settings (Section 2.5); and
- identify how automating the process will help (Section 2.6).

2.1 What is Bone Age?

To fully understand BAA, we must firstly understand bone age. A person's bone age should not be considered the same as their chronological age. Where chronological age refers to the amount of time a person has been alive, bone age describes the current state of skeletal development of a person.

BAA involves using various factors such as the size and shape of the epiphysis (see Figure 2.1) of a bone to measure the development from immature to mature. Each individual epiphysis in the hand is known as an ossification centre. The process of ossification is the conversion of soft tissue into hard tissue to form the epiphysis, which then fuses to the bone. The actual development of a patient is then compared to standard forms of skeletal maturity to estimate bone age. A single BAA gives the paediatrician information about the patient's skeletal development at that time. If further assessments are performed, the progress of the patient's skeletal development can be ascertained. Bone age within 10% of the patient's chronological age is considered to be normal by clinicians [GR04]. Although the exact factors that give normal skeletal maturation are not known, it is thought that genetics, hormones and environmental factors play an important part.

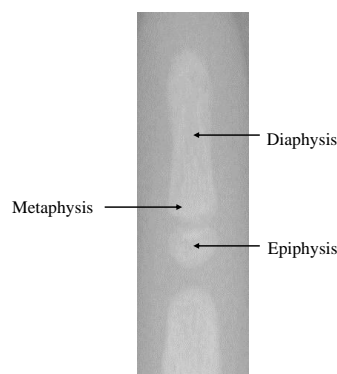


Figure 2.1: An example radiograph of the distal phalanx of the middle finger, with epiphysis, metaphysis and diaphysis labelled.

2.2 Why is Bone Age Assessment Performed?

There are three main reasons why BAA is performed by paediatricians, these are:

1. Diagnosis of health issues;
2. Prediction of final adult height; and
3. Verifying age claims

2.2.1 Diagnosis and Management of Health Issues

By performing a single BAA paediatricians can diagnose whether a patient is suffering from a growth and/or puberty disorder [HJT07, TWM⁺75]. The two types of growth disorder that can be diagnosed are Primary and Secondary. Primary growth disorders generally occur due to prenatal damage or a genetic defect. In this type of skeletal deficiency, bone growth and height are affected rather than skeletal maturity.

Secondary growth disorders are generally caused by other factors e.g. nutritional or metabolic. With this type of skeletal deficiency, skeletal maturity and height are both affected. However, if found early enough and treatment applied it is possible for a patient to reach full adult height [GR04].

Diagnosing the correct growth disorder can be difficult in cases where skeletal development has been affected less than height. Treatment for a growth deficiency can be monitored, by performing subsequent BAA and therefore, gives an indication of whether treatment is working [GR04].

2.2.2 Prediction of Final Adult Height

Another use of BAA is for predicting a patient's final height [TWM⁺75]. There are various factors that can affect a person's final height. However, under normal

conditions it is accepted that adult height is mainly hereditary and methods based upon this have been presented [RC02].

Another common method to calculate adult height is to use the current height of the patient and the heights from earlier in their life. The main drawback of using this method is that every child develops differently, with some obviously developing earlier than others. Therefore if the degree of development is known, which BAA gives us, a final height prediction for a patient is potentially more accurate. Various methods that use BAA to predict final adult height have been proposed [RWT75, TWMC75]. All of these methods are based on children that have grown up healthily, and therefore are more accurate/reliable for healthy children. The described methods have 95% confidence intervals of 7 to 9cm [GR04]. The method proposed by Tanner *et al.* [TWMC75] uses Equation 2.1 to predict final height, where p refers to the predicted final height of the person (cm), α refers to the height coefficient, h is the patient's current height (cm), β is the age coefficient, a refers to the patient's chronological age in years, γ is the bone age coefficient, b is the bone age in years and c is a constant.

$$p = \alpha \times h - \beta \times a - \gamma \times b + c \quad (2.1)$$

As females and males develop at different rates to each other, there are different coefficients and constants tables for both. For females, pre and post menarche tables also exist.

2.2.3 Verifying Age Claims

Since the 1980's there has been an increase in the amount of people applying for asylum in European Countries [ZGFP03]. This has lead to a need for methods to verify age, as this can be of great importance to local authorities [HoCoHR07] for various reasons such as voting, legal responsibility, enforcement of juvenile criminal law or general criminal law.

The groups most affected by this are those that do not know their age or have invalid age documents, and those suspected of giving a false age. The task to calculate the person's age is very difficult and a common way to determine age in these people is to use a combination of BAA, dental assessment and physical assessment [BW08, IC06, Spa95, SOR⁺03, vRLR⁺01]. These methods have been found to be more accurate when used as in combinations than individually [BW08, SOR⁺03]. However, the use of BAA for this task is controversial as it is introducing the patient to radiation which may be unnecessary [BW08, IC06, Sel12].

2.3 X-ray Acquisition Process

The discovery of X-rays, by Röntgen in 1895, created a new field of clinical radiology which allowed new medical techniques that were previously unimaginable. Röntgen performed various experiments with his new rays, testing them on card and various metals. On December 22nd 1895 [SB10], Röntgen took a series of X-rays of his wife's hand. One of these is shown in Figure 2.2. On 1st January 1896 at the University of Freiburg, Röntgen presented his “new rays” and the experiments that he performed [Sta96].

X-rays are electromagnetic waves that are outside of the visible spectrum, due to them having shorter wavelengths and thus more energy. Due to the shortness of the wavelengths, X-rays can be thought of acting as a particle as well as a wave. Therefore can be referred to in terms of energy as well as length.

The process of generating X-ray photons is performed in a vacuum (X-ray) tube. The X-ray tube is made up of: a glass casing, a cathode/filament which emits the electrons, an anode which the electrons collide with, and finally the radiation window in the tube from where the X-ray photons are emitted, this is usually made from Beryllium due to the fact that it absorbs a very small amount of the X-ray photons that penetrate through it.

In order to create the X-ray photons, a current is applied to the filament/cathode.



Figure 2.2: An X-ray of the hand of the wife of Wilhelm Conrad Röntgen [Spi00], taken on December 22, 1895.

As the cathode heats up, negatively charged electrons are emitted, these electrons are then accelerated towards the positively charged anode. Between the cathode and the anode the electrons are focussed towards the focal point of the anode. As with standard image processing the size of the focal point has an affect on the resulting image, with the larger the focal point, the less sharp the image. The X-ray photons are created when the electrons collide with the anode, and are then emitted from the vacuum tube through the radiation window located below the anode.

X-rays react differently with materials depending on the speed to which the electrons are accelerated and the material in question. For medical diganostics such as BAA, the acceleration voltages of the electrons lies in the range 25-150 kV [Buz08]. This gives the opportunity to view objects that are covered by a surface opaque to visible light e.g. the bones in the hand.

In order to create an X-ray image in a medical diagnostic procedure, the emitted X-rays penetrate different types of material (tissue) in the human body. Where each

of these have different levels of attenuation which effects the intensity of the X-ray. The basic equation for the attenuation of an X-ray through a certain material can be calculated using the Beers-Lambert [KC01] law:

$$I = I_0 e^{-\mu\eta}. \quad (2.2)$$

Where I refers to the final intensity, I_0 is the intensity at the source, η is the length of the path through the material, and finally μ is the linear attenuation coefficient of the material penetrated. The main type of interaction between X-rays and material when taking a hand radiograph, is for the X-rays to penetrate through the tissue. As well as this there are another two types of interaction that occur when the X-rays are penetrating through the different tissues of the human body:

- Photoelectric Interaction: this is where the energy of the emitted X-ray photon is absorbed by an electron, and thus totally absorbed by the tissue [KC01].
- Scatter: this is where the emitted X-ray photon is deflected from its original course. This can be either Compton or coherent scatter, where the former results in a slight loss of energy of the photon [KC01] and the latter does not have any loss in energy.

The final stage of acquiring the X-ray images is the detection of the X-ray photons after they have penetrated the object in question. This is done in one of two ways. The first way is to create an analog image, this is done by converting the X-ray photons into visible light performed using a scintillator. After the photons have been converted to visible light, the image is acquired using photographic film. However, since the rise of digital imaging and the move to more film free hospitals, the use of semi-conductors that detect the X-ray photons and thus create a digital image has increased.

2.4 A Brief History of Manual Bone Age Assessment

A diagrammatic view of the history of BAA is shown in Figure 2.3. In 1897, Behrendsen [Beh97] published one of the first studies of the developmental process of bone formation (ossification) in the hand. This study looked at various radiographs from newborn to the age of twenty. This study was not gender specific and was made up predominantly from radiographs of living subjects. A remark is made that the radiograph's of the younger subjects were harder to obtain due to their constant movement. As the majority of studies are performed using the hand we have provided a radiograph of the hand and wrist which has all the bones numbered in Figure 2.4, the name of each bone, along with its group can be seen to the corresponding number in Table 2.1, as these will be referred to throughout.

Pryor [Pry07] undertook a study of ossification in 1907. The study consisted of 360 radiographs of children aged ten and younger, with 300 of them below the age of seven. The children were from 225 families, ten families in the study having two or more children. From the study, Pryor made seven observations. These were:

1. The ossification process starts sooner than previous studies had suggested.
2. The ossification process for females is faster than that of the males.
3. The order that the carpal bones ossify is different to that which previous studies had suggested.
4. The ossification process in the first child of family starts at a younger age than subsequent children.
5. The ossification takes place at the same time in both hands.
6. The fusion of the epiphysis starts earlier than previously had been suggested.
7. The order of the ossification of the carpal bones is an inheritable trait.

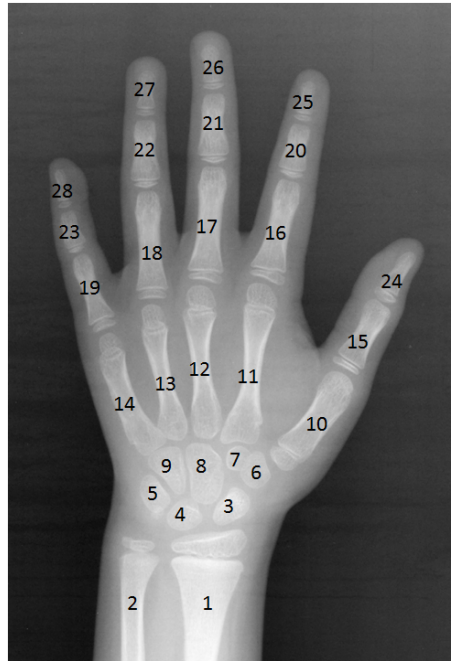


Figure 2.4: A radiograph of the hand with all bones numbered, the corresponding names can be seen in Table 2.1.

Table 2.1: The corresponding names to the bones numbered in Figure 2.4.

Number	Name	Group
1	Radius	Radius
2	Ulna	Ulna
3	Scaphoid	Carpals
4	Lunate	
5	Triquetrum	
6	Trapezium	
7	Trapezoid	
8	Capitate	
9	Hamate	
10	Metacarpal I	Metacarpals
11	Metacarpal II	
12	Metacarpal III	
13	Metacarpal IV	
14	Metacarpal V	
15	Proximal Phalanx I	Proximal Phalanges
16	Proximal Phalanx II	
17	Proximal Phalanx III	
18	Proximal Phalanx IV	
19	Proximal Phalanx V	
20	Middle Phalanx II	Middle Phalanges
21	Middle Phalanx III	
22	Middle Phalanx IV	
23	Middle Phalanx V	
24	Distal Phalanx I	Distal Phalanges
25	Distal Phalanx II	
26	Distal Phalanx III	
27	Distal Phalanx IV	
28	Distal Phalanx V	

This was followed by a more thorough study by Rotch [Rot09] in 1909. The main aim of this study was to understand in more detail the “anatomic changes” that take place in the early development of a child and to help solve problems associated with this development. In the study an analysis of 1000 cases was undertaken, 200 children of varying ages were chosen as being examples of normal development. These children had an X-ray of their hand and wrist taken and their skeletal development (known as a Röntgen record) and chronological age (years and months) noted. Compared to previous studies such as [Beh97] which used dead and living subjects, this study uses only living subjects and states that the differences seen in the order of ossification in previous studies is due to this mixture of subjects. This study observes and agrees with the order of ossification stated by Pryor [Pry07], which also only used living subjects. The main conclusions, in relation to bone age, were:

1. that when identifying problems connected with a child’s development, assume that age does not just refer to age but refers to bone age, chronological age and physiologic age;
2. there is a need for an index for use by physicians to analyse bone age and therefore fitness for school or physical work. Height, weight, teeth, birth certificates and statements of parents and/or guardians are inadequate for this purpose;
3. the belief that bone age and physiologic age will probably have some correlation and will be more important than the use of chronological age to solve problems in early development;
4. the use of the appearance of teeth to classify development in children is unreliable;
5. the human skeleton is the most appropriate part of the body to build an index of development. The best part of the skeleton to use is the joints, due to the epiphyses that appear there as they are connected to a child’s growth

and are early places to spot infection of growth diseases. The carpal bones and the bones of the wrist and hand are the optimal joint as they give more evidence of the state of development and are easy to acquire radiographs of. The development of the hand and wrist is best measured by classifying them into stages. This can only be achieved through the technology of X-rays; and

6. skeletal development is not the same as muscular development and the two may not be correlated.

After this study by Rotch and the statement of the apparent need for an “*Anatomic Index*”, various methods for BAA were proposed.

2.4.1 Atlas Methods

In 1898 Poland [Pol98] was the first person to use the idea of creating an atlas of radiographs that display normal development and therefore the ossification process in the hand and wrist. The atlas is made up of 19 radiographs from age one to 17 years, with a mixture of females and males. Accompanying each radiograph is a description of the skeletal development at that time. The most common application of this method is to compare a patient’s radiograph to each standard, then to choose the closest match and use this as the bone age estimate.

From 1926 to 1936, Todd [Tod37] undertook a thorough study of skeletal maturation and through a selection process produced his atlas of the hand and wrist. In this study over 4000 children were assessed and more than half of this number attended 12 or 13 assessments. Since previous work by Pryor [Pry07] and Rotch [Rot09] had agreed that females and males do not develop at the same rate, standard images for males and females were given. There are 40 male standard images ranging from three months to 18 years and nine months. For females there are 35 standards, taken from three months to 16 years and three months. For both genders an assessment was taken every three months until the patient was 15 months old. Assessments were taken every six months thereafter. In order to calculate the stan-

dard radiograph for a certain age, all the X-rays of that age are taken and ordered by the rate of maturity. A central group of this ordered list is then separated from the outliers. From this central group the image that most accurately represents the mode is chosen. Todd had planned for this to be the first of a series of six atlases. In the study, radiographs of six joints were taken. These were the shoulder, elbow, hand/wrist, hip, knee and foot/ankle, with an atlas planned for each. Unfortunately due to Todd's death in 1938, this was the only atlas he published.

Greulich and Pyle (GP) published the first edition of their "*Radiographic Atlas of Skeletal Development of the Hand and Wrist*" in 1950 [GP50]. This atlas was based upon the study conducted by Todd [Tod37]. The atlas published by Todd only used radiographs in the study up to 1936. However, the study continued and the GP atlas used radiographs up until the termination of the study in 1942. The children who took part were assessed every three months in the first year of life, every six months until the age of five and annually thereafter. The only exception to this was during puberty (where skeletal development is known to be accelerated), therefore standard radiographs are also provided at the age of 13 and a half for females and 15 and a half for males. As a result, this study has less standards than that of Todd [Tod37]. The authors claim there is no need for references every six months after the age of five.

A second edition of the Greulich and Pyle atlas was released in 1959 [GP59]. For this edition the standard radiographs were replaced with radiographs of new patients that showed the same level of skeletal development. Four new male standards were added, since more development occurs between two of the standards than previously claimed in [GP50]. A new female standard was also added for the age of three years old. The atlas consists of 31 standard radiographs of males from newborn to the age of 19 years and 27 standard radiographs of females from newborn to the age of 18 years. Standards are also given for females at the ages of 28 and 50. Along with each standard, there is a piece of text describing the development. Each stage of skeletal maturity for each bone is also described.

As well as the hand/wrist joint, atlases of other joint areas were published. Hoerr *et al.* [HPF62] in 1962 published an atlas of the foot and ankle. This was seen as the third in the series of skeletal atlases released from the Western Reserve University of Cleveland, Ohio after the atlases by Todd [Tod37] and Greulich and Pyle [GP50, GP59]. The radiographs used to make up the atlas were again taken from Todd's study from 1931 to 1942, and the authors state that the second atlas to be published by Todd was going to show the skeletal development of the foot/ankle joint. The reason for producing atlases of different joints is due to the development rates of each joint being uncorrelated. In the foot and ankle, the allowed variation in the shape of regularly occurring bones is larger than anywhere else in the body. However, the authors state it is still possible to calculate the skeletal maturity of a patient. In total 30 standard plates are used to show the skeletal development, with each being assigned a male and female skeletal age. As with the GP atlas of hand and wrist each standard has a piece of text summarising development and every bone has each of its developmental stages described.

The next atlas to be released by the group at the Western Reserve University of Cleveland, displayed the skeletal development of the knee [PH69]. The knee was used as it has fewer bones than the other joints and was therefore thought to be easier to use. Again the radiographs were supplied from Todd's study [Tod37]. In all there are 30 plates in the atlas with each having a description about the skeletal development. The first plate is made up of 6 neonatal radiographs and therefore the skeletal ages are given in fetal weeks. The second plate is made up of two male and female images within the first two weeks of birth. Plate three is of a radiograph taken after a month of life. This only contains a front view of the knee. All plates after this point include a front and side view of the patient's knee. The skeletal age of the final plate is at least 19 years for males and at least 16 years for females.

Pyle *et al.* [PWG71] released an adjusted atlas based on their previous work in 1959 [GP59], to be used in conjunction with the National Health Survey in the U.S.A. The survey was commissioned in 1956 by the U.S public health service. In

total the survey ran for 7 years between 1963 and 1970. Assessments took place at 40 locations across the U.S.A. The survey was split into three programs and the atlas was used for the second and third parts of the program. The atlas contained 26 standard plates of the hand and wrist with skeletal ages ranging from 3 to 216 months for males, 3 to 192 months for females.

In 1976, De Roo and Schröder released their “*Pocket Atlas of Skeletal Age*” [dRS76]. The authors stated that although many atlases had already been published, they were difficult to use if needed quickly. It is also stated that radiographs on previous works are not clear. The work published is in a “pocket-size” form and therefore can be carried around. The radiographs used in the atlas are generated from a large set of children. A new technique was used in developing the radiographs and this is claimed by the authors to make clearer images than those seen in previous publications. There are 31 reference images for males. These range from newborn to 18 years of age. A radiograph was taken every one and a half months up to the age of three months. They were then taken every three months up to the age of two years, bi-annually up to the age of seven and annually thereafter. There are 26 standard images for females in the same range as the males. These were taken at three monthly intervals until age two, six monthly until age four and annually thereafter. Along with each standard image there is a short piece of text stating the current amount of development.

Brodeur *et al.* [BSG81] published a skeletal atlas on the elbow joint in 1981. They discuss how the elbow has a complex maturation due to it having the most secondary centres of ossification. This is followed by a description of its maturation process. It is stated that the maturation process is unlike that of the wrist as it is not an orderly process. However, there is a pattern to the process that is “*reasonably reliable*”. The atlas is made up of 66 reference radiographs for males and the same amount for females, both ranging from birth to the age of 16 years and six months. Each reference is made up of two images; one where the arm is straight and the other where the elbow is bent. As with the other atlases there is a statement with

each reference describing the development. There are reference images for each four months for the first year and at six month intervals after this. After eight months the reference plate at each age is split into a high normal and a low normal reference.

One of the latest skeletal atlases to be released is that of Gilsanz and Ratib [GR04] in 2004. The data collected for their atlas was from the study of the Childrens Hospital Los Angeles. The participants in the study were healthy children or adolescents that were from families of European descent and had no history of chronic illness or taking regular medication. The studies were approved by the local Institutional Review Board (IRB) and each patient's parent signed consent forms. To make the atlas, 522 radiographs were used, half of which were male. The radiographs of each sex were split into 29 different age groups ranging from eight months to 18 years. Two independent radiologists ordered the radiographs in rate of maturity on six different groups of bones. These were: proximal phalanges, middle phalanges, distal phalanges, metacarpals, carpals, radius and ulna. The middle or average radiograph was selected as the amount of skeletal development for that particular group of bones at that age. However, this leads to more than one radiograph being selected as the average at one age. In order to overcome this problem the radiographs containing the average bones for that age group were merged to make an idealised image. There were four image processing steps used to merge the radiographs: firstly, all radiographs were resized to 800 x 800 pixels and the background made uniformly black; secondly, the contrast was optimised; thirdly, the radiographs were processed using an unsharp filter to enhance the edges and finally, the images were merged by replacing bones using translation, rotation and warping. These images were then resized again to 240 x 240, so that they can be used with PDAs and other hand held equipment. The authors tested the use of their atlas against the GP atlas [GP59] by having two independent radiologists assess each radiograph using both. The results of the test showed that there was strong correlation between using the two atlases and that there was no statistical difference.

2.4.2 Oxford Method

The Oxford method was proposed by Acheson [Ach54], with the method named after the institution Acheson was working for. In the paper the other methods of assessing skeletal maturity are described and the disadvantages analysed. The new technique is then described. This method assesses each individual bone and assigns a score dependant on where it is in the development process. The development process for each bone is broken down into different stages. These stages are based upon the skeletal maturity indicators given in [GP59]. An example of how to use the method on the knee and the hand/wrist joints is presented. The authors address the issue of whether scores should be weighted for different bones. By performing a test, where the different bones are treated equally and the different joints independently. The results of this test show that the ossification centres of the hand/wrist area need to be weighted differently where as the ossification centres in the knee do not. It is also proposed that future studies should investigate whether the scores from different joint areas can be combined, based upon a percentage system.

In 1975, Tanner *et al.* [TWM⁺75] published a revised version of the method they had described previously [TW62]. The revised method shall hence be referred to as the TW2 method and the original as TW1. The method proposed is similar to that of Acheson [Ach54] in that it awards scores dependant on stages of the hand/wrist, with each bone having eight or nine stages. The stages for the TW2 are the same as that in TW1. However, the scores have been updated. Another change is that the TW2 method differentiates between males and females. The authors state that females are always advanced in skeletal maturity when compared to males of the same chronological age and in fact complete growth two years earlier. The new system has two separate ways of calculating maturation. The first is to use the radius, ulna and short bones (RUS) (the short bones cover the metacarpals and phalanges of fingers one, three and five). The second uses the carpal bones. The RUS method is found to outperform the carpal bone technique and is easier to use. Unlike the Acheson method each stage does not assign one point. Instead they

designed the scores “*in such a way as to minimise the overall disagreement between the bones*”. Once all bones have been awarded a score, these scores are summed and the bone age is calculated. The authors state that their method is appropriate for all populations. The mean score for a certain bone age is different for different populations, but this is to be expected.

The TW3 method was published in 2001 [TWH⁺01]. The maturity stages and scores remained the same as the TW2 method. The main difference is that the centile charts used to convert the Skeletal Maturity Score (SMS) (sum of bone ratings) into bone age are now updated to adapt to the modern population. The authors also state that other centile charts should be used when assessing other populations. Between the publication of the TW2 and TW3 method various studies have been undertaken [TWH⁺01]. These show that when a single observer has been asked to rate a radiograph twice, they get the same stage rating in 90% of instances. When different observers rate the same radiograph, the same rating was given 75-85% of the time, dependant on the study. However, when different stages were given they were adjacent stages. The authors also suggest that BAA is something that a computer can do better than a human.

One of the latest BAA methods to be published is that of Hseih *et al.* [HLW⁺11]. The Grouped-TW method (GTA), was developed in order to simplify the TW approach in addition to making it more efficient. It is stated that the GTA method takes the advantages of the GP [GP59] and TW3 [TWH⁺01] methods. Instead of using 13 bones like TW3 RUS method, three groups of three bones are used. Each bone in each group has its SMS score calculated. The bones in the group are then summed and the bone age calculated using centile charts like the TW method. This gives three age estimates. Each estimate is then weighted and summed together to give the final bone age.

In Section 2.5, we discuss the Oxford based TW and atlas based GP techniques in more detail, because these techniques are used most frequently for BAA [Spa95].

2.4.3 Planimetry Method

In 1928, Baldwin *et al.* [BKB28] published a method based upon the measurements of the epiphyses and carpal areas in a hand/wrist radiograph. These measurements were taken by putting the radiograph on an illuminating box and using a planimeter or sliding callipers. The study looked at 1300 radiographs taken between 1918 and 1928, children were aged from birth to 17 and a half years. The radiographs were split into different age groups. There were monthly assessments between birth and six months, bi-annually between six months and one year six months, then annually to seventeen years-six months. Records were then made of age, gender, number and name of carpal bones, number and location of epiphyses, if fusion of an epiphysis had begun and if the child was over one year, two diameters of the wrist are also recorded.

Cameriere *et al.* [CFMC06] published a version of the planimetry method for BAA in 2006. This involved calculating the ratio of the area of the carpal bones alone to the same area combined with epiphyses of the radius and ulna. The study was conducted on 150 Italian children aged between 5 and 17. Linear regression was performed to find the relation between age and this ratio. The results from testing show that the method has a standard error estimate of 1.19 years.

2.4.4 Numerical Method

Sontag *et al.* [SSA39] were the first to propose the use of a numerical method. Numerical methods investigate the number of ossification centres at various ages. A study from 1932 to 1939 was undertaken. This consisted of taking radiographs of the joints of the left side of a patient's body. Assessments were undertaken at monthly intervals up to twelve months and then at six monthly intervals up to five years, with 149 children. An investigation of 67 ossification centres was undertaken. An ossification centre was counted as soon as it had appeared on an X-ray, with the mean and standard deviation calculated at each age interval. There was found to

be a different rate of ossification centres appearing for males and females. Due to this difference, a curve for each gender was calculated. These could then be used to calculate the skeletal development of a patient. The advantage of using such a method is that it is not subject to the same errors as if only one joint is used. In addition, the method is efficient and more objective than other methods and is simple to use. However, the disadvantages are that it does not rate the development of an epiphysis after its appearance and the method is not applicable to children over the age of five.

In 1946, Elgenmark [Elg46] proposed a similar method to that of [SSA39]. The author remarks that the previously proposed methods do not work on the Swedish population and that a new method that produces satisfactory results is required. The data comes from a study at the Samariten Children's Hospital in Stockholm, Sweden which was performed between 1942 and 1945, with 429 males and 423 females assessed. The children were in the same range as in [SSA39]. However the ages for assessment were more frequent, with children being assessed every three months between the ages of one and three. The joints of the right hand side of the body were assessed and in 59 cases both sides of the body were assessed. In this study 68 ossification centres were studied to find their appearance. A comparison is made on cases where both sides are assessed and it is found that they do not mature uniformly. It is also found that there is a higher correlation between appearance of ossification centres and height than with age. Therefore it is proposed that height and number of ossific centres should be used to determine skeletal development.

These methods require the patient to be exposed to a large amount of radiation. Since our knowledge of the effects of prolonged exposure to radiation have increased, the use of such methods has decreased.

2.4.5 Age of Appearance Lists

Garn *et al.* [GRS67] propose the Age of Appearance list method for BAA. They undertook a study and found that there are six main body parts involved in skeletal

development. These are: hand, foot, elbow, knee, shoulder and hip. Radiographing the hip exposes the patient to a lot of radiation and is the least contributory of the six joints, therefore this joint is not needed for BAA. The order that the various ossification centres appear through out the body was investigated and ranked in order of appearance. They state that such ranking would be appreciated as it shows which ossific centres need most attention and that this forms the basis of a point additive system if the highest ranking centres were selected. The top 20 ossific centres were selected for males and females, as has been stated before males and females develop differently and hence the same ossific centres do not appear on both lists. However, the top ranking ossification centres of both only contain bones from the hand, foot and knee. Therefore, three radiographs are needed instead of the six to eight that would be needed if all the joints were used. This makes the solution more efficient and the patient is exposed to less radiation. However, this is still more radiation than using other methods and hence it is rarely used.

2.5 How Bone Age Assessment is Currently Performed

BAA is currently performed in hospitals worldwide on a daily basis. This procedure is undertaken by obtaining a radiograph of the patient's left hand. It is widely accepted that the hand is a good indicator of skeletal maturity [CFMC06, HLW⁺11, Rot09, TWH⁺01]. This is for three main reasons: firstly, it has many ossification centres in a small area; secondly, due to the small area, the patient is exposed to minimal radiation; and finally, it is an easy area to radiograph. An examination of the skeletal development of the hand is then undertaken using one of two methods: the Atlas method of Greulich and Pyle (GP) [GP50, GP59] or the Oxford style method proposed by Tanner and Whitehouse (TW) [TW62, TWM⁺75, TWH⁺01].

The bone age determined from the method used is then compared with the chronological age to determine if the skeletal development is at the expected rate. If

there is a significant difference between the patient's bone age and chronological age then the paediatrician will diagnose the patient with a growth or puberty disorder [HJT07, HLW⁺11].

2.5.1 The Greulich and Pyle Method

The atlases published by Greulich and Pyle were described in Section 2.4.1. The most common way the method is implemented is for the clinician to check the patient's radiograph against each of the example radiographs in the atlas. When comparing against each radiograph certain features of the skeletal development are checked. Such features are, firstly, the development of the epiphysis and secondly, the presence of certain carpal bones. Once the example radiograph that the clinician believes to show the skeletal development closest to that of the patient has been decided upon, they assign the age of the radiograph as the patient's bone age.

The main disadvantages of using this method are that it is subjective and therefore it is harder to reproduce a diagnosis [Ach54, BEK⁺99]. The use of an atlas method assumes that the ossification process happens in an orderly fashion among all people, however this may not be the case [Ach54]. Another criticism is the long intervals between standards. However this cannot really be addressed now that the effects of prolonged exposure to radiation are well documented. A study in 1990s U.S.A found that for certain areas of the modern population, the atlas was not a good representation [OIAB96], although a more recent study was performed in the Netherlands and the atlas is found to be still valid [vRLR⁺01].

2.5.2 The Tanner and Whitehouse Method

In contrast to the atlas based method of Greulich and Pyle [GP59], the Tanner and Whitehouse (TW) method [TWH⁺01] grades a selection of bones dependant on the method being used. The various methods published by Tanner and Whitehouse were discussed in Section 2.4.2. Here is a list of the various methods proposed and

the bones analysed:









- TW2 20 Bones: radius, ulna, phalanges and metacarpals of fingers one, three and five, carpals;
- RUS Bones: radius, ulna, phalanges and metacarpals of fingers one, three and five;
- Carpals: carpals.

Each bone has various stages and each stage has various criteria for a bone to be at a certain stage of skeletal development all these criteria must be met. Table 2.2 shows an example of the criteria needed for each stage of skeletal development for the distal phalange of the middle finger. Also shown are visual representations of what each stage should look like. Each stage for every bone has a numerical score once a score has been obtained for each bone, these are summed together to give the SMS. This is converted into a bone age using a centile chart.

Each bone at each stage has a certain score associated with it. Generally each set of bones (e.g. metacarpals) have similar scores associated with them. The most heavily weighted bones are those of the wrist, the radius and ulna. In this work we concentrate on segmenting the phalanges of the middle finger, this is for three reasons. Firstly, as these are finger bones they should be easier to segment than the bones of the wrist and palm, secondly, it gives the ability for the work to be compared to previously proposed methods [ACA09, NvM⁺03, TKJP09], and finally, the middle finger is the most heavily weighted of the fingers used. In future editions of ASMA, adding more bones to the system will be investigated. However, at this stage of development, these three bones are used to investigate that the concept works.

The advantages of using such a method in comparison to the GP method are that it overcomes the subjectivity and results are more reproducible [BEK⁺99]. It does not have a strict order of ossification. A centile chart can be calculated for any

Table 2.2: The Tanner-Whitehouse stages of distal phalange III [TWM⁺75].

Stage	Image	Description
B		The centre is just visible as a single deposit of calcium, or more rarely as multiple deposits. The border is ill-defined.
C		The centre is distinct in appearance and disc-shaped, with a smooth continuous border.
D		The maximum diameter is half or more the width of the metaphysis.
E		The epiphysis is as wide as the metaphysis. The central portion of the proximal border has grown toward the end of the middle phalanx, so that the proximal border no longer consists of a single convex surface; no differentiation into palmar and dorsal surfaces, however, can yet be seen.
F		Palmar and dorsal proximal surfaces are distinct, and each has shaped to the trochlear articulation of the middle phalanx. The palmar surface appears as a projection proximal to the thickened white line representing the dorsal surface.
G		The epiphysis caps the metaphysis.
H		Fusion of epiphysis and metaphysis has now begun.
I		Fusion of epiphysis and metaphysis is completed.

population which makes the method usable for any population. The TW method does not have discrete intervals and therefore removes some of the restrictions of the GP method. However, rating individual bones is time consuming and so the GP method is used more often.

2.6 The Need For Automation

Tanner and Whitehouse believed BAA would be a task that a computer could undertake with more accuracy than a human assessor [TWH⁺01]. We believe there are many (potential) advantages of having an automated system. These are:

- More accurate results as it will overcome the subjectivity problem associated with the GP method and hence giving paediatricians more confidence in their diagnosis.
- A more efficient process than the TW method.
- It will give paediatricians more effective use of their time.
- As diagnosis are more accurate and efficient, it will save money.

Previous automated BAA systems have been proposed by [ACA09, Eff93, MSC⁺00, NvM⁺03, PPKGC03, Tho02, TKJP09] and are discussed further in Chapter 3.

2.7 Summary

In this chapter we have discussed:

- what bone age is and how it differs from chronological age;
- the main uses of bone age assessment and thus why it is important;
- a timeline of manual bone age assessment that covers the main discoveries of the procedure and the many different methods proposed to implement it;

- the methods that are currently used to for bone age assessment, along with the advantages and disadvantages of using these methods; and
- the need to automate the process.

Chapter 3

Automated Bone Age Assessment

In the last chapter we reviewed BAA and its relation to clinical radiology. There are many potential advantages to automated bone age assessment (ABAA) (see Section 2.6) and the task seemingly lends itself to being automated [TWH⁺01]. Hence, many ABAA algorithms have been proposed. The aims of this chapter are as follows:

- to describe various image processing techniques and classification algorithms (Sections 3.1 and 3.2);
- to discuss the previously proposed methods for ABAA (Section 3.3);
- to investigate why none have gained widespread acceptance (Section 3.4); and
- to discuss the pitfalls of using Active Appearance Models (AAMs) for ABAA (Section 3.5).

3.1 Image Processing Algorithms

Before describing the previously proposed methods for ABAA, it is necessary to understand the techniques used as part of the estimation process. Hence, in this section a variety of image processing algorithms that are used in ASMA and have been used in previously proposed ABAA systems are discussed.

3.1.1 Active Appearance Models (AAM)

Active Shape Models (ASM) [Coo00] and Active Appearance Models (AAM) [CET01] have commonly been used for a wide variety of image processing applications [MB04, TMTM12]. However, the use of ASMs has decreased since the introduction of AAMs, as the AAM incorporates intensity as well as shape features. Hence we only describe AAMs. To use AAMs to extract an object of interest, the model must first be trained over a set of manually annotated images. An AAM is made up of two independent models. The first describes the variation in shape and is a point distribution model (PDM), and the second models the variation in appearance.

The PDM is created by placing k landmarks representing (x, y) co-ordinates along the outline of the object we wish to model. The landmarks are then normalised for translation, rotation and scale and Principal Components Analysis (PCA) is applied, to give a compact model of shape of the form:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{S}\mathbf{b}_s, \quad (3.1)$$

where \mathbf{s} refers to a set of landmarks for an image, $\bar{\mathbf{s}}$ is the mean shape, \mathbf{S} is the set of eigenvectors that define the allowed variation of the shape, and \mathbf{b}_s are the shape parameters. Each eigenvector in \mathbf{S} represents a different mode of variation of shape and appear in the matrix in decreasing order of their variation.

AAMs also encompass a model of the intensity variation within the shape. The labelled training images used to construct the shape model are warped to the mean shape, $\bar{\mathbf{s}}$, with the pixel intensities then concatenated into a vector. PCA is then applied to these shape normalised intensity vectors. This provides a compact model of appearance variation of the form:

$$\mathbf{a} = \bar{\mathbf{a}} + \mathbf{A}\mathbf{b}_a, \quad (3.2)$$

where \mathbf{a} refers to a shape normalised image, $\bar{\mathbf{a}}$ is the mean appearance, \mathbf{A} is the

set of eigenvectors that define the allowed variation of appearance, and \mathbf{b}_a are the appearance parameters.

Obviously there will be correlation between the shape and appearance model. In order to remove this we concatenate the parameters of both models:

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_a \end{pmatrix}. \quad (3.3)$$

Where \mathbf{b}_s and \mathbf{b}_a , are calculated through rearranging Equations 3.1 and 3.2 respectively. \mathbf{W}_s refers to the weight matrix, which allows for the shape and appearance parameters to be compared directly. Each element of the weight matrix $\mathbf{W}_{i,i}$ is calculated as the RMS change in appearance \mathbf{a} per unit change of the i^{th} shape parameter in \mathbf{b}_s . PCA is then applied to the combined parameters \mathbf{b} to form the model:

$$\mathbf{b} = \mathbf{Q}\mathbf{c}, \quad (3.4)$$

where \mathbf{Q} is the set of eigenvectors, and \mathbf{c} are the parameters that control both the shape and appearance of the model.

Once the model has been built, a further iterative algorithm is used to fit the outline to new instances. A variety of these have been proposed. For this work we use the Inverse Compositional AAM proposed by Matthews and Baker [MB04]. As the system is fully automated, a guess at the initial set of landmarks is made based upon the size of the image. The landmarks and appearance are then warped to the mean shape, with the error between the warped image and $\bar{\mathbf{a}}$ calculated. The landmarks are then updated using gradient descent. This process is continued until the error between the warped image and mean appearance converges.

3.1.2 Otsu Thresholding

The Otsu thresholding technique [Ots75] has been used previously for segmenting hand radiographs [BKZ08, Zie09]. The method uses the probability distribution, \mathbf{p} , of the pixel intensities in the image \mathbf{I} , to calculate the optimal threshold i^* , to split the image into foreground and background. The mean pixel intensity, μ_T , is calculated using:

$$\mu_T = \sum_{j=0}^{255} j\mathbf{p}_j. \quad (3.5)$$

A pixel intensity i , is classified as background with a given probability, $\omega(i)$, using the cumulative probability distribution:

$$\omega(i) = \sum_{j=0}^i \mathbf{p}_j. \quad (3.6)$$

The mean intensity of pixels up to level i , $\mu(i)$, and the between class variance of intensities $\sigma_B^2(i)$, are calculated using Equations 3.7 and 3.8 respectively.

$$\mu(i) = \sum_{j=0}^i j\mathbf{p}_j. \quad (3.7)$$

$$\sigma_B^2(i) = \frac{(\mu_T\omega(i) - \mu(i))^2}{\omega(i)(1 - \omega(i))} \quad (3.8)$$

The optimal threshold, i^* , is calculated using:

$$i^* = \arg \max_{0 \leq i \leq 255} \sigma_B^2(i), \quad (3.9)$$

which provides the division of the image into background and foreground regions.

3.1.3 Canny Edge Detector

The Canny edge detector [Can87] is a multistage algorithm that combines differential filtering, non-maximal filtering, and thresholding with hysteresis, and has been used previously in the context of segmenting hand radiographs [LBTS05, MMCD⁺05]. To summarise, the Canny algorithm:

1. smoothes the image \mathbf{I} with a Gaussian filter;
2. estimates the gradient magnitude $\|\nabla I\|$ and direction θ (See Equations 3.10 to 3.13) at each pixel and quantises the gradient directions to be one of $\{0, 45, 90, 135\}$ degrees;
3. performs non-maximal suppression by switching off candidate pixels that are not locally maximum in the direction of the gradient.
4. identifies definite edge pixels as those with a gradient magnitude above a global high value threshold T_{high} , and switches off pixels that have a gradient magnitude below a global low threshold T_{low} ; and
5. checks the pixels with gradient magnitude between T_{high} and T_{low} to determine if there is a path that connects them to a definite edge pixel. Those that are connected to a definite edge form the edge, otherwise they do not.

$$\|\nabla I\| = \sqrt{\frac{\delta I^2}{\delta x} + \frac{\delta I^2}{\delta y}}, \quad (3.10)$$

$$\theta = \tan^{-1} \left(\frac{\frac{\delta I}{\delta y}}{\frac{\delta I}{\delta x}} \right) \quad (3.11)$$

$$\frac{\delta I}{\delta x} = \mathbf{I}_{x,y} - \mathbf{I}_{x+1,y}, \quad (3.12)$$

$$\frac{\delta I}{\delta y} = \mathbf{I}_{x,y} - \mathbf{I}_{x,y+1}, \quad (3.13)$$

3.2 Classification Algorithms

In this work we investigate the use of machine learning classifiers for a variety of tasks in the ABAA process. Here we describe a variety of classifiers. Each classifier shall be discussed and how it would handle classifying the following example.

Given the training set \mathbf{T} :

	\mathbf{a}_1	\mathbf{a}_2	\cdots	\mathbf{a}_{n-1}	\mathbf{c}
\mathbf{t}_1	$\mathbf{T}_{1,1}$	$\mathbf{T}_{1,2}$	\cdots	$\mathbf{T}_{1,n-1}$	\mathbf{c}_1
\mathbf{t}_2	$\mathbf{T}_{2,1}$	$\mathbf{T}_{2,2}$	\cdots	$\mathbf{T}_{2,n-1}$	\mathbf{c}_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
\mathbf{t}_m	$\mathbf{T}_{m,1}$	$\mathbf{T}_{m,2}$	\cdots	$\mathbf{T}_{m,n-1}$	\mathbf{c}_m

Where each row $i = 1 \dots m$ refers to a training sample \mathbf{t}_i , each column $j = 1 \dots (n - 1)$ refers to an attribute \mathbf{a}_j , and the final column $j = n$ refers to the class value \mathbf{c}_i of the training sample. Each attribute \mathbf{a}_j can be either discrete or continuous and has a set of values \mathbf{v} associated with it.

Some query sample \mathbf{q} needs to be classified, where \mathbf{q} is a vector of length n with the same attributes as the training set \mathbf{T} and the class value $\mathbf{q}_n = \text{null}$.

3.2.1 k -Nearest Neighbour (k -NN)

The k -NN classifier is one of the oldest machine learning classifiers [FHJ52]. The advantages of the classifier are its speed and ease of use. It is often used as a baseline classifier for evaluation purposes.

The k -NN classifier is made up of two stages. The first stage is to calculate the distance \mathbf{d}_i from each training sample \mathbf{t}_i to \mathbf{q} , which can be defined as:

$$\mathbf{d}_i = \sqrt{\sum_{j=1}^{n-1} (\mathbf{q}_j - \mathbf{T}_{i,j})^2}. \quad (3.14)$$

Once the distance \mathbf{d}_i to each training sample \mathbf{t}_i has been calculated, the next

stage is to assign a class to the query sample \mathbf{q} . The class is assigned by finding the k closest training samples to \mathbf{q} . If the majority of the k closest training samples have been assigned to a given class c^* , \mathbf{q} is also assigned to class c^* .

In general, k is an odd number, as this avoids difficulties where there is no majority class. Here, we use the Euclidean distance measure which is the standard measure; however other measures such as the Manhattan distance measure can be used.

3.2.2 Naive Bayes

Naive Bayes is a simple classifier that has proven to be fast and effective in many areas of computer science such as machine learning [Ces90], data mining [LL98] and information retrieval [Lew98]. Naive Bayes makes the assumption that all attributes \mathbf{a}_i are independent, hence why it is called Naive.

The Naive Bayes classifier is based on Bayes Theorem. To build the classifier, the probability distribution $p(\mathbf{a}_j|\mathbf{c}_k)$ for each attribute \mathbf{a}_j given each class value \mathbf{c}_k is calculated, along with the probability of each class $p(\mathbf{c}_k)$.

The class assigned c^* to the query sample \mathbf{q} is then calculated as follows:

$$c^* = \arg \max_{\mathbf{c}_k \in \mathbf{c}} p(\mathbf{c}_k|\mathbf{q}), \quad (3.15)$$

where:

$$p(\mathbf{c}_k|\mathbf{q}) = p(\mathbf{c}_k) \prod_{j=1}^{n-1} p(\mathbf{a}_j = \mathbf{q}_j|\mathbf{c}_k). \quad (3.16)$$

3.2.3 C4.5 Decision Tree

The C4.5 tree was proposed by Quinlan in [Qui93]. C4.5 is a greedy algorithm and produces a top-down tree. The advantages of this classifier are that it is good for

explaining classifications, as you can follow the route along the tree and therefore gain understanding of the dataset. The classifier can also handle missing data. The classifier works in two stages, firstly, building the tree from the training set \mathbf{T} , and then classifying any query sample \mathbf{q} .

There are three basic operations of the C4.5 tree to build a classifier, the first is to select the attribute a^* to split on, a^* is calculated using:

$$a^* = \arg \max_{\mathbf{a} \in \mathbf{T}} (\text{gainratio}(\mathbf{T}, \mathbf{a})). \quad (3.17)$$

Where a^* is the attribute with maximal gain ratio. The gain ratio criterion uses the information gain and the entropy $H(\mathbf{X})$ of any potential split, where:

$$\text{gainratio}(\mathbf{T}, \mathbf{a}) = \frac{\text{infogain}(\mathbf{T}, \mathbf{a})}{H(\mathbf{T})}, \quad (3.18)$$

$$\text{infogain}(\mathbf{T}, \mathbf{a}) = H(\mathbf{T}) - \sum_{v \in \mathbf{v}} \frac{|\mathbf{T}_v|}{|\mathbf{T}|} H(\mathbf{T}_v), \quad (3.19)$$

and

$$H(\mathbf{x}) = - \sum_{i=1}^n p_i \log_2 p_i. \quad (3.20)$$

\mathbf{T}_v refers to the subset of \mathbf{T} where $\mathbf{a}_i == \mathbf{v}_j$ and $|\mathbf{T}|$ is the cardinality of the set. Entropy was introduced by Shannon [Sha48], it measures the uncertainty associated with a random variable \mathbf{x} , using the probability p_i of each possible value x_i of \mathbf{x} . Information gain measures the expected reduction in entropy due to splitting on attribute \mathbf{a} .

Once a^* has been calculated, a^* becomes the parent node and each of the possible values \mathbf{v}_j of a^* are child nodes if a^* is discrete, however if a^* is continuous the best split point s is calculated using information gain and two child nodes are created, these are $\mathbf{v} \leq s$ and $\mathbf{v} > s$. This process is repeated on each of the child nodes

where the subset \mathbf{T}_v is used, and continued until one of the stopping criteria are met (the second operation). Here are three stopping criteria for C4.5:

- The child node is empty, $\mathbf{T}_v = \emptyset$,
- There are no more potential features to branch on, or
- all of \mathbf{T}_v has the same class value.

The third and final operation of the C4.5 tree whilst building the classifier from the training set is to prune the tree. With the C4.5 tree this is done by backtracking and investigating if removing a branch decreases the accuracy of the classifier.

The query sample \mathbf{q} is then classified by following the branches of the tree that the attributes of \mathbf{q} conform with. The class label c^* given at the final node in the tree that \mathbf{q} reaches, is the class label assigned.

3.2.4 Support Vector Machines (SVM)

SVMs were introduced by Cortes and Vapnik [CV95], and have performed well on many real-world problems such as spam categorisation [DWV99] and image classification [CHV99].

The standard SVM expects that the attributes \mathbf{a} are linearly separable based on class \mathbf{c} by some function $f(\mathbf{t})$, where:

$$f(\mathbf{t}) = \mathbf{w} \cdot \mathbf{t} + b, \quad (3.21)$$

\mathbf{w} refers to a normal vector to $f(\mathbf{t})$ and b refers to the offset of $f(\mathbf{t})$ from the origin along \mathbf{w} . A simple method to train the SVM is as follows, for each training sample, if the $\mathbf{c}_i(\mathbf{w} \cdot \mathbf{t}_i) \leq 0$ holds true then update $\mathbf{w} \leftarrow \eta \mathbf{c}_i \mathbf{t}_i$. This process is repeated until all samples are classified correctly.

The query sample \mathbf{q} is then assigned class c^* depending on which side of $f(\mathbf{t})$ it lies. Here a simple training algorithm for a SVM classifier has been shown, usually

a more advanced algorithm is used to calculate the hyperplane $f(\mathbf{t})$ that has the maximum margin between the classes. If the attributes \mathbf{a} are not linearly separable based on class \mathbf{c} , a kernel function (such as quadratic or radial basis) can be used to transform the data so that it is linearly separable in the new feature space. This means that different kernels are useful for different tasks, hence why we use linear, quadratic and radial basis function kernels in this work.

3.2.5 Random Forest

Random Forest [Bre01] is a tree based ensemble technique that diversifies through random attribute selection.

A Random Forest classifier is made up of a set number p random trees. For each tree created a random subset of training samples \mathbf{S} from \mathbf{T} is used to train the tree. The tree is then built by selecting a random subset of attributes \mathbf{b} and calculating the best split from \mathbf{S} using the GINI index [BFOS84]. The stopping criteria for the tree are the same as with C4.5.

In order to predict the class c^* of \mathbf{q} , the class of \mathbf{q} is predicted by each tree in ensemble. The class with the majority of predictions is then assigned to \mathbf{q} .

3.2.6 Rotation Forest

Rotation Forest [RKA06] is a tree based ensemble technique like Random Forest, where diversity is achieved through subspace transformation.

Rotation Forests are like Random Forests in that they are made up of an ensemble of p trees. Each tree is built on a subset \mathbf{S} of training set \mathbf{T} , PCA is then applied to all the attributes in each subset \mathbf{S} and hence p rotations of axes occur. In order to keep the whole variation of the subset, all principal components are kept. A C4.5 tree is then built for each of the newly transformed subsets.

The same method as with Random Forest is then used to classify the query

sample \mathbf{q} , where a vote for a class is taken from each tree and the class with the majority of votes is assigned.

3.2.7 Multilayer Perceptron

Multilayer Perceptrons [MP69] are a form of artificial neural network made up of several Perceptrons. A perceptron is a model of a neuron inside the brain. A single linear perceptron can be seen as being similar to a SVM, to train a linear perceptron, take each training sample \mathbf{t} , multiply it by some weight vector \mathbf{w} and add a bias b , this is known as the activation x and is calculated using:

$$x = b + \sum_{i=1}^{n-1} \mathbf{w}_i \mathbf{t}_i. \quad (3.22)$$

The output of the perceptron is a function of the activation $f(x)$. The major difference from SVMs is how it handles non-linearly separable data. Instead of using a kernel function, several linear perceptrons are used which creates decision regions. In order to train the perceptrons the weight vector \mathbf{w} is updated to minimise the error e this is done using the back propagation algorithm where \mathbf{w} is updated after each training sample \mathbf{t} . \mathbf{q} would then be assigned a class c^* calculating which decision region it falls into.

3.3 A Brief History of Automated Bone Age Assessment

In this section we shall firstly discuss the work by Thodberg *et al.* (Section 3.3.1) and then the work done by Pietka *et al.* (Section 3.3.2) as these are the most referenced contributors to the field. After this we shall discuss other methods (Section 3.3.3) that have been proposed in chronological order. Table 3.1 shows a list of research groups that have contributed to the field, and the techniques used.

Table 3.1: Table showing research groups that have contributed to the field of ABAA.

Authors	Basic Techniques	References
Thodberg <i>et al.</i>	AAM based	[Tho02, TR03, TKJP09, MDS ⁺ 09, Tho09, vRLT09, TJC ⁺ 09]
Adeshina <i>et al.</i>	AAM Based	[ACA09]
Mahmoodi <i>et al.</i>	ASM based	[MSC ⁺ 00]
Niemeijer <i>et al.</i>	ASM based	[NvM ⁺ 03]
Efford	ASM Based	[Eff93]
Pietka <i>et al.</i>	Canny (Feature Based)	[PMGKH91, PKKH93, PH95, Pie95, PPG ⁺ 01, PGP ⁺ 01, PGPK ⁺ 04, ZGL07]
Manos <i>et al.</i>	Canny (No Age Assessment)	[MCRS93, MCRS94]
Sharif <i>et al.</i>	Modified Canny (No Age Assessment)	[SZC ⁺ 94]
Tristan-Vega and Arribas	Sobel (Feature Based)	[TA05, TVA08]
Michael and Nelson	Thresholding (Feature Based)	[MN89]
Morris and Walshaw	Thresholding (No Age Assessment)	[MW94]
Chang <i>et al.</i>	Thresholding (Feature Based)	[CHJT03]
Zielinski <i>et al.</i>	Otsu (No Age Assessment)	[BKZ08, Zie09]
De Luis Garcia <i>et al.</i>	Active Contours (No Age Assessment)	[DMFAAL03, MMCD ⁺ 05]
Han <i>et al.</i>	Active Contours (No Age Assessment)	[HLP07]
Pal <i>et al.</i>	Fuzzy Sets (Feature Based)	[PK83, PPK84, PP86]
Lehmann <i>et al.</i>	Watershed (No Age Assessment)	[LBTS05]
Kim and Kim	DCT / LDA (Feature Based)	[KK07]
Giordano <i>et al.</i>	Difference of Gaussian (Feature Based)	[GLM ⁺ 07, GSSL09, GSSL10]
Martin-Fernandez <i>et al.</i>	Warping to target image	[MMFAL03]

3.3.1 BoneXpert by Thodberg *et al.*

Thodberg *et al.* [Tho02, TKJP09] use Active Appearance Models (AAMs) [CET01] for their automated bone age assessment system (BoneXpert). This algorithm fits a model on 1559 manually labelled images, uses the model to find shape and intensity features, then regresses the features onto age. It consists of three layers: A, B and C. Layer A involves fitting separate AAMs to each of the 15 RUS bones separated into three separate age epochs (3 to 8.2 years, 8.2 to 13 years, and 13 to 18 years for boys). The BoneXpert bone age is constructed in layer B directly from the 45 separate models constructed in layer A. A linear regression of 10 shape, 10 intensity and 10 texture features onto chronological age for each short bone and epoch combination is performed, with a model selection stage used to reduce the number of dependent variables. The bone age estimates are averaged over the bones within each epoch, and this average bone age estimate is used as the independent variable for a further regression for each bone/epoch combination. This second regression is the final model for each bone/epoch combination. Thus for a new image, layer B produces 45 separate age estimates from the 45 separate linear regression models. Layer C involves fitting and using these models for a new image. To locate the bones, a large number of separate initialisations for each AAM are generated, and the AAM reconstruction with the best fit over each of the epochs is selected for each bone. Validity checks are performed for each predicted outline. Quality of each model is assessed with “*the residual error between the observed bone and the AAM-generated image*”, which is calculated using a “misfit” score [TR03].

For every bone, between 0 and 3 models can be approved as correct for the three different epochs. The linear models of stage B are applied to obtain between 0 and 3 bone age estimates for each bone. These estimates are combined with a weighted average, the weights being derived through the quality of measure of the associated AAM fitted outline. An average over all bones is then calculated to form the initial age estimate, B_0 . Individual bone age estimates more than 2.4 years from B_0 are discarded. If the number of remaining bones with at least one valid estimate is

below some user defined threshold, the entire image is rejected. Otherwise, a new weighted average composite average, B_1 , is computed. The weights for this average are the estimate of the probability of observing the predicted bone age, assuming a normal distribution with mean B_0 and standard deviation 1.4 (chosen through experimentation). Missing bones are assigned the age estimate B_1 .

The authors present results comparing their age estimates to the images in the GP atlas and describe and evaluate a mechanism for recreating TW scores. There is no detailed evaluation of how accurate the estimates B_1 are in comparison to chronological ages because the authors claim that *“bone age is a poor predictor of chronological age”* [TKJP09]. Other assessment criteria include the consistency of estimates for the same subject over time and the frequency of rejections.

When validated against the images in the GP atlas, BoneXpert estimates B_1 are on average 0.7 years larger than the chronological ages of the subject. The authors then perform a post hoc adjustment by subtracting 0.7 from each estimate. They justify this by citing differences in populations between the GP atlas subjects and those used to construct the model. After this adjustment, the BoneXpert predictions have a standard deviation of 0.42 years to the true age of the GP atlas ages. To recreate TW, the bone age estimates are mapped onto the TW scores using a training set of images with TW ratings assigned by a human operator. On a cross validation of 84 radiographs they report 68% agreement between BoneXpert and the human scorer, with 94% of the estimates within one stage of each other. For the phalanges, BoneXpert is in agreement with the human on approximately 70%-80% of bones.

BoneXpert has been released as a black box commercial system and assessments are charged at 10 euros per radiograph. Since releasing BoneXpert as a commercial system the authors have released a number of journal papers applying the system to various studies [MDS⁺09, Tho09, vRLT09] in order to validate it for clinical use.

In [MDS⁺09] they make adjustments to BoneXpert to agree with manual GP ratings of five raters. The adjustments made to BoneXpert were firstly, to add 119 extra hand radiographs over the age of 15 to the training data and, secondly, to

change the calculation of GP ages so that instead of subtracting 0.70 years from the BoneXpert age, a function of the BoneXpert age and gender is calculated and added to the original BoneXpert bone age. The new algorithm was tested on 1,097 hand radiographs, 14 of which were rejected by the system for a variety of reasons. The root mean square deviation (RMSD) stated is 0.72 years between BoneXpert and the manual raters.

Two studies are verified in [Tho09] using the updated version of BoneXpert described in [MDS⁺09]. The first study contains 531 radiographs in the age range 4–20, the RMSD compared to two ratings from manual raters is 0.71 years. However seven ratings were repeat-rated by new clinicians as the authors state that they were “*initially rated incorrectly*”. The radiographs that were rated again all had an original deviation 1.9 years, and so a question arises as to if these seven were incorrectly rated why did the author not seek to find out if the other radiographs were incorrectly rated. The second study and results are the same as presented in [MDS⁺09].

A validation study on Dutch children is performed in [vRLT09], again using the updated version of BoneXpert presented in [MDS⁺09]. Testing was performed on 405 hand radiographs with an RMSD of 0.71, this was performed on the same study as in [Tho09], however using less radiographs. Again the authors state seven ratings were repeat-rated by new clinicians.

Another proposed use of BoneXpert is for automatically predicting adult height [TJC⁺09]. Two predictions for adult height are calculated, one using information from parents and another using the mean population height. The first method uses bone age, and chronological age to calculate growth potential. This is used with current height to calculate a raw value of adult height that is then used with the mean of the parents height and/or mean population height to predict final adult height. The method using parents heights was tested on 231 radiographs with a Root Mean Square Error (RMSE) of 3.3cm for boys and 2.7cm for girls.

3.3.2 The Work by Pietka *et al.*

The group at the University of Southern California (Pietka *et al.*) first published a paper on assisting automated bone age assessment in 1991 [PMGKH91]. The first stage of the proposed method is to standardise the radiograph through removing the unexposed background using an algorithm to find the radiation field (the area of the image where X-ray photons were absorbed). The radiographs are then thresholded using the average grey value intensity to contain a hand silhouette. If necessary the image is rotated. A region of interest around the phalanges (PROI) is found by searching “*for a pair of lines*”. One of the lines should intersect at least three fingers and the other the centre of the hand silhouette. These are used to find the PROI. Once the PROI is found, a Sobel filter is used to detect edges and produces an edge map. The edge map is thresholded using an “*empirically determined value*”. The middle finger is located using the tip. The separation between the phalanges of the finger is then found and the lengths are determined as well as ratios between the bones. Using these measurements, the bone age is estimated using the method proposed in [GKP⁺72]. The algorithm is tested on 50 hand radiographs and successfully got measurements on 47 images with a Mean Squared Error (MSE) of 0.08mm when compared to manual annotations. Age estimates were made on 19 radiographs with a mean difference of 1.57 years and a standard error of 0.32 years.

The next paper published [PKKH93] investigates the use of a region of interest around the carpal bones (CROI). As with [PMGKH91], the first stage of the proposed method is to standardise the radiograph using the method proposed in [PMGKH91]. The next stage is to find the region of interest, this is performed by calculating the web between the thumb and index finger and the wrist by finding the narrowest part of the hand silhouette. A two step thresholding is performed to extract the bones from the CROI. This is done by firstly, calculating a threshold based on an analysis of the histogram, and, secondly, using a dynamic threshold method. As non-carpal bones maybe present, any objects that are present after thresholding that touch one of the edges of the CROI are removed. The binary mask is then

subjected to some morphological filters to remove other non-carpal objects. Eight features are extracted from each of the carpal bones. These are shown in Table 3.2. These are used to calculate which carpal bone is which and therefore calculate two ratio measures to assist in the age assessment. The proposed algorithm is tested on 30 hand radiographs between the ages on 0–9 for boys and 0–8 for girls for the detection of the carpal bones. A preliminary study of the importance of features is undertaken and finds area and perimeter found to be the most discriminatory. No age assessment is performed in this paper and the authors state they shall do this in future work.

Table 3.2: Features extracted from segmented carpal bones in [PKKH93].

Feature Number	Feature Name
1	Area
2	Perimeter Length
3	Compactness Ratio
4	Center of Gravity
5	Convexity Coefficient
6	Lengthening Ratio
7	Average pixel intensity
8	Average pixel discrepancy

A method to assess the state of fusion of the epiphysis of a bone is presented in [PH95]. This is done by firstly, performing wavelet decomposition on a region of interest (ROI). The output components of the wavelet decomposition are then subjected to a test to work out if the image is overexposed, if so the image is rejected as it “*would not give reasonable results*”. A quantitative measure based on the components and size of the ROI is used to classify which of the four stages of fusion (None, Early, Advanced and Complete) the ROI is in. Testing is performed on 90 hand radiographs with an accuracy of 83.3% on classifying the correct stage of fusion.

In [Pie95], features are extracted from the PROI and CROI of a hand radiograph and are used to get a phalangeal bone age (PBA) and a carpal bone age (CBA)

respectively. These are compared to clinicians estimations as well as to each other. Three features are extracted from the phalanges. These are length, width of metaphysis and epiphyses and stage of epiphyseal fusion (using the method proposed in [PH95]). Three features are also used for each of the carpal bones. Although it is not explicitly stated what these features are, it is assumed they are the three best performing features from the analysis in [PKKH93]. In order to classify bone age, a fuzzy classifier is used that *“has been developed by defining membership functions and a classification rule”*. If a bone is not extracted correctly, the features of that bone were not used in the classification. The algorithm is tested on 120 hand radiographs. It is found that the calculated PBA differs from the clinicians bone age by less than 6 months on 75% of cases and 63% using the CBA. Although no age range is given, it is assumed that the age range of the radiographs used is the same as that in [PKKH93] as the authors state that it becomes hard to extract them because *“at the age of 9–10 they start overlapping”* in reference to the carpal bones.

Pietka *et al.* publish an updated version of their algorithm in [PGP⁺01]. The updated version has the same stages as those proposed in [PMGKH91]. However the work performed in each stage is different. The paper mainly concentrates on the extraction of three Epiphyseal-Metaphyseal Regions of Interest (EMROI) along the phalanges of the middle finger and does not perform any bone age estimation. Firstly, the background is removed using a dynamic threshold as in [PMGKH91]. However a different method is used to find the tip of the middle finger, which involves covering the thresholded image with a grid and calculating the mean values of pixels. Step wedge functions are then used to find the tip. Once the tip has been found the axis for the finger is calculated and this is analysed to find the EMROIs. The EMROI is processed using a Sobel edge detector, from the resulting edge map three features are extracted. These are epiphysis width, diaphysis width and metaphysis width. These features are used to calculate two ratios. The algorithm is tested on 200 hand radiographs of boys in the age range 0–14 and girls in the age range 0–12. The accuracy of the extracted features is then investigated by making comparisons to manually marked radiographs. Features are most accurately esti-

mated from the distal phalange, with the proximal phalange being least accurate. The discriminatory power of the features is assessed through plots against patient age. The feature found to be most discriminative is the epiphysis diameter divided by the metaphysis diameter for females in the age range 4–10 and males in the age range 4–13.

Another publication by Pietka *et al.* [PPG⁺01] investigates locating six EMROI: the joints between distal phalanges - middle phalanges and middle phalanges - proximal phalanges of fingers two, three and four. A histogram analysis then takes place in order to build a model of the background. Which is used to remove the background. The hand object is then segmented using thresholding. The phalangeal axes are then found by scanning the mask horizontally and finding the midpoints of areas of high intensities (the same method as used in [PGP⁺01]). Next, the EMROIs are found using the method explained in [PGP⁺01]. Testing is performed on 130 hand radiographs between 1 and 18 years old. Hand extraction failed 1% of the time, phalangeal axes extraction failed 7% of the time and 3% of EMROIs were not extracted correctly.

In [PPKGC03], Pietka *et al.* describe a method that uses *c*-means clustering and Gibbs Random Fields to segment bones from a radiograph. Using the same EMROI as in [PPG⁺01] and extracted using the same method. Once the EMROIs are segmented, the bone is segmented using *c*-means clustering to preliminary separate bone from soft tissue and Gibbs Random Fields is used to finalise the segmentation. Next features are extracted from the EMROIs including the output from wavelet decomposition as in [PH95], measurements based on the features discussed in [PGP⁺01] and image based features. The system calculates bone age using a fuzzy classifier on each bone and a process of “*defuzzification*”. The system is tested on 231 hand radiographs and reports a chronological error of 0.94 years in boys and 1.13 in girls although the error measure used is not mentioned.

A Graphical User Interface (GUI) is proposed for assisting bone age assessment in [PGPK⁺04]. This extracts the same six EMROIs as in [PPG⁺01] and uses the

method to segment the bone from each EMROI presented in [PPKGC03]. Depending on the age of the patient, either two or three features are extracted. If the patient's age is below ten, the features extracted are the ratio between metaphyseal width and epiphyseal width, and, the ratio between height and width of the epiphysis. If over 10, the wavelet decomposition method described in [PH95] is used to extract features. After features are extracted the GUI displays EMROIs with the most similar features.

Zhang *et al.* [ZGL07] continue the groups work and propose an updated version of the algorithm presented in [PKKH93]. In the proposed method, the CROI is found and extracted using the methods proposed in [PKKH93] and [PPG⁺01]. The CROI is then smoothed using an anisotropic diffusion filter [PM90] and the bones segmented using Canny edge detection [Can87]. Non-bone objects are removed using the methods described in [PKKH93] as well as calculating the eccentricity of the bone. The Capitate (the largest carpal bone) is identified and the major axis of the bone is used to separate the CROI into different regions. In this paper only the Capitate and Hamate are used further. Four features are extracted from the bones. These are: ellipse diameter, eccentricity, solidity and triangularity. A fuzzy classifier was used to make age estimates based upon the extracted features. Testing is performed on 205 radiographs in the age range 0–5 for males and 0–7 for females. A segmentation accuracy of “*about 80%*” for children under two and “*just under 100%*” is stated for children over two years of age. The ages are compared to those of two clinicians graphically, although no error measurement is given.

3.3.3 Other Proposed Methods for Automated Bone Age Assessment

Pal and King [PK83] were one of the first groups of researchers to propose a method to segment hand radiographs. The first step of the algorithm is to equalise the histogram and then smooth the image. Fuzzy sets and a membership function are used to separate the image into regions, and then edge detection is used to

achieve the final segmentation of the image. The algorithm is demonstrated on one radiograph of the radius. Pathak *et al.* extend this work to perform bone age assessments of the radius in [PPK84]. The algorithm presented is a three stage hierarchical classifier that uses shape descriptors of the outline of the epiphysis. The algorithm is shown on a radiograph of the radius, with no accuracy of the method given. Pathak and Pal describe an extended version of the algorithm for classification of TW stages of the radius in [PP86] and test on the same radiograph used in [PPK84]. The classifier is built upon a six tuple fuzzy grammar and a seven tuple fractionally fuzzy grammar, and, the fractionally fuzzy grammar found to be the best performing.

Michael and Nelson presented their system HANDX in [MN89]. The proposed method labels pixels in a hand radiograph based upon their intensity, with the assumption made that there are three groups of pixels in a hand radiograph: background, soft tissue and hard tissue. Where each of these groups is assumed to have a normal distribution of pixel intensities within the histogram. After the pixels have been labelled, the histogram is then modified so that none of the distributions overlap. Using the modified radiograph the bone is segmented by thresholding the image, and, the resulting binary mask is labeled into six regions (palm and fingers 1–5). The proximal phalange of the third finger is then found using this anatomical information. An initial approximation of the bone is found using a “blob detector”. This is then refined using an adaptive contour. Three features are extracted from the bone: length, width and area. Testing is performed on two radiographs with the measurements presented. However, these are not compared to a manual rater and hence no accuracy can be calculated from this.

The method presented by Efford [Eff93] for automatically assessing skeletal maturity uses ASMs to segment the bone. The proposed method generates a hand silhouette by thresholding the radiograph. The resulting binary mask is then filtered using various morphological operators. A vertical line that intersects the phalanges of the middle finger and a horizontal line that intersects the metacarpals are then

calculated, in order to locate the other bones in the hand and wrist. The shape of the silhouette is analysed to ensure that it is as expected for a normal hand. This is performed using chain codes with 11 landmarks: five fingertips, four between fingers and two for the wrist. The bones then are segmented using ASMs. Active Contours (AC) and Fast Fourier Transforms (FFT) are also used but the authors state that these methods are found not to perform as well as ASMs. In order to assess skeletal maturity, each criteria from the TW2 stages is changed into code, features are then extracted from the test image and compared to see if the criteria for a TW2 stage are met e.g. metaphysis and epiphysis are the same width. No results of the proposed system are presented.

Manos *et al.* proposed a system to segment hand radiographs in [MCRS93, MCRS94]. The proposed algorithm is split into two threads. The first uses the Canny edge detector to find the edges of the bones. The second thread smooths the image using an edge-preserving smoothing technique and then splits the image into regions based on the pixel intensities. The regions are merged if the grey-levels of neighbouring regions was similar. At this point the edges found in the first thread are used in conjunction with the resulting image from the second thread to find common boundaries and use these to further merge regions. Once this is completed the regions are then labelled as bone or background based upon a set of rules. In [MCRS93] and [MCRS94], the algorithm is tested on 14 and 10 hand radiographs respectively, although neither paper publishes a result on the accuracy of the algorithm.

Morris and Walshaw propose a method for the segmentation of the finger bones in [MW94]. The image is split into a number of subregions and the pixel values in the subregion are forced to a mean pixel intensity of 128 and standard deviation of 60. The image is then thresholded and each region identified. Metacarpals are identified by taking the central horizontal row of the radiograph and using peak detection to identify the central point. The axes of the finger are then identified by calculating the line between each of the metacarpal points and the centre of the

wrist. These are extended down the fingers and ROIs are marked around each of the identified bones. The edges of the identified bones are then thinned and the phalanx and epiphysis (if applicable) are identified. The algorithm is tested on six radiographs from age 2 to near adult with two result images shown although no accuracy of the algorithm is given.

Sharif *et al.* [SZC⁺94] propose a method to segment the bones from hand radiographs by firstly equalising the intensities over the radiograph due to “*Dense bones in abundant soft tissue such as metacarpals and carpals absorb more radiation than sparse bones in thin soft tissue such as distal phalanges and, therefore, result in an image intensity profile that decreases towards the ends of the fingers.*” This is done using an equalisation function. The bones are then extracted using a modified version of Canny edge detection. Testing is performed on one hand radiograph with a figure showing the resulting image with and without equalisation. However, no accuracy results are given.

Mahmoodi *et al.* [MSC⁺00] use ASMs to segment bones and extract shape descriptors from the segmented contour (e.g. Epiphysis-to-Metaphysis ratio). The features are then regressed against age. The accuracy over certain age groups on 57 images is reported as 82% for males and 84% for females.

Chang *et al.* propose a fully automated method for bone age assessment using phalangeal features in [CHJT03]. Firstly, the hand silhouette is located. This is performed by thresholding the image on the mean pixel intensity. The middle finger is then located using a similar method to that of [PGP⁺01] and the bone segmented using Canny edge detection. Three features are segmented from each of the phalanges as well as the length of the middle finger and the principle changes in the shape of the phalanx and epiphysis over time. Another set of features based on Discrete Cosine Transform (DCT) coefficients is also extracted and compared to the shape features. Testing is performed on 917 radiographs and a correct segmentation occurs in 89.86% of cases. Classification of bone age is performed using a back propagation neural network. The DCT features outperform the shape based features

with results of 83.86%, 76.54% (females), and 79.05%, 78.84% (males) within 1.5 years of chronological age respectively.

Martin-Fernandez *et al.* [MMFAL03] propose a method to align a target radiograph to a template radiograph in order to automate the GP method of bone age assessment. The proposed method consists of two stages. Firstly, locating landmarks in relevant areas of the hand radiograph. Secondly, extracting features using the landmarks in the target image and a number of template images in order to find the closest match and hence calculate the bone age. It is stated in the paper that the landmarking stage is performed manually. After the landmarks are labelled, a wire-model of the hand is built. This is used to create five finger masks and five metacarpal masks. The second stage is to align the regions found in the target image to some template images using a global registration method followed by a partial and fine registration method. The method is tested on two radiographs and the results suggest that the alignment of a target image to template image improves after each registration phase.

De Luis-Garcia *et al.* [DMFAAL03] use a thinning algorithm along with ACs to segment bones from a hand radiograph. The algorithm they present is split into two main steps. Firstly, calculation of the location of the phalanges and therefore the initial position for the AC. Secondly, the use of ACs to find the shape of the bone for segmentation. As with [BKZ08, Zie09], the background is removed and a binary mask is created. A thinning algorithm is applied and the branches of the fingers are interpolated into straight lines with the exception of the thumb. Seeds are then placed at certain distances on the vector. The metacarpals are found by calculating the circumference of four points on the vectors that represent the fingers. The circumference is analysed to find the five metacarpals. To find the phalanges of the thumb, concentric circumferences with increasing radius are drawn. These seeds are the initial position of the AC. In [DMFAAL03], the equation for external energy is an inflation energy function. This forces the AC to grow until the image force draws the AC to the edge. The method is tested on 59 radiographs, with a 91.5%

accuracy of finding correct seeding locations and 73.9% success rate of correctly identifying the contour.

Niemeijer *et al.* [NvM⁺03] propose a method to automate skeletal age assessment that uses ASMs [Coo00] to segment the distal phalange of the third finger. A separate model for each TW stage (E–I) is trained. For new data, these models are used to extract the phalange, and the similarity of the fitted bone to the training bones in the model space is measured, with a nearest neighbour, maximum correlation and linear discriminant approach. The system is evaluated on 71 images by comparing the predicted TW stage against the TW ratings of clinicians. The second clinician gave the same stage as the first clinician on 80.3% of the rated bones and was within one stage 100% of the time. The results presented from the proposed system assigned 73.2% of bones with the correct TW stage and 97.2% within one TW stage.

Lehmann *et al.* propose a method to segment a hand radiograph in [LBTS05]. The first step of the method is to apply the Canny edge detection algorithm [Can87], followed by the watershed transform [VS91] in order to segment the image. This results in over-segmentation, which is overcome by region merging using a nearest-neighbour graph to find adjacent regions followed by a hierarchical attributed region adjacency graph. The region merging is complete when the radiograph is one region. Local, regional, global and hierarchical information are combined in the region merging process. To test the algorithm 10 radiographs are used. These are compared to manual segmentations of each image. The proposed algorithm is found to have a segmentation overlap of 85.1% pixels. However, the dataset used for testing is small and the stopping criteria are not discussed.

Munoz-Moreno *et al.* [MMCD⁺05] propose a method to segment bones from a hand radiograph. Firstly, a Gaussian filter is applied to the image. The method then splits into two threads that are run in parallel. The first thread applies edge detection using the Canny algorithm [Can87] to the radiograph. The morphological operation of dilation is then applied to ensure that all edges are closed contours.

After the edge detection, the watershed transform [VS91] is applied. In order to solve the over-segmentation problem a threshold is applied to detect what is bone and non-bone. The threshold is calculated as a proportion of the overall image area. The aim of the second thread is to find the axes along the fingers. This is performed using the method proposed by De Luis-Garcia et al. [DMFAAL03]. An adaptive threshold is used to remove the background from the radiograph. A thinning algorithm is used to get a rough estimate of the finger axes. These axes are refined, using a vector to approximate the original axes. The vectors are then repositioned by analysing the normal to the vector to find the centre of each finger.

A coarse-to-fine bone segmentation algorithm is presented by Han *et al.* in [HLP07]. The algorithm is split into three stages: Metaphyseal region segmentation; Model based ROI locating; and Epiphyseal region segmentation. The Metaphyseal Region Segmentation is performed by applying a Sobel filter in horizontal and vertical directions. The watershed transform is then applied, the opening morphological operator is applied to the resulting image. Ellipse region fitting techniques are used to find the epiphyseal ROIs. Finally Active Contours are used to find the outline of the epiphysis.

Kim and Kim [KK07] propose a method to perform bone age assessment that used DCT and Linear Discriminant Analysis (LDA). Nine EROI are segmented, these are the epiphysis for each of the phalanges of fingers two three and four. Each EROI is then rescaled to a size of 200×200 . The DCT coefficients are then calculated and the most discriminative used along with the LDA transform matrix. A feature vector is calculated using the DCT coefficients projected onto the LDA transform matrix. This allows a bone age for each EROI to be calculated and then an average bone age used for the final assessment. Three potential averages are investigated, the average from all nine EROIs, the average of seven EROIs (discarding the youngest and oldest) and using the median age. The method is tested on 393 radiographs and a leave-one-out cross validation performed. The median average performs best with an average error of 0.60 years and variance of 0.40 years, although the error

measure used is not stated.

Tristan-Vega and Arribas propose a semi automated method for bone age assessment in [TA05, TVA08]. The case study uses 158 radiographs, with 30 removed due to poor quality. The proposed method extracts radius and ulna. The segmentation algorithm begins with some manually spaced landmarks and an initial segmentation line calculated by interpolating the landmarks. A ROI around the initial segmentation is then calculated and a Sobel edge detector used on this area to find the true bone edge. Morphological operators are used on the output edge mask and an adaptive clustering performed on the pixels inside the bone edge. A potential method for automated landmarking is discussed, however, is found not to be robust enough and hence is not used. From the segmented bone 89 features are extracted and then put through an LDA feature selection process. A neural network is used to classify TW3 stages, with three different ensemble voting methods used (Majority Voting, Ensemble Average and Perceptron Average). Each of the voting methods perform similarly in the leave one out cross validation with Majority Voting performing best with a mean error of 0.94 years and a maximum error of 3.21 years.

Zielinski *et al.* [BKZ08, Zie09] propose a method for segmenting hand radiographs to help aid diagnose patients with arthritis. The input radiograph is first dilated and Gaussian filtered. The new image then has an adaptive threshold applied to it using the Otsu algorithm [Ots75]. This results in a binary mask, which is thinned using various structuring elements to form branches. The pixels on the branches are then analysed to extract the branches that approximate the fingers. Vectors are used in the locations of the branches to form approximations of the finger axes. The grayscale intensities from the vectors are then analysed to find the joints between bones e.g. distal phalange to middle phalange.

Adeshina *et al.* [ACA09] propose an ABAA system that uses AAMs. 170 images from patients between 5 and 20 years of age (87 male and 83 female) are manually annotated with 330 landmarks points which are then used to fit a more detailed shape using a non-rigid registration algorithm. An AAM model is trained on the

whole data set and a linear regression model from the AAM features to chronological age fitted. Different regression models are used for male and female patients. The paper compares the difference between single AAMs for each bone and combined sets of bones, e.g. the carpals. The combined models slightly outperform the individual bone models. Using leave one out cross validation, the average Mean Absolute Error (MAE) for the single bone models is 1.47 ± 0.08 years against chronological age for females and 1.26 ± 0.07 for males. The reported performance of the models using the 13 RUS bones had MAE of 0.80 ± 0.09 and 0.93 ± 0.08 for females and males respectively. The algorithm is essentially a simpler version of BoneXpert (with the addition of a registration phase) that relies on an AAM for the outline and feature extraction. Whilst the age estimation is evaluated on unseen data through cross validation, there is no discussion as to the accuracy of the outline detection algorithm on unseen data.

Giordano *et al.* [GSSL10] propose a system similar to that of Pietka *et al.*, the system uses eight EMROIs and a CROI. The proposed method consists of three stages preprocessing, extraction of ROIs and assessment. Firstly the background is thresholded to get a binary mask of the hand using a local mean and standard deviation. The thumb, third and fifth fingers are then extracted using a wedge function similar to that used by Pietka *et al.* in [PGP⁺01]. Once extracted the grey-level profile of the axis of each finger is taken into account and the EMROIs found using a Difference of Gaussian (DoG) filter presented in the authors earlier work [GLM⁺07, GSSL09]. The bone is then segmented from the EMROI by using another DoG filter and thresholding, a method that uses Gibbs random fields is used to remove any unwanted artefacts and fill any holes inside the foreground. Thirteen shape features are then extracted from each EMROI. In order to classify the TW2 stage of an EMROI a 1-NN classifier is used. In order to segment the CROI, the same wedge functions that are used to extract the EMROIs are used again, the CROI is located by finding the web between the index finger and the thumb, a second point is then located by tracing across the radiograph from the background/soft tissue edge. This forms the CROI, a derivative difference of gaussian (DrDoG) filter is

then applied to the CROI. The bones in the CROI are then segmented by applying Canny edge detection, with a filling algorithm applied to the edge map. Each of the carpals is identified from its location within the CROI. TW2 stage assignment is performed using ACs and ASMs. The final extraction of a CROI is performed using ACs, with the model of each carpal built using the GP Atlas standards. A TW stage is assigned based on the minimum Mahalanobis distance between the query bone AC and the ASM model of a particular stage. The method is tested on 106 radiographs, against two raters, the EMROI achieved a correct TW stage accuracy of 81.2% and 83.9% and within one stage accuracy of 90.6% for each respective rater. The CROI achieves TW stage accuracies of 60.4% and 65.1%, and, a within one stage accuracy of 86.8% for both raters. Finally the MAE of the proposed method against the raters is 0.67 and 0.25 years.

3.4 Why is There a Lack of Widespread Acceptance of Automated Bone Age Assessment Systems?

Many systems for ABAA have been proposed. However, from all of those proposed, only BoneXpert has been released as a commercial system. It is used in 16 hospitals and 15 hospitals are testing it [Tho12]. None of these are in the United Kingdom. It is also noteworthy that BoneXpert is not allowed for clinical use in the U.S.A and instead has the status of an investigational device. So this raises the question why have none of these systems gained widespread acceptance?

There are two main reasons for this: firstly, lack of verification, and, secondly, lack of transparency.

The majority the proposed methods for ABAA give results for accuracies for the assessment, however they do not give an accuracy for each stage (e.g. hand or bone segmentation). It is vital to any automated bone age assessment system that

it can handle a bad segmentation because if it does not it will give an incorrect assessment. When papers have stated the accuracy of segmentation they have not taken into consideration how a bad segmentation can be caught. The most obvious way to handle such an exception is to use some form of classification. There are two ways that a bad segmentation can be handled. It can be rejected or another attempt at segmentation with some new parameters could be attempted. This lack of verification could potentially result in an incorrect assessment and therefore affect the diagnosis being given. Clinicians may then have less confidence in any proposed system.

The second problem relates to how the automated assessment is performed. Many of the proposed methods use features derived from AAMs/ASMs or through wavelet decomposition. Whilst these have been proven to produce good results, experience with clinicians has indicated that they would prefer to have some knowledge as to why an assessment is given e.g. due to small epiphysis width. Hence it would be more beneficial to clinicians if a system was based upon features derived from the shape and texture of the bone which can easily be translated. This problem again causes the clinicians to have low confidence in any potential system as they are unable to know what exactly is causing a diagnosis. All of the methods that have been previously proposed suffer from at least one of these two problems. In order for an automated bone age assessment to gain worldwide acceptance from hospitals and governments, it would seem that both of these problems need to be addressed.

3.5 The Pitfalls of Using AAMs for Automated Bone Age Assessment

Bone ageing consists of three stages: locating the relevant bones; deriving discriminatory features from the bones; and regressing these features onto age (or constructing a classifier to recreate the TW stage). The majority of research in this field uses ASMs or AAMs to combine the first two stages of locating the bones and

deriving features. ASMs/AAMs have been shown to be highly effective in a range of image processing applications [MB04, TMTM12]. We have experimented with ASMs/AAMs [DTL⁺12, DTTB11], but have ultimately rejected the approach for a methodology that is closer to that of Pietka *et al.* [PGP⁺01, PPKG03]. Whilst AAMs obviously can perform segmentation well, there are several reasons for not pursuing this approach. Firstly, AAMs require the manually labelling of “landmark” points in a training set of images. The bones of the hand are fairly simple shapes that do not have many natural landmarks and hence the placement of landmarks can be highly variable between subjects. Secondly, using the model to segment a new image requires a starting template close to the correct position. We found that even with the hand outline the AAM would fit a hand shape that was in fact an outline of the carpal bones. This tendency to fit a valid shape in the wrong location makes automated validation of the process difficult (as shown in Chapter 4). Thirdly, the requirement of training data means that the model is only representative of the population from which the training data is sampled. This makes it hard to develop models tailored to specific demographics without labelling a whole new sample of images. Fourthly, the standard use of AAMs is to capture variation with a homogeneous population in order to use this to detect whether new images are outliers or members of a different population. With hand images, there is a wide variation between the members of the population, and the variation is continuous. BoneXpert overcomes this by splitting the population into three age groups, but this requires three times as much training data and introduces complexities into the predictive stage. Finally, the features the AAM derives do not necessarily have any direct clinical interpretation, and hence make it harder to use the model to explain the relationship between physical characteristics and age estimates.

3.6 Summary

In this chapter we have discussed:

- a variety of image processing techniques and machine learning classification algorithms;
- methods that have previously been proposed for automated bone age assessment;
- why none of these methods have gone on to gain widespread acceptance; and
- the pitfalls of using Active Appearance Models for automated bone age assessment.

Chapter 4

ASMA Stages A & B: Hand Segmentation and Classification

The research in this chapter was published in [DTTB11, DTL⁺12].

This chapter discusses stages A and B of the proposed ASMA system (See Figure 1.1). The aims of these stages are to segment the hand from the radiograph and then verify if the segmented outline is correct. This chapter addresses two problems associated with these tasks:

1. what is the best method for extracting the outline (Sections 4.2 and 4.3); and
2. how to evaluate whether a given outline is in fact a correct outline of a hand (Sections 4.4 and 4.5).

In order to address the first problem, four candidate algorithms for extracting the hand outline are assessed, these are: Otsu thresholding [Ots75], Canny edge detection [Can87], Active Appearance Models (AAM) [CET01], and contouring. These are briefly described in Section 4.2. Despite the fact that the first three candidates have previously been used to extract hand outlines [BKZ08, LBTS05, Tho02], our experience is that the variability in intensity across images, low contrast between the background, flesh, and bone, and variability in hand size and shape

mean that the extraction of the outline is non-trivial, and that none of the algorithms assessed are consistent enough for our requirements. Figure 4.1 shows two examples of incorrect outlines. In these examples the algorithm has found the internal outline of the metacarpals or phalanges. We have observed several other types of error, such as over extended regions or cropping of individual fingers. Therefore, an ensemble method that combines the outlines formed from a range of transformed images (see Section 4.3) is introduced, which is found to create much better outlines (as assessed in Section 4.5).

Clearly, an incorrectly segmented outline will compromise any subsequent steps of bone segmentation and age modelling. Since we wish for ASMA to be a fully automated bone age assessment system, we are required to solve the second problem and hence, find an automated means of classifying whether an outline is correct. The dataset of images used for this work is described in Section 4.1. Every image had a outline segmented automatically (Sections 4.2 and 4.3), that was manually labelled as correct or incorrect depending on the quality of the segmentation. Features were then extracted from the training images and a range of classification algorithms (Section 4.4) were evaluated using the testing set. The results are presented and analysed in Section 4.5.

4.1 Dataset Used

A dataset of 1370 images collected from the Children’s Hospital Los Angeles [GZS⁺07] is used to verify each stage of the ASMA system. All of the images in the dataset were captured on film and then digitised using a laser film scanner and recorded in a database with the following information: date of birth, date of examination, chronological age, gender, ethnicity, tanner index, trunk height, height, weight and the age ratings of two different clinicians using the Greulich-Pyle method. Each image in the dataset was stored as an eight-bit JPEG image. JPEG encoding is a lossy encoding and thus, could lead to difficulties with the segmentation stages of

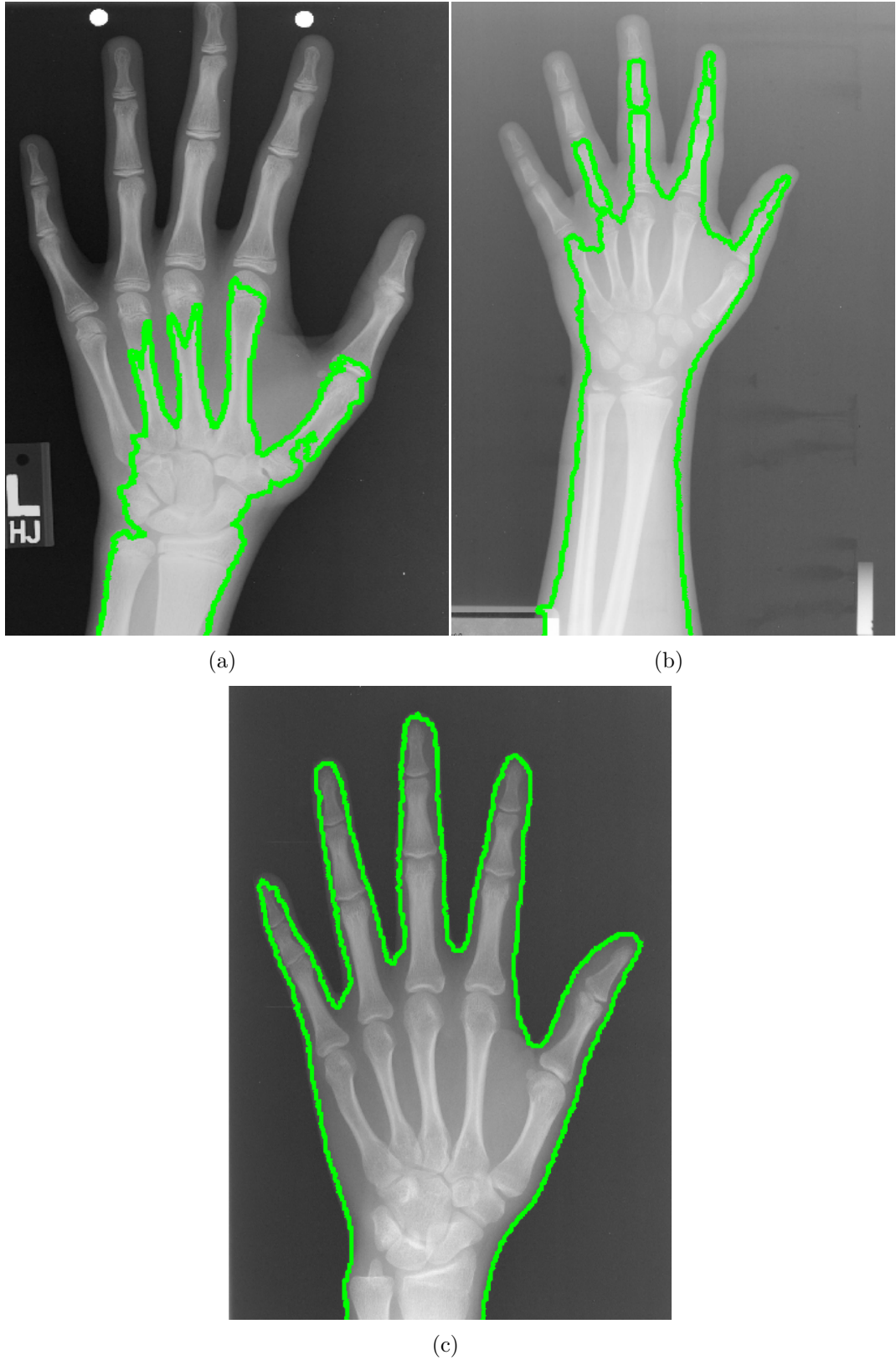


Figure 4.1: (a) and (b) Two examples of incorrectly located hand outlines, and (c) a hand outline correctly segmented. All outlines created using contouring algorithm.

ASMA as some image information will have been lost. All of the images in the dataset are in the age range 0 – 18 years. The amount of radiographs of each age, gender and ethnicity are shown in Tables 4.1–4.3. In this chapter a training set of 1000 images and a test set of 370 images are used.

Table 4.1: The amount of radiographs of each age in the complete dataset.

Age	0	1	2	3	4	5	6	7	8	9
No of Images	20	40	40	40	40	71	63	69	72	68

Age	10	11	12	13	14	15	16	17	18
No of Images	105	103	116	113	100	80	80	79	71

Table 4.2: The amount of radiographs of each gender in the complete dataset.

Gender	No of Images
Female	689
Male	681

Table 4.3: The amount of radiographs of each ethnicity in the complete dataset.

Ethnicity	No of Images
Asian	334
African-American	354
Caucasian	323
Hispanic	359

4.2 Outlining a Hand

Extensive research has been performed on the segmentation of hand radiographs [BKZ08, LBTS05, Tho02, Zie09]. The majority of this work concentrates on the direct segmentation of the bones and uses the problem of finding the outline merely as a motivational example. However, we consider the seemingly easier problem of

finding the hand outline as the most sensible first step for ASMA. Once we have obtained an outline that we are confident is correct, the position of the bones is highly constrained and thus much easier to detect. We investigate the use of three commonly used algorithms for outlining and a contouring algorithm which we believe has not been used for this purpose before.

4.2.1 Active Appearance Models

Active Shape Models (ASM) [Coo00, Eff93, MSC⁺00, NvM⁺03] and Active Appearance Models (AAM) [CET01, Tho02, TKJP09, ACA09] (described in Section 3.1.1) have commonly been used for the segmentation of bones from radiographs. However, the use of ASMs for this task has decreased since the introduction of AAMs, as the AAM incorporates intensity as well as shape features. For fitting the model, we use the Inverse Compositional AAM proposed by Matthews and Baker [MB04].

AAMs are a powerful method for tasks where the user wants to classify objects by shape. However, the object being modelled needs to be well defined, hence, a large set of example images of the object are required for a large amount of variation to be modelled.

Some of the advantages of using AAMs are: firstly, the model can incorporate the knowledge of an expert from the annotation of the training examples (e.g. knowing the difference between a bone in Stage E and a bone in Stage F of the TW method) and secondly, AAMs are able to model the variation of shape and texture in a compact representation and only needs the knowledge of the object gained from the training set.

4.2.2 Otsu Thresholding

The Otsu method [Ots75] for thresholding (described in Section 3.1.2) has been used previously for the use of segmenting the hand from radiographs [BKZ08, Zie09]. Unlike the AAM method, the Otsu method is fully automated as it avoids the need

for a person to label potentially hundreds of images.

However, the disadvantages of such a method are, it does not make use of any shape and/or appearance information. This could mean that the foreground selected may not be the desired object, due to an unknown artefact affecting the probability distribution of the image. Also, there may be more than one outline in the binary mask after thresholding. In order to address this problem, the assumption is made that the largest object in the radiograph is the hand.

4.2.3 Canny Edge Detection

The Canny edge detector [Can87] (described in Section 3.1.3) has been used previously in the context of hand radiograph segmentation by [LBTS05, MMCD⁺05]. As with the Otsu method, the Canny edge detector is a fully automated solution to image segmentation, requiring no hand annotation of images.

A major disadvantage of using this method is that it detects edges, not necessarily outlines. This could cause a problem if there are no strong edges around the object to be segmented. The Canny edge detector identifies multiple edges in an image, and along with the Otsu algorithm the assumption is made that the longest edge represents the hand outline.

4.2.4 Contour Algorithm

To the best of our knowledge, this method has not been previously used in this context, tasks this method has been applied to successfully include weather analysis [HGH⁺76] gesture recognition [ST07], and road sign recognition [PDMPC94]. The contour algorithm used here takes an input radiograph \mathbf{I} with intensity range 0–255. n equally spaced contour levels l_1, l_2, \dots, l_n between the minimum and maximum pixel intensities are calculated, where $l_1 < l_2 < \dots < l_n$. A simple example input can be seen in Figure 4.2(a), which would have contour levels at 50, 100 and 150 (assuming $n = 3$). For each contour level l_i , pixels that have an edge intersecting

l_i are found. In Figure 4.2(b) we show the edges in the example at $l_i = 50$.

The first edge is calculated and highlighted (Figure 4.2(c)). The t intercept and contour point (p_x, p_y) are calculated using Equations 4.1 – 4.3. Where, l_i refers to the contour level, z_0 and z_1 are the pixel intensities either side of the edge, (x_0, y_0) and (x_1, y_1) are the co-ordinates respectively of the pixels and the t intercept refers to the point between the two pixels where the contour level intercepts. A visualisation of this is shown in Figure 4.3.

$$t = \frac{l_i - z_0}{z_1 - z_0} \quad (4.1)$$

$$p_x = x_0 + (t(x_1 - x_0)) \quad (4.2)$$

$$p_y = y_0 + (t(y_1 - y_0)) \quad (4.3)$$

Based on this example the calculations would be:

$$t = \frac{50 - 0}{75 - 0} = 0.\dot{6} \quad (4.4)$$

$$p_x = 1 + (0.\dot{6} \times (2 - 1)) = 1.\dot{6} \quad (4.5)$$

$$p_y = 1 + (0.\dot{6} \times (1 - 1)) = 1 \quad (4.6)$$

The contour point $\mathbf{p} = (p_x, p_y)$ is recorded as the first point of the contour \mathbf{c} . The edges connecting to the current edge are then checked to see if they intersect level l_i , as shown in Figure 4.2(d).

If there is a connecting edge that intersects level l_i as shown in Figure 4.2(e), the process is repeated with the t intercept, contour point \mathbf{p} being calculated (see

0	75	0
125	200	160
75	125	155

0	75	0
125	200	160
75	125	155

(a) A simple example input image **I**. (b) Edges at $l_i = 50$ highlighted.

0	75	0
125	200	160
75	125	155

0	75	0
125	200	160
75	125	155

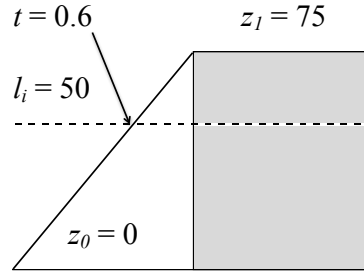
(c) The first edge at $l_i = 50$ highlighted. (d) The first edge at $l_i = 50$ recorded (bold line) and the connecting edges highlighted (dashed line).

0	75	0
125	200	160
75	125	155

0	75	0
125	200	160
75	125	155

(e) The first edge at $l_i = 50$ recorded (bold line) and $l_i = 50$ are highlighted (bold the connecting edge where the contour level intercepts (dashed line)).

Figure 4.2: An example of a contour **c** being found at $l_i = 50$, on a simple input image **I**.

Figure 4.3: An example of the t value calculation.

equations 4.7 to 4.9), and \mathbf{p} being concatenated onto the contour \mathbf{c} (Figure 4.2(f)).

$$t = \frac{50 - 0}{125 - 0} = 0.4 \quad (4.7)$$

$$p_x = 1 + (0.4 \times (1 - 1)) = 1 \quad (4.8)$$

$$p_y = 1 + (0.4 \times (2 - 1)) = 1.4 \quad (4.9)$$

The contour \mathbf{c} is terminated when any of the following happen:

- All connecting edges are not intercepted by l_i .
- The contour returns to an edge it has already visited.
- The contour leaves image \mathbf{I} (as in the example shown).
- All connecting edges have been marked by other contours.

After each contour \mathbf{c} at each contour level l_i has been found, the set of contours \mathbf{S} is returned.

As with the Canny and Otsu methods, the assumption is made that the largest contour represents the outline of the hand. A simple version of the contouring algorithm is given in Algorithm 4.1.

Algorithm 4.1 Contour outlining algorithm

Input: An image \mathbf{I} .**Output:** An outline \mathbf{o} .

```

1. Find  $n$  contour levels.
for each contour level  $l_i$  do
    2.1. Find all the pixels in  $\mathbf{I}$  with an edge crossing the contour level  $c_i$ .
    while all crossing points have not been marked in a contour do
        2.2.1 Find the first crossing pixel  $\mathbf{p}$ , start new contour  $\mathbf{c}$ .
        while stopping criteria are not met do
            2.2.1.1 Scan surrounding points, for edges crossing  $l_i$ .
            2.2.1.2 If crossing point available and stopping criteria not met, add to  $\mathbf{c}$ .
        end while
        2.2.2 Add  $\mathbf{c}$  to set of contours  $\mathbf{S}$ .
    end while
end for
3. Calculate  $\mathbf{o}$ , the longest single contour as the hand outline.
return  $\mathbf{o}$ 

```

4.3 Ensemble Algorithm

There are two main factors that make finding a hand outline difficult:

1. the background/hand division we are attempting to find can be obscured by the hand/bone division, which is often more pronounced; and
2. the distributions of pixel intensities vary greatly from image to image. This is caused by differences in the machine used, deterioration of the bulb over time, and the fact that the energy emitted is non-uniform across the bulb (commonly referred to as the Heel Effect; an example is shown in Figure 4.4).

One way of overcoming the first problem is to rescale an image \mathbf{I} to extenuate the background/hand division using a power transform, $\mathbf{I}^\gamma + c$ (where c is a constant that offsets the intensities back in the range $0 - 255$). However, the second problem of variation in the distribution of intensities means the optimal γ value is image dependent. Hence we propose an ensemble approach. This involves creating twenty rescaled images with scaling factor $\{\gamma_1 = 0.1, \gamma_2 = 0.2, \dots, \gamma_{20} = 2.0\}$ (An example

of the rescaled images is shown in Figure 4.5). All of the rescaled images are then independently segmented by one of the algorithms discussed in Section 4.2, returning a list of (x, y) co-ordinates which makes up the hand outline \mathbf{o}_i for the image rescaled with γ_i .

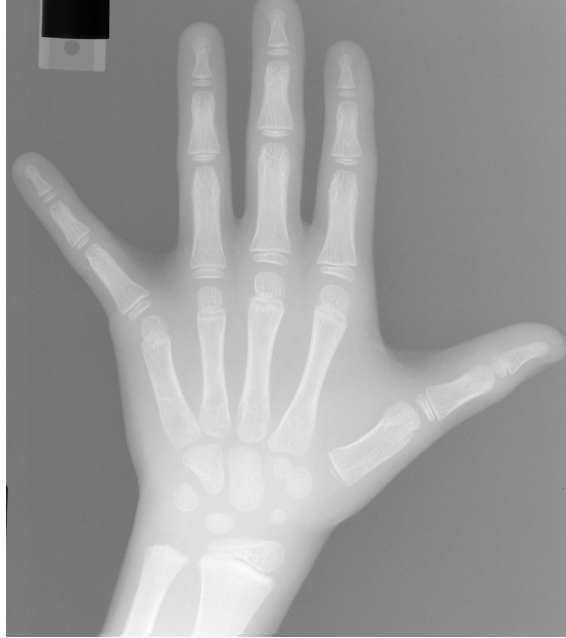
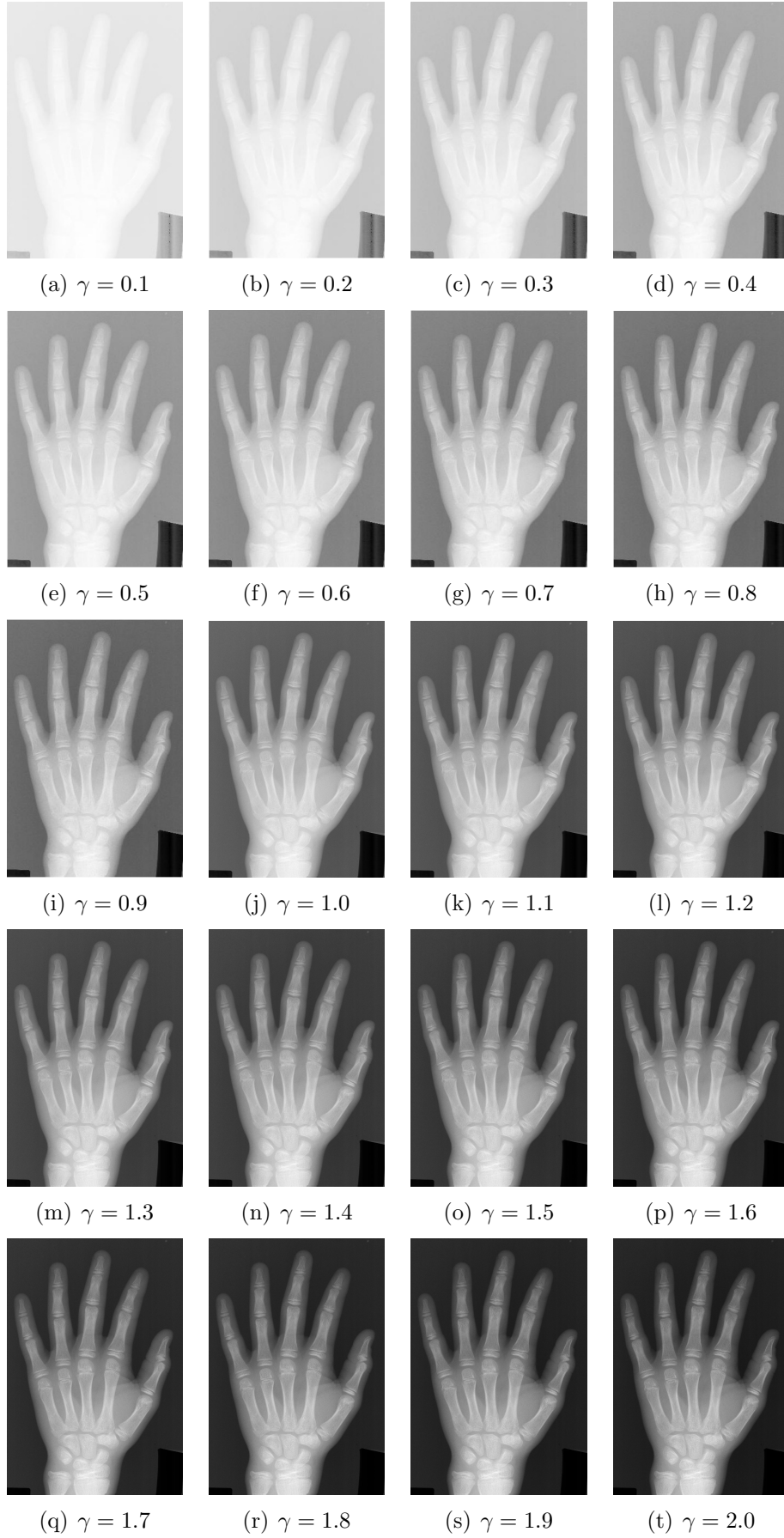


Figure 4.4: An example radiograph of the hand where the heel effect is visible. Notice that the background pixels on the right hand side of the image are brighter.

Once the twenty outlines $\{\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_{20}\}$ have been created, the problem is to choose one. Unlike traditional classification ensembles [ZZCL12], the best outline cannot be selected by voting. Instead, we propose two separate novel selection methods. The first method compares the shape of each automatically generated outline to a set of idealised manually labelled outlines (Section 4.3.1). The second method uses a test statistic based upon the difference in intensity distributions of the radiograph inside and outside the automatically calculated outline (Section 4.3.2).

4.3.1 Dynamic Time Warping Outline Selection

Dynamic Time Warping (DTW) measures the similarity of two one-dimensional series, and has become popular in the field of time series data mining [DTS⁺08].

Figure 4.5: γ corrected hand radiographs.

Given a point-wise Euclidean distance matrix $\mathbf{M}(\mathbf{q}, \mathbf{t})$ of the size $n \times m$ between some query series $\mathbf{q} = \{q_1, \dots, q_n\}$ and a training series $\mathbf{t} = \{t_1, \dots, t_m\}$, where $\mathbf{M}_{i,j} = \sqrt{(q_i - t_j)^2}$. A warping path $\mathbf{w} = \{(a_1, b_1), (a_2, b_2), \dots, (a_k, b_k)\}$ is any set of index pairs that define a traversal of matrix \mathbf{M} that obey three conditions: firstly, $(a_1, b_1) = (1, 1)$; secondly, $(a_k, b_k) = (n, m)$; and finally, $0 \leq a_{k+1} - a_k \leq 1$ for all $k < n$ and $0 \leq b_{k+1} - b_k \leq 1$ for all $k < m$. The distance for any path \mathbf{w} is:

$$D_{\mathbf{w}}(\mathbf{q}, \mathbf{t}) = \sum_{i=1}^k \mathbf{M}(a_i, b_i). \quad (4.10)$$

The DTW distance between series is the total distance of the warping path \mathbf{w}^* through \mathbf{M} with the minimum total distance. Let \mathbf{W} be the space of all feasible paths. The DTW path \mathbf{w}^* can be calculated as:

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbf{W}} (D_{\mathbf{w}}(\mathbf{q}, \mathbf{t})). \quad (4.11)$$

To apply DTW to hand outlines we first need to transform each outline into a one-dimensional series. This is achieved by computing the Euclidean distance of each pixel along the outline to the midpoint of the wrist. Hence for any outline $\mathbf{o} = \{(x_i, y_i), \dots, (x_n, y_n)\}$, the associated one-dimensional series is defined as:

$$\mathbf{q} = \{q_i = \sqrt{(x_m - x_i)^2 + (y_m - y_i)^2} | 1 \leq i \leq n\}. \quad (4.12)$$

Where:

$$x_m = \frac{x_n - x_1}{2}, \quad (4.13)$$

$$y_m = \frac{y_n - y_1}{2}. \quad (4.14)$$

Examples of converting good and bad outlines into a one-dimensional series are

shown in Figures 4.6 and 4.7 respectively. These show the difference between one-dimensional series and hence why DTW should be applicable for this task. For any given image, we create 20 candidate series from our outlines. We wish to select the series that most resembles a correct hand outline using DTW to measure similarity. Of course, there is a wide variation in possible correct outlines. In order to get a range of ground truth candidates we took the 59 idealised radiographs that are presented in [GR04]. These range in age from 8 months to 18 years. We manually outlined these images to form a set of correct outlines, $\{\mathbf{t}_1, \dots, \mathbf{t}_{59}\}$. Our selected outline is the outline \mathbf{o}_k that has the minimum median DTW distance to our set of correct outlines, i.e. $\min DTW(\mathbf{o}_k, \mathbf{t}_j)$. An example of DTW applied to two one-dimensional point series of hand outlines is shown in Figure 4.8.

4.3.2 Likelihood Ratio Outline Selection

An alternative approach to DTW is to select an outline based on the intensity distribution of the original radiograph both inside and outside of the outline. Given an outline \mathbf{o} , the set of points inside the outline of the original radiograph \mathbf{A} , the set of points outside the outline be \mathbf{B} and where, $|\mathbf{A}| = n_a$ refers to the number of pixels inside the outline and $|\mathbf{B}| = n_b$ the number of pixels outside the outline. An image has intensity values in the range 0 – 255. The number of points in set \mathbf{A} with intensity k shall be referred to as a_k and the number of points in set \mathbf{B} with intensity k as b_k . The histograms of intensity occurrences both inside and outside of the outline can be formed and these can be used to calculate the intensity distributions from the relative frequencies,

$$p_{a_k} = \frac{a_k}{n_a}, \quad (4.15)$$

$$p_{b_k} = \frac{b_k}{n_b}. \quad (4.16)$$

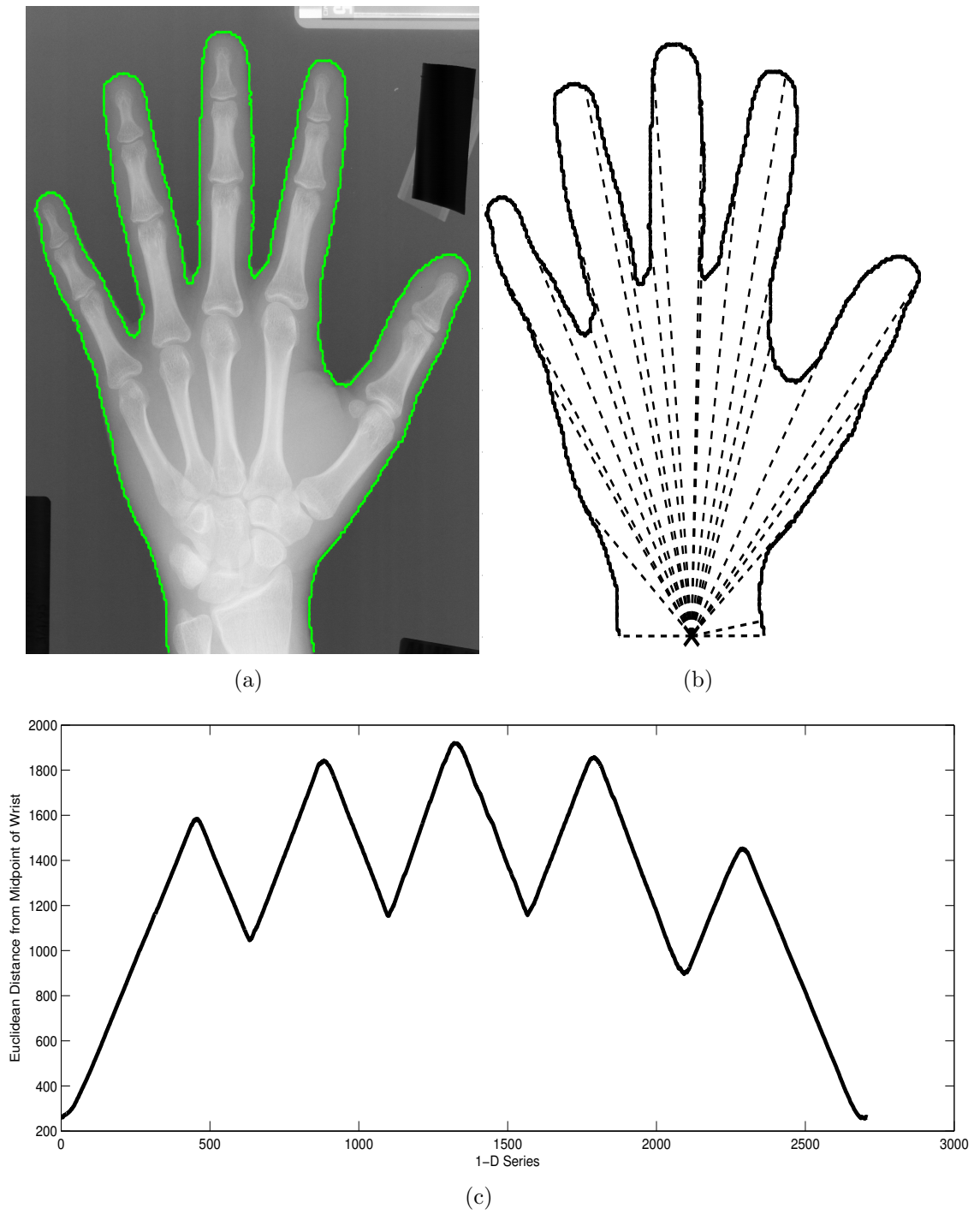


Figure 4.6: An example of a good hand outline from a radiograph (a) being converted into a one-dimensional series (b) and (c).

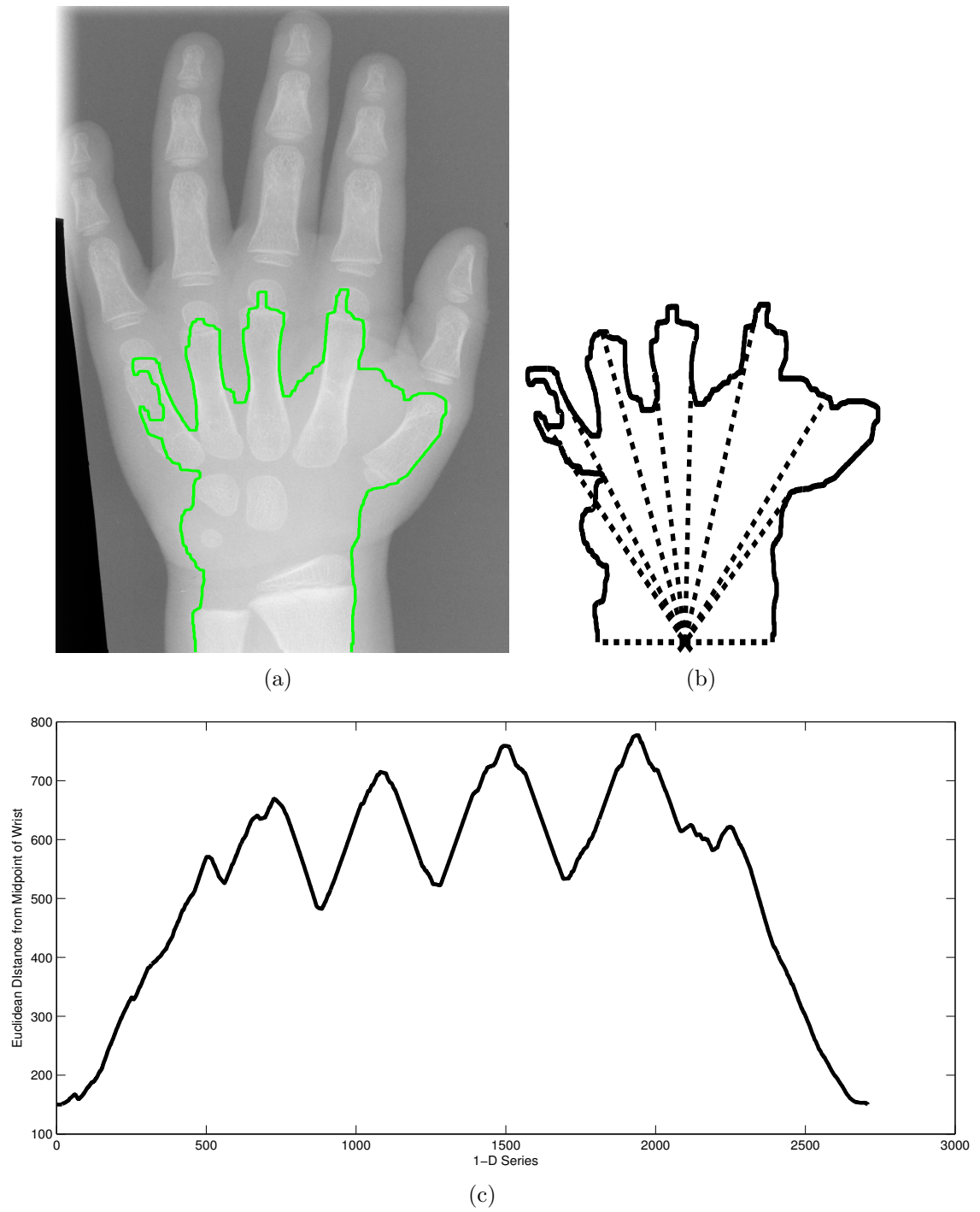


Figure 4.7: An example of a bad hand outline from a radiograph (a) being converted into a one-dimensional series (b) and (c).

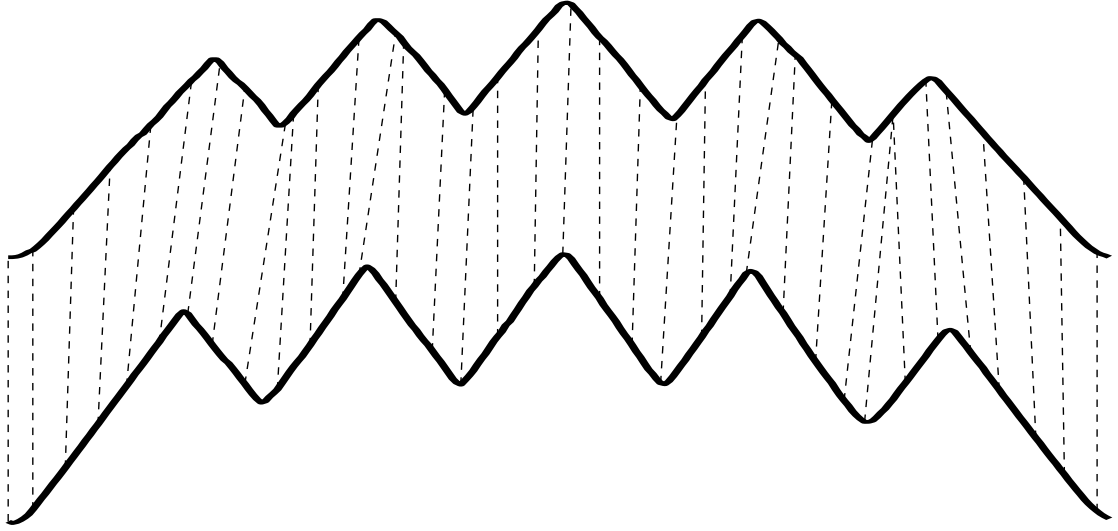


Figure 4.8: An example of DTW between two one-dimensional series of hand outlines.

The correct outline \mathbf{o}_k is calculated as the outline where the intensity distribution within the outline is maximally different from that outside of the outline. To do this we use the likelihood ratio statistic for the test of the null hypothesis that the distributions are equal. Under this null, our probability estimates are:

$$p_k = \frac{a_k + b_k}{n_a + n_b}. \quad (4.17)$$

and the test statistic is given by the log of the likelihood ratio,

$$d_L(\mathbf{o}, \mathbf{I}) = \log(L(\mathbf{A}, \mathbf{B})) = - \sum_{i=0}^{255} \left[p_{a_i} \log \left(\frac{p_{a_i}}{p_i} \right) + p_{b_i} \log \left(\frac{p_{b_i}}{p_i} \right) \right] \quad (4.18)$$

Our likelihood ratio selection criteria is to choose the outline that satisfies $\mathbf{o}_k = \min d_L(\mathbf{o}, \mathbf{I})$.

4.4 Classification of Validity of an Outline

One of the main priorities of ASMA is to minimise the requirement for human intervention in bone age assessment. The ability to automatically detect whether an outline is a valid hand is crucial, because the later stages of bone extraction and age prediction will be compromised if the calculated hand outline is incorrect. There are two main causes of errors in outlining: firstly, it may be caused by the inaccuracy of the outlining algorithm (none of the approaches described in Section 4.2 are 100% robust against the sources of variation described previously) or secondly, by problems with the original image (such as fingers overlapping). Hence, our priority is to avoid incorrect outlines being passed on to the next stage in the process.

The classification task is to predict whether an outline is a valid hand given the outline and the image. We first produced 1000 hand outlines using a mixture of the methods described in Section 4.2. Three volunteers manually labelled the training data as correct or incorrect. Since the priority at this stage is to make sure we do not progress with an incorrectly labelled image, an outline is labelled as correct only if all three human subjects classify it as correct. The training set has 638 positive cases and 362 negative cases.

4.4.1 Transformation

In order to classify images, some form of feature extraction is needed [ECGFS12, HCZ12, ZHC12]. For the classification of hand outlines features are extracted in two stages. Firstly, we adopt two fundamentally different representations and, secondly, we derive features from these representations through transformation. As with the ensemble of outlines, one representation is based on shape and the other intensity.

For intensity features, each segmented image was transformed into two separate intensity distributions, one for outside the outline and one for inside the outline. These two distributions were concatenated to form an instance for each image. Image intensities range from 0 – 255, but using all of these values may obscure the

true differences between the distributions. Hence, five separate datasets were created with distributions derived from quantisation: Intensity(256) (frequencies for all possible values inside and out, hence 512 features), Intensity(128) (merge every two intensity values), Intensity(64), Intensity(32) and Intensity(16) (merge intensities 0-15,16-32 etc. to give 32 features).

The shape features are derived by transforming the outline onto a one-dimensional series using the same method as described in Section 4.3.1. Each series is then smoothed using a median filter of length 51, z -normalised to remove the scaling effect of age variation and resampled to ensure that each is the same length as the shortest series (2709 attributes). The approach of using a one-dimensional series to represent objects segmented from medical images has been proposed in related literature [EHC⁺11, HCZ10]. Several standard transformations were applied to the one-dimensional series:

- **Principal Component Analysis (PCA) Transforms.** PCA forms a linear transform to an alternative set of orthogonal vectors. We tried two forms of PCA. The first, PCA1, found the components on the whole training set. The second, PCA2, performed the transform on the positive cases only, then used the components to define features for both the positive and negative cases. For both PCA methods we created two data sets. The first contains all the components and the second retained the components that explain 95% of the variation (10 and 14 components respectively).
- **Fast Fourier Transforms (FFT).** The FFT can be used to capture phase independent information in a series. We used three versions of FFT: FFT (Full) retains all the transformed Fourier terms; FFT (14) kept only the first fourteen Fourier terms; and PS further transforms the FFT into the Power Spectrum (by squaring and adding the real and complex Fourier terms).
- **Autocorrelation Function (ACF).** The ACF describes how a series is correlated with itself over a range of different intervals. Thus the k th term of the ACF

measures the correlation between each point and the point preceding it by k places. Here all possible intervals ($k = 1 \dots 2708$) to calculate ACF terms are used.

Another approach to classifying hand outlines is to derive a number of descriptive features from the one-dimensional series and then use these summary features for classification. Extracting bespoke features from time series and one-dimensional ordered series for classification is a common approach in the literature [LBCSA11]. Whilst running the segmentation algorithms on radiographs a common observation is that an outline is often incorrect because it misses a finger, or incorrectly finds a partial bone outline instead of the hand outline. Hence features are extracted that relate to the peaks and troughs in the series, which should relate to the finger tips and webs.

There are two main stages in the implementation of the feature extraction. Firstly the peaks and troughs in the one-dimensional series that correspond to the tips of the fingers and the webs of the hand are detected (Algorithm 4.2). Secondly, these landmarks are used to compute a number of features to represent each hand outline in the data set.

Algorithm 4.2 makes the assumption that the initial slope of the series \mathbf{q} is positive. A window of size r iteratively moves across \mathbf{q} from the first position to $|\mathbf{q}| - r$. For each possible starting location of the window, the total sum of all points within the window is computed and compared to the sum of the previous window. If the gradient of the slope was previously observed to be positive, the difference between this sum and the last sum must be positive for this property to remain true; if the difference is negative then the gradient of the line must have changed at some point within the window. The local maximum of the window is identified and extracted as a finger tip and added to the set of finger tips. The line direction is updated to a negative slope and the algorithm continues. Conversely, if the line was previously heading in a negative direction, the difference between this window and the previous window must be negative for this to remain true. If the difference

Algorithm 4.2 Find Finger Tips And Webs

Input: One-dimensional Hand Outline \mathbf{q} Window Size r **Output:** Set of tip locations \mathbf{f} Set of web locations \mathbf{w}

1. Assume that the initial slope of \mathbf{q} is positive, $s = \text{true}$.
 2. Set the sum of the last window, $g = 0$.
 - for** each index i up to $|\mathbf{q}| - r$ **do**
 - 3.1 Calculate the sum of the current window, $h = \sum_{j=i}^{i+r} q_j$.
 - 3.2 Calculate the difference, $d = h - g$.
 - if** $s = \text{true}$ and $d < 0$ **then**
 - 3.3.1 Find local maximum of current window, $m = \max(q_i, \dots, q_{i+r})$.
 - 3.3.2 Add m to set of tip locations \mathbf{f} .
 - 3.3.3 Set slope to negative, $s = \text{false}$.
 - else if** $s = \text{false}$ and $d > 0$ **then**
 - 3.4.1 Find local minimum of current window, $n = \min(q_i, \dots, q_{i+r})$.
 - 3.4.2 Add n to set of web locations \mathbf{w} .
 - 3.4.3 Set slope to positive, $s = \text{true}$.
 - end if**
 - 3.5 Set the sum of the last window as sum of current window, $g = h$.
 - end for**
 - return** \mathbf{f}, \mathbf{w}
-

is now detected to be positive, the local minimum of the window corresponds to the location and this location is extracted and added to the set of webs, and the direction of the line is updated. The algorithm continues processing the hand outline q until it reaches the end of the series, where the sets of finger tip and web positions are returned. A graphical illustration of Algorithm 4.2 is shown in Figure 4.9.

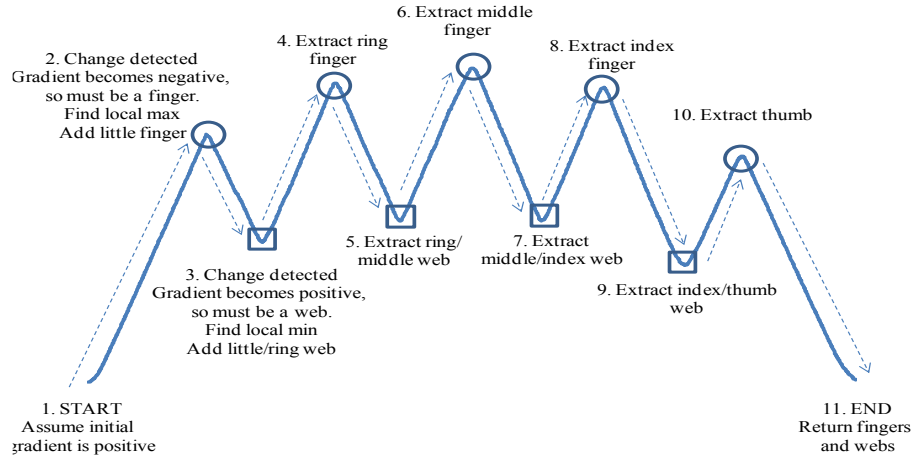


Figure 4.9: A graphical illustration of Algorithm 4.2.

Once Algorithm 4.2 has been performed, the second phase is carried out to transform the finger and web landmarks into a set of features. In total 14 features were extracted. These fall into four distinct categories: the number of landmarks found (number of fingers, number of webs), relative finger tip positions to the index finger tip (thumb to index, middle to index, ring to index, little to index), relative web positions to the thumb/index finger web (index/middle, middle/ring, ring/little), and the ratio of finger height to wrist width (thumb/wrist, index/wrist, middle/wrist, ring/wrist, little/wrist).

The extraction of these features also allowed for a simple classification rule to be enforced: when given a hand outline to extract features from, if the number of finger tips observed by Algorithm 4.2 is not equal to five or the number of webs extracted is not equal to four, the hand outline is classified as incorrect, because all hands in the dataset used have five fingers and four webs.

4.4.2 Classifiers Used

We conducted our classification experiments on the 15 datasets (Raw, FFT (Full & 14 attributes), Power Spectrum (PS), Auto Correlation Function (ACF), PCA1 (100% & 95% variation), PCA2 (100% & 95% variation), Intensity Distribution (16,32,64,128,256 levels), Extracted Features) with ten different classifiers. The classifiers used are the WEKA [HFH⁺09] implementations of k NN [FHJ52] (where k is set through cross validation), Naive Bayes [Lew98], C4.5 tree [Qui93], Support Vector Machines [CV95] with linear, quadratic and radial basis function kernels, Random Forest [Bre01] (with 30 and 100 trees), Rotation Forest [RKA06] and Multilayer Perceptron [MP69].

4.5 Results

There are four stages to our experimentation. Firstly, we evaluate classifiers on our training set of outlines and choose a subset of classifiers to use in testing. Secondly, we apply our outlining algorithms to 370 test images. Thirdly, we assess the outline outputs with the classifiers. Finally, we manually assess the outlines and comment on the suitability of the classifiers.

4.5.1 Classifying Outlines

The training set has 638 positive cases and 362 negative cases. The raw (normalised) data has 2709 attributes. All classifiers are assessed through a ten fold cross validation, with the mean classification accuracy shown in Table 4.4. Generally, building classifiers on transformed data did not improve on the accuracy of those built on the raw data. However, Random Forest with 100 base classifiers achieves the highest overall accuracy of 93.5% using all components of PCA2. A classification accuracy of over 90% is sufficient at this point in the development cycle, hence we continue to use a Random Forest classifier trained on data transformed by PCA2 (100%).

Table 4.4: Ten fold cross validation accuracy of hand segmentation classification (%).

	k -NN	NB	C4.5	SVML	SVMQ	SVMR	RanF30	RanF100	RotF30	MLP
Raw Data	87.7	83.1	85.8	89.2	89.6	88.2	89.5	89.7	90.8	78.4
FFT (Full)	74.1	70.7	79.4	79.4	80.2	84.1	80.2	80.5	85.6	72.3
FFT (14)	66.5	66.7	66.9	66.8	66.9	64.5	62.0	63.6	67.3	67.2
PS	72.8	71.0	76.9	80.0	76.4	71.2	77.4	77.7	84.8	78.3
ACF	80.7	73.4	81.1	85.6	88.5	82.9	86.1	86.4	89.3	79.1
PCA1 (100%)	36.2	80.9	79.7	45.9	46.6	64.2	81.9	81.9	81.7	72.1
PCA1 (95%)	87.3	82.6	85.0	86.3	86.4	63.9	88.6	89.0	89.0	89.5
PCA2 (100%)	80.5	84.8	85.9	86.7	80.3	76.0	93.3	93.5	93.1	81.1
PCA2 (95%)	86.2	84.1	85.1	84.5	85.6	64.3	88.4	87.8	87.0	88.7
Intensity(16)	76.2	67.8	74.6	77.2	73.6	63.8	81.6	82.6	80.1	77.0
Intensity(32)	77.7	66.8	75.0	76.5	75.9	66.5	81.3	82.9	79.4	76.4
Intensity(64)	78.0	67.1	72.7	78.1	75.8	73.1	81.3	83.2	80.5	74.1
Intensity(128)	77.9	66.9	73.7	77.8	75.0	73.9	81.2	82.4	80.5	68.7
Intensity(256)	76.9	68.1	74.9	78.4	77.0	73.6	82.7	82.7	80.4	65.5
Features	88.2	75.0	86.0	83.0	86.6	69.0	89.4	89.5	90.0	89.1

4.5.2 Generating Test Outlines

The second stage of the experimentation involves forming hand outlines on 370 separate testing images. We used the four methods described in Section 4.2, then run ensembles of the Canny, Otsu and contour algorithms on rescaled images. The AAM is trained on a manually labelled set of 30 images.

4.5.3 Testing the Outlines

We trained a Random Forest classifier with 100 base classifiers using all 1000 of the PCA2 training data, then used this classifier to label the outlines generated by our ten outlining techniques as correct or incorrect. Table 4.5 shows the percentage of correct outlines for each outlining algorithm, as determined by the Random Forest classifier. Firstly, these results suggest that the AAM technique is the best performing outlining scheme (85% correct) and that Canny is the worst, failing to form a single correct outline. We investigate these results further in Section 4.5.4. Secondly, ensembling with the likelihood ratio method actually makes Otsu and contour worse. When coupled with the fact that the intensity based classifiers performed poorly (see Table 4.4), this implies that the intensity information is too noisy to use to distinguish outlines. Finally, Table 4.5 demonstrates that ensembling with DTW improves the performance of both the Otsu and the contour outlining algorithm.

Table 4.5: Percentage of the 370 outlines classified as correct by the Random Forest classifier (100).

	AAM	Canny	Contour	Otsu
Non-Ensemble	85.68%	0.00%	25.51%	13.78%
Ensemble (DTW)	N/A	0.00%	77.30%	45.41%
Ensemble (LLR)	N/A	0.00%	1.08%	6.48%

4.5.4 Manual Assessment of Outlining Algorithms

Table 4.5 suggests AAM is the best outliner. In order to explore these results further, we manually labelled the test outlines of AAM and contour ensemble (DTW) as correct or incorrect. This demonstrated that whilst the AAM algorithm usually finds a valid hand shape, this outline is often not in the correct location. Figure 4.10 gives two examples of this phenomenon.

The fitting procedure used by AAMs actually constrains the model to a hand like shape, so the problem for the AAM is finding the location for the constrained outline. Errors occur with the AAM when the search algorithm becomes stuck in a local optimum. In fact, manual inspection revealed that only 187 test images (50.54%) were correctly outlined by AAM. Table 4.6 shows the confusion matrix of the Random Forest classifier against our manual labelling for AAM outlines. The classifier made 140 false positive classifications.

Table 4.6: Confusion matrix for Random Forest on AAM outlines.

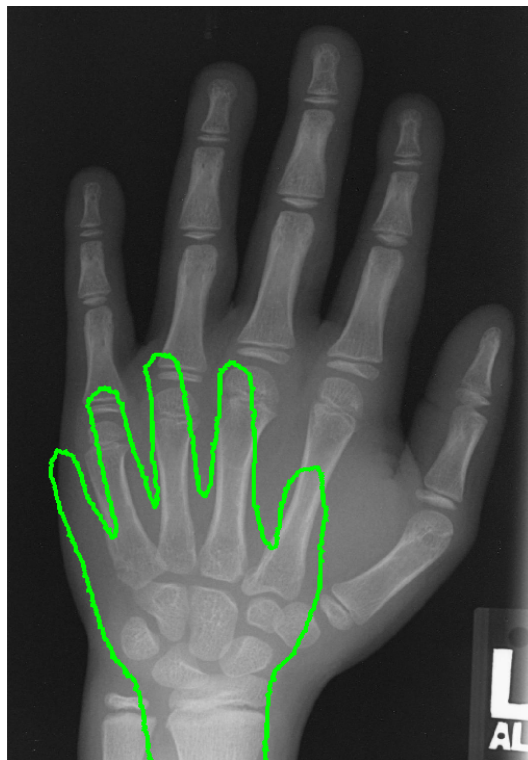
		Actual	
		1	0
Classified	1	177	140
	0	10	43

Since we are primarily concerned with minimising false positives, this presents a serious problem. We could increase the training set size for AAM and potentially improve performance, but there is a strong likelihood that errors of this nature will still occur. Alternatively we could alter our classification scheme to use a measure derived from the image intensity rather than the outline. However, the intensity based classifiers achieved a maximum 82% accuracy. This indicates that discrimination by intensity distributions is harder than discrimination by shape.

The Canny outliner was classified as getting no outlines correct. A visual inspection revealed this was an overly pessimistic scoring, but nevertheless that Canny performed poorly. We believe that this is caused by two factors. Firstly, Canny is an edge detector rather than an outline detector. Whilst a human may classify the



(a)



(b)

Figure 4.10: Two examples of AAM finding an incorrect outline.

Canny output as good, since it broadly looks correct, it does not generally form a continuous outline. Secondly, although the intensity difference between the hand and the background is obvious to the human eye, the actual intensity differentials at the boundaries are not great.

In contrast, a visual inspection of the contour ensemble (DTW) outlines reveals that it correctly found 320 outlines (86.46%). Table 4.7 shows that the Random Forest made only 21 false positive classifications and was cautious about labelling, with over twice the number of false negative as false positives. This is actually desirable, as our primary concern is to stop incorrect outlines proceeding to the bone extraction stage.

Table 4.7: Confusion matrix for Random Forest on contour ensemble (DTW) outlines.

		Actual	
		1	0
Classified	1	265	21
	0	55	29

Our primary conclusion from these experiments is that the contour ensemble (DTW) is the most appropriate outlining algorithm for hand images, and that Random Forest classifiers using a PCA transformation are the most appropriate way of automatically classifying outlines as correct or not.

4.6 Conclusions

This chapter discussed stages A and B of the proposed ASMA system with a novel ensemble algorithm for outlining radiographs and a classification scheme to automatically detect whether an outline is correct introduced. The main findings of the chapter are:

- of the two voting schemes used in the ensemble, DTW outperforms the likelihood ratio test;

- when used in conjunction with a contouring outline algorithm, the ensemble (DTW voting scheme) extracted correct outlines from over 80% of images. The only other contender in terms of accuracy is AAMs, but this method is not suitable for this project due to the errors made by the method being hard to detect automatically; and
- the most effective classifier assessed is a random forest applied to hand outlines transformed into principal components.

Chapter 5

ASMA Stages C, D & E: Bone Segmentation, Feature Extraction and Bone Segmentation Classification

The work in this chapter is an extended version of research published in [DTB12].

In this chapter we cover stages C–E of the proposed ASMA system (See Figure 1.1. This involves finding three Regions of Interest (ROIs) around the phalanges of the middle finger, segmenting the bone from each ROI, extracting features from the bone segmentation and finally classifying whether the segmentation is correct. As the main priority for ASMA is to be a fully automated bone age assessment system, any bad segmentations need to be rejected. This chapter investigates:

1. how to find a ROI around each of the phalanges of the middle finger (Section 5.1);
2. given an ROI how to segment the hard tissue from soft tissue/background (Section 5.1);

3. how to extract features from the segmentation (Section 5.2); and
4. evaluating whether a given bone segmentation is a correct segmentation and investigating which method is best for classification (Section 5.3).

In order to address the first problem we investigate the use of the one-dimensional hand series and a peak/trough detection algorithm to find the tip and webs of the middle finger. Then, given the location of the middle finger we locate the ROIs based on the known anatomy of the finger.

After extracting each ROI, the next problem is how to segment the hard tissue. There are various problems associated with this process including the unknown number of areas of hard tissue, hard tissue overlapping due to bent fingers, etc. A novel technique for bone segmentation which uses Gaussian pyramids in conjunction with Canny edge detection [Can87] is proposed. This ensures that the level of detail is not too fine and hence unnecessary artefacts that could lead to an incorrect segmentation are eliminated.

Stage E is described in two parts. Firstly, we describe 25 features derived through an investigation of the Tanner-Whitehouse stages. Once extracted, the features can be used for two different purposes within ASMA: classification of bone segmentation and predicting bone age. Secondly, the methods used to extract the features are summarised.

As with the hand segmentation, the bone segmentation stage is not 100% robust. Figure 5.1 shows two examples of incorrect and one example of correct segmentations. In the incorrect examples the algorithm has connected an unfused phalanx and epiphysis, and also, not segmented the epiphysis. Several other types of error have also been observed, such as over/under extended regions and the phalanx not being found. Any incorrectly segmented outline will compromise any subsequent ASMA steps, and since the main priority of ASMA is to be a fully automated bone age assessment system, we are required to find an automated means of classifying whether a segmentation is correct. All of the images that had a correct

hand outline from Chapter 4 and are in the age range 2 – 18 years are used in this chapter. Restriction to this age group is common with previously proposed ABAA systems [ACA09, TKJP09] since it has been shown that all BAA is unreliable on patients under the age of 2 years [BEK⁺99]. Each image had the phalanges of the middle finger segmented automatically (Section 5.1). Features were then automatically extracted from the segmentation (Section 5.2), if the segmentation had passed through each of these stages without being rejected due to failing a set of rigid rules, it was manually labelled as correct or incorrect depending on the quality of the segmentation. The images were then split into testing and training data. A range of representations of the segmentations were then extracted from the training images and a set of classification algorithms (Section 5.3) were evaluated. The best classifier and representations were compared using the previously unseen testing set. The results are presented and analysed in Section 5.4.

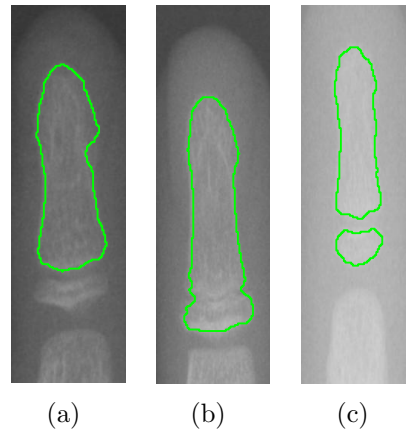


Figure 5.1: (a) and (b) two examples of incorrect bone segmentations, and (c) a correct bone segmentation.

5.1 Bone Segmentation

Once a correct hand outline has been obtained and verified as correct, the next stage of ASMA is bone segmentation. The bone segmentation stage of ASMA is split into two parts. Firstly, finding the location of the three ROIs, and, secondly,

the segmentation of areas of hard tissue from the ROI.

5.1.1 Locating the ROIs

Given that both a two-dimensional and one-dimensional hand outline have been extracted from stages A and B of ASMA, it makes sense to use these for locating the middle finger. Algorithm 4.2, was used to find features from the one dimensional hand outline for classification purposes in Chapter 4. The output of the algorithm is a list of indexes of the tips and webs from the outline. These can be used here to denote the area that encloses the middle finger, using the index of the third tip, along with the indexes of the second and third webs on the two-dimensional outline. Figure 5.2 shows an example of this process.

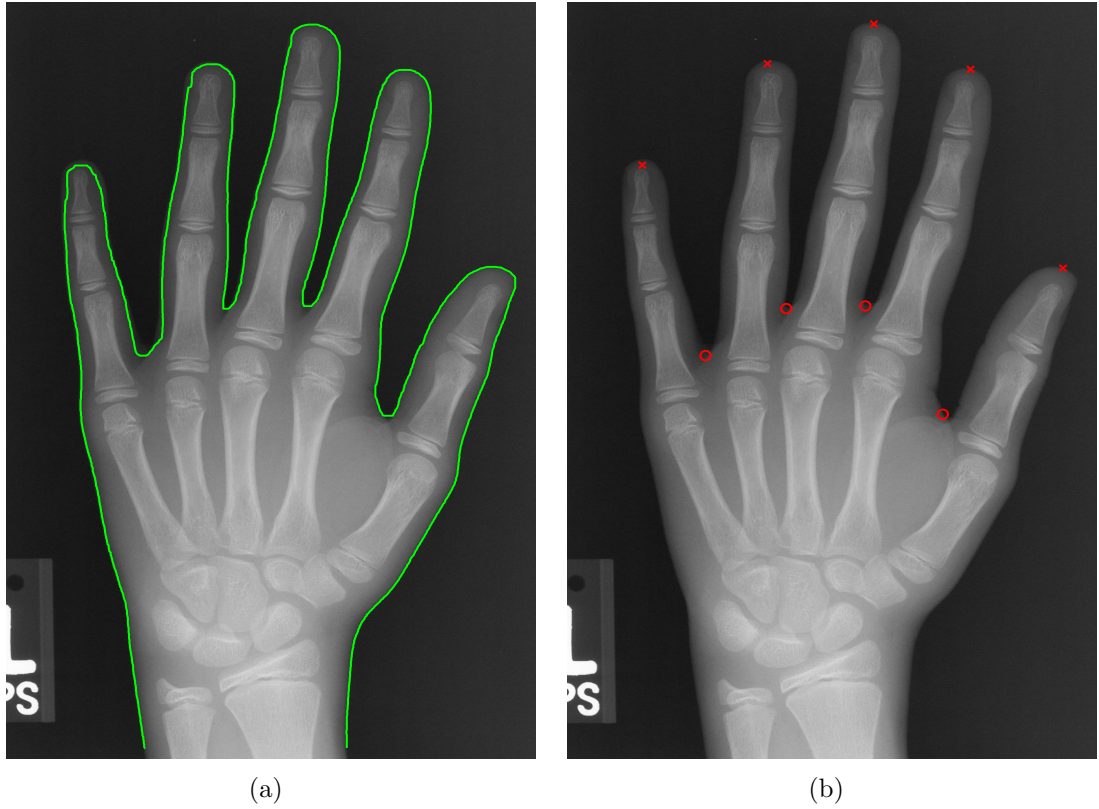


Figure 5.2: An example radiograph labelled with (a) the correct hand outline extracted from stages A and B of ASMA, and (b) the finger tips and webs found using Algorithm 4.2.

Once the tip and webs have been found, we can calculate an axis along the middle finger. The first point of the axis is the tip of the finger, and the final point of the axis is defined as the midpoint between the two webs. Linear interpolation is then used to calculate the rest of the axis. Since the one and two dimensional versions of the outline were already obtained during stages A and B, it is more intuitive to calculate the axes with these, than it is to use wedge functions [PGP⁺01, PPG⁺01].

We locate the ROI based upon the known anatomy of the finger. In order to do this the size of each ROI is estimated as a ratio of the length of the finger axis, and, in order to overcome the problem of overlapping bones, the ROIs overlap the neighbouring ROIs. An example of the process of calculating the axes and ROIs is shown in Figure 5.3.

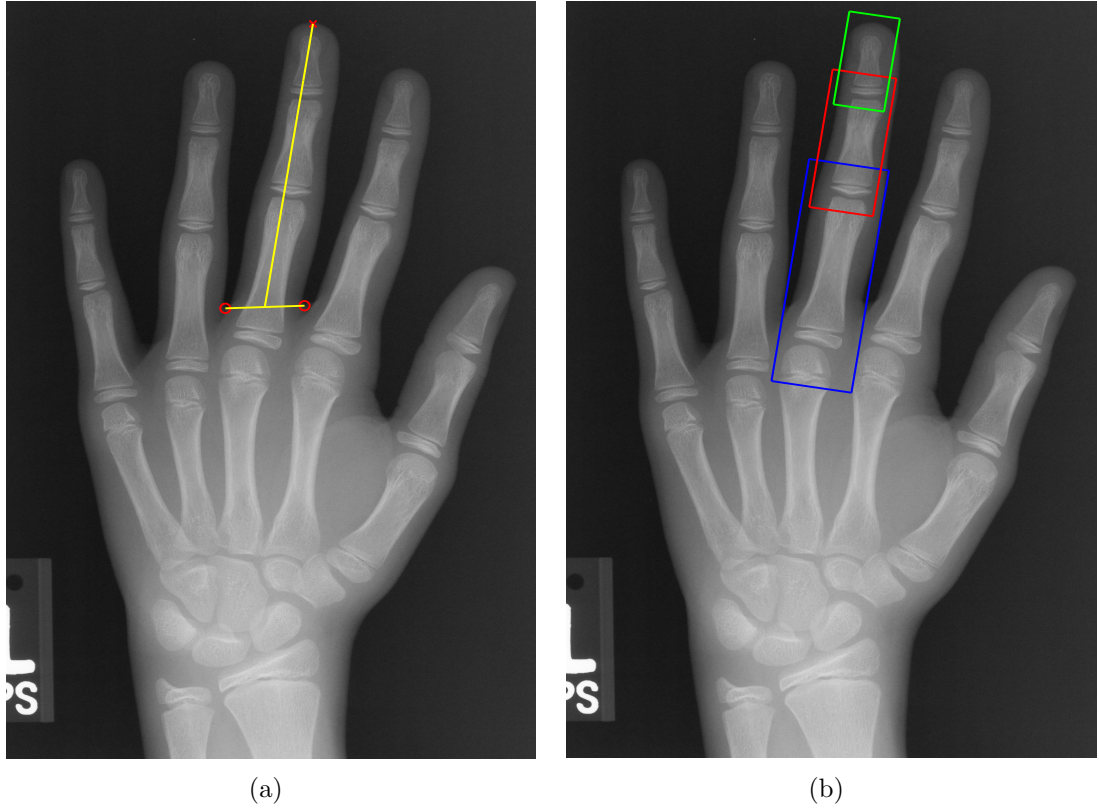


Figure 5.3: An example radiograph labelled with (a) the axes of the middle finger, and (b) the ROI around each of the phalanges of the middle finger.

The final stage of the process is to make all of the ROIs a uniform size. Each ROI

is warped into a rectangular shaped mesh (500×150 pixels) using a piecewise affine warp to create a final ROI for each of the three bones. The warp was performed for two main reasons. Firstly, to standardise rotation and scale of all types of a particular bone and hence make segmentation easier, and, secondly, to allow for a universal bone segmentation algorithm rather than a specialised function for each bone. Figure 5.4 shows a ROI for each bone after the warp.

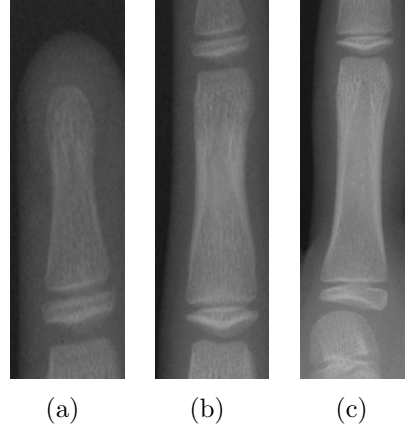


Figure 5.4: An example ROI of (a) distal, (b) middle and (c) proximal phalanges.

5.1.2 Segmenting Hard Tissue from ROI

Segmenting the bone from the ROI should be easier than finding the hand outline, since the background has been removed and the bone shapes are simpler. However, there are still problems associated with bone segmentation. Although the phalanx is known to be located in the ROI, there are still many unknowns, such as whether the epiphysis is present or not, where in the ROI the phalanx is located. Another problem arises from the fact that during the ossification process some parts of the epiphysis can be brighter than others, leading to only part of the epiphysis being segmented.

In order to overcome the problem of the high resolution detail the use of Gaussian Pyramids [AAB⁺84] was investigated. A Gaussian Pyramid hierarchy of images works by taking some input image \mathbf{I} , as level 0 of the pyramid. The image is blurred

using a Gaussian filter, and, then down-sampled by removing every other row and column. The resulting down-sampled image \mathbf{I}_1 , is level 1 of the pyramid. This process is then repeated n times to produce n layers of the pyramid. The algorithm is given in Algorithm 5.1.

Algorithm 5.1 Gaussian Pyramid Algorithm

Input: Image \mathbf{I}

 Number of Levels n
Output: Set of downsampled images \mathbf{J}

 1. Set input image \mathbf{I} as lowest level of pyramid \mathbf{J}_0 .

for each level of pyramid $\{i = 1 \dots n\}$ **do**

 2.1 Set image as the i^{th} level of pyramid $\mathbf{J}_i = \mathbf{J}_{i-1}$.

 2.2 Smooth image \mathbf{J}_i with gaussian filter.

 2.3 Remove every other row and column \mathbf{J}_i .

end for
return Set of downsampled images \mathbf{J} .

Down-sampling the image improves the segmentation, but it is crucial not to down-sample too much, since the distinction between the phalanx and epiphysis can be lost. We found that using the Gaussian pyramid algorithm, in conjunction with a Canny edge detection [Can87] at each level of the pyramid, performed well. At level n of the pyramid the only objects that should be left in the ROI are the hard tissue with the fine detail removed. Therefore edges found at level n form the basis for the edge to use at level $n - 1$. After this process was completed, any edges that were connected to the sides of the ROI box, or were not complete loops, were removed. For this work a five level pyramid is used, with a 3×3 Gaussian filter where $\sigma = 0.5$, and Canny edge detection using the default MATLAB parameters.

The output of the process should either be a binary mask with one or two outlines that form complete loops, depending on the presence of the epiphysis or not. The region inside the largest remaining loop is labelled as the phalanx. If a second loop is present it is labelled the epiphysis. In the rare case of no complete loops being present, the ROI is rejected (although this does not stop an age estimation being made for the patient as the other bones can be used). If more than two complete loops are found to be present, the area inside the loops is calculated, and the smallest

loop is discarded. This process is repeated until there are two loops left. The output of this process is a binary mask, an example of which is shown in Figure 5.5.

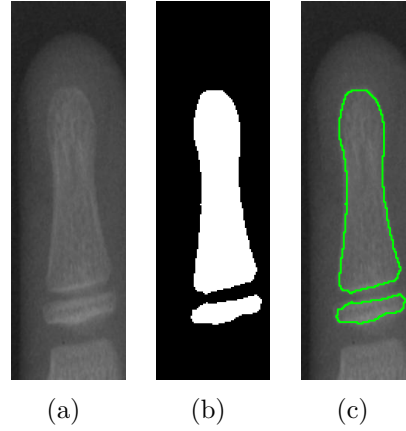


Figure 5.5: (a) An example of a distal phalanx ROI, (b) resulting binary mask, and (c) outline of segmentation imposed onto ROI.

5.2 Feature Extraction

Stage D of ASMA is to extract features from the segmentation. The extracted features have two potential uses. Firstly, they can be used as attributes for the classification of the bone segmentation (stage E), and, secondly, for use for predicting age either by classification for TW stages or in a regression model (stage F).

The goal of stage D is to be able to automatically extract the features that best capture the variability described in the text for classifying TW stages (see Tables 2.2, 5.1 and 5.2). From the text describing each of the phalanges stages, 25 features were derived as potentially being of use for the ASMA system. These features are shown in Table 5.3.

Currently only shape features are used. Although it is plausible that intensity features will help in the finer distinctions between the stages, shape features are clearly the most discriminatory.

From Tables 2.2, 5.1 and 5.2, the most important feature is the stage of epiphyseal

Table 5.1: The Tanner-Whitehouse stages of middle phalange III [TWM⁺75].




















Stage	Image	Description
B		The centre is just visible as a single deposit of calcium, or more rarely as multiple deposits. The border is ill-defined.
		
		
C		The centre is distinct in appearance and disc-shaped, with a smooth continuous border.
		
		
D		The epiphysis is half or more the width of the metaphysis.
		
		
E		The central portion of the proximal border has thickened and grown towards the end of the adjacent phalanx, shaping to its trochlear surface.
		
		
F		The epiphysis is as wide as the metaphysis.
		
		
G		The epiphysis caps the methaphysis.
		
		
H		Fusion of epiphysis and metaphysis has now begun.
I		Fusion of epiphysis and metaphysis is completed.

Table 5.2: The Tanner-Whitehouse stages of proximal phalange III [TWM⁺75].









Stage	Image	Description
B		The centre is just visible as a single deposit of calcium, or more rarely as multiple deposits. The border is ill-defined.
C		The centre is distinct in appearance and disc-shaped, with a smooth continuous border.
D		The epiphysis is half or more the width of the metaphysis.
E		The proximal border of the epiphysis is concave and distinctly thickened.
F		The epiphysis is as wide as the metaphysis and follows closely its shape, although it does not yet cap it at the edges.
G		The epiphysis caps the methaphysis.
H		Fusion of epiphysis and metaphysis has now begun.
I		Fusion of epiphysis and metaphysis is completed.

Table 5.3: Features derived from Tanner-Whitehouse stages.

Feature Number	Feature Name
1	Epiphysis
2	Phalanx Ellipse Height
3	Phalanx Ellipse Width
4	Phalanx Height
5	Phalanx Width
6	Phalanx First Quartile Width
7	Phalanx Third Quartile Width
8	Metaphysis (Phalanx Ninety Percentile) Width
9	Phalanx Eccentricity
10	Phalanx Width to Height Ratio
11	Phalanx Roundness
12	Phalanx Area to Perimeter Ratio
13	Phalanx First Quartile to Width Ratio
14	Phalanx Third Quartile to Width Ratio
15	Phalanx Metaphysis to Width Ratio
16	Epiphysis Ellipse Height
17	Epiphysis Ellipse Width
18	Epiphysis Height
19	Epiphysis Width
20	Epiphysis Eccentricity
21	Epiphysis Distance to Phalanx
22	Epiphysis Width to Height Ratio
23	Epiphysis Roundness
24	Epiphysis Area to Perimeter Ratio
25	Epiphysis Width to Metaphysis Ratio

development and hence the first feature derived is the presence of the epiphysis. This is a binary variable and is estimated by counting the number of foreground regions from the binary mask extracted during bone segmentation. The other features are summary measures of the phalanx (features 2 - 15), and, if the epiphysis is present, summary measures of the epiphysis (features 16 - 25). From the extracted features, features 16 to 25 should model the stage of development of the epiphysis during the early-mid stage of development, thus should be more influential at predicting TW stages during this development phase; whilst phalangeal features should be of more importance at the beginning and end of skeletal development. Obviously the

warping process into the rectangular mesh performed during the bone segmentation stage will have a detrimental effect in the later stages of ASMA. As the height and width features extracted would be incorrect. In order to stop this problem occurring the binary mask is warped back to the landmarks of the original ROI. An example of this process is shown in Figure 5.6.

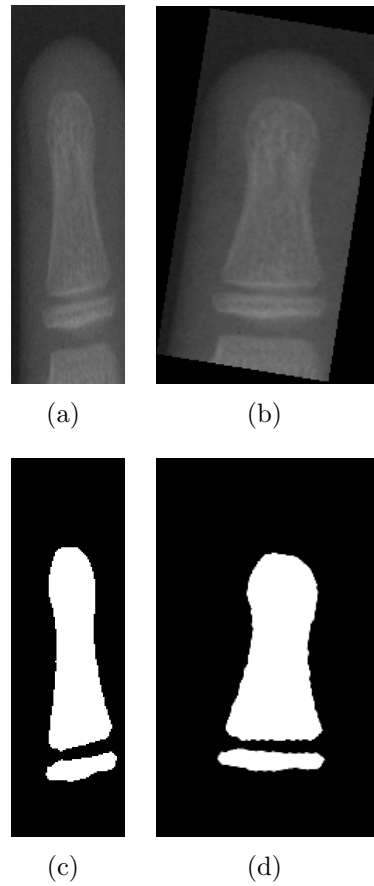


Figure 5.6: (a) and (b) ROI before and after unwarp process, (c) and (d) bone segmentation of (a) before and after unwarp process.

Basic size descriptors such as height and width are obviously going to be reasonably indicative of age, although they are of less use than one might first think because the size of the image of the hand does not directly map to the size of the actual hand. This is because the focus of the radiograph machine is adjusted so that the hand image is approximately the same size independent of actual size.

The obvious way to find the height and width of the phalanx and epiphysis is to

find the length of the vertical line down the centre of the bone for height and the length of the horizontal line across the middle of the vertical for width. However, this assumes the bones are vertically aligned, which is often not the case, since fingers are often not straight. In order to calculate an estimate for height and width, we fit an ellipse to both the phalanx and epiphysis (if present).

A standard way of fitting an ellipse to an image is to use the Hough transform [Bal81]. The algorithm creates an ellipse \mathbf{e} with features: central co-ordinate (x, y) , the length of the major axis a , the length of the minor axis b and the angle of the ellipse θ . The error e between \mathbf{e} and the bone segmentation binary mask \mathbf{S} is calculated using the cost function:

$$e = c - \beta d. \quad (5.1)$$

Where c is the number of pixels that are in the ellipse \mathbf{e} and are identified as hard tissue in \mathbf{S} (foreground pixels), d is the number of pixels that are in the ellipse \mathbf{e} and are not identified as hard tissue in \mathbf{S} (background pixels), and β represents a penalty value. The optimal ellipse \mathbf{e}^* is calculated as the ellipse \mathbf{e} with the maximal value of e , and is the output of the algorithm. An algorithmic description is given in Algorithm 5.2.

However, as shown in Algorithm 5.2 the transform is inefficient, due to the number of nested loops, and possible permutations of an ellipse. It is faster therefore to fit the transform on a lower resolution mask and gain an initial approximation of the ellipse. Hence, as with the bone segmentation, we make use of Gaussian pyramids. An ellipse is fitted at the lowest resolution with the Hough transform. This ellipse is then used as the initial starting point for the transform at the next highest level, which is refined by adjusting the parameters a small amount to find the best fitting ellipse \mathbf{e}^* . The axes of the ellipse generated at the highest resolution of the mask are used to calculate an estimate for the height and width of the phalanx and epiphysis (features 2, 3, 16, and 17). An example of the calculated ellipses with

Algorithm 5.2 Elliptical Hough Algorithm

Input: Bone Segmentation **S****Output:** Best fitting Ellipse **e***

```

for each pixel  $(x, y)$  in the segmentation S do
  for each possible major-axis value  $a$  do
    for each possible minor-axis value  $b$  do
      for each angle  $\theta = \{0 \dots 179\}$  do
        1.1  $\mathbf{e} = [x, y, a, b, \theta]$ .
        1.2  $e = \text{calculateError}(\mathbf{e}, \mathbf{S})$ .
        if  $e$  is maximal error then
          1.3.1 Set best fit ellipse  $\mathbf{e}^* = \mathbf{e}$ .
        end if
      end for
    end for
  end for
end for
return Best fit ellipse  $\mathbf{e}^*$ .

```

major and minor axes is shown in Figure 5.7(a). For the phalanx and epiphysis ellipses, Gaussian pyramids of four and two levels are used respectively.

In order to ensure that the ellipse has fitted well, we interpolate the line of each axis through the whole mask to gather the correct height and width of the phalanx (features 4, and 5) and epiphysis (features 18, and 19) if present, as shown in Figure 5.7(b).

The new interpolated major axis of the phalanx is used to calculate the width of the phalanx at various places along its length. The objective of extracting these features is to capture the change from a fairly straight sided bone in the early stages, to one that narrows in the middle in the latter stages (this phenomena most obviously occurs in the distal phalange, see Table 2.2). In addition to finding the width at the middle of the axis (feature 5), we also find the first quartile, third quartile width and metaphysis width (features 6-8), shown in Figure 5.7(c). The ratio of each of these widths in relation to the middle width is calculated (features 13 - 15). If the epiphysis is present, the relationship between the width of the metaphysis and epiphysis can also be captured (feature 25).

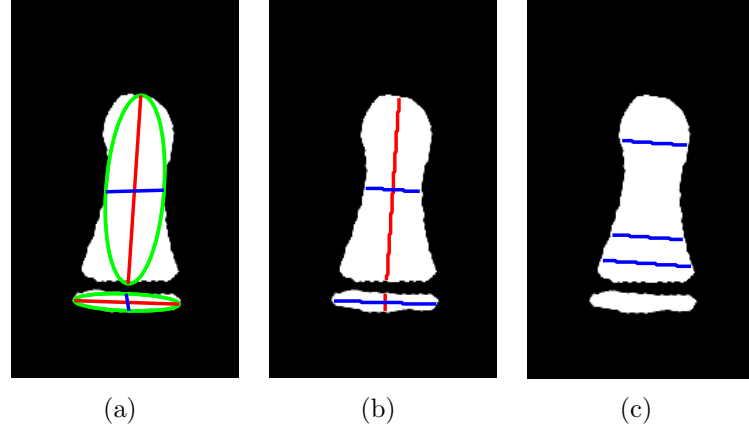


Figure 5.7: Shows various features calculated from a bone segmentation (a) the best fitting ellipse for both the phalanx and epiphysis, along with the major (red) and minor (blue) axes, (b) the interpolated height (red) and width (blue) of the phalanx and epiphysis, and (c) the various widths calculated along the phalanx.

A further way of capturing the progressive change of shape of the phalanx and epiphysis is to measure how circular the phalanx and epiphysis (if present) are. This can be done using two different methods. Firstly, by calculating the eccentricity of the ellipses used to model the hard tissue in the ROI. The eccentricity of an ellipse is a standard measure of its roundness (features 9, and 20), which can be calculated using:

$$ecc = \sqrt{1 - \left(\frac{g^2}{f^2}\right)} \quad (5.2)$$

Where the length of the semi-major axis and semi-minor axis are referred to as f and g respectively. Another method to capture the change in shape is by calculating the roundness h of the phalanx and epiphysis themselves (features 11 and 23), using Equation 5.3, where k is the area of the phalanx/epiphysis and p is the length of the perimeter.

$$h = \frac{4\pi k}{p^2} \quad (5.3)$$

This measure was derived by substituting the equation for the radius of a circle (Equation 5.4) into the equation for area (Equation 5.5), and re-arranging to form Equation 5.6. Which holds true for a circle. However, the bony tissue is not circular and so the value of roundness h will change over the course of skeletal development.

$$r = \frac{p}{2\pi} \quad (5.4)$$

$$k = \pi \frac{p^2}{(2\pi)^2} \quad (5.5)$$

$$1 = \frac{4\pi k}{p^2}. \quad (5.6)$$

Another key discriminatory characteristic is the distance from epiphysis to the phalanx. For ASMA, this is quantified as the Euclidean distance between the mid-points of the two (feature 21).

5.3 Bone Segmentation Classification

As with the segmenting of the hand from the radiograph, the bone segmentation approach described in Section 5.1 is not 100% robust against the sources of variation. Hence we require a mechanism for validating whether any given segmentation is correct. This is performed in two phases. The first phase uses fixed rules based on the segmentation and feature set. The second phase uses a pre-trained classifier to predict whether the outline is correct or not.

5.3.1 Rejection Rules

The approach of the ASMA system is very conservative, in that at any stage of the process the image can be rejected (shown in Figure 1.1). The main reason for this is

to ensure that no bad segmentations or features go on to the next stage and hence have a detrimental effect on age prediction. In order to stop this occurring a check needs to be taken at every stage that can reject any bad segmentations or feature extractions.

Obviously a bad bone segmentation / feature extraction stage for one or two of the three phalanges does not mean that the whole radiograph should be rejected. As age predictions are based on segmentations deemed acceptable.

The rejection rules used are relatively simple. For the bone segmentation stage, a ROI is rejected if no areas of hard tissue are found. This removes the most obviously incorrect segmentations. The second rule is aimed at ensuring that the extracted features are correct. This is extremely important to ASMA, as both stage E and F are reliant on this. The feature extraction relies quite heavily on the elliptical hough transform calculating the correct ellipse, as many of the other features are dependent on it. The easiest method to check if this has been done correctly is to check the length of the axes and compare them to the interpolated height and width of the bone. If a large difference occurs in any of the heights and widths for both the phalanx and epiphysis, the feature extraction phase is rejected.

5.3.2 Training a Classifier

In Section 5.4 we describe the experimental evaluation of classifiers for bone segmentations. We look at a variety of different representations for the segmentations and use the classifiers described in Section 3.2.

For the training data, we represent the segmentations in three different ways. The first representation is to use the features extracted in Section 5.2. Although extracted features were not the optimal representation for the hand segmentations, it is expected that they will perform better here. The features were derived from the TW standard stages for each of the phalanges and hence the correct segmentations should provide a good model. Also the features extracted from the hand outlines

were from the one-dimensional series whereas the features used here were taken from the actual segmentation.

Secondly, the segmentations are represented as a one dimensional series of the outline. The one-dimensional series is obtained by calculating the Euclidean distance of each pixel along the outline of a piece of hard tissue its midpoint. Obviously, for these series to be comparable and, hence, for an optimal chance of classification, all the series need to start in the same position. This is not as easy as with the hand outlines as there is no obvious starting point for the series. As the width for both the phalanx and epiphysis were calculated in the feature extraction stage, we know the first point on the outline of the segmentation that was intersected when calculating the width and use this as the starting point of the series and move around the hard tissue in a clockwise direction. If an epiphysis is present the series is concatenated to the phalanx series. Clearly, the length of outlines will vary. To simplify the classification the outlines were resampled to ensure each was the same length as the shortest series (80 attributes, 50 phalanx, and 30 epiphysis). If no epiphysis is present the phalanx series would be followed by 30 zeros. An example of this process is shown on a good and a bad outline in Figures 5.8 and 5.9 respectively.

The final representation of the data uses the intensity distribution of the segmentation, by concatenating the intensity distribution within the segmentation and the intensity distribution outside of the segmentation.

The classification task is to predict whether a bone segmentation is valid or not. We first produced 600 segmentations of each of the phalanges of the middle finger to use as training data and labelled these as correct or incorrect. The training set of each phalanx has 300 instances where the epiphysis is not present and 300 where it is present.

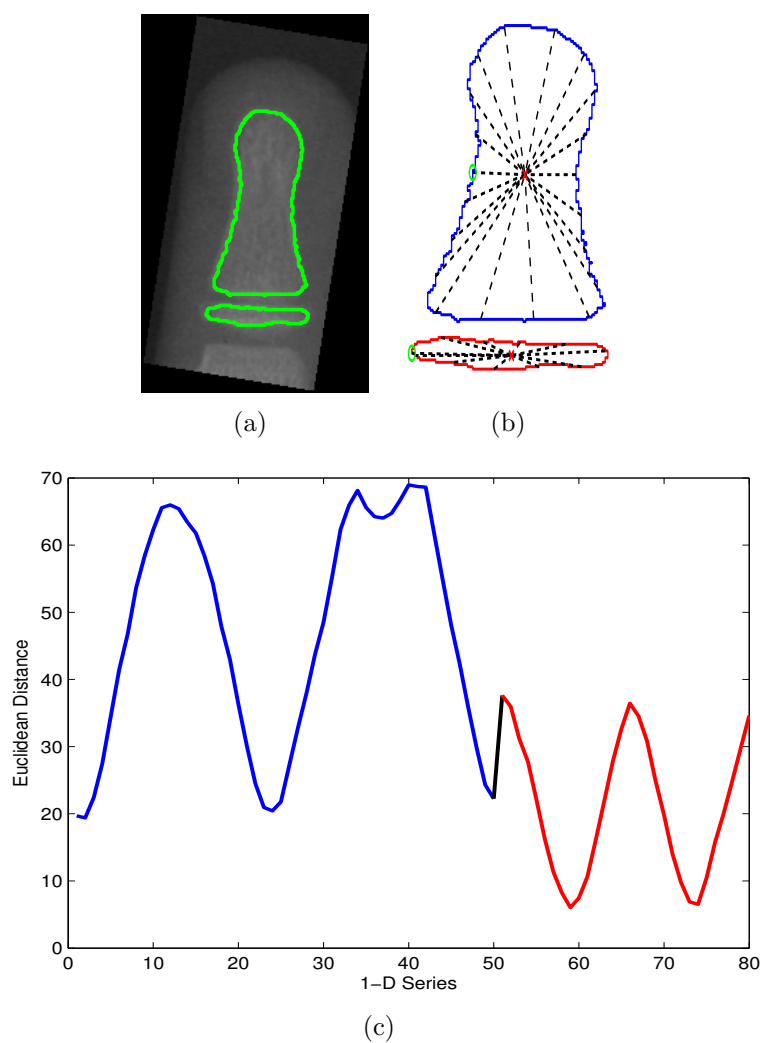


Figure 5.8: An example of a good bone segmentation from a ROI (a) being converted into a 1-D series (b) and (c).

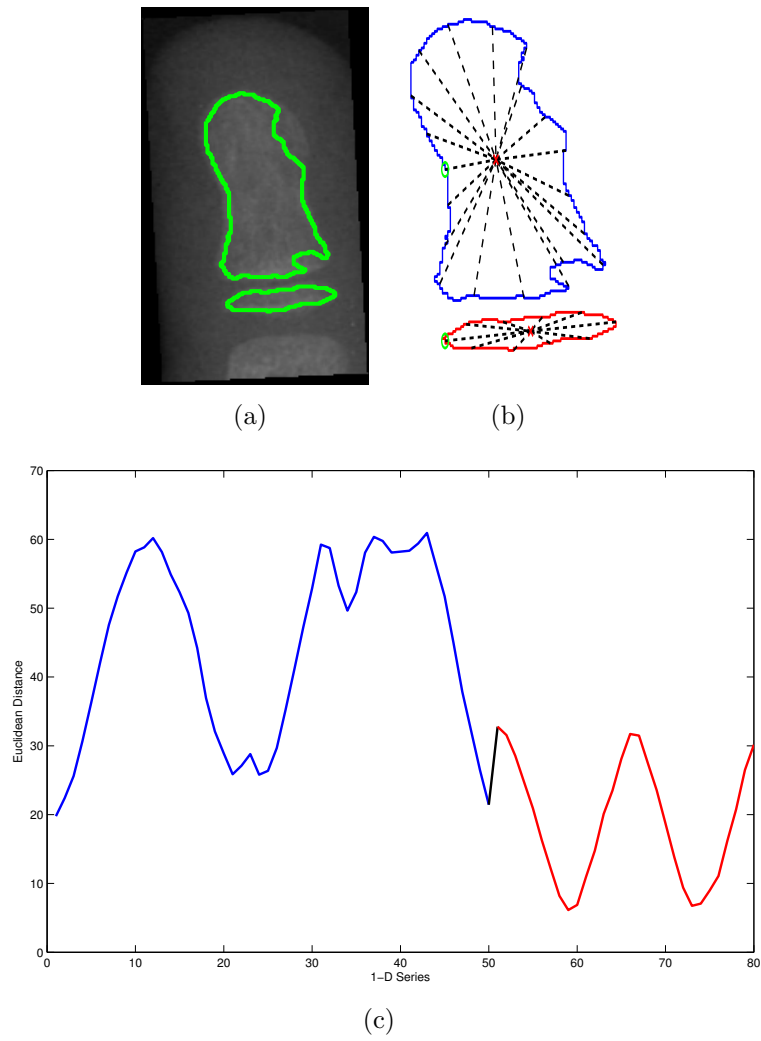


Figure 5.9: An example of a bad bone segmentation from a ROI (a) being converted into a 1-D series (b) and (c).

5.4 Results

There are three stages to our experimentation. Firstly, we evaluate classifiers on our training set of outlines and choose a subset of combinations of classifiers/representations to use in testing. Secondly, we assess the selected classifiers on the testing set of outlines. Finally, we manually assess the outlines and comment on the suitability of the classifiers.

In order to classify a segmented bone, each segmentation must first be manually labelled as correct or incorrect. Each set of segmentations of a certain phalanx was split into training and testing data. The number of instances for each is shown in Table 5.4.

Table 5.4: Number of instances for bone segmentation classification. In brackets is the number of good segmentations /number of bad segmentations.

	No. of Instances	No. of Training	No. of Testing
Distal	876	600(378/222)	276(161/115)
Middle	891	600(388/212)	291(166/125)
Proximal	891	600(406/194)	291(199/92)

5.4.1 Classifying Segmentations

In the training stage the training instances of each bone are classified using ten separate classifiers on the three segmentation representations. We conducted our classification experiments using the WEKA [HFH⁺09] implementation of k NN [FHJ52] (where k is set through cross validation), Naive Bayes [Lew98], C4.5 tree [Qui93], Support Vector Machines [CV95] with linear, quadratic and radial basis function kernels, Random Forest [Bre01] (with 30 and 100 trees), Rotation Forest [RKA06] and Multilayer Perceptron [MP69].

Table 5.5 shows the ten fold cross validation accuracy (on all three bones) of ten classifiers. From the results, it is observed that the two best performing represen-

tations are the extracted features and one-dimensional series. It is also noteworthy that the more complex classifiers (quadratic support vector machine, random forest and rotation forest) perform better with the extracted features representation than the one dimensional series, with the opposite being true of the simpler classifiers. The SVM with a quadratic kernel (SVMQ) achieves the best results with accuracies of 86.61% and 86.28% on the extracted features and one-dimensional series respectively.

Table 5.5: Overall cross-validation bone segmentation accuracy (%).

	Features	1-D Series	Intensity
k-NN	81.61	82.11	60.28
Naive Bayes	65.89	66.72	60.72
C4.5	80.22	80.17	59.44
SVML	83.50	76.72	62.33
SVMQ	86.61	86.28	58.61
SVMR	68.22	70.39	65.11
Random Forest (30)	84.11	83.50	66.72
Random Forest (100)	84.39	84.22	67.28
Rotation Forest (30)	85.17	83.44	64.39
Multilayer Perceptron	84.39	83.00	57.33

Tables 5.6–5.8 show the results of the ten-fold cross validation for the distal, middle and proximal phalanges respectively. From Table 5.6 we can see that the SVMQ performs the best, achieving an accuracy of 88.00% on the extracted features. On the majority of classifiers used, the extracted features performed best, followed by the one-dimensional series and then the intensity distributions. The only two classifiers where the one-dimensional series performs better than the extracted features are Naive Bayes and SVM with Radial Basis Function Kernel (SVMR). However, these are the two worst performing classifiers for both datasets.

The results of the middle phalanx (Table 5.7) indicate that the one-dimensional series is the optimal representation for this bone, although the extracted features representation also performs well. The SVMQ on the one-dimensional series is the best performing classifier/representation combination achieving an accuracy of 85.17%,

Table 5.6: Distal phalanx cross-validation bone segmentation accuracy (%).

	Features	1-D Series	Intensity
k-NN	82.17	80.83	58.67
Naive Bayes	72.33	73.83	62.00
C4.5	81.67	79.17	59.33
SVML	86.67	79.33	61.50
SVMQ	88.00	84.67	64.83
SVMR	72.33	76.00	63.00
Random Forest (30)	85.33	83.17	70.33
Random Forest (100)	85.67	83.50	71.83
Rotation Forest (30)	85.83	83.67	68.83
Multilayer Perceptron	86.00	82.17	50.00

with same classifier achieving an accuracy of 83.00% on the extracted features.

Table 5.7: Middle phalanx cross-validation bone segmentation accuracy (%).

	Features	1-D Series	Intensity
k-NN	79.17	79.83	58.83
Naive Bayes	61.67	61.33	61.00
C4.5	77.17	78.17	61.50
SVML	79.17	70.67	62.17
SVMQ	83.00	85.17	55.83
SVMR	64.67	64.67	64.67
Random Forest (30)	80.50	80.67	63.50
Random Forest (100)	80.50	82.17	64.67
Rotation Forest (30)	82.33	79.33	62.17
Multilayer Perceptron	80.67	81.67	61.00

The proximal phalanx results shown in Table 5.8 conform with the results shown from the middle phalanx, where the SVMQ is best performing classifier, achieving accuracies of 89.00% and 88.83% on the one-dimensional series and extracted features respectively. This is the highest accuracy of the individual bones, with the middle phalanx having the lowest.

Given the results shown in Tables 5.5–5.8, it is obvious that the best classifier to continue with is the SVMQ classifier. However, the choice of best representation

Table 5.8: Proximal phalanx cross-validation bone segmentation accuracy (%).

	Features	1-D Series	Intensity
k-NN	83.50	85.67	63.33
Naive Bayes	63.67	65.00	59.17
C4.5	81.83	83.17	57.50
SVML	84.67	80.17	63.33
SVMQ	88.83	89.00	55.17
SVMR	67.67	70.50	67.67
Random Forest (30)	86.50	86.67	63.33
Random Forest (100)	87.00	87.00	65.33
Rotation Forest (30)	87.33	87.33	62.17
Multilayer Perceptron	86.50	85.17	61.00

for the data is more difficult. The extracted features slightly outperform the one-dimensional series overall, but the one-dimensional series outperforms the features on the middle and proximal phalanges. Due to this, both representations of the data will be extracted on the test sets of images, where a further comparison can take place.

The intensity distribution representation performed poorly. One potential reason for this result is that the ROI may contain some hard tissue of the bone below/above. If this is the case, it means that any incorrect segmentation where the epiphysis is missing or the hard tissue is not completely segmented could potentially be classified as a correct segmentation due to the intensity distributions of both being similar.

5.4.2 Performance on Test Outlines

After the performance on the training data, the next stage of testing was to test the best performing SVMQ classifier on unseen test data. In order to do this the two best performing representations (extracted features and one-dimensional series) of the test segmentations were extracted. The number of instances in each testing set is shown in Table 5.4. The results of the SVMQ on the testing data are shown in Table 5.9, as we would expect there is a slight decrease in the overall accuracy

in comparison to the training data. Interestingly, for the middle phalanx there is a slight increase in accuracy for the extracted features and the same for the proximal phalanx on the one-dimensional series. Also of interest is that the one-dimensional series only outperforms the extracted features on the proximal phalanx. This indicates that different representations may be better suited for certain bones. It would appear that the most consistent representation is the extracted features though with all accuracies in the range of 84 – 86%.

Table 5.9: Bone segmentation accuracy of SVMQ on unseen data (%).

	Features	1-D Series
Distal	84.42	74.64
Middle	85.91	82.82
Proximal	84.54	90.03
Overall	84.97	82.63

5.4.3 Manual Assessment of Bone Segmentation Classification

The final stage of testing is to manually assess the performance of the SVMQ on the previously unseen test data. The confusion matrices for each bone using the extracted features are shown in Table 5.10, and, the one-dimensional series representation in. Table 5.11. The confusion matrices show that the classifier performs to a similar level for each phalanx on the extracted features.

Table 5.10: Confusion matrices of the SVMQ classifier on the features dataset.

(a) Distal				(b) Middle				(c) Proximal			
		Actual				Actual				Actual	
Classified		1	0	Classified		1	0	Classified		1	0
		146	28			153	28			192	38
	0	15	87		0	13	97		0	7	54

Table 5.11: Confusion matrices of the SVMQ classifier on the one-dimensional series dataset.

(a) Distal				(b) Middle				(c) Proximal			
		Actual				Actual				Actual	
Classified	1	135	44	Classified	1	142	26	Classified	1	192	22
	0	26	71		0	24	99		0	7	70

The main differences between the two representations are on the distal and proximal phalanges, where the extracted features are better than the one-dimensional series on the distal and vice-versa for the proximal. Again, this indicates that perhaps different representations are better for different bones. However, at this stage of development of ASMA it would be preferable to have one method for bone segmentation classification and that the feature extraction representation is the best for this purpose as it is more consistent than the one-dimensional series representation.

5.5 Conclusions

This chapter discussed stages C, D and E of ASMA. We have presented a novel algorithm for segmenting bones, discussed features importance in bone age assessment, described methods used to extract features and specify a classification scheme to automatically detect whether a segmentation is correct. The main findings of the chapter are:

- the best performing classifier was the SVM with a quadratic kernel; and
- that of the three representations used, the extracted features and one-dimensional series perform the best. However, the extracted features representation slightly outperforms the one-dimensional series overall in both cross-validation and on unseen test data, therefore is the choice for stage E of ASMA.

Chapter 6

ASMA Stage F (Part One): Classification of Tanner-Whitehouse Stages

The work in this chapter is an extended version of research published in [DTB12].

In this chapter we cover the first part of stage F of ASMA (See Figure 1.1). The final stage of any automated bone age assessment system is the ability to estimate bone age. There are three potential ways for this to be done: Firstly, classify each bone according to the TW stages, calculate the SMS and hence the bone age; secondly, classify according to the closest GP standard; and finally, regress onto chronological age. ASMA performs the first and third of these tasks. This chapter describes how ASMA classifies the TW stages of the three bones segmented by the method outlined in Chapter 5. Specifically, addressing the following questions:

1. What are the most important features for assessing skeletal maturity? (Section 6.1.1)
2. Are different features more/less important throughout the skeletal development process? (Section 6.1.2)

3. Do features interact with each other to better predict Tanner-Whitehouse stages? (Section 6.1.3)
4. Are there simple rules for skeletal maturity assessment? (Section 6.1.3)
5. Can we build classification systems in order to recreate the TW system? (Section 6.2)

The first four questions are addressed through an exploratory analysis of the features (Section 6.1). This involves using the information gain metric to find how discriminatory each feature is. This is done, firstly, across all TW stages, and secondly, on individual TW stages. The interaction of features is explored through a C4.5 tree classifier built on the segmentations which are used to derive simple rules for TW stages.

To answer question five, we perform an experimental evaluation of different classifiers. The same classifiers are used as those in Chapters 4 and 5. A ten fold cross validation classification experiment on the training data for each of the bones is performed in order to find the best classifiers. Once the best classifiers are found, we then evaluate the classifiers performance on the test data. We believe that by building ASMA in such a way, clinicians will have a greater understanding of how a decision was made by this system in comparison to systems built on ASMs or AAMs [NvM⁺03, TKJP09]. Thus they will have greater confidence in the resulting estimates.

6.1 Exploratory Analysis of Features

A benefit of using a feature based approach to ABAA is it allows an exploratory analysis of how the classifications are formed. All of the bone segmentations used in Chapter 5 that were labelled as correct are used in this analysis. For each individual bone, all segmentations were assigned a TW stage (D–I). The number of instances for each bone is shown in Table 6.1.

Table 6.1: Number of instances for exploratory analysis of features.

	No. of Instances
Distal	539
Middle	554
Proximal	605

6.1.1 Overall Information Gain of Features

The first stage of our exploratory analysis is to measure the importance of each feature over the whole skeletal development process. To do this we calculate the information gain for each feature \mathbf{f}_i . For discrete valued attributes e.g. epiphysis presence, this is straight forward. However, the other 24 features that are extracted are continuous. A standard way of calculating the information gain is to discretise the value on some split point s so that $\mathbf{f}_i \leq s = 0$ and $\mathbf{f}_i > s = 1$. To represent each feature fairly, the maximal information gain g^* of each feature is used:

$$g^* = \max_{s \in \mathbf{s}}(\text{info gain}(\mathbf{f}_i, s)), \quad (6.1)$$

where s is a split point in the set of all possible split points \mathbf{s} , and information gain is calculated using Equation 3.19. In order to calculate g^* , a vector of length n that models feature \mathbf{f}_i , is taken as input and sorted. This gives an orderline \mathbf{o} . After \mathbf{o} has been calculated, the set of split points \mathbf{s} can be calculated. Each split point is calculated as:

$$\mathbf{s}_j = \frac{\mathbf{o}_j + \mathbf{o}_{j+1}}{2}, \quad (6.2)$$

where $j = 1, \dots, n - 1$. The information gain g for each split point is then calculated and the maximal information gain g^* is found.

Tables 6.2, 6.3 and 6.4 show the ranking of features based on the information gain metric for the distal, middle and proximal phalanges respectively. In each of

the tables, column 1 shows the ranking of the 15 phalanx based features, and column 2 shows the top 15 features when the epiphysis is present.

6.1.1.1 Distal Phalange

Table 6.2 shows the results for the distal phalange. From the results, it is observed that the most important feature is the presence of the epiphysis. By investigating the TW descriptions (see Table 2.2) it is clear that the TW stages can be broken into two groups by this feature. The next two most important features (features 8, and 15), both relate to the metaphysis width. Again, through the skeletal development shown in Table 2.2, we can see that the metaphysis does alter over time, especially once fusion of the epiphysis has commenced. It is also noticeable how this change in the metaphysis relates to the change of width in the centre of the phalanx over the development period. Interestingly, it would seem that the higher on the phalanx the width is taken, the less discriminatory that width is. This indicates that even when the epiphysis is not present the majority of development is in the lower half of the phalanx.

When the epiphysis is present for the distal phalanx, it is obvious that the features relating to this are the most discriminatory, with nine out of ten of the features appearing in the top 15 ranks. Features that relate to the width of the epiphysis (features 17, 19, and 25) are the most important, if we look at the TW stages where the epiphysis is present, it is clear that the epiphysis gets wider over the development process. The only feature extracted from the epiphyseal area not to feature in the top 15 features when the epiphysis is present is the epiphysis height. As there is not as much difference in the height observed in the TW stages as there is in the width. The features relating to phalanx height (features 2, and 4) achieve similar rankings in both columns, whilst the features relating to the metaphysis width (features 8, and 15) are less discriminatory when the epiphysis is present. The reason for this is that when the epiphysis is present, the measure between the metaphysis width and the epiphysis width (feature 25) is more discriminatory.

Table 6.2: Features ranked by information gain for distal phalange III. The first column shows the ranks of the phalanx features on all images. The second column shows the top 15 ranks of all features on images where the epiphysis is present. The information gain is given in brackets.

Phalanx feature ranks	Epiphysis+Phalanx feature ranks
1 (0.9589)	19 (0.5716)
8 (0.8639)	17 (0.5133)
15 (0.7772)	25 (0.4858)
2 (0.7163)	4 (0.4558)
4 (0.7091)	2 (0.4367)
14 (0.5437)	21 (0.4342)
7 (0.5277)	12 (0.3123)
9 (0.4804)	23 (0.3045)
12 (0.4660)	3 (0.2565)
10 (0.4595)	8 (0.2422)
11 (0.2817)	20 (0.2379)
3 (0.2526)	24 (0.2350)
5 (0.2076)	7 (0.2343)
6 (0.1518)	22 (0.2066)
13 (0.0424)	16 (0.2039)

6.1.1.2 Middle Phalange

The results on the middle phalange are shown in Table 6.3. These results indicate that as with the distal phalange, the presence of the epiphysis (feature 1) is the most discriminatory feature. Interestingly, the features relating to the phalanx height (features 2, and 4) are the next most important, as these features were not expected to perform well due to the size of the radiograph not being relative to the size of the patients hand. The features relating to the metaphysis width (features 8, and 15) are less discriminatory than for the distal phalanx. One possible reason for this is that the middle phalange is more tubular than the distal phalange, and hence there is less variation throughout development (feature 8) as well as when in comparison to the width of the phalanx (feature 15). This is confirmed by the example images used for each TW stage [TWH⁺01], shown in Tables 2.2 and 5.1. The results on the middle phalange show that width features become less discriminatory the higher they are

on the phalanx e.g. phalanx width is less discriminatory than metaphysis width. All of the phalanx based width to width ratios (features 13, 14, and 15) perform poorly, with all three in the bottom four ranks of the phalanx based features. As with the metaphysis features this is probably due to the fact that the middle phalanx is a more tubular bone.

The same five features are the best performing when the epiphysis is present for the middle phalange as for the distal phalange, although they appear in a slightly different order with epiphysis to metaphysis width being the most discriminatory. Only seven of the ten epiphysis based features appear in the second column of Table 6.3. As with the distal phalange, the epiphysis height is not in the top 15 most discriminatory features. The other two epiphysis based features are the epiphysis eccentricity, and height to width ratio (features 20, and 22). Again, this is probably due to the change in epiphysis height over skeletal development being less than the change seen in epiphysis width.

Table 6.3: Features ranked by information gain for middle phalange III. The first column shows the ranks of the phalanx features on all images. The second column shows the top 15 ranks of all features on images where the epiphysis is present. The information gain is given in brackets.

Phalanx feature ranks	Epiphysis+Phalanx feature ranks
1 (0.9058)	25 (0.6600)
4 (0.8247)	19 (0.6579)
2 (0.8239)	17 (0.6223)
11 (0.5905)	2 (0.4715)
12 (0.5897)	4 (0.4629)
8 (0.5731)	21 (0.4535)
9 (0.5364)	11 (0.4504)
10 (0.4401)	12 (0.3512)
7 (0.3981)	9 (0.2931)
5 (0.3134)	10 (0.2743)
3 (0.3078)	8 (0.2648)
15 (0.2383)	7 (0.2598)
6 (0.2163)	23 (0.2539)
13 (0.1597)	24 (0.2424)
14 (0.0138)	16 (0.2411)

6.1.1.3 Proximal Phalange

The results showing the maximal information gain for each feature on the proximal phalange are shown in Table 6.4. The proximal phalange like the middle phalange is more tubular than the distal. Hence, the best performing features shown in column one of Table 6.4 are the presence of the epiphysis (feature 1), and features relating to phalanx height (features 2, and 4). The metaphysis width (feature 8) is the fourth most discriminative feature. This is a higher ranking than on the middle phalange but lower than the distal phalange. If we look at the images and TW stages relating to this bone (shown in Table 5.2) and compare them to the middle phalange stages (shown in Table 5.1), it can be observed that although the proximal phalange is more tubular than the distal, the change in metaphysis width is more apparent than with the middle phalange.

As with the distal and middle phalange, the same five features are the most discriminatory when the epiphysis is present. However, only six of the epiphysis based features appear in column two. With the features based on epiphysis height (features 16, and 18), and the measures of roundness of the epiphysis (features 20, and 23) not included in the top 15. A possible reason for the roundness measures not appearing is due to the fact that the change in shape of the epiphysis is not as prevalent on the proximal phalange as it is on the distal phalange (see Table 2.2).

6.1.1.4 Overall

Across all three of the bones investigated, various similarities in the rankings of the features have been observed. For all of the bones used, the most important feature is the presence of the epiphysis. This splits the TW stages into two distinct groups. Features measuring phalanx height (features 2, and 4) are important, but become less so when the epiphysis is present, when features that relate to the width of the epiphysis (features 17, and 19) become more discriminatory. The features that measure the metaphysis width (features 8, and 15) are found to be more important for

Table 6.4: Features ranked by information gain for proximal phalange III. The first column shows the ranks of the phalanx features on all images. The second column shows the top 15 ranks of all features on images where the epiphysis is present. The information gain is given in brackets.

Phalanx only feature ranks	Epiphysis+Phalanx feature ranks
1 (0.9402)	25 (0.5718)
2 (0.8313)	19 (0.5191)
4 (0.8303)	17 (0.4797)
8 (0.6390)	2 (0.3709)
11 (0.5519)	4 (0.3628)
12 (0.4971)	21 (0.3511)
15 (0.4724)	24 (0.3112)
9 (0.4719)	11 (0.2899)
10 (0.4707)	12 (0.2301)
7 (0.3379)	9 (0.2114)
3 (0.2778)	8 (0.1965)
14 (0.2603)	3 (0.1792)
5 (0.1980)	10 (0.1734)
6 (0.1623)	7 (0.1609)
13 (0.0641)	22 (0.1518)

the distal phalange than the middle or proximal because the middle and proximal phalanx are more tubular than the distal phalange. In general, the worst performing phalanx based features tend to be those relating to the width of the top half of the phalanx (features 3, 5, and 6). As well as the first-quartile, third-quartile and metaphysis width to phalanx width ratios (features 13, 14, and 15) (with the exception of the metaphysis to width ratio (feature 15) on the distal phalange). When the epiphysis is present, the features that relate to this tend to be the more discriminatory. The worst performing epiphysis based features are those related to the height of the epiphysis (features 16, and 18). This is due to the fact that there is not as much difference in these features over the development process as there are in others e.g. epiphysis width.

6.1.2 Information Gain based on Tanner-Whitehouse Stage

The second stage of our exploratory analysis is to calculate the information gain for each of the features for each individual TW stage.

The results displayed in Table 6.5 are for the distal phalange. With the phalanx height (TW stages D and E), epiphysis width (TW stages F, G and H), and, presence of the epiphysis being the three most important features for individual stages. Overall the worst three performing features are phalanx width, phalanx first quartile width and phalanx first quartile to width ratio.

Table 6.6 shows the results for middle phalange. For TW stages D, E and H, the features that measure phalanx height are found to be the most discriminatory. Features that relate to epiphysis width are the most discriminatory for TW stages F, and G, and for TW stage I the presence of the epiphysis is the most important feature. The phalanx third quartile to phalanx width ratio is the worst performing feature overall, with all of the phalanx width ratio based features performing badly.

The results for the proximal phalange are shown in Table 6.7. The most important features for TW stages D, E and H are again those that relate to phalanx height. For TW stages F, and G the most important feature is the epiphysis to metaphysis ratio. The presence of the epiphysis is the most discriminatory feature for TW stage I. The worst performing feature overall is the phalanx first quartile to width ratio.

Across all of the phalanges certain trends become clear. On the majority of TW stages the most discriminatory features are the same for all of the phalanges. This is especially true for TW stages F and G. For TW stages where this is not the case, the top ranked features are usually the same on the more tubular (middle and proximal) bones. Generally the worst performing features have been those based on phalanx widths, with the exception of metaphysis width. The epiphysis based features have been found to be less discriminatory in the earlier stages, become more discriminatory in the middle stages of development and then become less important in the latter stages, as would be expected.

Table 6.5: The information gain of every feature for each Tanner-Whitehouse stage of distal phalange III. The rank of each feature is given in brackets, where if two or more features are equally discriminatory, they are given the mean rank.

Feature Number	Stage D	Stage E	Stage F	Stage G	Stage H	Stage I
1) Epiphysis	0.0894 (17.5)	0.0894 (15)	0.2603 (10)	0.0813 (11)	0.0072 (24)	0.9116 (1)
2) Phalanx Ellipse Height	0.2292 (2)	0.1366 (2)	0.2158 (14)	0.0322 (15)	0.0239 (10)	0.6106 (5)
3) Phalanx Ellipse Width	0.1548 (7)	0.0438 (22)	0.0249 (23)	0.0134 (20.5)	0.0135 (16)	0.1644 (12)
4) Phalanx Height	0.2394 (1)	0.1388 (1)	0.2075 (15)	0.0304 (16)	0.0249 (8)	0.6111 (4)
5) Phalanx Width	0.1408 (8)	0.0183 (24)	0.0111 (25)	0.0071 (25)	0.0165 (13)	0.1053 (13)
6) Phalanx First Quartile Width	0.0923 (11)	0.0276 (23)	0.0168 (24)	0.0107 (22)	0.0180 (12)	0.0787 (14)
7) Phalanx Third Quartile Width	0.1610 (6)	0.0854 (16)	0.1517 (18)	0.0247 (17)	0.0090 (22)	0.4837 (7)
8) Metaphysis Width	0.1624 (5)	0.0993 (3)	0.2429 (13)	0.0545 (14)	0.0103 (19)	0.8253 (2)
9) Phalanx Eccentricity	0.1001 (9)	0.0698 (18)	0.1606 (17)	0.0134 (20.5)	0.0075 (23)	0.4219 (8)
10) Phalanx Width to Height Ratio	0.0959 (10)	0.0674 (19)	0.1310 (19)	0.0093 (24)	0.0100 (20)	0.4047 (9)
11) Phalanx Roundness	0.0716 (22)	0.0522 (20)	0.0721 (21)	0.0095 (23)	0.0115 (18)	0.2391 (11)
12) Phalanx Area to Perimeter Ratio	0.1932 (3)	0.0902 (11)	0.1053 (20)	0.0228 (18)	0.0139 (15)	0.3666 (10)
13) Phalanx First Quartile to Width Ratio	0.0075 (25)	0.0161 (25)	0.0270 (22)	0.0157 (19)	0.0062 (25)	0.0182 (15)
14) Phalanx Third Quartile to Width Ratio	0.0271 (24)	0.0489 (21)	0.1879 (16)	0.0616 (13)	0.0202 (11)	0.5238 (6)
15) Phalanx Metaphysis to Width Ratio	0.0701 (23)	0.0729 (17)	0.2465 (12)	0.0637 (12)	0.0301 (7)	0.7695 (3)
16) Epiphysis Ellipse Height	0.0894 (17.5)	0.0940 (5)	0.2634 (8)	0.1118 (5)	0.0308 (6)	N/A
17) Epiphysis Ellipse Width	0.0894 (17.5)	0.0919 (8)	0.3326 (3)	0.1417 (3)	0.0412 (2)	N/A
18) Epiphysis Height	0.0899 (12.5)	0.0904 (10)	0.2603 (10)	0.0963 (9)	0.0248 (9)	N/A
19) Epiphysis Width	0.0894 (17.5)	0.0930 (6.5)	0.3640 (1)	0.1477 (1)	0.0441 (1)	N/A
20) Epiphysis Eccentricity	0.0894 (17.5)	0.0899 (13)	0.2697 (7)	0.1010 (8)	0.0164 (14)	N/A
21) Epiphysis Distance to Phalanx	0.0894 (17.5)	0.0930 (6.5)	0.3102 (4)	0.1447 (2)	0.0367 (5)	N/A
22) Epiphysis Width to Height Ratio	0.0894 (17.5)	0.0899 (13)	0.2744 (5)	0.1021 (7)	0.0125 (17)	N/A
23) Epiphysis Roundness	0.1866 (4)	0.0899 (13)	0.2603 (10)	0.0845 (10)	0.0095 (21)	N/A
24) Epiphysis Area to Perimeter Ratio	0.0899 (12.5)	0.0909 (9)	0.2719 (6)	0.1247 (4)	0.0408 (3)	N/A
25) Epiphysis Width to Metaphysis Ratio	0.0894 (17.5)	0.0950 (4)	0.3424 (2)	0.1105 (6)	0.0374 (4)	N/A

Table 6.6: The information gain of every feature for each Tanner-Whitehouse stage of middle phalange III. The rank of each feature is given in brackets, where if two or more features are equally discriminatory, they are given the mean rank.

Feature Number	Stage D	Stage E	Stage F	Stage G	Stage H	Stage I
1) Epiphysis	0.1202 (18.5)	0.1984 (12.5)	0.0934 (11)	0.0561 (10.5)	0.0012 (25)	0.7975 (1)
2) Phalanx Ellipse Height	0.2948 (2)	0.2619 (2)	0.0570 (13)	0.0361 (14)	0.0580 (1)	0.6031 (3)
3) Phalanx Ellipse Width	0.1741 (9)	0.0992 (19)	0.0174 (21)	0.0108 (22)	0.0259 (13)	0.1726 (11)
4) Phalanx Height	0.3011 (1)	0.2631 (1)	0.0575 (12)	0.0378 (13)	0.0577 (2)	0.6151 (2)
5) Phalanx Width	0.1468 (12)	0.0957 (21)	0.0114 (23)	0.0104 (23)	0.0241 (14)	0.1661 (12)
6) Phalanx First Quartile Width	0.1662 (10)	0.0524 (23)	0.0107 (24)	0.0192 (18)	0.0087 (21)	0.0753 (14)
7) Phalanx Third Quartile Width	0.1835 (8)	0.0988 (20)	0.0127 (22)	0.0095 (24)	0.0293 (12)	0.2214 (10)
8) Metaphysis Width	0.2079 (5)	0.1675 (16)	0.0365 (17)	0.0164 (20)	0.0239 (15)	0.4561 (4)
9) Phalanx Eccentricity	0.2062 (6)	0.1599 (17)	0.0454 (15)	0.0265 (16)	0.0429 (4)	0.3765 (5)
10) Phalanx Width to Height Ratio	0.1946 (7)	0.1121 (18)	0.0348 (18)	0.0195 (17)	0.0337 (11)	0.3207 (8)
11) Phalanx Roundness	0.2578 (3)	0.1922 (14)	0.0383 (16)	0.0467 (12)	0.0411 (6)	0.3702 (6)
12) Phalanx Area to Perimeter Ratio	0.2438 (4)	0.1866 (15)	0.0198 (20)	0.0175 (19)	0.0387 (8)	0.3699 (7)
13) Phalanx First Quartile to Width Ratio	0.0179 (24)	0.0552 (22)	0.0507 (14)	0.0111 (21)	0.0149 (19)	0.1308 (13)
14) Phalanx Third Quartile to Width Ratio	0.0077 (25)	0.0073 (25)	0.0041 (25)	0.0083 (25)	0.0020 (24)	0.0085 (15)
15) Phalanx Metaphysis to Width Ratio	0.0380 (23)	0.0423 (24)	0.0243 (19)	0.0293 (15)	0.0171 (18)	0.2354 (9)
16) Epiphysis Ellipse Height	0.1202 (18.5)	0.2085 (6.5)	0.1345 (6)	0.0725 (7)	0.0093 (20)	N/A
17) Epiphysis Ellipse Width	0.1202 (18.5)	0.2047 (11)	0.1858 (3)	0.2115 (1)	0.0396 (7)	N/A
18) Epiphysis Height	0.1202 (18.5)	0.2072 (8)	0.1125 (7)	0.0685 (8)	0.0059 (23)	N/A
19) Epiphysis Width	0.1202 (18.5)	0.2085 (6.5)	0.1954 (2)	0.1965 (3)	0.0383 (9)	N/A
20) Epiphysis Eccentricity	0.1210 (13.5)	0.1984 (12.5)	0.0992 (9)	0.0769 (6)	0.0237 (16)	N/A
21) Epiphysis Distance to Phalanx	0.1210 (13.5)	0.2230 (4)	0.1688 (4)	0.1356 (4)	0.0420 (5)	N/A
22) Epiphysis Width to Height Ratio	0.1202 (18.5)	0.2060 (9.5)	0.0957 (10)	0.0625 (9)	0.0189 (17)	N/A
23) Epiphysis Roundness	0.1626 (11)	0.2277 (3)	0.1034 (8)	0.0561 (10.5)	0.0339 (10)	N/A
24) Epiphysis Area to Perimeter Ratio	0.1202 (18.5)	0.2060 (9.5)	0.1353 (5)	0.0780 (5)	0.0064 (22)	N/A
25) Epiphysis Width to Metaphysis Ratio	0.1202 (18.5)	0.2163 (5)	0.2286 (1)	0.1966 (2)	0.0486 (3)	N/A

Table 6.7: The information gain of every feature for each Tanner-Whitehouse stage of proximal phalange III. The rank of each feature is given in brackets, where if two or more features are equally discriminatory, they are given the mean rank.

Feature Number	Stage D	Stage E	Stage F	Stage G	Stage H	Stage I
1) Epiphysis	0.0258 (21.5)	0.3154 (13)	0.1554 (11)	0.0436 (11)	0.0017 (25)	0.8286 (1)
2) Phalanx Ellipse Height	0.0896 (1)	0.4028 (1)	0.1292 (13)	0.0243 (12)	0.0402 (1.5)	0.6720 (2)
3) Phalanx Ellipse Width	0.0624 (5)	0.1699 (21)	0.0254 (22)	0.0092 (22)	0.0250 (12)	0.1878 (12)
4) Phalanx Height	0.0858 (2)	0.4005 (2)	0.1363 (12)	0.0242 (13)	0.0402 (1.5)	0.6688 (3)
5) Phalanx Width	0.0440 (10)	0.1170 (23)	0.0175 (23)	0.0074 (23)	0.0175 (19)	0.1278 (13)
6) Phalanx First Quartile Width	0.0601 (7)	0.0991 (24)	0.0080 (25)	0.0066 (24)	0.0150 (21)	0.0823 (14)
7) Phalanx Third Quartile Width	0.0505 (9)	0.2050 (20)	0.0285 (21)	0.0139 (19)	0.0273 (9)	0.2223 (10)
8) Metaphysis Width	0.0616 (6)	0.3260 (5)	0.0752 (16)	0.0151 (18)	0.0299 (7)	0.4943 (4)
9) Phalanx Eccentricity	0.0567 (8)	0.2531 (18)	0.0740 (17)	0.0162 (17)	0.0259 (10)	0.3725 (7)
10) Phalanx Width to Height Ratio	0.0410 (12)	0.2639 (17)	0.0606 (18)	0.0135 (20)	0.0227 (15)	0.3675 (8)
11) Phalanx Roundness	0.0631 (4)	0.3042 (15)	0.0839 (15)	0.0203 (15)	0.0336 (5)	0.4195 (5)
12) Phalanx Area to Perimeter Ratio	0.0783 (3)	0.2902 (16)	0.0522 (19)	0.0175 (16)	0.0332 (6)	0.3656 (9)
13) Phalanx First Quartile to Width Ratio	0.0073 (25)	0.0319 (25)	0.0147 (24)	0.0055 (25)	0.0103 (24)	0.0486 (15)
14) Phalanx Third Quartile to Width Ratio	0.0140 (24)	0.1503 (22)	0.0370 (20)	0.0123 (21)	0.0253 (11)	0.1972 (11)
15) Phalanx Metaphysis to Width Ratio	0.0369 (13)	0.2068 (19)	0.0907 (14)	0.0221 (14)	0.0104 (23)	0.4062 (6)
16) Epiphysis Ellipse Height	0.0260 (18)	0.3154 (13)	0.1595 (8)	0.0554 (8)	0.0163 (20)	N/A
17) Epiphysis Ellipse Width	0.0263 (16.5)	0.3280 (3.5)	0.2371 (3)	0.1325 (2)	0.0370 (3)	N/A
18) Epiphysis Height	0.0270 (15)	0.3154 (13)	0.1562 (10)	0.0484 (9)	0.0143 (22)	N/A
19) Epiphysis Width	0.0263 (16.5)	0.3280 (3.5)	0.2418 (2)	0.1280 (3)	0.0352 (4)	N/A
20) Epiphysis Eccentricity	0.0258 (21.5)	0.3208 (7.5)	0.1787 (5)	0.0581 (6)	0.0241 (13)	N/A
21) Epiphysis Distance to Phalanx	0.0259 (19)	0.3190 (10)	0.2292 (4)	0.1119 (4)	0.0274 (8)	N/A
22) Epiphysis Width to Height Ratio	0.0258 (21.5)	0.3208 (7.5)	0.1716 (6)	0.0571 (7)	0.0229 (14)	N/A
23) Epiphysis Roundness	0.0430 (11)	0.3226 (6)	0.1570 (9)	0.0447 (10)	0.0208 (17)	N/A
24) Epiphysis Area to Perimeter Ratio	0.0271 (14)	0.3190 (10)	0.1709 (7)	0.0867 (5)	0.0179 (18)	N/A
25) Epiphysis Width to Metaphysis Ratio	0.0258 (21.5)	0.3190 (10)	0.2702 (1)	0.1393 (1)	0.0216 (16)	N/A

6.1.3 Feature Interaction

The final stage of our exploratory analysis is to examine how the features interact to classify TW stages. This is performed by training a C4.5 tree on each phalange. These trees illustrate simple decision rules that encapsulate the description of how to determine TW stages for each phalange. The resulting trees can be seen in Figures 6.1–6.3, for the distal, middle and proximal phalanges respectively. At this stage of the analysis we are more interested in investigating the interaction of the features than the final classifier. Hence a more aggressive pruning technique has been used where the minimum number of instances for a class node is five.

The root node on all of the trees is the epiphysis presence feature (feature 1). The features which relate to the width of the epiphysis (features 17, 19 and 25), are all used on the next three levels of each tree, although in differing arrangements. The use of these three features is not surprising, given that they are the top performing features for discriminating each bone, when the phalanx and epiphysis are present. As shown by the results in both of the earlier stages of the exploratory analysis (Sections 6.1.1 and 6.1.2). All of the trees tend to handle the classification of stages similarly by splitting them into three groups. The distal phalange handles it slightly differently, by splitting the groups as follows: firstly TW stages D, and, E; secondly, TW stages E–G, and finally, TW stages F–H. Where as the middle and proximal split them into: TW stages D, and E; TW stages E, and F, and finally, TW stages F–H. This difference is due to the less tubular nature of the distal phalange causing different feature interactions to the other phalanges.

One of the main purposes of this part of the exploratory analysis is to extract simple decision rules from the C4.5 trees for TW classification. In order to make ASMA a more transparent system for clinicians to use. Tables 6.8–6.10 show the simplest rule for classifying each TW stage on the distal, middle and proximal phalanges respectively.

The rules for classifying TW stages D and E, are relatively simple for all of the phalanges, with all of them only needing three features. The rules also carry

the same pattern in that the differentiation between the two stages is made on the last feature. As would be expected, given the results in the first part of the exploratory analysis (see Section 6.1.1), the features used in the rules are almost all epiphysis based. For TW stages F, G, and H, the rules are generally a bit more complex than for the earlier TW stages. The features used to make up the rules are generally epiphysis based, although phalanx based features are used at the end of the majority of the rules for these TW stages. Finally, TW stage I has the simplest rules, for the middle and proximal phalanges it is just dependant on the epiphysis not being present. The distal phalange rule is slightly more complex in that it is also dependant on the metaphysis width being larger.

Two standard measures used to assess rules are Confidence [AIS93] and Coverage [MM95]. Confidence measures the accuracy of a rule, whereas Coverage measures the proportion of instances of that particular class a rule covers. These were applied to the rules shown in Tables 6.8–6.10. Generally, the best rules were those for TW stage I, with all rules achieving values for Confidence and Coverage greater than 0.9. For the Confidence measure, all rules performed well, with values greater than 0.67. TW stages E, F, G, and H, achieved values for Coverage of less than 0.3. However, this would be expected as there are more rules for these TW stages.

Table 6.8: Simple decision rules extracted for classifying Tanner-Whitehouse stages of distal phalange III. Rules extracted from Figure 6.1.

TW Stage	Rule
D	IF epiphysis is present AND epiphysis width ≤ 52 AND epiphysis width to metaphysis ratio ≤ 0.97 THEN STAGE D
E	IF epiphysis is present AND epiphysis width ≤ 52 AND epiphysis width to metaphysis ratio > 0.97 THEN STAGE E
F	IF epiphysis is present AND epiphysis width > 52 AND phalanx height > 129 AND phalanx first quartile width to width ratio ≤ 1.47 AND metaphysis width to width ratio ≤ 1.39 AND epiphysis width to metaphysis ratio ≤ 1.04 THEN STAGE F
G	IF epiphysis is present AND epiphysis width > 52 AND phalanx height > 129 AND phalanx first quartile width to width ratio > 1.47 THEN STAGE G
H	IF epiphysis is not present AND metaphysis width ≤ 1.78 AND phalanx first quartile width to width ratio > 1.24 THEN STAGE H
I	IF epiphysis is not present AND metaphysis width > 1.78 THEN STAGE I

Table 6.9: Simple decision rules extracted for classifying Tanner-Whitehouse stages of middle phalange III. Rules extracted from Figure 6.2.

TW Stage	Rule
D	IF epiphysis is present AND epiphysis width ≤ 54 AND phalanx roundness > 0.65 THEN STAGE D
E	IF epiphysis is present AND epiphysis width ≤ 78 AND phalanx roundness ≤ 0.65 THEN STAGE E
F	IF epiphysis is present AND epiphysis width > 78 AND epiphysis ellipse width ≤ 75.98 THEN STAGE F
G	IF epiphysis is present AND epiphysis width > 78 AND epiphysis ellipse width > 75.98 AND phalanx height > 200 AND epiphysis ellipse height > 24.18 THEN STAGE G
H	IF epiphysis is present AND epiphysis width > 78 AND epiphysis ellipse width > 75.98 AND phalanx height > 200 AND epiphysis ellipse height ≤ 24.18 AND epiphysis width to height ratio ≤ 7.45 AND epiphysis ellipse height ≤ 12.33 THEN STAGE H
I	IF epiphysis is not present THEN STAGE I

Table 6.10: Simple decision rules extracted for classifying Tanner-Whitehouse stages of proximal phalange III. Rules extracted from Figure 6.3.

TW Stage	Rule
D	IF epiphysis is present AND epiphysis width to metaphysis ratio ≤ 1.04 AND epiphysis ellipse width ≤ 51.99 THEN STAGE D
E	IF epiphysis is present AND epiphysis width to metaphysis ratio ≤ 1.04 AND $51.99 < \text{epiphysis ellipse width} \leq 67.98$ AND epiphysis height > 18 THEN STAGE E
F	IF epiphysis is present AND epiphysis width to metaphysis ratio ≤ 1.04 AND epiphysis ellipse width > 67.98 AND epiphysis width > 101 THEN STAGE F
G	IF epiphysis is present AND epiphysis width to metaphysis ratio > 1.04 AND epiphysis width ≤ 113 AND phalanx third quartile width to width ratio > 1.19 THEN STAGE G
H	IF epiphysis is present AND epiphysis width to metaphysis ratio > 1.04 AND epiphysis width > 113 AND phalanx width to height ratio ≤ 0.24 AND phalanx area to perimeter ratio ≤ 31.49 THEN STAGE H
I	IF epiphysis is not present THEN STAGE I

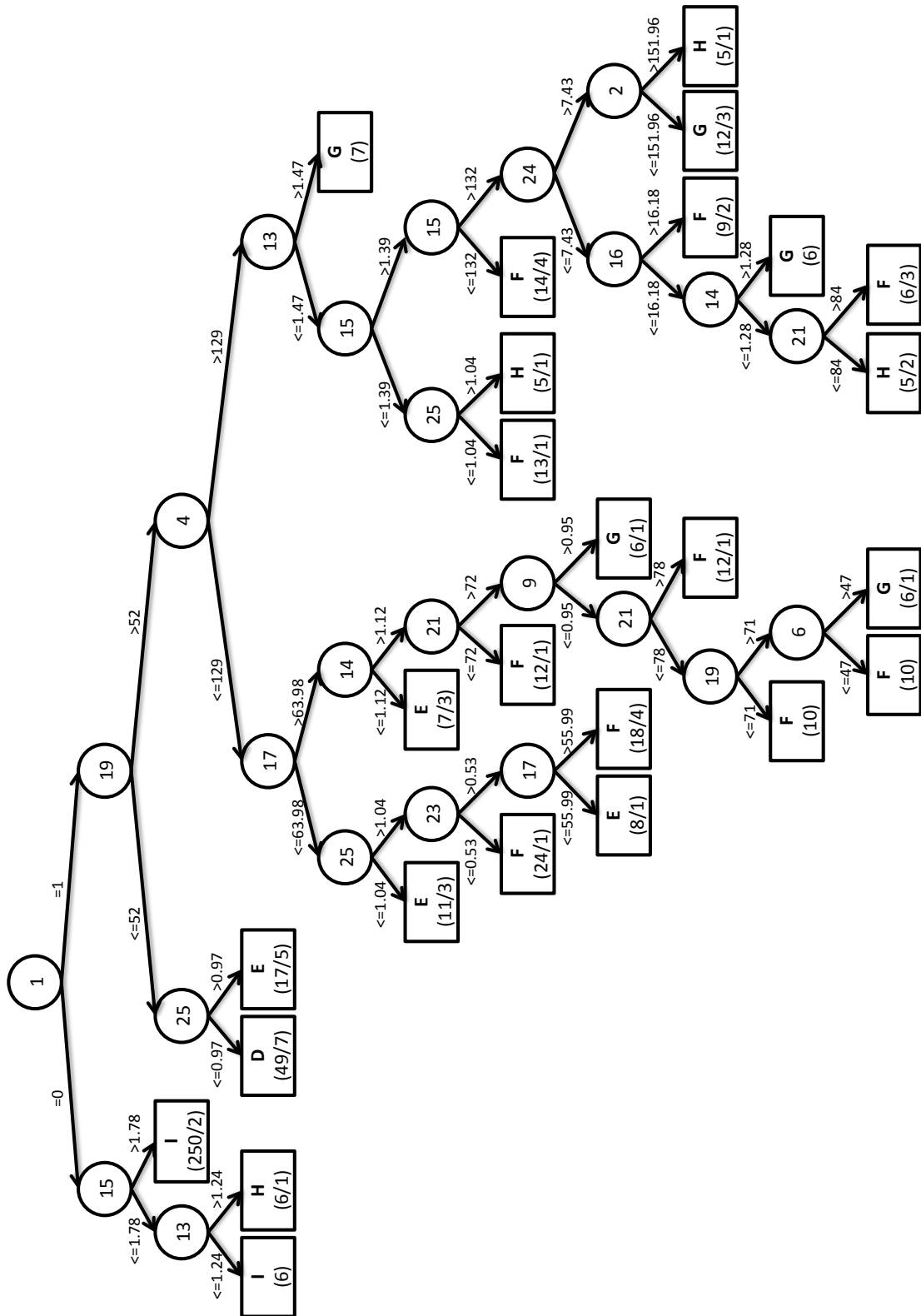


Figure 6.1: A C4.5 tree showing the interactions of the extracted features for Tanner-Whitehouse stage classification on distal phalange III.

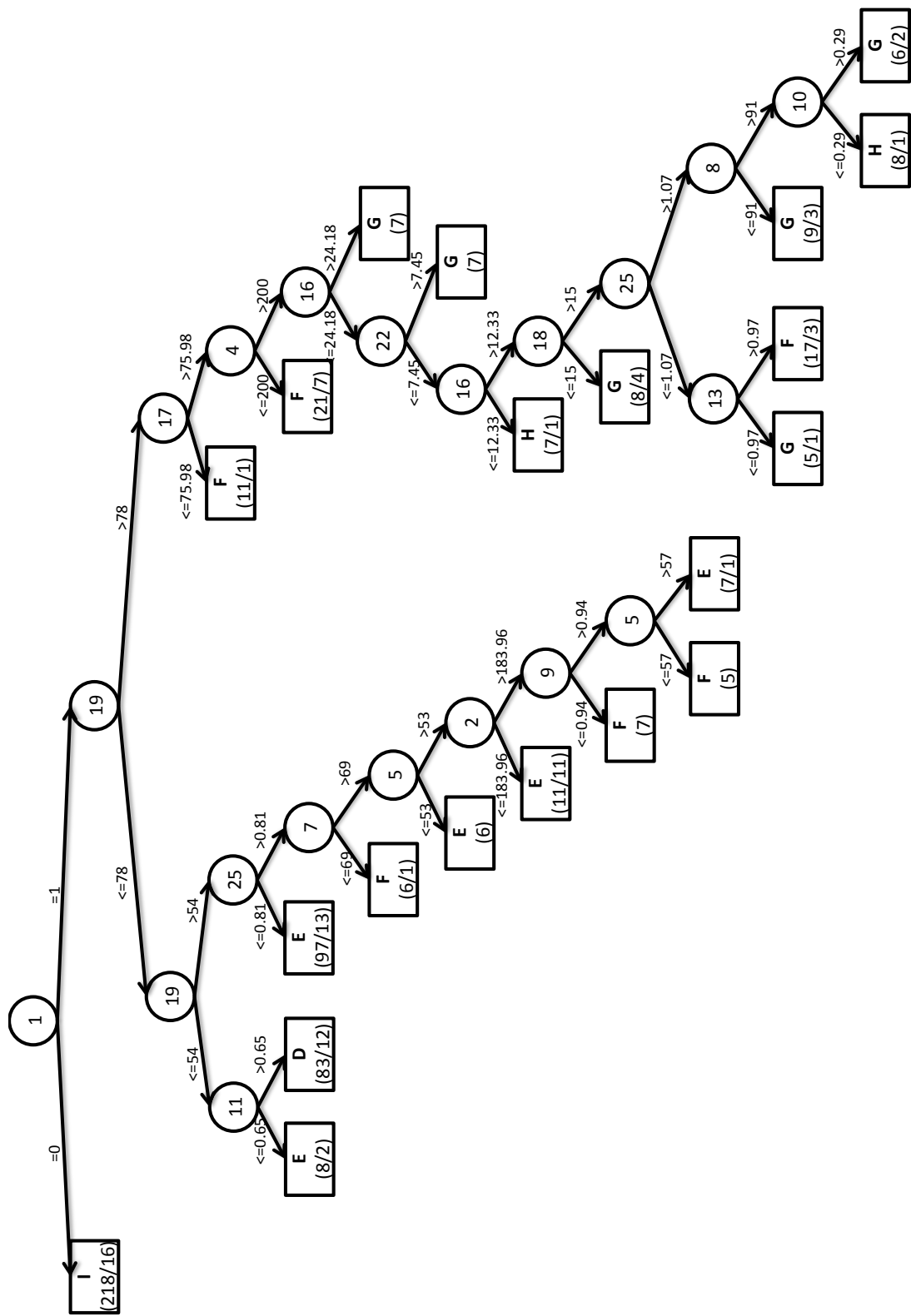


Figure 6.2: A C4.5 tree showing the interactions of the extracted features for Tanner-Whitehouse stage classification on middle phalange III.

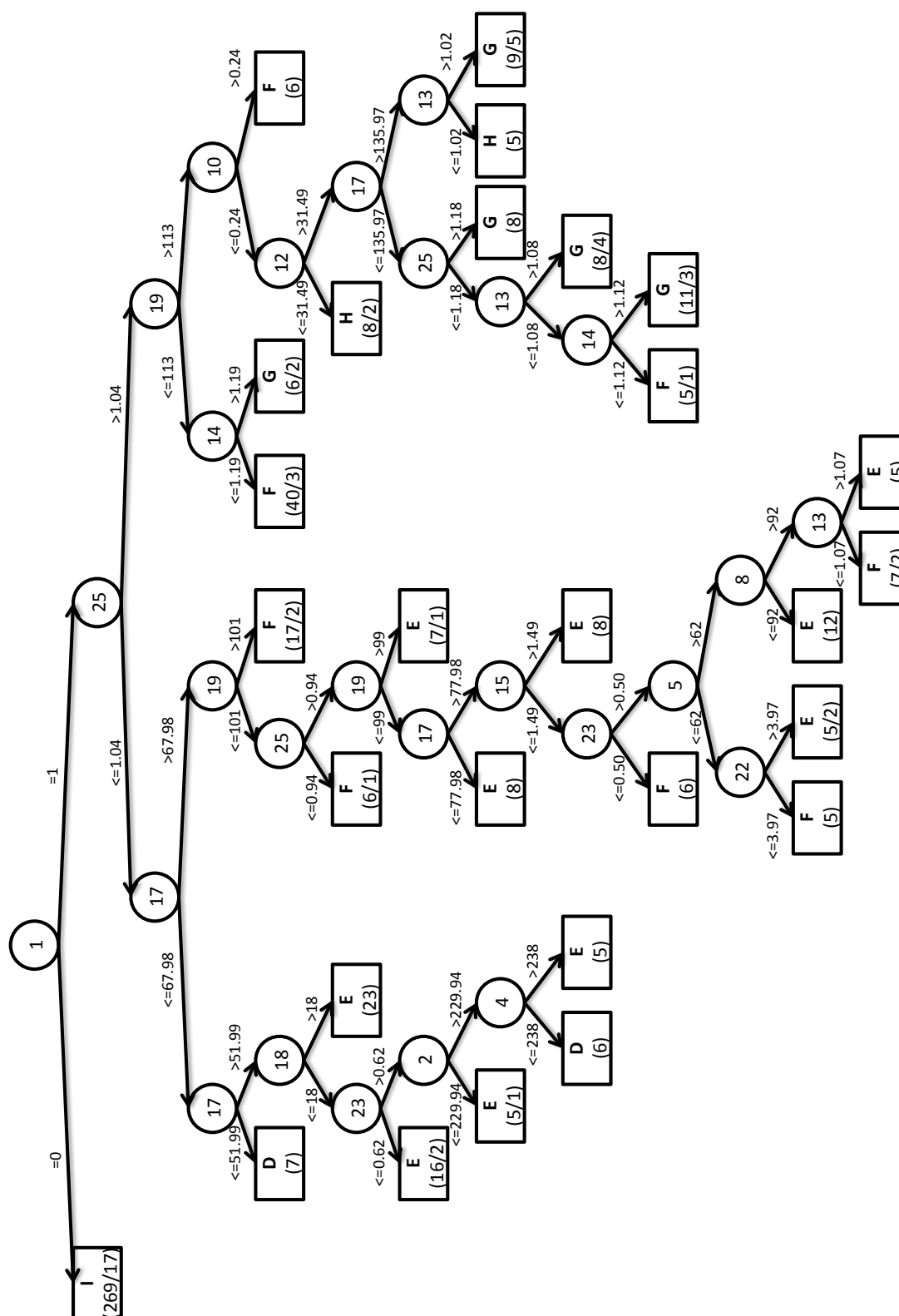


Figure 6.3: A C4.5 tree showing the interactions of the extracted features for Tanner-Whitehouse stage classification on proximal phalange III.

6.2 Classification of Tanner-Whitehouse Stages

After the exploratory analysis of features, the next task is the classification of Tanner-Whitehouse stages. There are three stages to our experimentation. Firstly, using the same classifiers as in Section 5.4, we conduct a ten fold cross validation classification experiment on the training data for each bone, in order to find the best performing classifiers to use in testing. Secondly, we evaluate the best performing classifiers on a previously unseen test set of features. Finally, we compare the results presented here to those of previous systems. Each set of segmentations of the three phalanges was split into training and testing data, the number of instances for each is shown in Table 6.11. To train each classifier we first produced a set of 400 bone segmentations for each phalanx of the middle finger. The remaining segmentations are used as testing data. As with previous publications [NvM⁺03, TKJP09] that have described ways of classifying TW stages we present the accuracy and within one stage accuracy.

Table 6.11: Number of instances for Tanner-Whitehouse classification.

	No. of Training	No. of Testing
Distal	400	139
Middle	400	154
Proximal	400	205

6.2.1 Cross-Validation of Tanner-Whitehouse Stages

The overall results from the ten-fold cross validation on the training data are shown in Table 6.12. The classifiers used are the WEKA [HFH⁺09] implementations of k NN [FHJ52] (where k is set through cross validation), Naive Bayes [Lew98], C4.5 tree [Qui93], Support Vector Machines [CV95] with linear, quadratic and radial basis function kernels, Random Forest [Bre01] (with 30 and 100 trees), Rotation Forest [RKA06] and Multilayer Perceptron [MP69]. As with the segmentation classification, the best performing classifier is the SVM with a quadratic kernel, which

correctly classifies over 80% correctly and gets nearly all bones within one TW category of the true category.

Table 6.12: Classification of overall Tanner-Whitehouse stage accuracy (%).

	Correct Stage	Within One Stage
k-NN	78.83	98.25
Naive Bayes	78.67	99.25
C4.5	76.00	97.58
SVML	80.83	98.17
SVMQ	82.58	99.25
SVMR	69.00	90.50
Random Forest (30)	81.08	98.67
Random Forest (100)	82.00	98.75
Rotation Forest (30)	80.42	98.67
Multilayer Perceptron	79.92	98.92

The results of the cross-validation on the individual phalanges are shown in Tables 6.13–6.15. The majority of the classifiers used achieve an accuracy in the region of 80–85% for the distal and proximal phalanges. The accuracies for the middle phalange are lower, with most classifiers achieving an accuracy in the range of 75–78%. The within one stage accuracies for all of the phalanges are in the region of 97–100%. For the distal and middle phalanges, the SVM with a quadratic kernel (SVMQ) is the best performing classifier. Achieving accuracies of 84.75% and 87.25% respectively, and calculating almost all of the instances within one TW stage. The best performing classifier for the middle phalange is the SVM with a linear kernel (SVML) which achieves an accuracy of 77.75%. In general, the more complex classifiers (SVMs, Ensemble Forests, MLP) tend to outperform the simpler classifiers (k-NN, C4.5 tree). Given the results in Tables 6.12–6.15, the classifier that will be used on the test data is SVMQ. As it was best performing best classifier overall (as shown in Table 6.12), as well as being the most consistent. The majority of the classifiers used here, performed well with the exception of SVM with radial basis function kernel (SVMR), which indicates that this kernel is not suitable for this problem.

Table 6.13: Classification of distal phalange III Tanner-Whitehouse stage accuracy (%).

	Correct Stage	Within One Stage
k-NN	81.00	97.00
Naive Bayes	80.00	98.75
C4.5	77.00	97.00
SVML	82.25	97.25
SVMQ	84.75	98.00
SVMR	73.75	89.50
Random Forest (30)	82.50	98.25
Random Forest (100)	83.00	98.00
Rotation Forest (30)	81.25	97.75
Multilayer Perceptron	82.75	98.00

Table 6.14: Classification of middle phalange III Tanner-Whitehouse stage accuracy (%).

	Correct Stage	Within One Stage
k-NN	75.25	98.75
Naive Bayes	77.25	99.25
C4.5	69.50	98.00
SVML	77.75	98.50
SVMQ	75.75	99.75
SVMR	61.00	89.50
Random Forest (30)	74.75	99.00
Random Forest (100)	77.25	99.50
Rotation Forest (30)	75.00	99.25
Multilayer Perceptron	72.75	99.00

6.2.2 Performance on Testing Data

The next stage is to evaluate the best performing SVMQ classifier on unseen test data. The number of instances used for each bone, in the testing data are shown in Table 6.11. The results are shown in Table 6.16. The results are broadly consistent with the results found in the cross-validation stage of testing. There is a decrease in the accuracy for the distal and proximal phalanges in comparison to the cross-

Table 6.15: Classification of proximal phalange III Tanner-Whitehouse stage accuracy (%).

	Correct Stage	Within One Stage
k-NN	80.25	99.00
Naive Bayes	78.75	99.75
C4.5	81.50	97.75
SVML	82.50	98.75
SVMQ	87.25	100.00
SVMR	72.25	92.50
Random Forest (30)	86.00	98.75
Random Forest (100)	85.75	98.75
Rotation Forest (30)	85.00	99.00
Multilayer Perceptron	84.25	99.75

validation results. For the middle phalanx there is a slight increase in accuracy. This also occurred on the bone segmentation classification undertaken in Section 5.4.2. As with the cross-validation results, the proximal phalange result is found to be the most accurate individual bone. The worst performing bone individually, is the distal phalanx. This was also the worst performing individual bone on the bone segmentation classification test data, discussed in Section 5.4.2.

Table 6.16: Tanner-Whitehouse stage accuracy of SVMQ on unseen data (%).

	Correct Stage	Within One Stage
Distal	74.82	98.56
Middle	75.97	100.00
Proximal	78.05	99.51
Overall	76.51	99.40

6.2.3 Comparison to Previously Published Work

To put the performance into context, it is worth comparing these results to those of previously published TW classifiers. Thodberg *et al.* [TKJP09] perform a cross validation on 84 images. Niemeijer *et al.* [NvM⁺03] split their data into a training

set of 119 images and a testing set of 71 images. Niemeijer *et al.* report an accuracy of 73.2% correct and 97.2% within one stage on the distal phalange. Our results and those of Thodberg *et al.* are presented in Table 6.17.

Table 6.17: Comparison of results with previously proposed method [TKJP09], with all percentages rounded to the nearest whole number.

	ASMA Cross Validation	ASMA Test	Thodberg <i>et al.</i> [TKJP09]
Distal	85% (99%)	75% (99%)	71% (96%)
Middle	76% (100%)	76 % (100%)	75% (98%)
Proximal	87% (100%)	78 % (100%)	77% (99%)

The data and experimental regime used to obtain these results are not the same, so we should be cautious in drawing any conclusions about relative performance. However, it seems that the three algorithms are broadly comparable, with approximately 75%-80% of cases correct and 95%-100% within one class. The results are also comparable to those of human raters as presented in [TWH⁺01], where different observers rating the same radiograph gave the same rating on 75-85% instances and were within one stage on all instances.

6.3 Conclusions

In this chapter, we have discussed the first part of stage F of the proposed ASMA system. Firstly, an exploratory analysis of the features extracted in Chapter 5 was undertaken. This was followed by an experimental evaluation of the use of classification schemes for use with Tanner-Whitehouse stages. The main findings of this chapter are:

- over the whole skeletal development process, the most important feature for all three of the bones used is the presence of the epiphysis (feature 1). However, when the epiphysis is present, the features that are most important relate to the width of the epiphysis (features 17, 19, and 25);

- when investigating the discriminatory power of features for individual stages, features relating to phalanx height (features 2, and 4) are the most important for the earlier stages of development, features relating to epiphysis width are most important during the middle stages of development, and that the presence of the epiphysis is the most important for the final stage; and
- the best performing classifier was the SVM with a quadratic kernel, achieving an overall accuracy was 76.51%, and a within one stage accuracy 99.40%. Which is as accurate as both previously proposed systems and human assessors [NvM⁺03, TKJP09].

Chapter 7

ASMA Stage F (Part Two): Regression Onto Chronological Age

In this chapter we investigate regression models with respect to chronological age. In the longer term we believe that this is the most appropriate method for ABAA, as it has multiple advantages over the alternatives: Regression allows a prediction on a continuous scale rather than predicting discrete stages like current methods (TW or GP). Also, the ability to build different regression models on different populations means the approach is far more customisable. This offers the possibility of avoiding one of the problems with current methods. In this chapter we discuss:

1. a piecewise regression based on the presence of the epiphysis (Sections 7.2 and 7.3);
2. feature selection, validation checks and transformations for each of the models (Section 7.2);
3. the predictive power of the models (Section 7.3); and,
4. the use of models built on different genders and ethnicities (Section 7.4).

For clarity and simplicity, we restrict ourselves to the family of linear regression models for modelling age as a function of the shape features extracted in Chapter 5. Linear regression has the advantage of producing models that are comprehensible and compact. In Section 7.1 a brief introduction to linear regression is given.

We perform a piecewise regression, where a separate model is built on the instances of a bone where the epiphysis is present and where it is not present. The models built when the epiphysis is present are based upon 24 extracted features (all features except epiphysis presence) and on the 14 phalanx based features when not present. All of the segmentations labelled as correct in Chapter 5 were used for the regression experiments.

The core model selection technique is described in Section 7.2. The model selection method used in ASMA is a forward selection technique that uses the Akaike information criteria [Aka76] as the basis for the stopping condition. Validation checks and diagnostics are performed to ensure that the basic regression assumptions hold.

In Section 7.3, we compare the results of performing regression on a single bone and then investigate the effects of combining bones and performing further regressions. We expect that as different types of bones are incorporated in to the regression, the error of the regression will decrease. We will compare the results from the regression to manual Greulich-Pyle results collected from two clinicians [GZS⁺07]. A leave-one-out cross validation is used for the regression experiments with the WEKA [HFH⁺09] implementation of Linear Regression.

Finally, in Section 7.4 we investigate models built on different genders and ethnicities, to see if models tailored to a certain population are more accurate than general models.

7.1 Linear Regression

Linear regression is a widely used statistical technique that attempts to model the relation between some response variable \mathbf{y} on some predictor (regressor) values \mathbf{X} ,

in the form:

$$\mathbf{y} = \beta \mathbf{X} + \mathcal{E}. \quad (7.1)$$

Where \mathbf{y} is a vector of length n . For ASMA this is chronological age. \mathbf{X} is a matrix of size $(n, k + 1)$:

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{X}_{1,2} & \cdots & \mathbf{X}_{1,k+1} \\ 1 & \mathbf{X}_{2,2} & \cdots & \mathbf{X}_{2,k+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{X}_{n,2} & \cdots & \mathbf{X}_{n,k+1} \end{pmatrix}, \quad (7.2)$$

where k refers the number of features being regressed onto. The first column of the matrix is made up of ones so that the intercept can be calculated, with the rest of the row \mathbf{X}_i being the set of features relating to \mathbf{y}_i . β is a vector of length $k + 1$, where β_1 refers to the intercept and β_j refers to the weight given to the j^{th} feature. The final term, \mathcal{E} is a vector of length n which contains the error for each instance i .

In order to perform the regression we need to calculate an estimate for the weight vector, $\hat{\beta}$. The most common way of doing this is to use the least squares method, where:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (7.3)$$

Once $\hat{\beta}$ has been calculated, we can then calculate the estimated values of the response variable using:

$$\hat{\mathbf{y}} = \hat{\beta} \mathbf{X}. \quad (7.4)$$

This allows us the ability to analyse the model by calculating the residual error \mathbf{e} :

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \quad (7.5)$$

Alternatively by substituting Equations 7.3 and 7.4 into Equation 7.5, we can also calculate the residuals as:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \quad (7.6)$$

or:

$$\mathbf{e} = \mathbf{y} - \mathbf{H}\mathbf{y}. \quad (7.7)$$

Where \mathbf{H} refers to the Hat matrix (sometimes called projection matrix). This gives us information about the leverage of instance i on the model and is important in the analysis of the model (Section 7.2), and is calculated as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (7.8)$$

7.2 Model Selection

The core model selection technique employed is a forward selection regression, with no variable interactions, using the Akaike Information Criterion (AIC) [Aka76] as the basis for the stopping condition. Forward selection is based on the premise that individual features are added to the model one at a time. This starts with a base model and adds the best candidate feature to the model until the stopping criteria is met.

The base model $\hat{\mathbf{y}}_{base}$ used in the model selection process here is:

$$\hat{\mathbf{y}}_{base} = \hat{\beta}_1, \quad (7.9)$$

where $\hat{\mathbf{y}}$ is the set of predicted ages and β_1 is the intercept. The next step of the forward selection method is to create a individual models $\hat{\mathbf{y}}_i$ with each feature \mathbf{X}_i added to the model:

$$\hat{\mathbf{y}}_i = \hat{\beta}_1 + \hat{\beta}_i \mathbf{X}_i. \quad (7.10)$$

The Akaike Information Criterion is then used to measure the goodness of fit of a model. AIC is defined as:

$$AIC = 2p - 2\ln(L), \quad (7.11)$$

where p refers to the number of parameters in the model (for the base model $p = 1$), and L refers to the maximised likelihood value. The first term of equation 7.11 can be thought of as a penalty term for fitting extra parameters to the model, with the aim being not to overfit the model. The second term can be thought of as a reward for the fit between the model and the regressand. The AIC of each model is then calculated, with the optimal model selected as:

$$\hat{\mathbf{y}}^* = \min_{i=1,\dots,k} AIC(\hat{\mathbf{y}}_i). \quad (7.12)$$

The AIC of the best fitting model is then compared to the AIC of the base model. If $AIC(\hat{\mathbf{y}}^*) < AIC(\hat{\mathbf{y}}_{base})$, the new model is a better fit and so the base model is updated. The process is repeated with the features that have not been used in the base model, being added to the model to find a better fit. This continues until one of the two stopping criteria are met. The first stopping criteria holds true, if $AIC(\hat{\mathbf{y}}_{base}) \leq AIC(\hat{\mathbf{y}}^*)$. This means that the base model has a better fit than would be achieved by adding extra features to the regression. The second stopping criteria is met, if a model using all of the features has been used. A version of the model selection algorithm can be seen in Algorithm 7.1.

Algorithm 7.1 Model Selection Algorithm

Input: Response variable \mathbf{y} Regressors \mathbf{X} **Output:** Best fitting model $\hat{\mathbf{y}}_{base}$

```

1. Set stopping criteria to false,  $sc = false$ .
2. Calculate base model  $\hat{\mathbf{y}}_{base}$ .
while  $sc == false$  do
  3.1 Calculate each possible model  $\hat{\mathbf{y}}_i$ , by adding an individual feature and weight
   $\hat{\beta}_i \mathbf{X}_i$  to the base model  $\hat{\mathbf{y}}_{base}$ .
  3.2 Calculate best fit model  $\hat{\mathbf{y}}^*$ , using AIC.
  if  $AIC(\hat{\mathbf{y}}^*) < AIC(\hat{\mathbf{y}}_{base})$  then
    3.3.1 Set base model as best fit model  $AIC(\hat{\mathbf{y}}_{base}) = AIC(\hat{\mathbf{y}}^*)$ 
  else
    3.4.1 Stopping criteria met,  $sc = true$ .
  end if
  if all features used in  $\hat{\mathbf{y}}^*$  then
    3.5.1 Stopping criteria met,  $sc = true$ .
  end if
end while
return Best fitting model  $\hat{\mathbf{y}}_{base}$ .

```

7.2.1 Piecewise Regression

In the exploratory analysis undertaken in Chapter 6, it was found that the most discriminatory feature is the epiphysis presence (feature 1), as well as the root node for all of the C4.5 trees. We construct two separate models for each phalange. Firstly, the instances where the epiphysis is detected, and secondly, the instances where the epiphysis is not detected. We denote the models for the proximal phalange as P_e, P_p , where P_e is the model constructed on data where the epiphysis is detected and P_p is the model constructed on instances where the epiphysis is not detected. Similarly, the models built on just the distal phalange are denoted D_e, D_p and the middle phalange are M_e, M_p .

7.2.2 Multiple Bone Models

In addition to examining regression models on single bones, we investigate ways of forming predictions from multiple bones. There are two obvious ways of doing this: we could either concatenate features and build the model on the expanded feature set, or we can produce a model for each bone and combine the predictions. We chose to combine estimates from individual bones. The main benefit of adopting this approach is that it is more flexible for cases when we cannot extract all the required bones. We denote models using the average of all the bones present as *DMP*.

7.2.3 Outliers of Models

Linear regression is particularly susceptible to outliers since they can exert excessive leverage on the model. Here we define an outlier as an instance with an absolute standardised residual > 2.5 . In Figures 7.1(a)(c)(e)-7.2(a)(c)(e), the predicted age against chronological age of the models is shown as well as the absolute standardised residuals against the predicted age. The absolute standardised residual of an instance \mathbf{r}_i is calculated as:

$$\mathbf{r}_i = \frac{\mathbf{e}_i}{s\sqrt{(1 - \mathbf{H}_{i,i})}}, \quad (7.13)$$

where e_i refers to the residual of instance i , $\mathbf{H}_{i,i}$ is the leverage term of the instance, s is the standard error:

$$s = \sqrt{\frac{SSR}{n - k - 1}}, \quad (7.14)$$

and SSR is the sum of squared residuals:

$$SSR = \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2. \quad (7.15)$$

There are several unusual observations in the data. These are shown in Table 7.1. The results indicate that there are far more outliers for the epiphysis models. This is probably due to the regression lines produced on the epiphysis being closer to the actual age than those produced on the phalanx models. A standard measure used to calculate if an instance is exerting excessive leverage on the model is Cook's distance:

$$\mathbf{d}_i = \frac{\mathbf{e}_i^2}{k+1} \left(\frac{\mathbf{H}_{i,i}}{1 - \mathbf{H}_{i,i}} \right). \quad (7.16)$$

This measure is used to make a comparison between the full model and the model built if instance i is removed. The standard measure to calculate if an outlier is exerting excessive leverage on the model is to use the 50th percentile of the $f_{k+1,n-k-1}$ -distribution. After examining the outliers identified, none of them are above $f_{0.5,k+1,n-k-1}$ and hence do not warrant further investigation. This indicates that none of these outliers are exerting excessive leverage on their respective models. However, as we have adopted a conservative approach throughout the development of ASMA, an investigation into the reason for the discrepancy of these images should be performed. The observed discrepancy between predicted and actual age may be caused by three factors. Firstly, the model may simply not capture all of the factors influencing age. Secondly, the child's development may be abnormal, meaning there is a genuine difference between bone age and chronological age. Thirdly, the checks we make to detect that the image processing has correctly captured the features may have failed. Examination of the images indicates that we have extracted the features correctly. Whilst the models would improve if we remove this data, it would also bias our evaluation of predictive power. Instead, we mitigate against outliers for the combined model *DMP* by ignoring any prediction that is less than or greater than 2 years of the other two predictions. This is a standard approach in the literature [TKJP09]. This approach is unsupervised and hence will not bias our assessment of predictive power. It is only used when we have features for all three bones, but will become more relevant when we include more bones in the model.

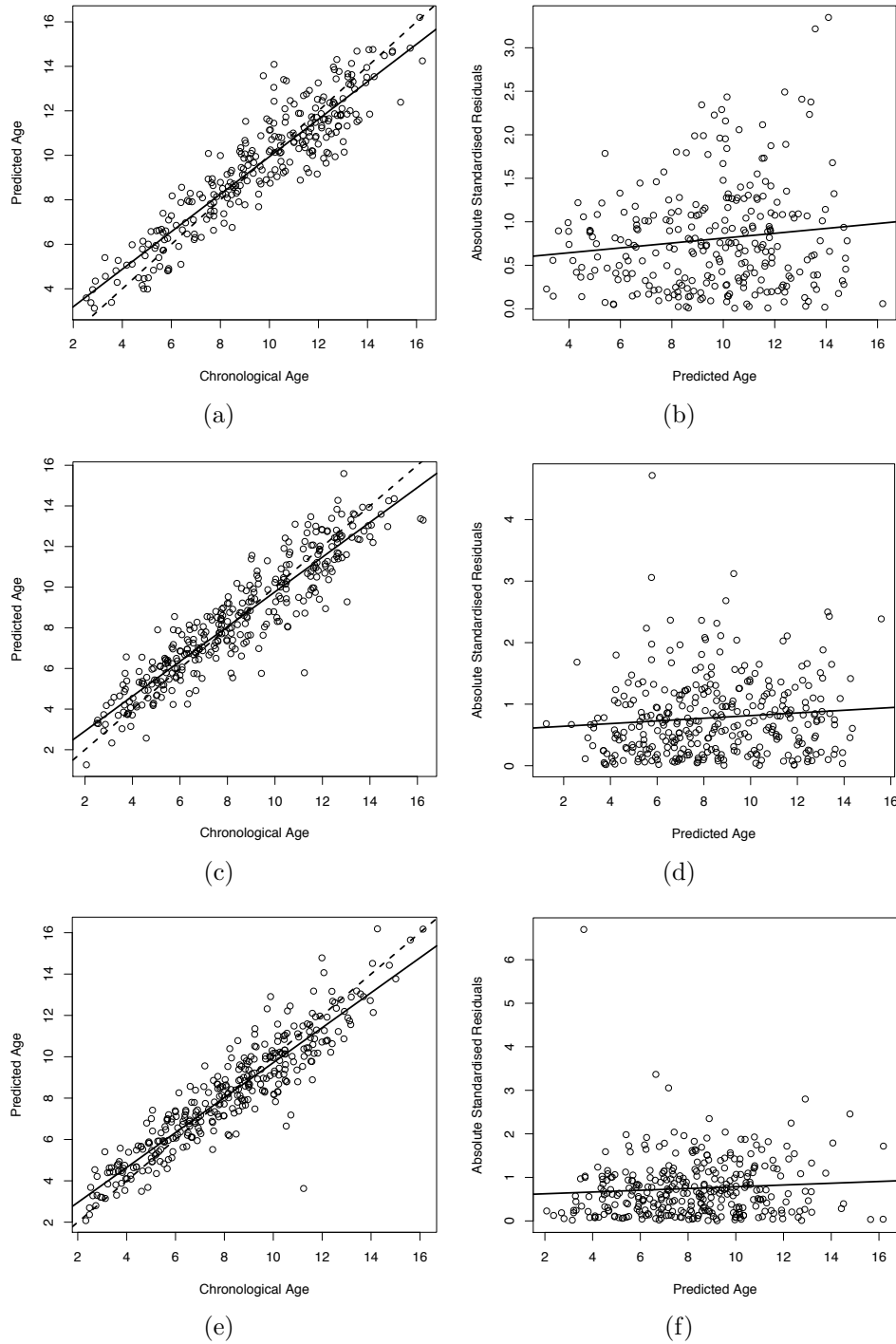


Figure 7.1: Epiphysis models. (a)(c)(e) Show the predicted age of each instance against the chronological age for D_e , M_e and P_e respectively. The dotted line is predicted = chronological. The solid line is the regression of predicted vs chronological. (b)(d)(f) Show the absolute standardised residuals against predicted age for D_e , M_e and P_e respectively. The solid line is the regression of absolute standardised residuals against predicted age.

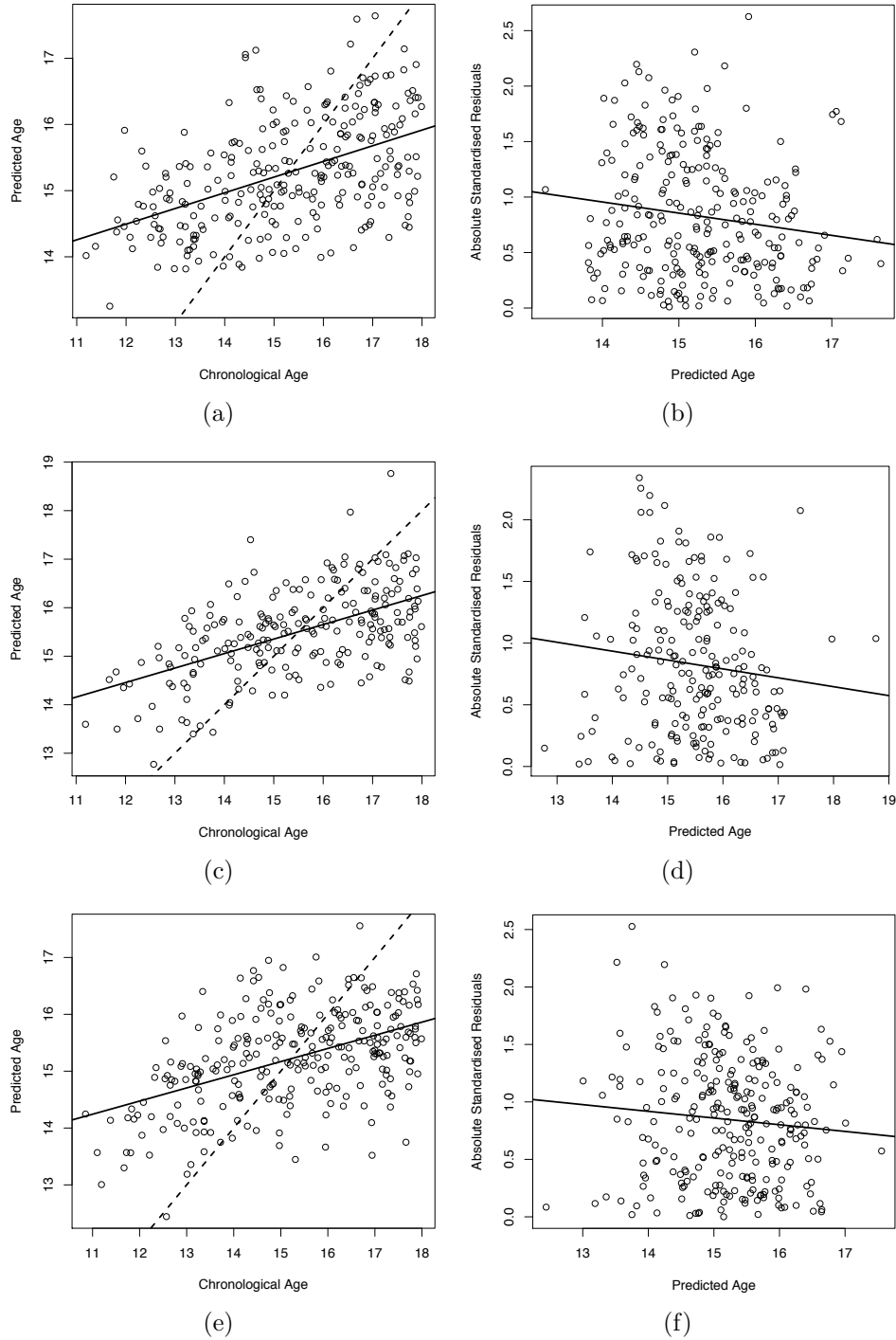


Figure 7.2: Phalanx models. (a)(c)(e) Show the predicted age of each instance against the chronological age for D_p , M_p and P_p respectively. The dotted line is predicted = chronological. The solid line is the regression of predicted vs chronological. (b)(d)(f) Show the absolute standardised residuals against predicted age for D_p , M_p and P_p respectively. The solid line is the regression of absolute standardised residuals against predicted age.

Table 7.1: Observations in individual models where absolute standardised residual > 2.5 . The threshold value for the model is the median value of the $f_{k+1,n-k-1}$ -distribution.

Model	$f_{0.5,k+1,n-k-1}$	Chronological Age	Predicted Age	Absolute Standardised Residual	Cook's Distance
d_e	0.87	10.19	14.09	3.35	0.14
		9.75	13.57	3.22	0.04
m_e	0.95	11.24	5.78	4.71	0.19
		13.04	9.28	3.12	0.01
		9.43	5.76	3.06	0.02
		12.13	8.94	2.68	0.03
p_e	0.94	11.24	3.63	6.69	0.25
		10.53	6.65	3.37	0.03
		10.72	7.18	3.06	0.02
		9.89	12.91	2.80	0.14
d_p	0.84	11.97	15.91	2.63	0.01
m_p	0.79	N/A	N/A	N/A	N/A
p_p	0.84	17.66	13.75	2.52	0.03

7.2.4 Heteroscedasticity of Models

One of the core regression assumptions is that the variance of the errors is constant. To check this assumption, we measure the correlation between the absolute values of the standard residuals and the response variable. If there is significant correlation, the assumption of constant variance is violated, and a transformation may be required.

Figures 7.1(b)(d)(f)-7.2(b)(d)(f), show plots of absolute standardised residuals against predicted age for each of the models. The fitted linear regression line between the variables demonstrates the correlation. Where, significant correlation can be observed in all of the models.

Changing variance, or heteroscedasticity, is commonly dealt with through a Box-Cox [BC64] power transform of the response:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda}, & (\lambda \neq 0), \\ \log(y_i), & (\lambda = 0). \end{cases} \quad (7.17)$$

The optimal value λ^* is defined as:

$$\lambda^* = \arg \max_{\lambda} \left(-\frac{1}{2} n \log \hat{\sigma}^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log y_i \right), \quad (7.18)$$

where $-5 \leq \lambda \leq 5$, and:

$$\hat{\sigma}^2(\lambda) = \frac{\mathbf{y}^{\lambda T} \mathbf{x}_r \mathbf{y}^{\lambda}}{n}, \quad (7.19)$$

and

$$\mathbf{x}_r = \mathbf{I} - \mathbf{H}. \quad (7.20)$$

For the epiphysis models, choosing λ^* in this way on the data sets yields a max-

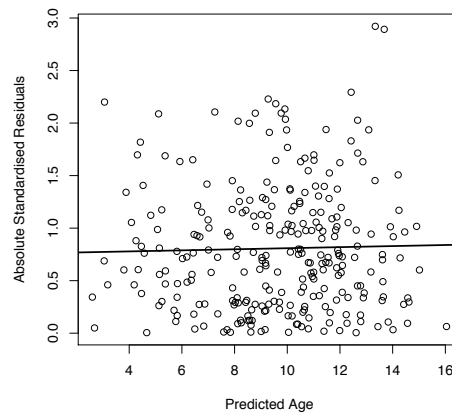
imum likelihood transform value in the range 0.5-0.7. For simplicity, we will use a power transform value of 0.67 for all experiments (a value significant for all the data folds we experimented with). Figure 7.3 shows the plot of absolute standardised residuals after transform, and demonstrates that the variance on all three bone models is now stabilised.

For the phalanx models, we do not see the same pattern in the residuals in Figure 7.2. Instead, we see a consistent over estimating at the lower age range and underestimating for higher ages. This pattern of error is generally indicative of lack of predictive power. The Box-Cox transform (after subtracting 10 from the response to remove the scale effect) yields λ^* values are in the range 1 to 1.2, indicating that transforming the response will not help.

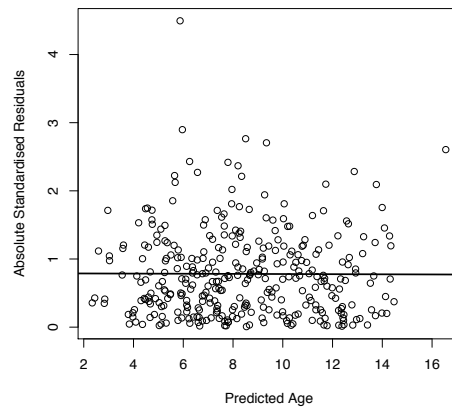
Further experimentation with generalised linear modelling and with regressor transformations using the Box-Tidwell procedure was also performed. Figure 7.4 shows the absolute standardised residuals plotted against the fitted values for each of the generalised linear models. Features 8 and 13 were the only features found to be significant when using the Box-Tidwell procedure for the models M_p and D_p respectively. The results of these transformed models are shown in Figure 7.5. As is shown in Figures 7.4-7.5, these transformations did not improve the untransformed models (shown in Figure 7.2). Hence, the phalanx data is not subjected to a transformation.

7.3 Predicting Age

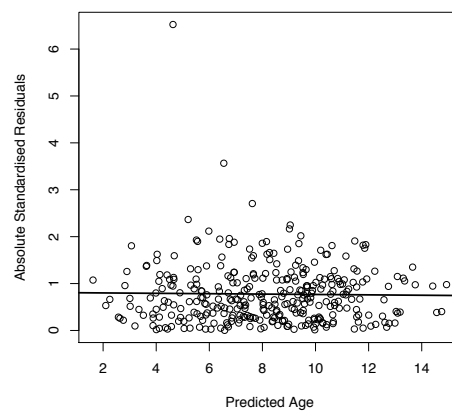
Table 7.2 shows the leave one out cross validation root mean square error (RMSE) of the linear models built on the individual bone features, and combinations of individual bone predictions, when the epiphysis is present. These results are presented in comparison to the RMSE of the scores given by clinicians using the GP system [GZS⁺07]. All models are constructed on age transformed by raising it to the power 0.67, but the RMSE scores are calculated by first transforming back to an age



(a)

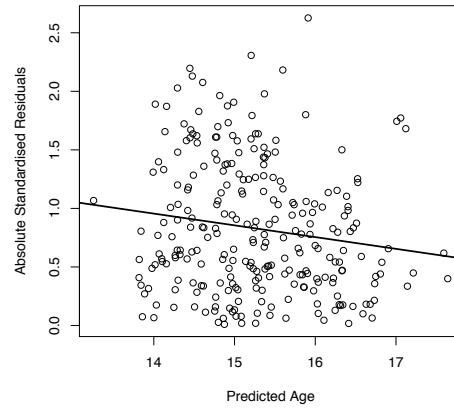


(b)

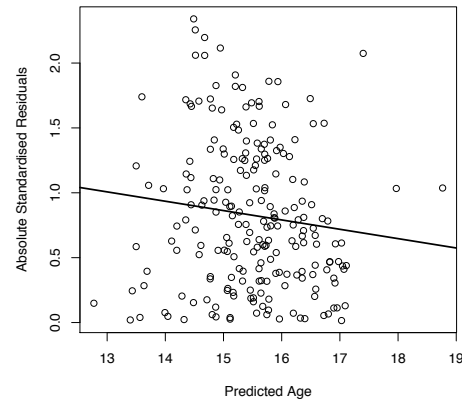


(c)

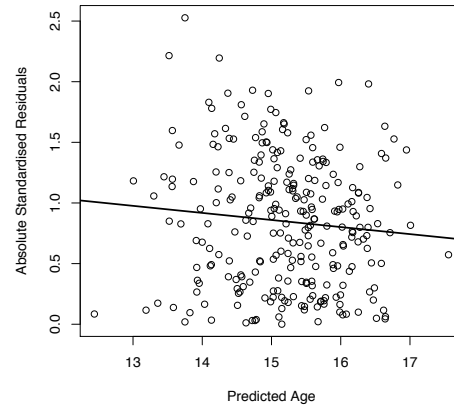
Figure 7.3: Absolute Residuals plotted against predicted age for the Epiphysis models after Box-Cox transform where $\lambda = 0.67$ (a) D_e , (b) M_e , and (c) P_e .



(a)



(b)



(c)

Figure 7.4: Absolute Residuals plotted against age based for the generalised linear models on the instances where the epiphysis is not present (a) D_p , (b) M_p , and (c) P_p .

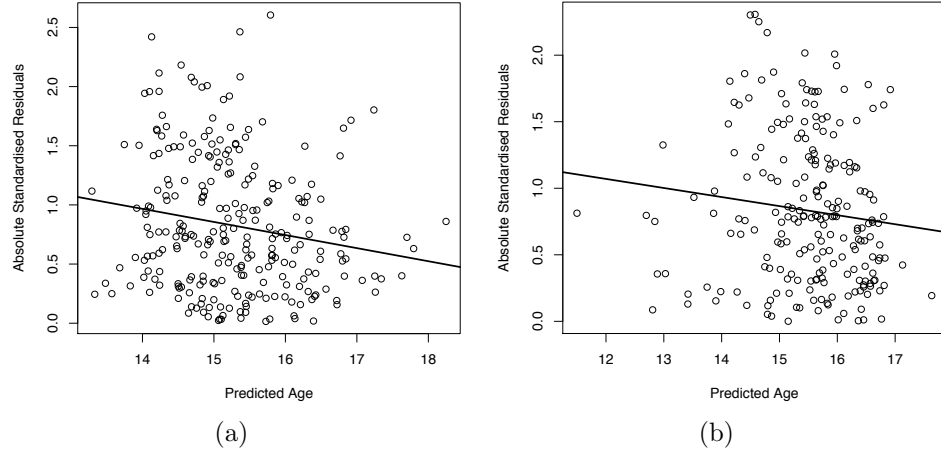


Figure 7.5: Absolute Residuals plotted against age after Box-Tidwell transformation of the regressors for the instances where the epiphysis is not present (a) D_p , and (b) M_p .

prediction. The results for the epiphysis bones are very encouraging. Increasing the number of bones in the model incrementally decreases the RMSE to the point where the three bone model is as accurate as expert human scorers. Figure 7.6 plots the predicted age against the actual age for the DMP epiphysis model for cases when we have all three bones. There is a slight bias of under predicting young subjects and over predicting older subjects, but the DMP explains approximately 90% of the variation in the response variable based on the coefficient of determination:

$$R^2 = \frac{SSR}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (7.21)$$

Table 7.3 shows the results for bones when the epiphysis is not detected. Although the error decreases as bones are added to the model, the combined DMP model is still less accurate than the human scorers. DMP only explains approximately 28% of the variation in age. Figure 7.7 plots the predicted age against the actual age for the DMP model for cases with no epiphysis where we have three bones. There is a consistent trend of overestimating the age of younger patients and underestimating older patients.

Table 7.2: RMSE for regression models where the epiphysis is detected. GP1 and GP2 are the RMSE for the two clinical estimates.

Model	Nos Cases	Regression	GP1	GP2
Single Bone Models				
D_e	275	1.24	0.89	0.86
M_e	335	1.27	0.85	0.92
P_e	334	1.12	0.87	0.86
Multiple Bone Models DMP				
At least 1 bone	566	1.19	0.87	0.89
At least 2 bones	294	1.03	0.86	0.87
3 bones	76	0.88	0.89	0.89

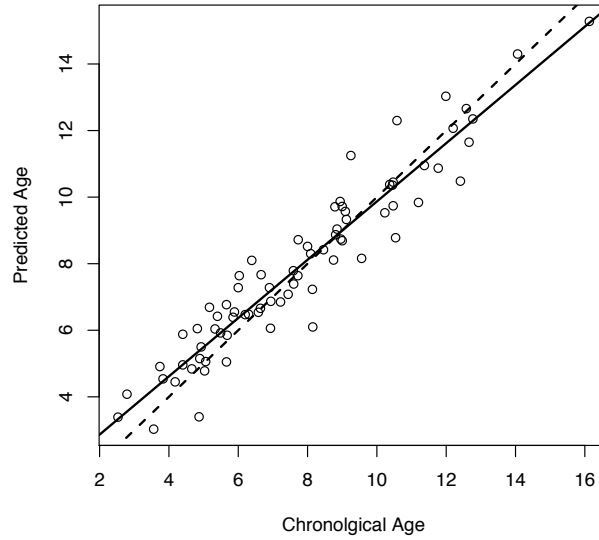


Figure 7.6: Predicted ages vs actual age for the epiphysis model DMP with three bones present. The dotted line is for predicted = chronological. The solid line is the regression of predicted vs chronological.

There are several reasons why the no epiphysis models are worse than the epiphysis models. Firstly, predicting age for subjects approaching full maturity is generally much harder. After development stops, bone features are no longer predictive of age, and the age at which development stops is highly variable. Another factor is that, based on the TW descriptors, intensity information is more important in distinguishing between almost fully mature bones. Including image intensity fea-

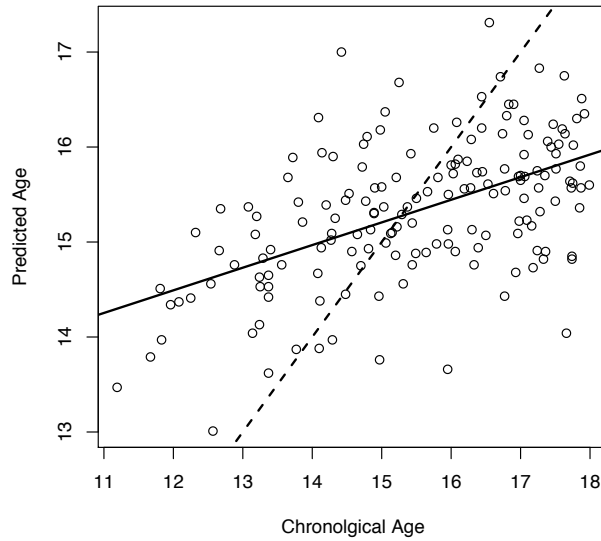


Figure 7.7: Predicted ages vs actual age for the no epiphysis model *DMP* with three bones present. The dotted line is for predicted = chronological. The solid line is the regression of predicted vs chronological.

Table 7.3: RMSE for regression models where the epiphysis is not detected. GP1 and GP2 are the RMSE for the two clinical estimates.

Model	Nos Cases	Regression	GP1	GP2
Single Bone Models				
D_p	261	1.56	1.24	1.29
M_p	217	1.48	1.22	1.24
P_p	267	1.6	1.19	1.24
Multiple Bone Models <i>DMP</i>				
At least 1 bone	320	1.53	1.22	1.26
At least 2 bones	257	1.48	1.24	1.28
3 bones	165	1.43	1.17	1.19

tures may reduce the error. Finally, the assumption of constant variance is clearly violated for all of the models, hence an alternative modelling technique may reduce the error.

The only published results we have been able to find that compare predicted age to actual age are in Adeshina *et al.* [ACA09], who report a Mean Absolute Error (MAE) for a Distal+Middle regression model of 1.26 for females and 1.28

for males. Table 7.4 presents the MAE for the comparable epiphysis models, non epiphysis models and the combined cases. These results demonstrate that even models built on a single bone without an epiphysis perform comparably to those reported in [ACA09], and when using cases with and without the epiphysis, ASMA performs much better. However, as with the results of the TW classification in Chapter 6, we need to be careful not to draw too many conclusions as different datasets are used by each proposed system.

Table 7.4: MAE for alternative bone combinations.

Model	Epiphysis	No-Epiphysis	Combined
distal	1.01	1.27	1.14
middle	0.98	1.23	1.08
DM	0.91	1.16	1.05
DMP	0.69	1.2	1.04

A comparison between the results of a model and the two GP raters can also be performed. For the instances in the three bone model *DMP* made from instances where the epiphysis is present, the RMSE between the results of the model and the two raters are 1.03 and 0.93 respectively, with a RMSE of 0.65 between the two sets of ratings on the same instances. The main reason for the inter-observer rating being lower than the comparisons with those predicted by the model is due to the fact that the populations of the radiographs used to assign the GP ratings are the same, where as the radiographs that the model is built upon are different.

7.4 Effects of Gender and Ethnicity

A linear model also offers a simple way of determining whether there are differences in age model between populations. In the long term we would like to build different regression models for ASMA based upon different local populations as this may lead to more accurate assessments. Here, we address the question of whether independent models for males D_m, M_m, P_m and females D_f, M_f, P_f , perform better than a general

model. We investigate this by splitting the instances where the epiphysis has been found into two different datasets depending on gender. As the no epiphysis models did not perform very well in Section 7.3, we do not use them during this section. As with the epiphysis models, all models are constructed on age transformed by raising it to the power 0.67, but the RMSE scores are calculated by first transforming back to form the age prediction.

In Table 7.5, we show the leave one out cross validation RMSE on each of the individual bones for both genders as well as overall D_g , M_g , and P_g . The numbers in brackets are the RMSE of the general models D_e , M_e and P_e on the same instances. We make use of this as it would be unfair to compare the gender based RMSE to the RMSE of the general models shown in Table 7.2. We can see that on the female models, the M_f model gives the same RMSE as the general model, with the D_f and P_f models performing worse. For the male models, we find that the distal D_m and middle M_m models perform better than on the same instances than the general model. With the proximal model P_m again performing slightly worse. The models D_g , M_g , and P_g are calculated using the weighted average of the female and male models, as this gives a fair indication of whether specific gender based models perform better than the general overall models shown in Table 7.2. If we compare the overall results of the gender specific models to the overall models we find that both of the D_g and M_g models produce better results than the general overall models D_e and M_e . The proximal P_g model performs worse than the overall model P_e . This is most likely be due to the outliers detected in Section 7.2 having more leverage on the individual gender models as these are built on fewer instances. We have also calculated the RMSE of combining the instances where all three bones are present DMP_g and compare it to the RMSE of the general model DMP , we can see that it performs slightly worse than the general model, again this could be caused by outliers exerting more leverage on the individual models. These results are promising though with the distal and middle phalange gender models outperforming their general model counterparts.

Table 7.5: RMSE for regression models based on gender. GP1 and GP2 are the RMSE for the two clinical estimates.

Model	Nos Cases	Regression	GP1	GP2
Female				
D_f	127	1.29 (1.24)	0.91	0.92
M_f	156	1.14 (1.14)	0.89	0.95
P_f	136	1.12 (0.99)	0.94	0.90
Male				
D_m	148	1.18 (1.25)	0.87	0.81
M_m	179	1.33 (1.38)	0.82	0.90
P_m	198	1.29 (1.20)	0.82	0.83
Overall				
D_g	275	1.23 (1.24)	0.89	0.86
M_g	335	1.24 (1.27)	0.85	0.92
P_g	334	1.22 (1.12)	0.87	0.86
Multiple Bone Model				
DMP_g	76	0.95 (0.88)	0.89	0.89

The other demographic variable we have available is ethnicity. We have built individual models for each ethnicity: Asian D_a, M_a, P_a , African-American D_{aa}, M_{aa}, P_{aa} , Caucasian D_c, M_c, P_c , and Hispanic D_h, M_h, P_h . As with the gender models, all models are constructed on the transformed response variable, $age^{0.67}$. With the RMSE scores calculated after transforming the estimated value back to form an age prediction.

The leave one out cross validation RMSE on each of the individual bones for all ethnicities as well as overall D_{eth}, M_{eth} , and P_{eth} , is shown in Table 7.6. As with Table 7.5, the values in brackets refer to the RMSE of the general model on the same instances. We can see that on the models made up of patients with Asian ethnicity, that the distal model D_a is the only model that performs better. For the African-American models both the distal D_{aa} and middle M_{aa} models perform better than the general models on the same instances. All of the other ethnicity based models perform worse than the general model on the same instances. The models D_{eth}, M_{eth} , and P_{eth} are calculated using the same method as with the gender

based models. If we compare the overall results of the ethnicity specific models to the overall models we find that the distal model D_{eth} achieves the same RMSE as the general model D_e . The M_{eth} , and P_{eth} models both perform worse than their overall model counterparts. As with the gender based models this is most likely to be due to the outliers detected in Section 7.2 having more leverage on the individual models. Again we calculate the RMSE of the multiple bone model DMP_{eth} from the ethnicity based models. We find that this gives us the same accuracy as the general model DMP , this is a promising result and means that further investigations into gender and ethnic models should be performed. Thus in future versions of ASMA we will investigate building models upon local populations.

Table 7.6: RMSE for regression models based on ethnicity. GP1 and GP2 are the RMSE for the two clinical estimates.

Model	Nos Cases	Regression	GP1	GP2
Asian				
D_a	76	1.24 (1.25)	0.95	0.94
M_a	78	1.22 (1.09)	0.96	0.98
P_a	69	1.10 (0.91)	0.91	0.84
African-American				
D_{aa}	82	1.21 (1.36)	0.83	0.79
M_{aa}	85	1.24 (1.32)	0.86	0.93
P_{aa}	97	1.53 (1.22)	0.89	0.85
Caucasian				
D_c	46	1.25 (1.09)	0.88	0.74
M_c	77	1.48 (1.37)	0.79	0.89
P_c	79	1.20 (0.99)	0.90	0.88
Hispanic				
D_h	71	1.28 (1.18)	0.89	0.93
M_h	95	1.56 (1.27)	0.81	0.90
P_h	89	1.51 (1.26)	0.78	0.86
Overall				
D_{eth}	275	1.24 (1.24)	0.89	0.86
M_{eth}	335	1.38 (1.27)	0.85	0.92
P_{eth}	334	1.36 (1.12)	0.87	0.86
Multiple Bone Model				
DMP_{eth}	76	0.88 (0.88)	0.89	0.89

7.5 Conclusions

In this chapter, we have discussed the final part of the proposed ASMA system. Firstly, we discussed linear regression along with the model selection technique used. A piecewise regression based on the presence of the epiphysis was then investigated. The regressors are the features extracted in Chapter 5 and the response variable is chronological age. The main findings of this chapter are:

- the more bones added to the model the more accurate it becomes;
- the best performing model was the *DMP* when the epiphysis is present, which achieves a lower RMSE than two manual raters using the GP method of bone age assessment, and also outperforms previously proposed ABAA systems that have performed regression onto chronological age; and,
- the results given in Tables 7.5 and 7.6 indicate that it would be promising to build models based on gender and ethnicity in the future

Chapter 8

Conclusions And Future Work

8.1 Conclusions

This thesis introduces the Automated Skeletal Maturity Assessment (ASMA) algorithm, a new fully automated method for bone age assessment, that consists of the following six distinct stages: a) segmenting the hand outline; b) classifying whether the hand outline is correct; c) segmenting the bones from within the outline; d) extraction of bone features; e) classifying whether the bone segmentations are correct; and finally, f) calculating bone age.

Although systems for the task of automated bone age assessment have been proposed before, none have gained widespread acceptance. We believe there are two main reasons for this: firstly, a lack of verification, and, secondly, a lack of transparency. With ASMA we have attempted to tackle both of these problems. The first problem is tackled by having conservative classification steps after each of the segmentation and feature extraction stages to ensure that no bad segmentations get through and thus result in a bad assessment. The second problem is tackled by deriving and extracting features that explain the variation in the Tanner-Whitehouse stages. Using these features means that the cause of an assessment can be found, which may give paediatricians a better understanding of the diagnosis and intended

course of treatment if necessary.

For stages A and B of ASMA (Hand Segmentation and Classification), we investigated the use of three commonly used methods (AAMs, Canny Edge Detection and Otsu Thresholding) and proposed a contouring algorithm that to the best of our knowledge, has not been used for this purpose before. A novel ensemble algorithm for outlining radiographs was described along with two voting schemes to find the best outline. We found that of the two voting schemes used, DTW outperforms the likelihood ratio test. Finally, we investigated the use of a variety of classifiers and transformations of outlines to automatically detect whether an outline is correct. We found that the ensemble algorithm, in conjunction with the contouring outline algorithm, extracted correct outlines from over 80% of images. We conclude that AAM is the only other contender in terms of accuracy, but is not suitable for this project because the types of mistake it makes are hard to detect automatically. The best performing classification/transformation combination for classifying if a hand outline is indeed correct was found to be the Random Forest classifier based on the hand outlines transformed into a set of principle components.

In order to perform stages C, D and E (Bone Segmentation, Feature Extraction, and Classification), a novel algorithm for the segmentation of the phalanges of the middle finger was presented. This involved using the hand outline obtained in stage A to find the tips and webs around the middle finger. A ROI around each bone is then calculated and the hard tissue segmented using Canny Edge Detection in conjunction with a Gaussian pyramid. The next stage is to extract features from the bone segmentations. We derive 25 shape based features from the Tanner-Whitehouse descriptions. The first of these is the binary feature variable based on the presence of the epiphysis. 14 features based on the phalanx and 10 on the epiphysis (if present). The methods used to extract the features are described along with a novel technique that combines the elliptical Hough transform and Gaussian pyramids. Finally, we investigate classification schemes to automatically detect whether a segmentation is correct. This is a two stage process. The first stage is a set of rules that will reject

any segmentations that do not conform to them. In the second stage we investigate classification schemes on a variety of classifiers with three different representations of the segmentations. We found that the optimal representation/classifier combination is the extracted features in conjunction with the SVM classifier using a quadratic kernel.

ASMA uses two separate methods to calculate bone age. Firstly, it builds classifiers to recreate the Tanner-Whitehouse method, and secondly, it regresses the feature set onto chronological age. In order to classify according to the descriptions given by Tanner and Whitehouse, we perform an exploratory analysis of the features extracted in stage D of ASMA. Using the information gain metric we rank features over the whole skeletal development process against individual Tanner-Whitehouse stages, and examine how they interact to form Tanner-Whitehouse classifications. We found that the most important feature for all three of the bones used is the presence of the epiphysis (feature 1). However, when the epiphysis is present, the features that are most important relate to the width of the epiphysis (features 17, 19, and 25). An experimental evaluation of the use of a variety of classification schemes for use with Tanner-Whitehouse stages was then undertaken. The best performing classifier was once again the SVM with a quadratic kernel. The overall accuracy of the classifier was 76.51%, with a within one stage accuracy 99.40%. We found ASMA to be at least as accurate as results published for previously proposed systems [NvM⁺03, TKJP09].

The final part of ASMA involves performing bone age assessments by regressing onto chronological age. The model selection technique used is forward selection with the Akaike information criteria as the stopping condition. We perform a piecewise regression where instances are split based on the presence of the epiphysis. The regressors used for the models were the features extracted in Chapter 5 and the response variable was chronological age. A leave one out cross validation was performed on both individual and multiple bone models. Unsurprisingly, the more bones added to the model, the more accurate it became. The best performing model

found was the *DMP* built on the features of three phalanges where the epiphysis was detected. This model achieves a lower RMSE than two manual raters using the GP method of bone age assessment. We also found that ASMA outperforms previously proposed ABAA systems that have constructed a regression onto chronological age. Finally, we investigated the use of gender and ethnicity independent models.

8.2 Future Work

One of the main advantages of the ASMA system is that it is a stage based system. This gives the ability to investigate ways to improve an individual stage without affecting the other stages in the process.

8.2.1 Stage A: Hand Segmentation

A possible area of further work for stage A is to investigate the use of other segmentation techniques such as semantic-based algorithms. Semantic-based techniques for image segmentation aim to group the pixels of an image into semantically meaningful sets, where the information conveyed by the pixels within a group is similar in some sense. Typically this involves some form of cluster analysis on the greyscale/colour pixel information, the image gradients, and/or the local texture information [IW06]. Alternatively it might involve a more sophisticated form of clustering and classification using decision trees [SJC08]. The advantage of these approaches is that they require no *a priori* knowledge and operate on the image data directly. However, there is no guarantee that the image segments align properly with real-world objects.

Another area of potential future work would be to investigate the Heel effect, to see if it is possible to create a local based solution, rather than the global technique that is currently used in the ensemble. If this were achievable, it may have the potential to lead to better segmentation accuracy.

8.2.2 Stage B: Hand Segmentation Classification

For stage B, the Random Forest / principle components classification scheme performs adequately at this stage of development. However, it may be possible to construct a better classifier, which lets no false positives through. One avenue of investigation would be to experiment with alternative transformations and classification schemes.

8.2.3 Stage C: Bone Segmentation

The obvious area of future work for stage C is to incorporate the phalanges of the other fingers, metacarpals, radius and ulna into ASMA, as this should produce better bone age estimates. As with the hand segmentation, a possible area of further work is to investigate other segmentation techniques. One of the most common types of errors found with the bad segmentations is that the epiphysis and phalanx are segmented as one region. Potential future work could be to investigate if it is possible to identify this type of error, and to rectify it so that the phalanx and epiphysis segmentation are distinct.

8.2.4 Stage D: Feature Extraction

Currently for stage D of ASMA, only shape features are derived. Future work would be to incorporate image intensity based features. This could potentially improve the latter stages of ASMA, since the features are used in all subsequent stages. However, one of the major problems with deriving image based features is that they may be harder to explain to clinicians and hence the system may lose some of its transparency.

8.2.5 Stage E: Bone Segmentation Classification

For stage E we could investigate the use of alternative transformations on the one-dimensional series and the use of ensemble classification schemes to try to improve performance. Whilst the SVMQ and extracted features combination works adequately, there are a larger number of false positives that get through this check than with stage B. It may be interesting to see if it is better to split the classification based on the presence of the epiphysis. This could also lead to errors where an epiphysis has been found which is not actually present, being identified and rectified.

8.2.6 Stage F: Classification of Tanner-Whitehouse Stages

Although the ideal classifier to create the most transparent solution is a tree based classifier such as a C4.5 tree. Due to it being traceable by clinicians who may as a result gain a greater understanding of a diagnosis. It would also be interesting to have the ability to classify all TW stages (B-I) for each bone. Future work will need to be done as more bones are incorporated into the system, as it may be the case that different classifiers work best for different bones.

8.2.7 Stage F: Regression onto Chronological Age

For the regression onto chronological age, we found that the gender and ethnicity independent based models performed well. This means that future versions of ASMA could be tailored for local populations, which may then incorporate features not currently used in manual and automated bone age assessment methods. This would require more data from a range of populations. By tailoring models to local problems, other research questions could be answered as sociological and environmental factors into skeletal development could be identified. Another interesting area of future work could be to incorporate a filter that detects when the skeletal development of a bone is complete. If a bone is identified with skeletal development as complete, there is no need to perform the regression on that bone and this result

could be given to the clinician.

This thesis has presented the ASMA method for automating bone age assessment, a task performed by paediatricians in hospitals worldwide on a daily basis. ASMA addresses problems that have been identified with previously proposed BAA methods, and when the results have been compared to both manual and automated systems, ASMA has been found to be at least as accurate.

Bibliography

- [AAB⁺84] E.H. Adelson, C.H. Anderson, J.R. Bergen, P.J. Burt, and J.M. Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984.
- [ACA09] S.A. Adeshina, T.F. Cootes, and J.E. Adams. Evaluating different structures for predicting skeletal maturity using statistical appearance models. In *Proc. MIUA*, 2009.
- [Ach54] R.M. Acheson. A method of assessing skeletal maturity from radiographs: A report from the oxford child health survey*. *Journal of Anatomy*, 88(Pt 4):498, 1954.
- [AIS93] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207–216, 1993.
- [Aka76] H. Akaike. An information criterion. *Math Sci*, 14(153):5–9, 1976.
- [Bal81] D.H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981.
- [BC64] G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.
- [Beh97] E. Behrendsen. Studien über die ossifikation der menschlichen hand vermittels des roentgenschen verfahrens. *Dtsch Med Wochenschr*, 23:433–435, 1897.
- [BEK⁺99] R.K. Bull, P.D. Edwards, P.M. Kemp, S. Fry, and I.A. Hughes. Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. *Archives of disease in childhood*, 81(2):172, 1999.
- [BFOS84] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

- [BKB28] B.T. Baldwin, H.G. Kelly, and L.M. Busby. *The University of Iowa Studies in Child Welfare: Anatomic Growth of Children: a Study of Some Bones of the Hand, Wrist, and Lower Forearm, by Means of Roentgenograms*. The University of Iowa, 1928.
- [BKZ08] A. Bielecki, M. Korkosz, and B. Zielinski. Hand radiographs pre-processing, image representation in the finger regions and joint space width measurements for image interpretation. *Pattern Recognition*, 41(12):3786–3798, 2008.
- [Bre01] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [BSG81] A.E. Brodeur, M.J. Silberstein, and E.R. Graviss. *Radiology of the pediatric elbow*. GK Hall Medical Publishers, 1981.
- [Buz08] Thorsten M Buzug. *Computed tomography: from photon statistics to modern cone-beam CT*. Springer, 2008.
- [BW08] J. Benson and J. Williams. Age determination in refugee children. *Australian family physician*, 37(10):821–824, 2008.
- [Can87] J. Canny. A computational approach to edge detection. *Readings in computer vision: issues, problems, principles, and paradigms*, page 184, 1987.
- [Ces90] B. Cestnik. Estimating probabilities: A crucial task in machine learning. In *Proceedings of the ninth European conference on artificial intelligence*, volume 1990, pages 147–9, 1990.
- [CET01] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681, 2001.
- [CFMC06] R. Cameriere, L. Ferrante, D. Mirtella, and M. Cingolani. Carpals and epiphyses of radius and ulna as age indicators. *International Journal of Legal Medicine*, 120(3):143–146, 2006.
- [CHJT03] C.H. Chang, C.W. Hsieh, T.L. Jong, and C.M. Tiu. A fully automatic computerized bone age assessment procedure based on phalange ossification analysis. In *IPPR Conference on Computer Vision, Graphics and Image Processing*, pages 463–468, 2003.
- [CHV99] O. Chapelle, P. Haffner, and V.N. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999.
- [Coo00] T.F. Cootes. An introduction to active shape models. *Image Processing and Analysis*, pages 223–248, 2000.

- [CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [DMFAAL03] R. De Luis-Garcia, M. Martín-Fernandez, J.I. Arribas, and C. Alberola-Lopez. A fully automatic algorithm for contour detection of bones in hand radiographs using active contours. In *Proceedings of the IEEE International Conference on Image Processing*, pages 421–424, 2003.
- [dRS76] T. de Roo and H.J. Schröder. *Pocket atlas of skeletal age*. Springer, 1976.
- [DTB12] L. M. Davis, B.J. Theobald, and A. Bagnall. Automated bone age assessment using feature extraction. *Intelligent Data Engineering and Automated Learning-IDEAL 2012*, pages 43–50, 2012.
- [DTL⁺12] L. M. Davis, B.J. Theobald, J. Lines, A. Toms, and A. Bagnall. On the segmentation and classification of hand outlines. *International Journal of Neural Systems*, 22(5), 2012.
- [DTS⁺08] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.
- [DTTB11] L. M. Davis, B.J. Theobald, A. Toms, and A. Bagnall. On the extraction and classification of hand outlines. *Intelligent Data Engineering and Automated Learning-IDEAL 2011*, pages 92–99, 2011.
- [DWV99] H. Drucker, D. Wu, and V.N. Vapnik. Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054, 1999.
- [ECGFS12] A. Elsayed, F. Coenen, M. García-Fiñana, and V. Sluming. Region of interest based image classification: A study in mri brain scan categorization. *Data Mining Applications in Engineering and Medicine*, pages 225–248, 2012.
- [Eff93] N.D. Efford. Knowledge-based segmentation and feature analysis of hand wrist radiographs. In *Proceedings of the Society of Photo-optical Instrumentation Engineers*, volume 1905, pages 596–608. Citeseer, 1993.
- [EHC⁺11] A. Elsayed, M. Hijazi, F. Coenen, M. García-Fiñana, V. Sluming, and Y. Zheng. Time series case based reasoning for image categorisation. *Case-Based Reasoning Research and Development*, pages 423–436, 2011.

- [Elg46] O. Elgenmark. *The Normal Development of the Ossific Centres During Infancy and Childhood: A Clinical, Roentgenologic, and Statistical Study*. Almqvist & Wiksell, 1946.
- [FHJ52] E. Fix and J.L. Hodges Jr. Discriminatory analysis-nonparametric discrimination: small sample performance. Technical report, DTIC Document, 1952.
- [GKP⁺72] S.M. Garn, P. Keith, A.K. Poznanski, J.M. Nagy, et al. Metacarpophalangeal length in the evaluation of skeletal malformation. *Radiology*, 105(2):375–381, 1972.
- [GLM⁺07] D. Giordano, R. Leonardi, F. Maiorana, G. Scarciofalo, and C. Spampinato. Epiphysis and metaphysis extraction and classification by adaptive thresholding and dog filtering for automated skeletal bone age analysis. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 6551–6556. IEEE, 2007.
- [GP50] W.W. Greulich and S.I. Pyle. *Radiographic Atlas of Skeletal Development of the Hand and Wrist: Based on the Brush Foundation Study of Human Growth and Development*. Stanford University Press, 1950.
- [GP59] W.W. Greulich and S.I. Pyle. *Radiographic atlas of skeletal development of the hand and wrist*. Stanford University Press, 1959.
- [GR04] V. Gilsanz and O. Ratib. *Hand bone age: a digital atlas of skeletal maturity*. Springer Verlag, 2004.
- [GRS67] S.M. Garn, C.G. Rohmann, and F.N. Silverman. Radiographic standards for postnatal ossification and tooth calcification. *Medical radiography and photography*, 43(2):45, 1967.
- [GSSL09] D. Giordano, C. Spampinato, G. Scarciofalo, and R. Leonardi. Automatic skeletal bone age assessment by integrating emroi and croi processing. In *Medical Measurements and Applications, 2009. MeMeA 2009. IEEE International Workshop on*, pages 141–145. IEEE, 2009.
- [GSSL10] D. Giordano, C. Spampinato, G. Scarciofalo, and R. Leonardi. An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones. *Instrumentation and Measurement, IEEE Transactions on*, 59(10):2539–2553, 2010.
- [GZS⁺07] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H.K. Huang. Bone age assessment of children using a digital hand atlas. *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, 31(4-5):322, 2007.

- [HCZ10] M.H.A. Hijazi, F. Coenen, and Y. Zheng. Retinal image classification using a histogram based approach. In *IEEE International Joint Conference on Neural Networks*, pages 3501–3507, 2010.
- [HCZ12] M.H.A. Hijazi, F. Coenen, and Y. Zheng. Data mining techniques for the screening of age-related macular degeneration. *Knowledge-Based Systems*, 29:83–92, 2012.
- [HFH⁺09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- [HGH⁺76] R.B. Husar, N.V. Gillani, J.D. Husar, C.C. Paley, and P.N. Turcu. Long-range transport of pollutants observed through visibility contour maps, weather maps and trajectory analysis. In *Preprint volume: Third Symposium on Turbulence, Diffusion and Air Pollution, American Meteorological Society, Reno, NV*, pages 344–347, 1976.
- [HJT07] C.W. Hsieh, T.L. Jong, and C.M. Tiu. Bone age estimation based on phalanx information with fuzzy constrain of carpals. *Medical and Biological Engineering and Computing*, 45(3):283–295, 2007.
- [HLP07] C.C. Han, C.H. Lee, and W.L. Peng. Hand radiograph image segmentation using a coarse-to-fine strategy. *Pattern Recognition*, 40(11):2994–3004, 2007.
- [HLW⁺11] C.W. Hsieh, T.C. Liu, J.K. Wang, T.L. Jong, and C.M. Tiu. A simplified rus bone age assessment procedure using grouped-tw method. *Pediatrics International*, 2011.
- [HoCoHR07] House of Lords House of Commons and Joint Committee on Human Rights. The treatment of asylum seekers, tenth report of session 2006/07, volume 1 report and formal minutes. <http://www.publications.parliament.uk/pa/jt200607/jtselect/jtrights/81/81i.pdf>, 2007.
- [HPF62] N.L. Hoerr, S.I. Pyle, and C.C. Francis. *Radiographic atlas of skeletal development of the foot and ankle: a standard of reference*. Thomas, 1962.
- [IC06] F. Introna and C.P. Campobasso. Biological vs legal age of living individuals. *Forensic anthropology and medicine*, pages 57–82, 2006.
- [IW06] D.E. Ilea and P.F. Whelan. Color image segmentation using a spatial k-means clustering algorithm. In *International Machine Vision and Image Processing Conference (IMVIP)*, 2006.

- [KC01] Richard A Ketcham and William D Carlson. Acquisition, optimization and interpretation of x-ray computed tomographic imagery: applications to the geosciences. *Computers & Geosciences*, 27(4):381–400, 2001.
- [KK07] H.J. Kim and W.Y. Kim. Computerized bone age assessment using dct and lda. *Computer Vision/Computer Graphics Collaboration Techniques*, pages 440–448, 2007.
- [LBCSA11] J. Lines, A. Bagnall, P. Caiger-Smith, and S. Anderson. Classification of household devices by electricity usage profiles. *Intelligent Data Engineering and Automated Learning-IDEAL 2011*, pages 403–412, 2011.
- [LBTS05] T.M. Lehmann, D. Beier, C. Thies, and T. Seidl. Segmentation of medical images combining local, regional, global, and hierarchical distances into a bottom-up region merging scheme. In *Proc. SPIE*, volume 5747, pages 546–555. Citeseer, 2005.
- [LEM⁺93] R.T. Loder, D.T. Estle, K. Morrison, D. Eggleston, D.N. Fish, M.L. Greenfield, and K.E. Guire. Applicability of the greulich and pyle skeletal age standards to black and white children of today. *Archives of Pediatrics & Adolescent Medicine*, 147(12):1329, 1993.
- [Lew98] D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *Machine Learning: ECML-98*, pages 4–15, 1998.
- [LL98] C.X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 73–79, 1998.
- [MB04] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.
- [MBP⁺01] S. Mora, M.I. Boechat, E. Pietka, H.K. Huang, and V. Gilsanz. Skeletal age determinations in children of european and african descent: applicability of the greulich and pyle standards. *Pediatric research*, 50(5):624–628, 2001.
- [MCRS93] G. Manos, A.Y. Cairns, I.W. Ricketts, and D. Sinclair. Automatic segmentation of hand-wrist radiographs. *Image and Vision Computing*, 11(2):100–111, 1993.
- [MCRS94] G.K. Manos, A.Y. Cairns, I.W. Rickets, and D. Sinclair. Segmenting radiographs of the hand and wrist. *Computer methods and programs in biomedicine*, 43(3):227–237, 1994.

- [MDS⁺09] D.D. Martin, D. Deusch, R. Schweizer, G. Binder, H.H. Thodberg, and M.B. Ranke. Clinical application of automated greulich-pyle bone age determination in children with short stature. *Pediatric radiology*, 39(6):598–607, 2009.
- [MM95] J.A. Major and J.J. Mangano. Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems*, 4(1):39–52, 1995.
- [MMCD⁺05] E. Muñoz-Moreno, R. Cárdenes, R. De Luis-García, M.Á. Martín-Fernández, and C. Alberola-López. Automatic detection of landmarks for image registration applied to bone age assessment. In *Proceedings of the 5th WSEAS International Conference on Signal Processing, Computational Geometry & Artificial Vision table of contents*, pages 117–122. World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA, 2005.
- [MMFAL03] M. Martin, M. Martin-Fernandez, and C. Alberola-Lopez. Automatic bone age assessment: A registration approach. In M. Sonka and J. M. Fitzpatrick, editors, *Medical Imaging 2003: Image Processing*, pages 1765–1776, San Diego, CA, USA, February 15–20 2003. SPIE Press. Proceedings of the SPIE 5032.
- [MN89] D.J. Michael and A.C. Nelson. HANDX: A model-based system for automatic segmentation of bones from digital hand radiographs. *IEEE Transactions on Medical Imaging*, 8(1):64–69, 1989.
- [MP69] M.L. Minsky and S. Papert. Perceptrons: an introduction to computational geometry. *Massachusetts Institute of Technology Press, Cambridge, Massachusetts*, 1969.
- [MSC⁺00] S. Mahmoodi, B.S. Sharif, E.G. Chester, J.P. Owen, and R. Lee. Skeletal growth estimation using radiographic image processing and analysis. *IEEE Transactions on Information Technology in Biomedicine*, 4(4):292–297, 2000.
- [MW94] D.T. Morris and C.F. Walshaw. Segmentation of the finger bones as a prerequisite for the determination of bone age. *Image and vision computing*, 12(4):239–245, 1994.
- [NvM⁺03] M. Niemeijer, B. van Ginneken, C. Maas, F.J.A. Beek, and M.A. Viergever. Assessing the skeletal age from a hand radiograph: automating the Tanner-Whitehouse method. In *SPIE Medical Imaging*, volume 5032, pages 1197–1205, 2003.
- [OIAB96] F.K. Ontell, M. Ivanovic, D.S. Ablin, and T.W. Barlow. Bone age in children of diverse ethnicity. *American Journal of Roentgenology*, 167(6):1395–1398, 1996.

- [Ots75] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11:285–296, 1975.
- [PDMPC94] G. Piccioli, E. De Micheli, P. Parodi, and M. Campani. Robust road sign detection and recognition from image sequences. In *Intelligent Vehicles' 94 Symposium, Proceedings of the*, pages 278–283. IEEE, 1994.
- [PGP⁺01] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H.K. Huang, and V. Gilsanz. Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal ROI extraction. *IEEE Transactions on Medical Imaging*, 20(8):715–729, 2001.
- [PGPK⁺04] E. Pietka, A. Gertych, S. Pospiech-Kurkowska, F. Cao, H.K. Huang, and V. Gilzanz. Computer-assisted bone age assessment: graphical user interface for image processing and comparison. *Journal of Digital Imaging*, 17(3):175–188, 2004.
- [PH69] S.I. Pyle and N.L. Hoerr. *A radiographic standard of reference for the growing knee*. CC Thomas, Springfield, Ill., 1969.
- [PH95] E. Pietka and H.K. Huang. Epiphyseal fusion assessment based on wavelets decomposition analysis. *Computerized medical imaging and graphics*, 19(6):465–472, 1995.
- [Pie95] E. Pietka. Computer-assisted bone age assessment based on features automatically extracted from a hand radiograph. *Computerized medical imaging and graphics*, 19(3):251–259, 1995.
- [PK83] S.K. Pal and R.A. King. On edge detection of x-ray images using fuzzy sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 5(1):69–77, 1983.
- [PKKH93] E. Pietka, L. Kaabi, M.L. Kuo, and H.K. Huang. Feature extraction in carpal-bone analysis. *Medical Imaging, IEEE Transactions on*, 12(1):44–49, 1993.
- [PM90] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629–639, 1990.
- [PMGKH91] E. Pietka, M.F. McNitt-Gray, M.L. Kuo, and H.K. Huang. Computer-assisted phalangeal analysis in skeletal age assessment. *Medical Imaging, IEEE Transactions on*, 10(4):616–620, 1991.
- [Pol98] J. Poland. *Skiagraphic atlas showing the development of the bones of the wrist and hand: for the use of students and others*. Smith, Elder, 1898.

- [PP86] A. Pathak and S.K. Pal. Fuzzy grammars in syntactic recognition of skeletal maturity from x-rays. *Systems, Man and Cybernetics, IEEE Transactions on*, 16(5):657–667, 1986.
- [PPG⁺01] E. Pietka, S. Pospiech, A. Gertych, F. Cao, H.K. Huang, and V. Gilsanz. Computerized automated approach to the extraction of epiphyseal regions in hand radiographs. *Journal of Digital Imaging*, 14(4):165–172, 2001.
- [PPK84] A. Pathak, S.K. Pal, and R.A. King. Syntactic recognition of skeletal maturity. *Pattern recognition letters*, 2(3):193–197, 1984.
- [PPKGC03] E. Pietka, S. Pospiech-Kurkowska, A. Gertych, and F. Cao. Integration of computer assisted bone age assessment with clinical PACS. *Computerized Medical Imaging and Graphics*, 27(2-3):217–228, 2003.
- [Pry07] J.W. Pryor. The hereditary nature of variation in the ossification of bones. *Anat Rec*, 1:84–88, 1907.
- [PWG71] S.I. Pyle, A.M. Waterhouse, and W.W. Greulich. *A radiographic standard of reference for the growing hand and wrist*. Press of Case Western Reserve University; distributed by Year Book Medical Publishers, Chicago, [Cleveland], 1971.
- [Qui93] J.R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.
- [RC02] R.G. Rosenfeld and P. Cohen. Disorders of growth hormone/insulin-like growth factor secretion and action. *Pediatric endocrinology. Second edition*. Philadelphia: Saunders, pages 211–88, 2002.
- [RKA06] J.J. Rodriguez, L.I. Kuncheva, and C.J. Alonso. Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630, 2006.
- [Rot09] T.M. Rotch. A study of the development of the bones in childhood by the roentgen method with the view of establishing a developmental index for the grading of and the protection of early life. *Tran. Assoc. Am. Phys*, 24:603, 1909.
- [RWT75] A.F. Roche, H. Wainer, and D. Thissen. The rwt method for the prediction of adult stature. *Pediatrics*, 56(6):1026, 1975.
- [SB10] A. Schmeling and S. Black. An introduction to the history of age estimation in the living. *Age Estimation in the Living*, pages 1–18, 2010.
- [Sel12] K. Sellgren. Child x-rays cause for concern, says michael gove. <http://www.bbc.co.uk/news/education-17823856>, 2012.

- [Sha48] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.
- [SJC08] J. Shotton, M. Johnson, and R. Cipolla. Semantic text on forests for image categorization and segmentation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- [SOR⁺03] A. Schmeling, A. Olze, W. Reisinger, M. König, and G. Geserick. Statistical analysis and verification of forensic age estimation of living persons in the Institute of Legal Medicine of the Berlin University Hospital Charité. *Legal Medicine*, 5:S367–S371, 2003.
- [Spa95] C. Spampinato. Skeletal bone age assessment. *University of Catania, Viale Andrea Doria*, 6:95125, 1995.
- [Spi00] Martin S Spiller. Introduction to radiology. http://doctorspiller.com/Dental%20radiology/X_Ray_Course.htm, 2000.
- [SSA39] L.W. Sontag, D. Snell, and M. Anderson. Rate of appearance of ossification centers from birth to the age of five years. *Archives of Pediatrics and Adolescent Medicine*, 58(5):949, 1939.
- [ST07] Y. Shi and T. Tsui. An FPGA-based smart camera for gesture recognition in HCI applications. *Computer Vision–ACCV 2007*, pages 718–727, 2007.
- [Sta96] A. Stanton. Wilhelm Conrad Röntgen On a New Kind of Rays: translation of a paper read before the Würzburg Physical and Medical Society, 1895. *Nature*, 53:274–276, 1896.
- [SZC⁺94] B.S. Sharif, S.A. Zaroug, E.G. Chester, J.P. Owen, and E.J. Lee. Bone edge detection in hand radiographic images. In *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, pages 514–515. IEEE, 1994.
- [TA05] A. Tristan and J.I. Arribas. A radius and ulna skeletal age assessment system. In *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, pages 221–226. IEEE, 2005.
- [Tho02] H.H. Thodberg. Hands-on experience with active appearance models. In *Proceedings of SPIE*, volume 4684, page 495, 2002.
- [Tho09] H.H. Thodberg. An automated method for determination of bone age. *Journal of Clinical Endocrinology & Metabolism*, 94(7):2239–2244, 2009.

- [Tho12] H.H. Thodberg. November 2012: Bonexpert licensed to 16 hospitals. <http://www.bonexpert.com/news/42-november-2012-bonexpert-licensed-to-16-hospitals>, 2012.
- [TJC⁺09] H.H. Thodberg, O.G. Jenni, J. Caflisch, M.B. Ranke, and D.D. Martin. Prediction of adult height based on automated determination of bone age. *Journal of Clinical Endocrinology & Metabolism*, 94(12):4868–4874, 2009.
- [TKJP09] H.H. Thodberg, S. Kreiborg, A. Juul, and K.D. Pedersen. The Bonexpert method for automated determination of skeletal maturity. *IEEE Transactions on Medical Imaging*, 28(1):52–66, 2009.
- [TMTM12] S.L. Taylor, M. Mahler, B.J. Theobald, and I. Matthews. Dynamic units of visual speech. In *ACM/Eurographics Symposium on Computer Animation (SCA)*, pages 275–284, July 2012.
- [Tod37] T.W. Todd. *Atlas of skeletal maturation*. Mosby, 1937.
- [TR03] H.H. Thodberg and A. Rosholm. Application of the active shape model in a commercial medical device for bone densitometry. *Image and Vision Computing*, 21(13):1155–1161, 2003.
- [TVA08] A. Tristan-Vega and J.I. Arribas. A radius and ulna tw3 bone age assessment system. *Biomedical Engineering, IEEE Transactions on*, 55(5):1463–1476, 2008.
- [TW62] J.M. Tanner and R.J. Whitehouse. *A New System for Estimating Skeletal Maturity from the Hand and Wrist: With Standards Derived from a Study of 2,600 Healthy British Children*. International Children’s Centre, 1962.
- [TWH⁺01] J.M. Tanner, R.J. Whitehouse, M.J.R. Healy, H. Goldstein, and N. Cameron. *Assessment of skeletal maturity and prediction of adult height (TW3) method*. Saunders, 2001.
- [TWM⁺75] J.M. Tanner, R.J. Whitehouse, W.A. Marshall, M.J.R. Healy, and H. Goldstein. *Assessment of skeletal maturity and prediction of adult height (TW2 method)*, volume 16. Academic Press London, 1975.
- [TWMC75] J.M. Tanner, R.H. Whitehouse, W.A. Marshall, and B.S. Carter. Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height. *Archives of disease in childhood*, 50(1):14, 1975.
- [vRLR⁺01] R.R. van Rijn, M.H. Lequin, S.G.F. Robben, W.C.J. Hop, and C. van Kuijk. Is the Greulich and Pyle atlas still valid for Dutch Caucasian children today? *Pediatric radiology*, 31(10):748–752, 2001.

- [vRLT09] R.R. van Rijn, M.H. Lequin, and H.H. Thodberg. Automatic determination of greulich and pyle bone age in healthy dutch children. *Pediatric radiology*, 39(6):591–597, 2009.
- [VS91] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE transactions on pattern analysis and machine intelligence*, 13(6):583–598, 1991.
- [ZGFP03] R. Zetter, D. Griffiths, M.S. Ferretti, and M.M. Pearl. An assessment of the impact of asylum policies in europe 1990–2000. *Home Office Research, Development and Statistics Directorate*, 2003.
- [ZGL07] A. Zhang, A. Gertych, and B.J. Liu. Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones. *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, 31(4-5):299, 2007.
- [ZHC12] Y. Zheng, M.H.A. Hijazi, and F. Coenen. Automated disease/no disease grading of age-related macular degeneration by an image mining approach. *Investigative Ophthalmology & Visual Science*, 53(13):8310–8318, 2012.
- [Zie09] B. Zielinski. Hand Radiograph Analysis and Joint Space Location Improvement for Image Interpretation. *Schedae Informaticae*, 17(-1):45–61, 2009.
- [ZZCL12] Y. Zhang, B. Zhang, F. Coenen, and W. Lu. Highly reliable breast cancer diagnosis with cascaded ensemble classifiers. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.