Efficient Data Representations for Eddy Current and Ultrasonic Applications

Fredrik Lingvall

28th February 2000

Abstract

Efficient data representation and choice of suitable basis (representation) are the main issues addressed in the thesis. Three separate applications are presented: two are in the field of non-destructive testing, while in the third methods for reconstructing room temperature distributions based on ultrasonic measurements are considered. The need for compact representation arises from the limited amount of data available for the classification of material defects and for temperature estimation, respectively. The main goal that was common for the first two projects was developing software tools of self-learning type, suitable for automatic classification of defects in multilayer aluminum aircraft structures and welds in steel material, respectively. Different NDT methods were used in both cases, eddy current (EC) for the aircraft structures and ultrasonics (US) for welds. A compact data representation was necessary in both cases, due to the low number of examples available for training the classifiers. This was accomplished by compressing the high dimensional data vector, obtained from the measurements, using various truncated bases, such as: wavelets, Fourier, and principal component bases. Efficient data representation was also a crucial part of the third project. The aim was to reconstruct a 2D-temperature distribution in many points of a room, based on a limited number of measurements (time of flight of an ultrasonic wave). To achieve a satisfactory performance strong prior knowledge regarding the reconstructed surface was necessary. The prior knowledge was incorporated by expressing the temperature distribution using suitable base functions. Computer simulations revealed that the principal component basis (specific for the measured data) clearly outperformed other more general base function sets (for instance, wavelets), which confirmed the importance of a suitable data representation.

Acknowledgements

Many persons have supported and encouraged me to complete this thesis, and would hereby like to thank them all. I would especially like to thank my supervisor Dr. Tadeusz Stepinski for all his help and encouragement. I would also like to express my gratitude to Dr. Mats Gustafsson and Tomas Olofsson for their advises and the many valuable discussions. I am also grateful to Prof. Anders Ahlén and Lars Ericsson for reading (and improving) the manuscript, and to Dr. Ping Wu for teaching me about ultrasonics.

I would also like to acknowledge D. Simonet and P. Auge from AEROSPATIALE Suresnes for providing the EC data, and Eider Martinez for helping me with the ultrasonic measurements (and lifting the heavy steel blocks...).

I also acknowledge all the people at the Signals and Systems Group who makes Magistern such a fun place to work at.

Last, but not least, I would like to thank my parents and my brothers for their help and support. Thank you!

Contents

Chapter 1

Introduction

1.1 Background

Non-destructive testing (NDT) does, as the name implies, denote methods for non-destructive detection of material flaws, or characterization of material properties, by means of techniques that do not impair functionality of the inspected material. NDT methods have found many industrial applications and they are vital for such fields as, nuclear power plants, car industry, aircraft and space industry, where the risk of failure is associated with serious consequences. NDT techniques are attractive for production lines since they make possible inspection of all produced parts at relative low cost to assure their quality. Their use in maintenance routines has made possible extending life length of nuclear and aerospace structures.

In many NDT applications it is impossible, due to the large number of inspections, for a human operator to perform satisfactory. Boredom, fatigue etc. will influence the probability of detection to a large extent. The results between different operators may also vary, resulting in inconsistency, which clearly is unsatisfactory [1]. Therefore, it is desirable to automatize the inspection so that the operator has to intervene only when a defect is detected. An automatic system does not have the above mentioned disadvantages provided that it is properly designed.

The performance of an NDT inspection depends on several factors associated with the procedure used for gathering NDT data and the way of processing the acquired data. Here, we will address the following issues that

affect the data processing:

The quality of the NDT measurements Generally, NDT methods are based on indirect measurement using ultrasonics, radiography, eddy current, etc. The measurement quality can be characterized by the amount of information relevant for the NDT inspection as well as the amount of noise and disturbances that deteriorate the results. NDT procedures usually yield only a partial information about material properties, for instance, only a limited information amount is produced in the form of a radiographic image, an ultrasonic B-scan, or an eddy current (EC) scan. If a specimen is inspected for defects, then the geometry of detected flaws is usually interesting. The limited information provided by NDT should be sufficient for determining at least critical features, such as, crack length, diameter of an inclusion, amount of porosity, etc.

Interpretation of the measurements The fact that a relevant information is contained in the measurements does not guarantee the success of the NDT inspection. The measurements have to be interpreted before reasonable decisions are made. The interpretation can be performed either by a human operator or by a machine, or (perhaps most commonly) by a combination of those. Since the measurement procedures often provide the required information about material in the form that is not well suited for human (or machine) interpretation, this information has to be transformed and presented in a form facilitating its interpretation. Learning how to interpret NDT measurements may be a difficult and long process for a human operator that may take several years to master [1]. An algorithm for automatic analysis of NDT data also requires some "training". Generally, machine training can be accomplished in two different ways: 1) A set of explicit rules can be formulated, based on which the machine (i.e. computer program) makes its decisions, and 2) the machine can be trained on a set of known examples to develop the ability of making correct decisions based on this experience (to interpolate based on the seen examples). The first approach is known as an knowledge-based system (KBS) or as an artificial intelligence (AI) approach [2]. The second approach is applied in learning systems, for instance, neural networks.

Note that, information representation is an important issue in both types of machine learning. For the KBS approach, a human designer has to transform the available information in to clear and unique de-

cision rules that cover all situations that can occur. The self learning machine based interpreter has parameters that have to be determined during training (for instance, weight coefficients in neural network). Obviously, if a large number of parameters has to be estimated from the training examples, then a sufficient number of examples has to be available to achieve good generalization performance [3, 4]. Thus, a good representation in learning systems is both compact and informative, since it reduces the amount of examples required for training. In other words, only a limited number of relevant features should be used for training automatic self learning systems aimed for interpretation of NDT data.

From the discussion above one can conclude that two key issues that have to be considered in relation to processing NDT data are, information *representation* and *classification*. Unfortunately, a limited amount of data for testing and evaluation of algorithms is a quite common situation in NDT applications. The main reason for that is the fact that manufacturing realistic artificial defects is in general expensive, difficult, and time consuming.

This thesis reports three different projects where classification and representation are of vital importance. The first two projects are in the field of NDT while the third is concerned with a method for mapping room temperature based on the measurement of the ultrasound velocity in air.

The main goal that was common for the first two projects was developing software tools of self learning type, suitable for automatic classification of defects in multi-layer aluminum aircraft structures and welds in steel material, respectively. The difference lied in the NDT methods used—EC was used for the aircraft structures and ultrasonics (US) for steel. In both cases manufacturing realistic defects with known geometry was associated with considerable expenses which implied a limited number of training examples. As a result of this constraint the way of data representation, or feature extraction, became an important issue. Efficient data representation was also a crucial part of the third project. The aim was to reconstruct a 2D temperature distribution in many points of a room based on a limited number of measurements (time of flight of an ultrasonic wave). To solve this problem strong prior assumptions regarding the reconstructed surface were necessary to achieve a satisfactory performance of the reconstruction. This means that an efficient compact representation was a common issue in all three projects.

From the above discussion it is apparent that when solving such problems we face a trade off between the amount of available *a priori knowledge*

and the amount of training data/number of measurements. Consider the following example as an illustration. Assume that a US method is used to examine flaws in a metal object. The shape of an ultrasonic response from a particular irregular flaw (e.g., crack) depends on the transducer location respective to the flaw (insonification angle). Thus, flaw rotation would obviously change the US response. Assume then that a US measurement can be performed that contains an amount of information which is sufficient for the flaw classification regardless of its orientation. Then, if the flaw rotation is totally unknown, then the classifier used for the flaw classification has to be trained for every possible flaw orientation. It is obvious that, if it is known *a priori* that the orientation of the flaw is limited to some angle interval, then the number of training examples can be limited substantially.

A similar situation occurs in reconstruction of a function in many points from few measurements. Since the number of measurements is lower than the number of parameters to estimate there are many solutions that are consistent with the measurements. Then, in order to find a unique solution prior assumptions must be made regarding the solution.

For self-learning classifiers (neural networks) the above mentioned trade off issue results in a problem where a number of parameters has to be determined from the available data and/or from prior knowledge. If the number of parameters in the algorithm is large a considerable amount of data is required, otherwise, the performance for new (unseen) examples will be poor. Consider for example, a situation where the data is high-dimensional—1000-dim is not uncommon in ultrasonics—and a simple linear classifier is used. This means that at least 1000 classifier parameters have to be estimated. This implies that, at least in theory, several thousands of examples must be available to obtain a reliable classifier. If the number of examples is too low, then the classifier will learn the training examples "by heart", but it will perform poor on unseen examples (poor generalization). To avoid this problem the number of parameters must be reduced or more data must be aquired.

The process of limiting the number of parameters in pattern recognition applications is known as *feature extraction.* The question is then how to compress data, or extract features, without loosing the information relevant for our application. The perhaps most common solution used in signal processing, is to express the data by means of Fourier series. That is, the data is expressed by a finite sum of sine waves. However, typical NDT signals, for instance, ultrasonic pulses (wavelets) found in pulse-echo ultrasonics, have a compact support and sinusoids (which have infinite support) may not be

well suited for the task. The issue of choosing a suitable basis (representation) is a common topic in all the three projects treated here. Although, the reasons for looking for a compact representation are different in the projects, the methods used for obtaining the compact representation are very similar.

As discussed above, the particular choice of classifier (or estimator) depends on the amount of *a priori* information available about the measurements and the measurement process. A relevant question is whether linear algorithms can solve the problem or the solution should be searched in the class of nonlinear algorithms. A general answer is very difficult to give, but *neural networks*, which have the ability to "learn" nonlinear mappings from examples, offer an interesting option when little is known *a priori* about the task.

1.2 Outline of the Thesis

The topic in Chapter 2 is research performed as a part of the Brite Euram III project *Cost Reduction by Advanced Non-destructive Inspection of Aeronautical Structures* (CANDIA). The research is concerned with the automatic detection of cracks in the lower layer of lap-joints in aluminum airspace structures by means of EC. Traditionally the inspections have been performed manually by means of a PC-based portable EC system that enables both evaluating and saving complex-valued EC patterns. The presence of defects is detected by the operator who is expected to analyze all responses to individual rivets saved in the computer. Since one single aircraft (AirBus) may contain several thousands of rivets that are to be inspected there is a considerable risk of human faults. The automatic system discussed in Chapter 2 should eliminate this problem by indicating to the operator only those rivets that produce patterns deviating from normal.

The measurement situation is, however, complicated by the fact that the cracks are located relatively deep (the magnitude of the eddy currents drops off fast with depth), and that there are strong inferences due to the rivets. The rivets may be made of materials with different conductivity which also complicates the task. The data acquisition process is not perfect either, the scanning line may not be perfectly aligned with the rivet row, and some data points is sometimes missing.[1] Using a tailor-made, low frequency probe made defect detection possible. The probe was designed in such a way

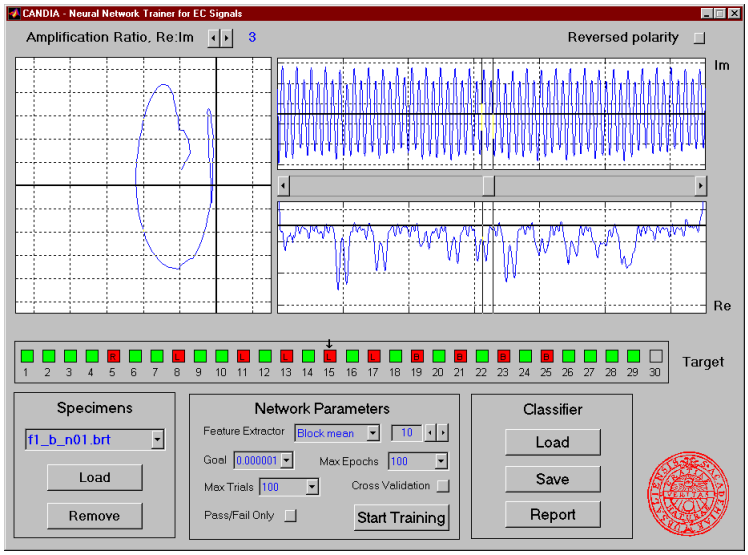
[1] There were some spikes found in the EC data which were due to insufficient acquisition speed of the hardware used.

that the interference due to rivets and the responses from typical cracks were separated maximally in the complex plane (different phase angles) to achieve good suppression of the rivet interference.

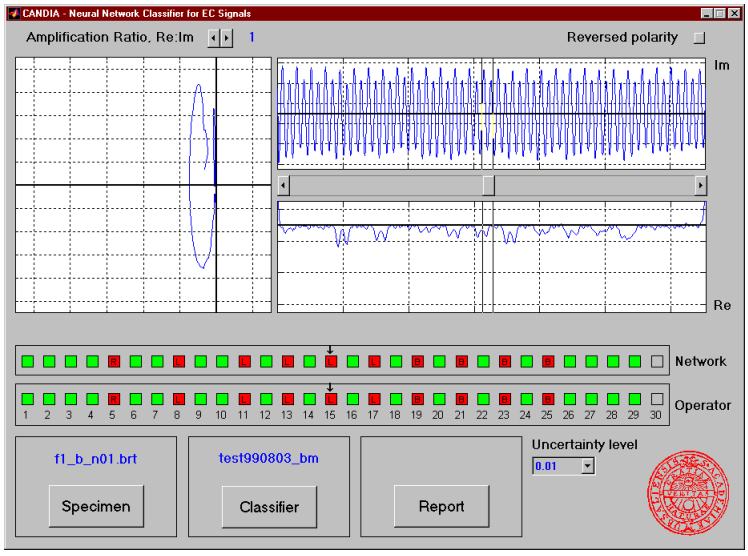
In Chapter 2 we present an automatic classifier of EC signals consisting of a pre-processor and a self learning classifier based on neural networks. The pre-processor has two functions, signal conditioning and feature extraction. Pre-processing was employed to reconstruct missing samples in the signal and to improve robustness of the feature extraction algorithms. Feasibility of four different feature extraction schemes was examined for this particular application: discrete wavelet transform, Fourier transform, principal component analysis (PCA) and a simple block averaging procedure. The classifier was implemented as a standard multi-layer perceptron (MLP) neural network which was chosen due to its ability to compensate for non-linear effects.

The system has been implemented with a user friendly graphical interface enabling both the training and the classification of the MLP. Examples of some interface images are shown in Figure 1.1.

Chapter 3 treats a project sponsored by the Swedish Neuclear Power Inspectorate (SKi) concerned with the characterization of various types of defects in welded steel structures. This is the fourth and last part of the project, using measurements performed on realistic defects implanted in V-welded carbon steel blocks. This was the main novelty compared to the earlier studies which (basically) only treated simulated and artificial defects [5, 6, 7]. The research reported in Chapter 3 is concerned with more realistic flaws, like, sharp cracks, volumetric defects and porosity that where implanted in welds at different positions. The basic task was to characterize the defects based on ultrasonic measurements and to classify them in to two main groups, soft (volumetric, porosity etc.) and sharp (cracks) defects. The ultrasonic measurements where performed using traditional pulse-echo techniques with direct angle beam and angle beam reflected from the back surface. Defect characterization can be regarded as an inverse problem in ultrasonics, based on a number of field measurement around the flaw (scatterer), we have to reconstruct the scatterer itself. In theory, to solve this problem we have to perform field measurements in all positions on a sphere surrounding the scatterer. This is, however, impossible due to practical constraints and we have a limited set of data in the form of B- and D-scans. Our research was based on the hypothesis that, despite the limited information available, it would be possible to distinguish different type of defects using their ultrasonic signatures. This goal was achieved for the simulated and



(a) Training



(b) Classification

Figure 1.1: Graphical user interface.

artificial defects. However, the main problem encountered for the realistic defects was variability of their ultrasonic signatures, depending on their individual features, such as, their geometrical form, position and orientation.

Due to the reasons mentioned above, the number of available defects was very limited (totally 36). The low number of available examples implied that an implicit and powerful data compression/feature extraction must be performed, in order to limit the number of parameters to be estimated in the classifier. Additionally, we had to normalize defect signatures with respect to their depth. By training on a very large number of examples the MLP classifier could, in principle, be trained to compensate for all orientation and position variations but this could not be achieved using the examples that where available.

Being aware of this severe limitation we tried to refine methods for data pre-processing, normalizing and feature extraction to simplify the classification task. These topics are discussed in Chapter 3 where selected measurement results are also presented.

In Chapter 4 we present a selection of methods suitable for reconstructing temperature distributions in a room based on ultrasonic measurements. The project was sponsored by IMRA SA, Sofia Antipolis, France.

The presented measurement principle uses the fact that sound velocity is a function of air temperature. Thus, by transmitting US pulses along known paths, and measuring the time of flight (TOF) between the transmitter and receiver, it is possible to estimate the temperature distribution in a room. In other words, the temperature map is reconstructed from a set of straight line projections—TOF measurements. This problem is similar to computed tomography (CT) widely used in medicine. As mentioned earlier, it is also of interest to find a suitable low dimensional model of temperature distribution in a particular room (office). The reason for that is the limitation imposed by the number of ultrasonic sensors. A low number of measurements (projections) results in a strongly underdetermined problem—many parameters are to be estimated from a few measurements. We investigated several approaches to alleviate this problem. The solution can be achieved by introducing *a priori* knowledge about the temperature distribution. For example, band limiting, or low pass filtering, is a simple way of incorporating prior knowledge. If no assumptions are made about the temperature distribution and the projections are few, then a unique solution will not exist, and hence, there will be many (infinitely many) temperature distributions that are consistent with the measurements.

As mentioned above, the temperature distribution can be expressed using a suitable reduced basis. Then the number of parameters to estimate is fewer which may lead to a unique solution. In Chapter 4 we present a selection of reconstruction schemes, such as, regularized least squares and truncated singular value decomposition methods, as well as the traditional CT methods, like the filtered back-projection algorithm. We also investigate some recursive methods, like algebraic reconstruction techniques (ART) and the multiplicative ART (MART) algorithm. We address separately the basis selection issue and we propose using principal component analysis (PCA) for this purpose.

1.3 Contributions

Parts of the material have been published as reports or in the following journals and conferences:

Chapter 2 Fredrik Lingvall and Tadeusz Stepinski, "Automatic Detection of Defects in Riveted Lapjoints using Eddy Current", In 7th European Conference on Non-Destructive Testing, Copenhagen 26–29 May 1998.

Fredrik Lingvall and Tadeusz Stepinski, "Automatic detecting and classifying defects during eddy current inspection of riveted lap-joints", NDT&E International, Vol 33, No 1 January 2000.

Chapter 3 Fredrik Lingvall and Tadeusz Stepinski, "Ultrasonic Characterization of Defects—Study of Realistic Flaws in Welded Carbon Steel" SKi report 1999.

Tadeusz Stepinski and Fredrik Lingvall, "Automatic defect characterization in in Ultrasonic NDT", accepted for the 15th World Conference in NDT, Rome 15–21 October, 2000.

Chapter 4 Fredrik Lingvall, Mats Gustafsson and Tadeusz Stepinski "Temperature Mapping with Ultrasonic Amenity Sensor—Simulation Results", Signals & System Group Dept. of Material Science Uppsala University.

1.4 Abbreviations

ANN Artificial Neural Network

ART Algebraic Reconstruction Technique

CLI Conventional Line Integral

CT Computed Tomography

DFT Discrete Fourier Transform

DWT Discrete Wavelet Transform

EC Eddy Current

FBA Filtered Backprojection Algorithm

FD Fourier Descriptors

ICA Independent Component Analysis

LS Least Squares

MART Multiplicative Algebraic Reconstruction Technique

M-SMART Modified Simultaneous Multiplicative Algebraic Reconstruction Technique

MRE Minimizing Reconstruction Error

NDE Nondestructive Evaluation

NDT Nondestructive Testing

ON Orthonormal

PC Principal Components

PCA Principal Component Analysis

PDF Probability Density Function

PSF Point Spread Function

R-LS Regularized Least Squares

ROI Region of Interest

SIRT Simultaneous Iterative Reconstruction Technique

SMART Simultaneous Multiplicative Algebraic Reconstruction Technique

SSE Sum Squared Error

SVD Singular Value Decomposition

TOF Time of Flight

TOFD Time of Flight Diffraction

TSVD Truncated Singular Value Decomposition

US Ultrasonic

UT Ultrasonic Testing

Chapter 2

Inspection of Aircraft Lap-joints using Eddy Current

2.1 Introduction

Inspection of riveted lap-joints in aeroplanes is a task which is a part of the regular maintenance procedure for aircrafts. The lap-joint inspection is usually performed manually using eddy current (EC) techniques. The procedure consists of scanning the rivet lines (lap-joints) with an EC probe connected to a PC, where the complex eddy current data is stored. The operator(s) then analyze the reponse from every rivet and determines if the response deviates from the normal. This process is time consuming, and therefore expensive, since a typical commercial aircraft (AirBus) contains tenths of thousands rivets. There exists also a considerable risk of human faults during analyzis of the EC data, due to the large number of rivets involved. Therefore, it is desireble to have tools that could detect, and perhaps classify, the abnormal responses so that the operator only need to analyze those rivets containing possible defects.

The topic of this chapter is the design software which has the desired properties described above. The goal was to develop a signal processing tool for automatic detection and classification of defects in riveted aeronautical structures by means of eddy current (EC) testing. The proposed method employs an effective pre-processing followed by a data compressing feature extractor that significantly reduces the data volume fed to a neural network

classifier. Section 2.2 describes data acquisition and selection of examples for training. Section 2.3 discusses the pre-processing employed and Section 2.4, the feature extraction methods used in the comparison. In Section 2.5 the neural network classifier is described, and finally the last two sections present classification results and a discussion.

2.2 EC Data

The used EC data was acquired by AEROSPATIALE Suresnes using a single frequency EC instrument. A deep penetrating probe, specially developed for this application, was used to detect cracks located in the lower layer of the lap-joints. The defects were manufactured mechanically as a result of a large number of fatigue cycles with forces simulating loads in real aircraft. Due to the direction of the applied forces the cracks appeared only along the rivet line and were located on the left and/or the right side of the rivets, see Figure 2.1.

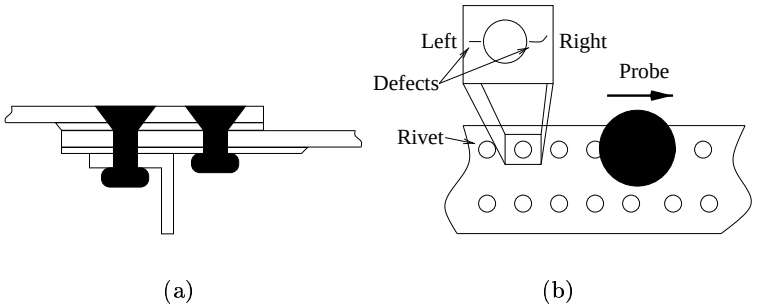


Figure 2.1: (a) Cross section of a lap-joint and (b) Defect location and probe movement in a lap-joint.

The EC inspection was performed using a simple mechanical scanner guiding the probe along the rivet line. The scanner was operated by hand and the EC data was digitized and stored in a personal computer. The probe was operated at a very low frequency (in the range of 1 kHz) to achieve a sufficient penetration depth. Figure 2.2 shows example data acquired during inspection of one rivet line. The strong periodic component in the signal interfering detection of the cracks originates from the rivets. The multi-coil EC probe was designed in this way that its response of each rivet consists of

two full periods with different amplitudes. The periodic component enables quite accurate determination of the probe position relative the inspected rivet.

The probe is designed in such a way that the rivet response and the defect response for the operating frequency are almost perpendicular in the impedance plane which can be seen in Figure 2.2(b).

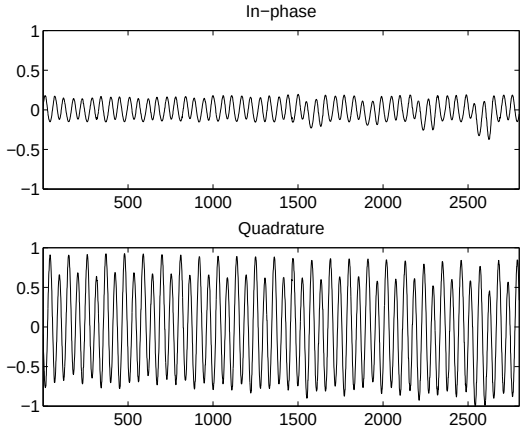
Totally 46 complex valued EC data vectors (each one containing serval rivets) containing signals corresponding to different sizes of defects were available for the tests. This data set has been split into three smaller subsets used for different purposes. The first subset was used for training the classifier, the second was used for its evaluation and the third contained all measurements with defects that were considered as large. The last group was not used at all because of two reasons: 1) large defects can easily be detected with simple thresholding and 2) the used EC instrument saturated for very large defect amplitudes. From those 46 complex data vectors approximately 900 examples were extracted (one example for each rivet). Most of these examples came from defect-free rivets, only 133 signals were defect responses. Thus the number of EC data available for training the neural network was relatively low which had to be taken into consideration when choosing the classifier architecture. This was one of the main reasons for including a powerful feature extraction algorithm processing the EC data before the classification.

2.3 Pre-processing

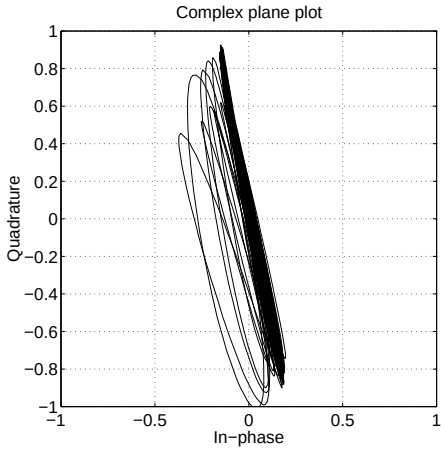
A number of pre-processing steps were required before feature extraction and classification were possible. The EC signal had to be normalized in a proper way. This was accomplished by using the rivet response in data, since it should be the same regardless which lap-joint the data came from. This is due to the fact that the distance between the rivets is the same (22 mm) in all samples and the data were acquired with the same probe (same EC instrument).

2.3.1 Median Filtering

The EC signal was first filtered in a moving window including a three-point median filter. This step was required only because of the imperfections of the acquisition device used for data digization which introduced large amplitude



(a)



(b)

Figure 2.2: Example EC data: (a) in-phase and quadrature component and (b) complex plane plot.

sparse spikes in the signal.

2.3.2 Rotation of EC Signals

As mentioned earlier the defect and rivet responses were separated in phase as much as possible by choosing a proper probe and by using a suitable operating frequency. However, this does not mean that the direction of the defect responses in the complex plane was known accurately. Generally, the absolute direction of the EC signals depends on the phase setting of the instrument and is often known only approximately. A standard procedure, applied before the inspection is to rotate the EC signal so that the defect responses lie along the in-phase direction while the disturbance (here the rivet response) in the quadrature direction (or vice versa). To achieve maximum suppression of the rivet responses a fine tuning procedure has been applied to the digitized signals. The method used to rotate the EC pattern was based on the observation that the dominating part of the signal energy was due to the periodic rivet response. That is, the complex EC signal denoted by a column vector $\mathbf{x}$ was rotated ϕ radians so that the energy in the imaginary component of the signal was maximized according to Eq. (2.1).

$$\max_{\phi} \parallel \mathrm{Im}\{\mathbf{x}e^{i\phi}\} \parallel^2 \tag{2.1}$$

Figure 2.3 shows an example EC signal before and after the rotation.

2.3.3 Normalization

Amplitude normalization, performed for all data sets, was based on the assumption that the responses from the rivets should have equal amplitude. Mean value of the positive part of the rotated signal quadrature was used as a robust amplitude estimate for the normalization. In other words a mean value of a half wave rectified rivet component was used as a measure of the EC signal amplitude.

One drawback of this method was that presence of large defects also affected the amplitude of the quadrature component and in this way degraded performance of the normalizing procedure. Therefore large defects were detected by thresholding and then these respective parts of the signal were removed before normalization.

The in-phase component after the normalization usually had a small dc-bias. This bias was removed by subtracting a value corresponding to

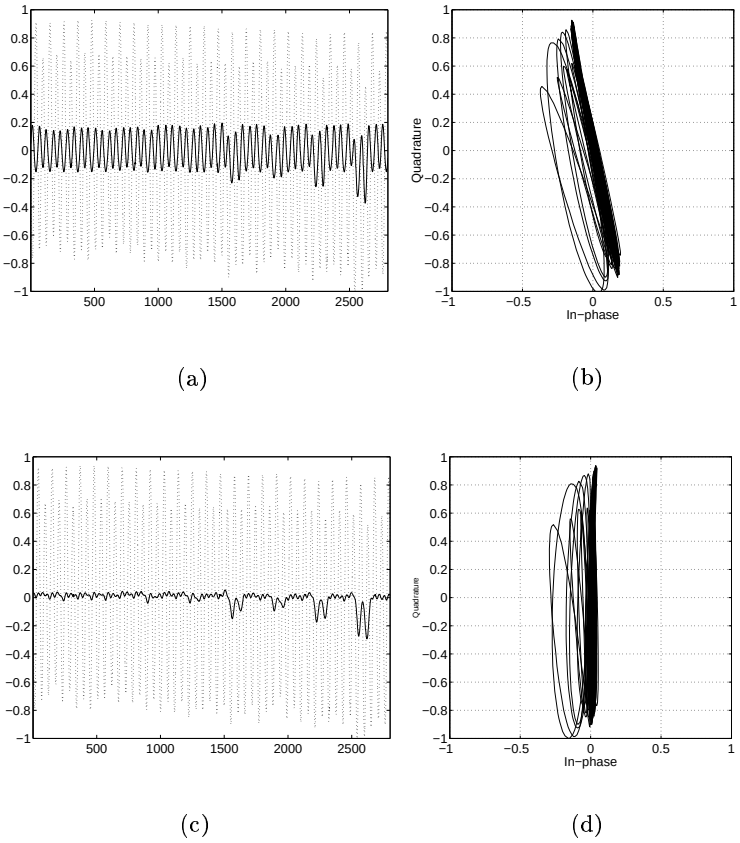


Figure 2.3: EC signal (solid - real component, dotted - imaginary component) as a function of probe position, (a) the original signal, (c) the rotated signal. The same signal as a complex valued contour, (b) the original signal, and (d) the rotated signal.

"the most likely amplitude" in the signal. The most likely amplitude value was found by estimating smoothed histogram of the signal and finding the amplitude corresponding to its maximum.

2.4 Feature Extraction

Feature extraction denotes in pattern recognition a procedure of mapping the original measurements into a relatively short vector representing features relevant for the classification[4]. The main purpose is mainly reducing the number of parameters (inputs) in the classifier. This was a vital operation in our case since the number of examples was rather limited. A large number of inputs to the neural network when only a few training examples are available results in a substantial risk for over-fitting. This means that the classifier "learns" the training examples well but performs poor on unseen data (poor generalization performance).

All data reduction schemes chosen for our application can be described by a linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{2.2}$$

where $\mathbf{A}$ is a "compression matrix" and vector $\mathbf{x}$ represents the measurements (assumed to have zero mean). The reason for choosing a linear method, which should be always tried first, is the existence of well established methods for selecting the transformation matrix $\mathbf{A}$. Of course, in many practical applications the relevant signal features can be obtained only by non-linear mappings of the measurements. In this application it is not clear whether the classification task requires really "non-linear" feature extractor. Our solution is to use a linear feature extractor and a non-linear classifier.

Care must be taken when choosing the matrix $\mathbf{A}$ to preserve the features in data that are relevant for defect classification. For example, position is an important feature if the defects location is to be preserved.

Four types of feature extractors, applied to a windowed signal, have been investigated here, block mean values, Fourier descriptors, a discrete wavelet transform, and a method based on principal component analysis (PCA).

2.4.1 Analyzing Window Centering

All feature extraction schemes were applied in a window precisely centered around each rivet. Precise centering was very essential and all subsequent processing relied on it. The classification was performed for signal features extracted for a single position of the analyzing window with respect to each rivet.

The window centering was based on the observation that the rivet response is a periodical sinusoidal-like signal with two positive peaks and one of the peaks is significantly larger than the other one (this feature resulted from the probe design). Thus, the rivet response (imaginary component of the EC signal) includes the information required to position the analyzing window around each rivet, which is illustrated in Figure 2.4. Since defect

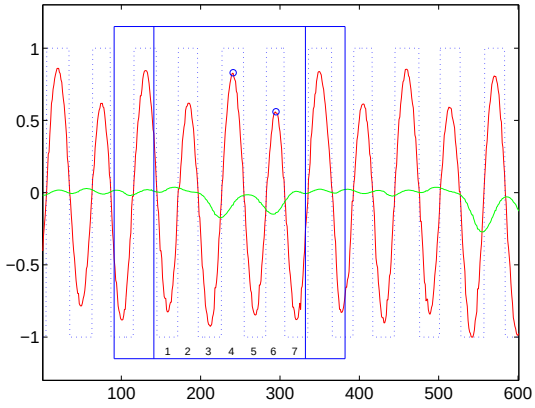


Figure 2.4: Centering of the analysis window using quadrature component of the EC signal. Horizontal axis shows sample number.

responses from the adjacent rivets overlap the window width was increased slightly (by 50 samples) after the centering. Also, the signal vector in this window was down-sampled from about 300 samples to a constant length of 128 samples. This was required by the subsequent feature extractors (e.g., wavelets) that operate on input vectors of dyadic length (power of two).[1] Furthermore, the data vector was windowed by a soft rectangular window (with sigmoid-like flanks) to remove the edge effects.

[1] A length of 256 could also have been used, but this did not improve the classification much so a length of 128 was used as a compromise.

2.4.2 Block mean

This is a very simple technique consisting in splitting the analysis window into a number of intervals (blocks) and calculating a mean signal value for each block. If the measurement vector $\mathbf{x} = \begin{bmatrix} x(1) & x(2) & \cdots & x(N) \end{bmatrix}$ is divided in M blocks ($N = 128$ here), then the elements of the feature vector $\mathbf{y} = \begin{bmatrix} y(1) & y(2) & \cdots & y(M) \end{bmatrix}$ can be expressed as

$$y(i) = \frac{1}{M} \sum_{j=(i-1)N/M+1}^{iN/M} x(j) \qquad i = 1, \cdots, M. \tag{2.3}$$

As Eq. (2.3) indicates blocks uniformly distributed within the analysis window were used in our case, this was the simplest solution, see the discussion in Section 2.7 for comments.

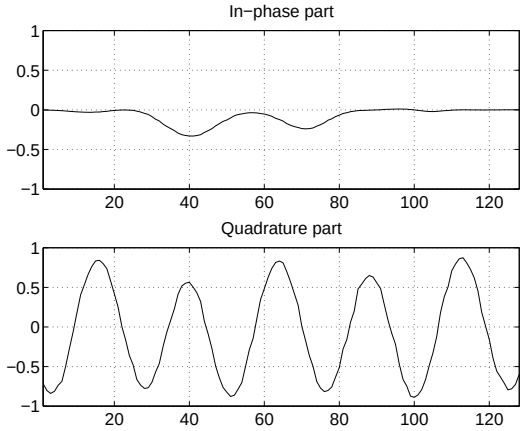
2.4.3 Fourier Descriptors

This is a standard technique used in image processing for recognition of different contours in digital images. The idea is to expand the contour of an object in Fourier series and use a limited number of Fourier coefficients, called Fourier Descriptors (FDs), as features for recognizing the object [8, 9, 10]. In our case the complex valued EC Lissajous patterns, defined by their real and imaginary components were used as an object contour, see Figure 2.5. The expansion was performed in the analysis window using the discrete Fourier transform (DFT).

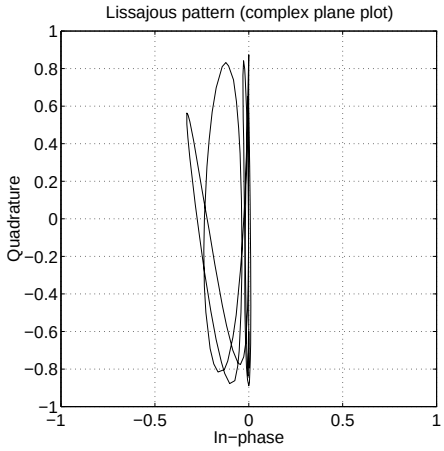
2.4.4 Wavelets

The *discrete wavelet transform* (DWT) forms an orthonormal basis which has several interesting features for this application. One of the most important features of DWT is fact that the basis consists of local functions with different positions and scales[11]. This feature is very useful in our case since it enables determining particular scales where the EC signal has significant energy. For instance, energy at small scales is mostly due to noise, so by removing small scale coefficients in y both noise reduction and data compression can be achieved.[2]

[2]For more sophisticated methods for signal de-noising using wavelets see for example[12].



(a)



(b)

Figure 2.5: EC response to an example "left defect": (a) in-phase and quadrature component, and (b) contour plot (Lissajous pattern).

The basis functions in the DWT consist of local functions with different positions and scales. All basis functions are constructed from the same template function $\psi(n)$, called *mother wavelet*, using the formula

$$\psi_{j,k}(n) = 2^{j/2}\psi(2^j n - k) \qquad n = 1, 2, \ldots, N. \tag{2.4}$$

Using these basis functions any vector $\mathbf{x}$ can be expressed as the linear combination

$$x(n) = \sum_{j,k} w_{j,k}\psi_{j,k}(n). \tag{2.5}$$

The wavelet coefficients are then given by the inner product

$$w_{j,k} = \mathbf{x}^T \boldsymbol{\psi}_{j,k}. \tag{2.6}$$

If the basis functions (vectors) $\boldsymbol{\psi}_{j,k}$ are collected in a matrix, and the coefficients $w_{j,k}$ in a vector, then a linear transformation of the form defined by Eq. (2.2) is obtained. There exists a very efficient computation scheme enabling evaluation of the inner product Eq. (2.6) using only $\mathcal{O}(N)$ operations (for comparison the FFT algorithm is an $\mathcal{O}(N \log N)$ algorithm). However, number of computations is really not a critical issue in this application.

The mother wavelet chosen for this application is the *Coiflet 2* wavelet, which is a rather smooth wavelet suitable for modeling the used EC signals. Figure 2.6 shows the first (largest scale) basis functions of the *Coiflet 2* mother wavelet.

2.4.5 Principal Component Analysis

The basis functions used for DWT are constructed without using any particular knowledge about the analyzed data. For each specific application one should of course select the mother wavelet which seems to fit the analyzed signal best—which is exactly what has been done in the previous section.

Another approach is to construct a basis which, given a data set, yields the best compression. In other words, the basis which results in the smallest error (in the least square sense) using only m of the N coefficients in $\mathbf{y}$ is chosen for the reconstruction of x, where $m < N$.

Using geometrical interpretation, the coordinate system is rotated so that the axes in the new system point in the directions of the largest variances of the analyzed data ($\mathbf{x}$ can be regarded as a stochastic vector). These

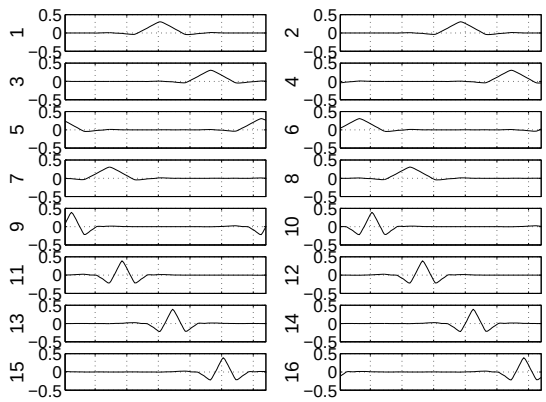


Figure 2.6: The first 16 basis functions of the Coiflet 2 mother wavelet in the DWT used for compressing EC data.

axes are referred to as *principal axes*. The idea behind compression is to represent the data using only a limited number of axes for which the variance is sufficiently large. The basis functions which fulfill this criterion are eigenvectors of the covariance matrix of $\mathbf{x}$ [4]. This method has several attractive features; all elements in $\mathbf{y}$ are uncorrelated (independent if x is Gaussian) which means that the covariance matrix of y is diagonal with the eigenvalues on the diagonal. If the eigenvectors are sorted in a descending order according to their eigenvalues, and only the m first ones are used, then the best possible representation is obtained with only m basis functions for all m orthonormal basis vectors. This procedure is known as *principal component analysis* (PCA).[3]

The procedure can be summarized as follows:

1. Calculate the sample covariance matrix of $\mathbf{x}$.
2. Perform eigenvector decomposition.
3. Sort the eigenvectors $\boldsymbol{\phi}_i$ according to their eigenvalues λ_i
4. Use the $\boldsymbol{\phi}_i$ with largest eigenvalues (variances). That is,

$$\mathbf{y} = \boldsymbol{\Phi}^T \mathbf{x}. \tag{2.7}$$

[3]The procedure of expanding a vector $\mathbf{x}$ using the eigenvectors is also known as the *discrete Karhunen-Loève expansion.*

The reconstruction is then defined as

$$\mathbf{x} = \sum_{i=1}^{m} y_i \boldsymbol{\phi}_i = \Phi y = \mathbf{A}^T \mathbf{y}. \tag{2.8}$$

were $\boldsymbol{\phi}_i$ are the columns in $\boldsymbol{\Phi}$.

Using the procedure above one obtains an intelligent system which from examples—in our case EC data—extracts an optimal set of orthogonal basis functions.

It should be mentioned that these basis functions do not have to be local like in the DWT case, in fact they have rather global character in our case. Figure 2.7 shows the 8 first basis functions obtained using PCA for EC data.

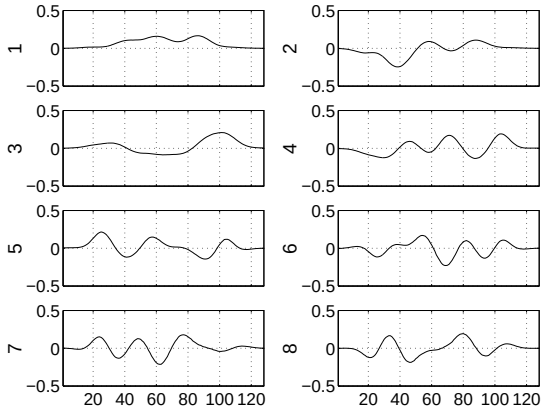


Figure 2.7: The first 8 basis functions obtained by principal component analysis of the EC data.

2.5 Classification

The classifier was implemented as a two-layer artificial neural network (ANN). The ANN had two outputs: one for defects on the right side of the rivets, and one for defects on the left side. After preliminary simulations it appeared that using only 3 neurons in the hidden layer was sufficient which confirmed high efficiency of the feature extraction schemes and implied a fast training due to a low number of coefficients in the ANN (a desirable feature if the amount of data is low).

The data set was (as described earlier) divided into two subsets: one set was used for training, and the second for cross-validation and evaluation. All data with large defects were removed from both the training and the evaluation sets.

2.6 Results

A large number of ANNs (1000 for each feature extractor) was trained using the AEROSPATIALE training data set. The optimization algorithm used to train the nets was very powerful (Levenberg-Marquardt [13]), but the risk of getting trapped in local minima seemed to be rather large, therefore a large number of networks, with random starting weight, was used during the training. A detection was flagged if one the outputs of the ANN exceeded the level 0.5.

All feature extractors had approximately the same performance on the evaluation data used if a sufficient number of coefficients was used. All the extractors resulted in 2–3 missed detections (of 68), 5–8 false detections (of 640) and misclassification of 6–9 detections (left instead of right crack for example)—it is difficult to say which feature extractor that has the best performance, due to the low amount of data available.

To illustrate the extractors performance defect responses from four adjacent rivets is presented in Figure 2.8. The rivet corresponding to the signal shown in Figure 2.8(a) was defect free, while the responses in the (b), (c) and (d) were due to rivets with cracks. The cracks in (b) and (c) were detected and the crack (d) was missed by all methods. By comparing the responses from the defect-free rivet (a) and the missed rivet (d), one can conclude that the magnitude of the respective signals is similar for both rivets. That is, the defect response in (d) has the amplitude in the order (or lower) of the noise level, which explains why this defect was missed.[4]

The number of coefficients required to accomplish the classifier performance described above varied between the different methods as shown in Table 2.1.

Figure 2.9 shows an example of reconstructions of an EC signal, containing a small disturbance, using the number of coefficients depicted in Table 2.1, for the DWT, PCA and DFT. The disturbance has dissapeard

[4]Note that the measurement vector is windowed by a soft rectangular window which explains why the EC response is zero at the beginning and the end of the vector.

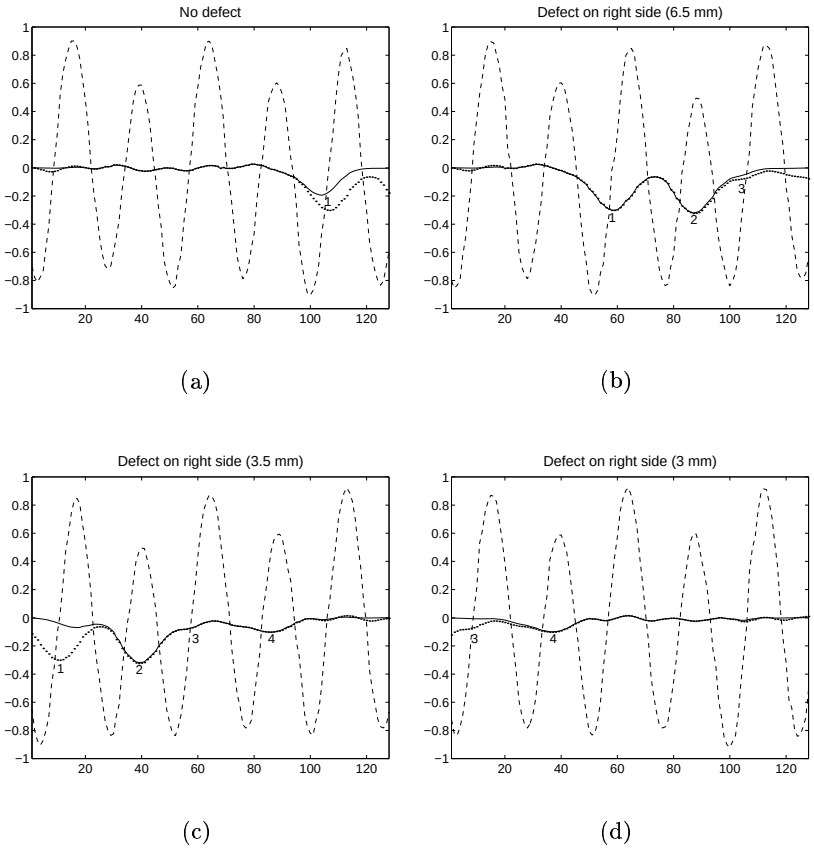


Figure 2.8: In-phase components from four adjacent rivets, (a) is the leftmost rivet and (d) is the rightmost one. Dotted curves show un-windowed data and the dashed curves is the rivet response signal (quadrature part of the EC signal). The rivet corresponding to the signal in (a) is defect-free, while the responses in the (b), (c) and (d) are coming from the rivets with a 6.5 mm crack, a 3.5 mm crack, and a 3 mm crack, respectively. There should be two "bumps" for each defect due to the probe design. The defects in (b) and (c) were detected while the defect in (d) was missed. Note that due to overlap of the analysis window the "bump" in the right hand side of the (a) marked "1"appears as the left bump in (b), and the right bump in (b) marked "2" appears as the left bump in (c) etc. Since neither "5" nor "6" bump can be distinguished from noise the defect in (d) was missed.

	Block mean	FD	Wavelet	PCA
# of Coeff.	12	12 (6 complex)	15	6
# of misses (false alarms)	2 (5)	3 (7)	3 (8)	3 (5)

Table 2.1: Classification results for the employed feature extraction methods. The performance seems (slighty) better for the block mean and the PCA methods, but it is difficult to draw any general conclusion from this low number of missed detections.

in the reconstructions, indicating that the robustness has been improved by performing proper feature extraction.

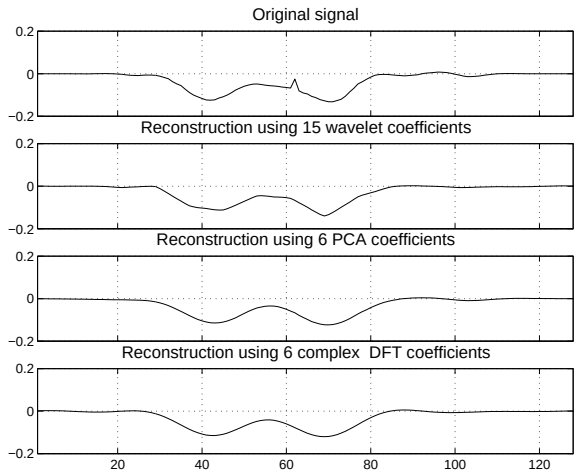


Figure 2.9: Reconstructions of an EC signal using reduced DWT (Coiflet 2), PCA and DFT bases. The original signal contain a small disturbance in the middle. This disturbance has disappeard when the recontructions is performed using truncated bases, indicating that the robustness has been improved.

2.7 Conclusions

The proposed method to detect and classify defects in lap-joints during EC inspection presented in the chapter appeared to perform well on the given data. Due to the relatively low amount of EC data available it is difficult

to perform a quantitative comparison of the classification performance of the different proposed pre-processing schemes. Generally, all of them performed well on the available data set. The main difference between them is the number of coefficients needed to achieve a satisfactory classification performance. Below some comments related to each pre-processing method are presented:

Block mean Taking simple averages in sections worked well. One peculiarity was though observed, namely that, increasing the number of blocks did not necessarily improve the classifier performance. This was probably due to the length and location of the blocks in the analyzing window. Blocks with the equal lengths was simply used, but if the block lengths and locations was chosen in a more sophisticated manner, then even better performance would probably achieved.

Fourier Descriptors This method is the only one that utilizes the full complex-valued EC signal. The data compressing performance was in the same order as the Block-mean and the DWT method.

Wavelet The wavelet method, was surprisingly, the method that needed the largest number coefficients. To achieve the desired performance the two largest scales were needed, which resulted in 15 basis functions (cf. Figure 2.6).[5] Better performance would probably be possible if an adapted wavelet basis[14], or a basis chosen from a WavePacket library[11, 15], was used.

PCA The PCA method achieved very good compression performance, only 6 coefficients were needed for the classification. This can be explained by using an adapted basis, deduced from the training examples. The use of an adapted basis is a very powerful method since it enables a very efficient data compression and a reliable elimination of large errors (outliers) at the same time. Outliers appear when the basis is not well matched to all measured data, which means that some part of data has features different from the majority, for instance due to signal saturation. Outlier detection can be performed by observing the reconstruction error. That is, the fact that the reconstruction error is large may indicate that the observed data probably originates from some other family than that the basis was constructed from. It is worth noting that due to the fact that PCA is based on discarding

[5] The two largest scales of the used DWT comprise 16 basis functions, but the last one was almost totally weighted out by the used windowing procedure.

the eigenvectors corresponding to the smallest eigenvalues does not guarantee correct representation of the significant features in all cases. It may namely happen that the discarded directions contain information essential for the classification. However, in this application PCA worked properly resulting in a very good classification performance.

Chapter 3

Characterisation of Defects in Welded Carbon Steel

3.1 Introduction

Characterization of defects is an important issue in many industrial applications. Knowledge of defect properties (geometries) can save substantial amounts of money due to the high costs involved in replacing parts in many applications. The location and size of a defect may, for example, not be critical for a particular application, and this knowledge can reduce maintenance costs substantially. In many applications it is also vital to be able to inspect (monitor) parts before a part actually breaks. Examples of such critical applications can be found in nuclear power plants and in the aircraft industry.

Ultrasonic inspection is one technique that enables defect monitoring, and defect sizing, of specimens non-destructively. However, the measurements typically obtained with ultrasonic (US) methods are complex to evaluate since the response depends on many factors, such as lobe characteristics of the transducer, the transducer bandwidth and center frequency, angle of inclination, depth and orientation of the defects etc. Today only very experienced operators are able to perform full evaluation of the results, especially for coarse grained materials (stainless steel). Therefore, it is desirable to have tools that can support operators in tasks, such as, flaw sizing and classification.

The goal of the research presented in this chapter has been to develop a software package that, by means of signal processing, could support an operator in making decisions regarding defect characteristics using pulse-echo ultrasonic measurements. The idea was to study US signatures from measurements, performed at our lab, from a carefully chosen selection of "real" defects. A suitable set of signal processing tools should then be developed based on the US characteristics of different flaw types, and the experiences gained from the measurements. Previous research [5, 6, 7], performed using simulated and artificial defects, has shown the feasibility of such an approach.

The selected defects were of types similar to those commonly encountered in real V-welds. The number of defects used was 36. A much larger number would have been desirable but the production cost for "real" defects is very high since the geometry of the used defects must be known. One must also be aware that, due to physical restrictions, it is only possible to insonify the defects from a limited number of views. The inspections are constrained to be performed from the front and the back surface of the test block. Normally, the upper rough surface of the weld also restricts inspection which limits the obtainable information even further.

The approach taken here consists in, first carefully analyzing the US responses, using different transducers, and then examining if there are features which are unique for individual classes of defects, and hence, could be used for characterization. The second step is to develop methods, or algorithms, for extraction of these features in a format suitable for a classifier. This is an important step due to the very low number of examples available. Note that the "classifier" can be either a human operator or a dedicated software. If the software approach is chosen, then there is an apparent need of incorporate strong *a priori* knowledge since the number of examples is very limited. A reasonable goal, which was adapted here, is to select two classes, one containing sharp defects (various types of cracks) and one including soft defects (slag inclusion and porosity). Note that, even though only two classes are used, there are still very few examples of each class available for training.

Section 3.2 includes a short description of the carbon steel blocks and the "natural" flaws that were implanted in them. The main part of this chapter, Section 3.3, describes the B-scan (and D-scan) measurements that have been performed and discusses the features extracted from the measurements. In Section 3.4 the algorithms for position estimation, region of interest selection, feature extraction, and depth normalization are presented. This section also contains a discussion concerning the different defect classes

proposed and a comparison of signal features between different flaw types. At the end of the section there is also a comparison of artificial flaws contra natural flaws. Finally, Section 3.5 gives the conclusions.

3.2 Realistic Test Blocks

Four blocks, each with 9 various flaws, were designed in collaboration with, and manufactured by Sonaspection International Ltd. All blocks have dimensions 42 mm × 400 mm × 600 mm and consist of two carbon steel plates, welded together (V-weld). The defect types and sizes manufactured in the blocks are summarized in Table 3.1. From Table 3.1 it can be seen that the

Flaw Type	Size in mm	No of flaws	Abbreviations
Root Crack	3	3	RC
Root Crack	6	3	RC
Lack of Side Wall Fusion	3	3	LOF
Lack of Side Wall Fusion	7	3	LOF
Side wall crack	3	3	SWC
Side wall crack	7	3	SWC
Center line crack	3	3	CC
Center line crack	6	3	CC
Slag	3–6	3	S
Porosity	6–10	3	P
Over Penetration	3–5	3	OP
Lack of Penetration	2–25	3	LOP

Table 3.1: Flaw list.

flaw population consists of 24 sharp flaws (various types of cracks and lack of fusion) and 12 soft type flaws (slag, porosity and over penetration). Closer analysis shows that there are three different types of cracks characterized by various sizes, angles and locations. The cracks were manufactured by mechanical fatigue and were implanted by semi-direct insertion (created before the welding process or at a pre-determined stage during welding). There are also natural sharp flaws in the form of lack of side wall fusion. This gives an idea of spread in the sharp flaw class which should also result in the variation of their ultrasonic signatures. The soft defects, on the other hand, should by their nature result in similar ultrasonic responses, independent of their orientation.

The test blocks have been subjected to careful ultrasonic inspection in our lab and all the defects were localized according to the reports from the manufacturer (copies of the block drawings are shown in Appendix 3.A).

3.3 Test Block Measurements

3.3.1 Transducers

The contact US inspection of the blocks has been performed using a mechanized scanner and a digital ultrasonic system based on a Saphir PC board. B-scans for each flaw were acquired and a flaw data base was created.

Two miniature screw-in transducers from Panametrics, with center frequencies 2.25 MHz (type V539-SM) and 3.5 MHz (type A545S-SM) were used in the (shear wave) contact inspection: Both transducers had nominal

f_0 [MHz]	B [MHz] (-6dB)	Angle [Degrees]	Producer
2.25	89%	45	Panametrics
2.25	89%	60	Panametrics
2.25	89%	70	Panametrics
3.5	58%	45	Panametrics
3.5	58%	60	Panametrics
3.5	58%	70	Panametrics

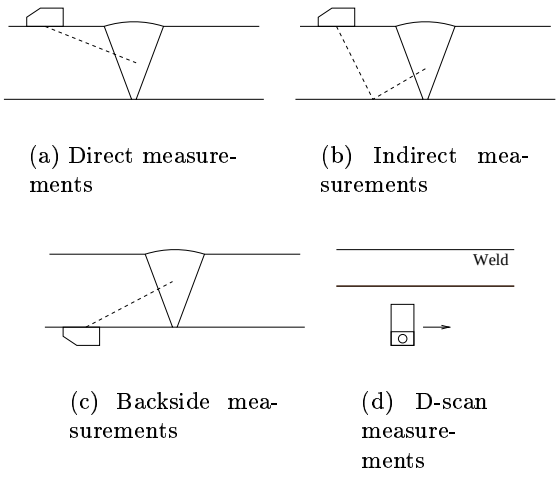
Table 3.2: Transducer list.

element size 0.5" (13 mm) and were assembled by screwing directly into miniature angle beam wedges type ABWM-5T, also from Panametrics. Six different angle beam transducer configurations, listed in Table 3.2, were created in this way. The advantage of this solution is obvious; by using the same active element the obtained transducers have very similar characteristics. Since ultrasonic response of a particular flaw is determined both by the flaw type and by the transducer characteristics it is essential for defect characterization to keep transducer characteristic as constant as possible.

3.3.2 Measurement Setup

Four different scanning methods were used. The aim was to make direct measurements from the top side of the steel blocks as shown in Figure 3.1(a).

However, for smaller angles direct measurements were obstructed by the upper surface of the V-weld. In such cases indirect measurements, or measurements from the backside were performed instead, which is shown in Figure 3.1(b) and (c). Another reason for making indirect (or backside)



(a) Direct measurements

(b) Indirect measurements

(c) Backside measurements

(d) D-scan measurements

Figure 3.1: Measurement setups.

measurements was low amplitude of the reflection obtained in the direct measurement. This was due to the unfavorable angle between the transducer main beam and certain flaws (sidewall cracks, for example).

The fourth scanning method, referred to as D-scans, is shown in Figure 3.1(d). In D-scans the probe is moved along the weld side-wise. D-scans are interesting because they reveal how the defect response varies along the defect. It is also interesting to see the response from the weld itself, both with and without a defect present. Typically the shape of the weld varies spatially and D-scan shows this variation rather clearly.

3.3.3 Measurements

The performed measurements are displayed in Table 3.3 and Table 3.4. The measurements consist of B- and D-scan data matrices and the total number of measurements are 2×133. The main part of the data comes from the welded steel blocks described above, but new measurements have also been performed, for comparison, on two aluminum blocks with artificial defects

also used in a previous project [7]. In addition to these measurements all flaws have also been subjected to manual inspections. The artificial flaws in Table 3.4 named `SBH` are side-drilled holes, and the ones named `S` are cracks (notches).[1]

3.3.4 Measurement Results

In this section a number of B-scans from each defect type is shown for illustration. They were selected so that both common features and feature variations are represented for each defect type. Note also, that some of the images contain responses from non-defect parts of the weld, like the top or bottom surface of the weld or the steel-weld junction. These echos are explained (if possible) when they are encountered. In the figure titles the name of the data files are given. An example is `p28b 1 3 5`, where `p28` means porosity flaw 28, `b` means backside measurements, `1` indicates that the flaw is located in test block PL4501, and `3 5` is the used transducer frequency (3.5 MHz). The B-scans presented in the following subsections are from measurements with the 3.5 Mhz transducer. The reason for showing the 3.5 Mhz transducer only is that the measurements is performed on carbon steel blocks with very little grain noise. This implies that the 2.25 MHz and the 3.5 MHz transducers should give similar results (which also was verified), with the exception that the 3.5 MHz transducer gives higher resolution due to the shorter wave length (both transducers have approximately the same bandwidth).

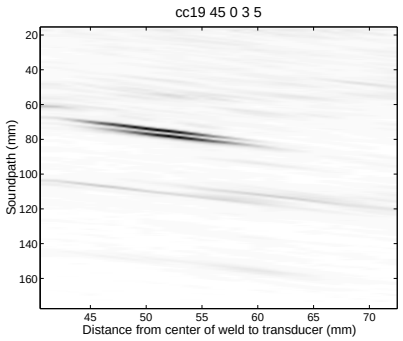
Center Cracks

The center cracks found in the steel test blocks were either vertical or slightly tilted. Figure 3.2 shows indirect measurements using the 45-degrees transducer. One can see that there are two rather strong peaks (too strong to be diffraction echos) in all three B-scans. An unambiguous explanation for the presence of the second echos are difficult to find, but one explanation may be that they originate from the structure of the cracks implanted into the weld. These two peaks do not always occur in signals from center cracks, an example is shown in Figure 3.2(a), where the crack has been scanned from the other side of the weld. Here only a single peak is seen. The double echos are also less pronounced if a higher angle probe is used. Figure 3.3 shows B-scans of the same defects obtained with a 60 degree transducer, and only

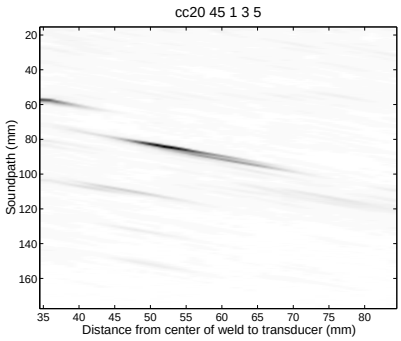
[1] For a more detailed description of the aluminum blocks see [7].

Defect	Direct			Indirect			Backside			D-scan		
	45	60	70	45	60	70	45	60	70	45	60	70
RC 1		•	•								•	
RC 2		•	•								•	
RC 3		•	•								•	
RC 4		•	•								•	
RC 5		•	•								•	
RC 6		•	•								•	
LOF 7					•		•	•			•	
LOF 8					•		•	•			•	
LOF 9					•							
LOF 10					•		•	•			•	
LOF 11					•		•	•		•		
LOF 12					•							
SWC 13					•		•	•	•		•	
SWC 14							•	•			•	
SWC 15					•		•	•			•	
SWC 16					•		•	•			•	
SWC 17				•			•	•		•		
SWC 18				•				•	•		•	
CC 19			•	•							•	
CC 20				•				•	•			•
CC 21				•				•	•		•	
CC 22				•					•			•
CC 23				•				•			•	
CC 24			•									•
S 25				•			•	•			•	
S 26				•				•	•		•	
S 27				•			•	•		•		
P 28				•			•				•	
P 29				•			•	•			•	
P 30				•			•	•		•		
OP 31		•									•	
OP 32		•									•	
OP 33		•									•	
LOP 34		•									•	
LOP 35		•	•								•	
LOP 36		•	•								•	

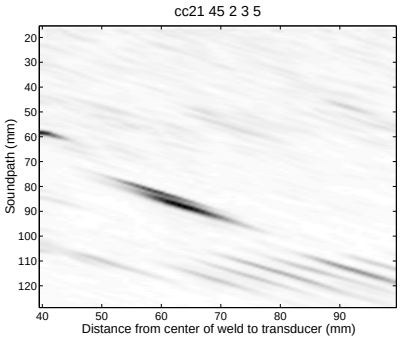
Table 3.3: B-scan and D-scan measurements made on the Steel-block welds with both 2.25 MHz and 3.5 MHz Transducers, using 45, 60, and 70 degree angle wedges.



(a) 3 mm crack, 10 mm from bottom surface



(b) 6 mm crack, 26 mm from bottom surface, tilted 2 degrees

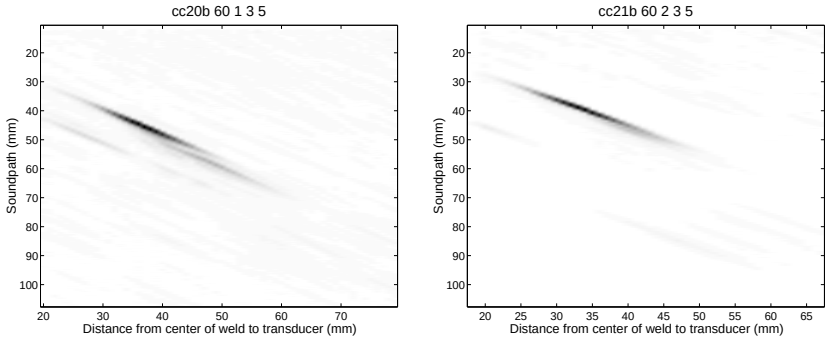


(c) 3 mm crack, 22 mm from bottom surface, 2 mm from center of weld, tilted 3 degrees

Figure 3.2: Indirect measurements from center cracks using the 45-degree transducer.

Defect	45	60	Block
S 4–7	•		B1
S 7–4	•		B1
SBH 1	•		B1
SBH 2	•		B1
SBH 3	•		B1
S1	•	•	B2
S2	•	•	B2
S3	•	•	B2
S4	•	•	B2
S5	•	•	B2
S6	•	•	B2

Table 3.4: B-scan measurements made on the aluminum blocks with artificial defects. The measurements were made with the 2.25 MHz and 3.5 MHz transducers.



(a) 6 mm crack, 26 mm from bottom surface, tilted 2 degrees

(b) 3 mm crack, 22 mm from bottom surface, 2 mm from center of weld, tilted 3 degrees

Figure 3.3: Backside measurements from center cracks using the 60-degree transducer.

one echo can be seen in Figure 3.3(b).

Sidewall Cracks

The sidewall cracks are located in the steel-weld junction and are, hence, tilted with the same angle as the weld (30 degrees). This makes it difficult to apply direct measurements, and all measurements are, therefore, performed from the backside or indirectly. Figure 3.4 and Figure 3.5 show backside measurements performed with the 45- and 60-degrees transducers. No double echos were noted for the sidewall cracks.

Lack of Fusion

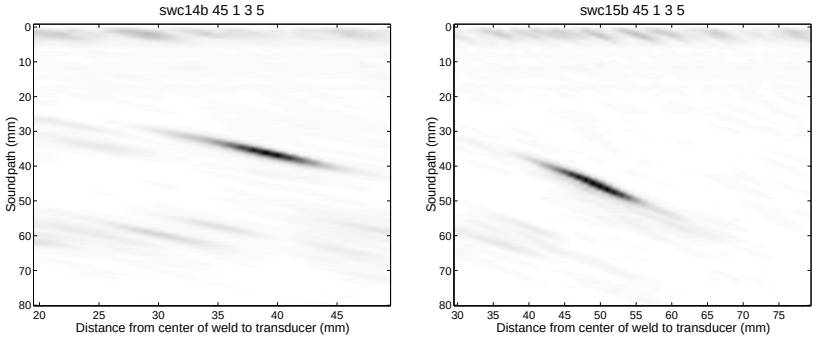
The lack of fusion (LOF) defects are also located in the in the steel-weld junction. This results in the same difficulty, as for the sidewall cracks, to make direct measurements. Hence, the measurements have been performed from the backside (or indirectly) here as well. Figure 3.6 and Figure 3.7 show measurements with the 45- and the 60-degree transducers, respectively. The echos seen around 60 mm for the 45-degree transducer originates from the top weld surface. Some of the LOF measurements also had two peaks (Figure 3.6(b) and (c)). However, the double echos were more separated than for the center cracks. The double echos were only seen when the 45-degree transducer was used.

Slag

The echos from the slag inclusions were rather distinct regardless of the transducer used. Figure 3.8 and Figure 3.9 show four examples using the 45- and 60-degree transducers. No direct measurements were performed, since the probe was obstructed by the weld surface (see Section 3.3.2).

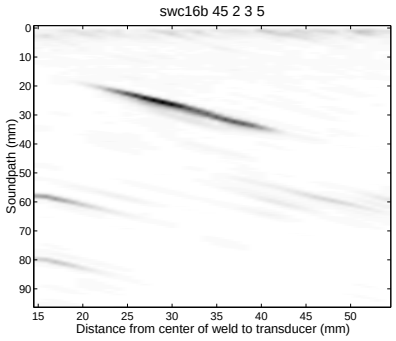
Porosity

Porosity can rather easily be separated from other types of defects due to the multiple echos encountered in the ultrasonic signal. Figure 3.10 and Figure 3.11 show 6 examples of B-scans acquired using both the 45- and 60-degree transducer from backside measurements. The echos observed at approximately 60 mm (in Figure 3.10) are, again, from the top surface of the weld. Note that the echos at approximately 25 mm in Figure 3.10c and



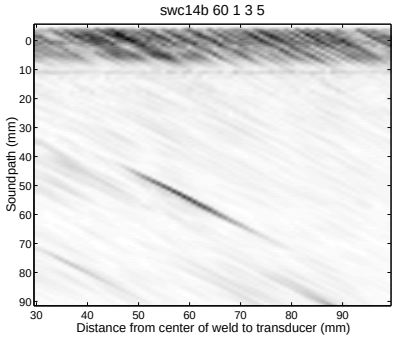
(a) 3 mm crack, 24 mm from bottom surface

(b) 7 mm crack, 29 mm from bottom surface

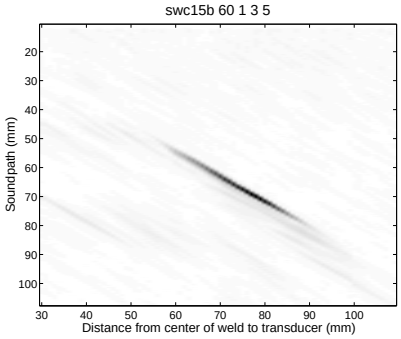


(c) 7 mm crack, 17 mm from bottom surface

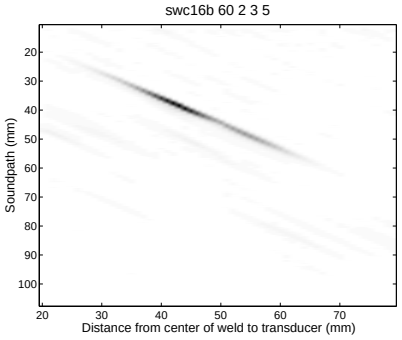
Figure 3.4: Backside measurements from sidewall cracks using the 45-degree transducer.



(a) 3 mm crack, 24 mm from bottom surface

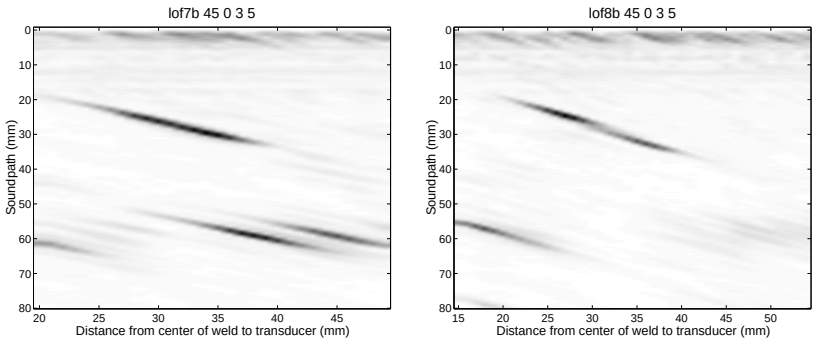


(b) 7 mm crack, 29 mm from bottom surface



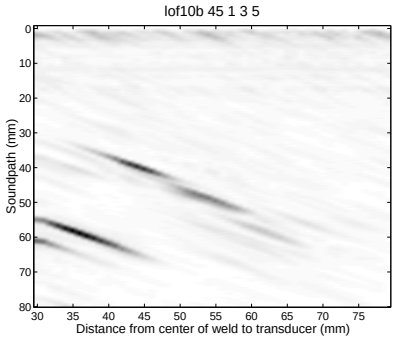
(c) 7 mm crack, 17 mm from bottom surface

Figure 3.5: Backside measurements from sidewall cracks using the 60-degree transducer.



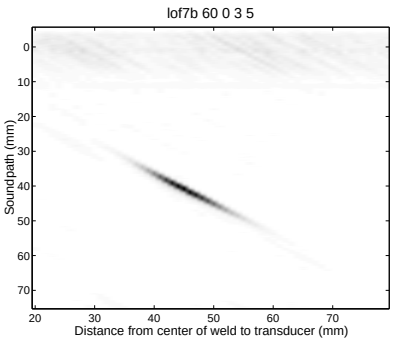
(a) 2.9 mm crack, 23.5 mm from bottom surface

(b) 6.9 mm crack, 29.4 mm from bottom surface

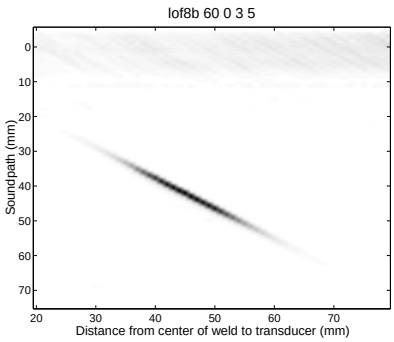


(c) 6.9 mm crack, 17 mm from bottom surface

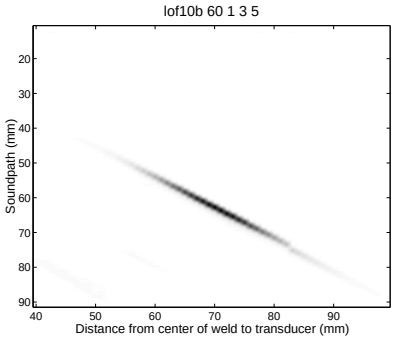
Figure 3.6: Backside measurements from lack of fusion using the 45-degree transducer.



(a) 2.9 mm crack, 23.5 mm from bottom surface

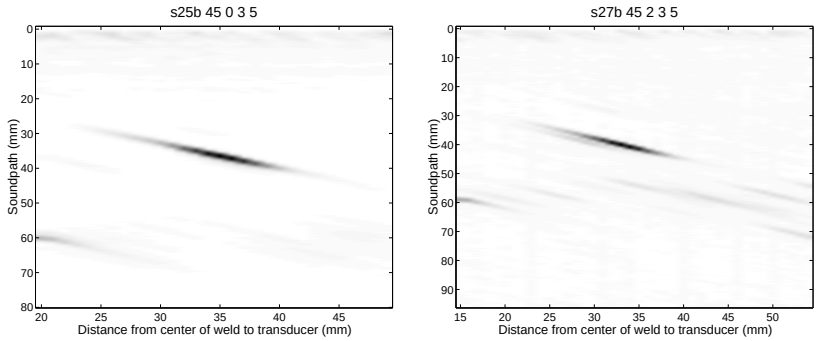


(b) 6.9 mm crack, 29.4 mm from bottom surface



(c) 6.9 mm crack, 17 mm from bottom surface

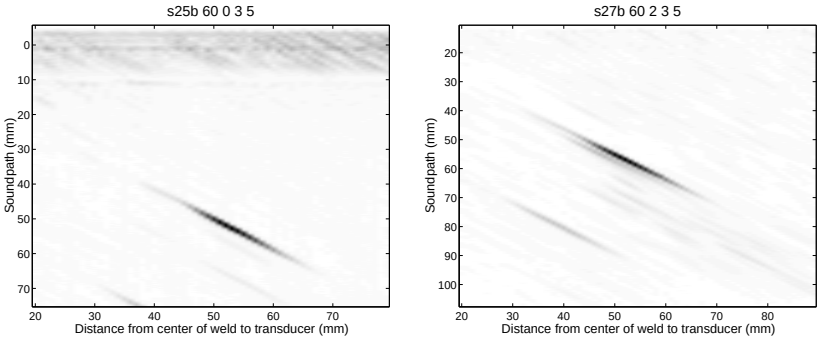
Figure 3.7: Backside measurements from lack of fusion using the 60-degree transducer.



(a) 3 mm slag, 25 mm from bottom surface

(b) 5 mm slag, 26 mm from bottom surface

Figure 3.8: Backside measurements from slag using the 45-degree transducer.

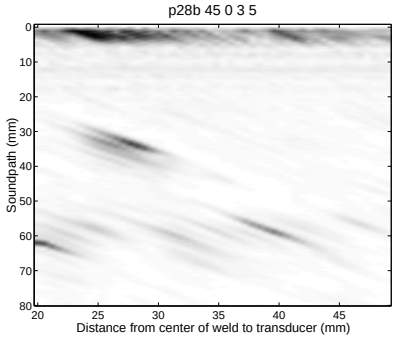


(a) 3 mm slag, 25 mm from bottom surface

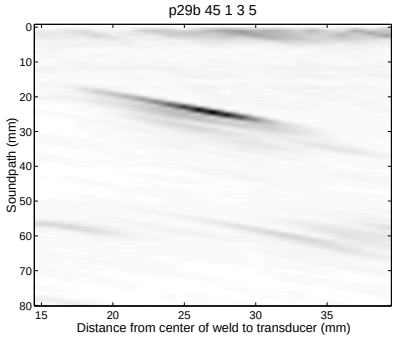
(b) 5 mm slag, 26 mm from bottom surface

Figure 3.9: Backside measurements from slag using the 60-degree transducer.

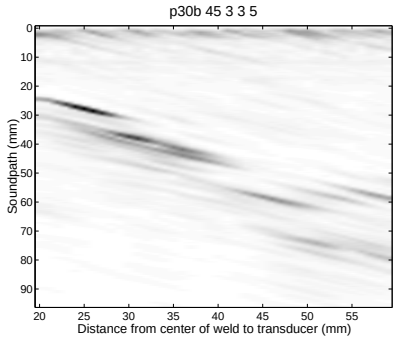
40 mm in Figure 3.11c probably come from a rather strong reflection due to the steel-weld junction in test block PL4503.



(a) 6 mm porosity, 22 mm from bottom surface



(b) 8 mm porosity, 15 mm from bottom surface

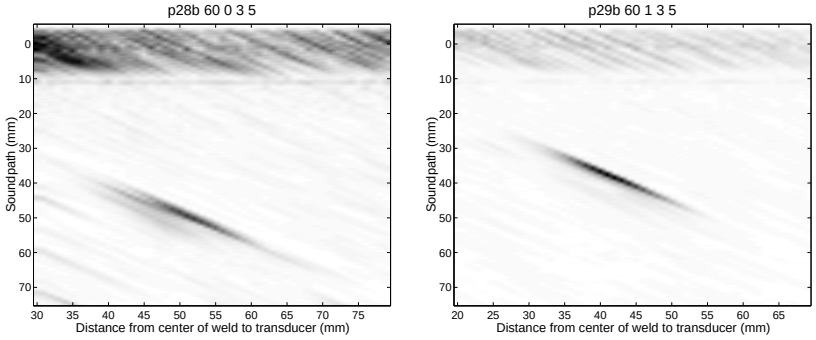


(c) 9 mm porosity, 26 mm from bottom surface.

Figure 3.10: Backside measurements from porosity using the 45-degree transducer.

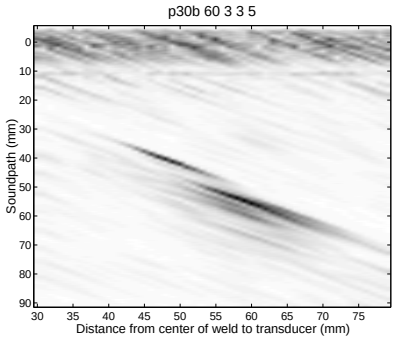
Root Crack

Root cracks result in strong reflections from the crack-bottom surface corner. In some cases, a small echo from the crack itself appears slightly before the crack-bottom echo. This "pre-echo" are somewhat difficult to see in the B-



(a) 6 mm porosity, 22 mm from bottom surface

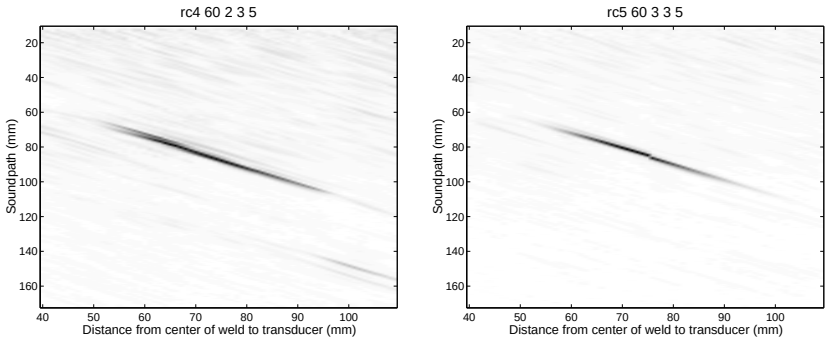
(b) 8 mm porosity, 15 mm from bottom surface



(c) 9 mm porosity, 26 mm from bottom surface

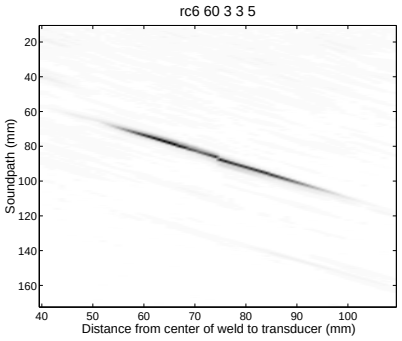
Figure 3.11: Backside measurements from porosity using the 60-degree transducer.

scans presented in Figure 3.12 and Figure 3.13, but will be more easily seen in A-scans presented later in Section 3.4.2 (Figure 3.29(c)). The origin of



(a) 3 mm crack, tilted 3 degrees

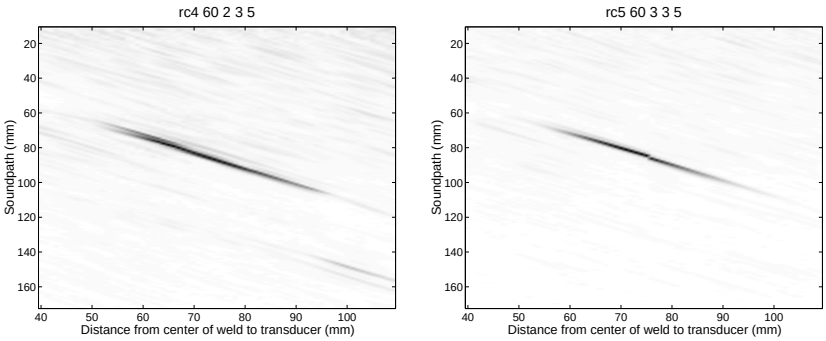
(b) 3.4 mm crack, tilted 17 degrees



(c) 7 mm crack, tilted 27 degrees

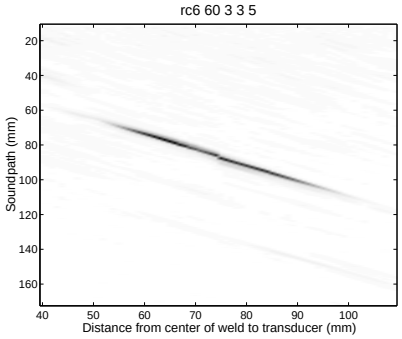
Figure 3.12: Direct measurements from root cracks using the 60-degree transducer.

these echos are unclear, but it may be due to the same phenomena as the double echos encountered in the signals from the center cracks.



(a) 3 mm crack, tilted 3 degrees

(b) 3.4 mm crack, tilted 17 degrees

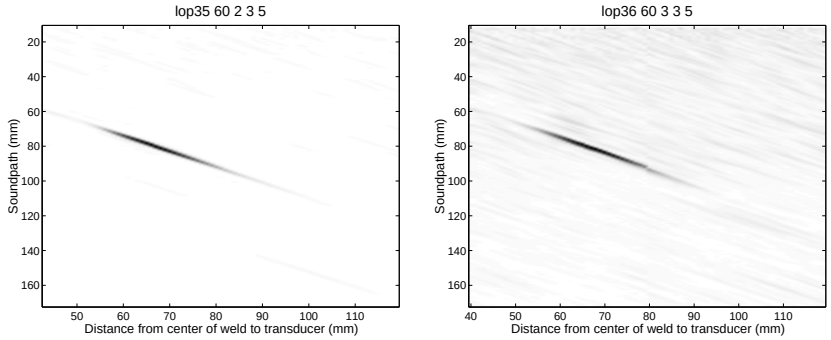


(c) 7 mm crack, tilted 27 degrees

Figure 3.13: Direct measurements from root cracks using the 60-degree transducer.

Lack of Penetration

The echos from lack of penetration defects look rather similar to root cracks, but they are more "distinct" since there is no echo from the top parts of the defect, only echos from defect-bottom surface corner is present. Figure 3.14 and Figure 3.15 show four examples using the 60- and 70-degree transducers (from direct measurements).



(a) 2 mm lack of penetration (b) 2.5 mm lack of penetration

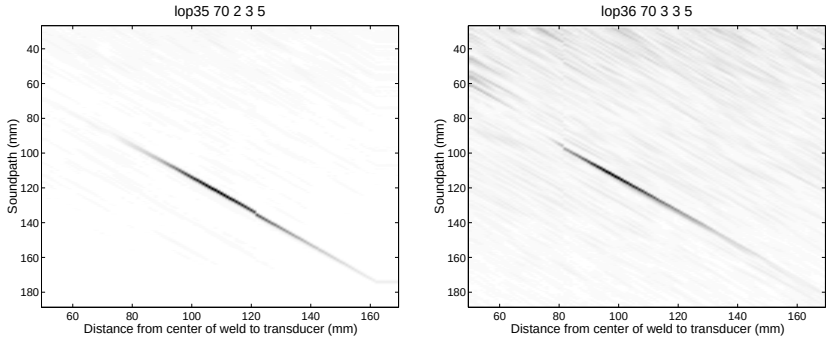
Figure 3.14: Direct measurements from lock of penetration using the 60-degree transducer.

Over Penetration

Over penetration can rather easily be distinguished from root cracks and lack of penetration since it is characterized by a proportionately long tail of small pulses after the main pulse (which comes from the bottom weld surface). Figure 3.16 shows three examples.

D-scans

D-scans provide information of how the response signal varies along a flaw. In particular, one can estimate the length (side-wise) of the flaw from these type of measurements. Two examples of D-scans are shown in Figure 3.17. Note the response from the bottom weld surface—shown as a horizontal line trough the D-scan—and how it is "shadowed" by the root crack in



(a) 2 mm lack of penetration (b) 2.5 mm lack of penetration

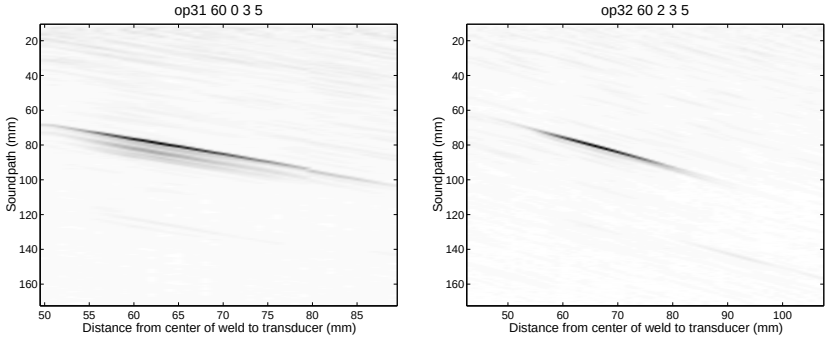
Figure 3.15: Direct measurements from lack of penetration using the 70-degree transducer.

Figure 3.17(a). Note also the typical ringings after the weld surface response, shown in Figure 3.17(b), that are characteristic for the over penetration.

3.4 Defect Characterization

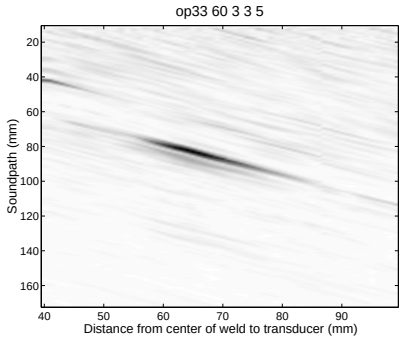
3.4.1 Signal Features and Feature Extraction

In order for a classification task to be successful there are a number of issues that must be considered. The number of available training examples must be compared to the complexity of the classifier (i.e. number of parameters in the classifier). If the training examples are few and the classifier is complex, then the classifier will perform well on training data and poor on unseen data [4, 3]. However, the parameters in the classifier can be reduced if the dimension of the input vector is reduced. The process of describing features in data in a compact way is known as feature extraction, and was briefly introduced in Chapter 1 and 2. In this application the number of training examples is very low and, hence, feature extraction is an essential part of the classification process. To succeed well the features must be descriptive, so that differences vital for classification are not lost. These aspects are illustrated in Figure 3.18, where a fictitious example characterized by only two features is shown. In this figure there are three classes present: one



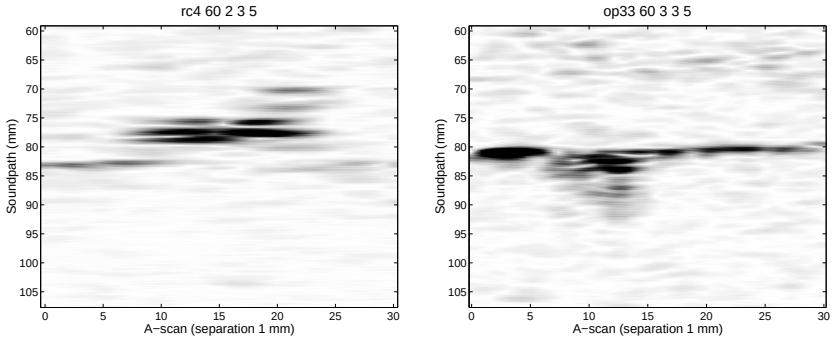
(a) 3 mm over penetration

(b) 4.5 mm over penetration



(c) 5 mm over penetration

Figure 3.16: Direct measurements from over penetration using the 60-degree transducer.



(a) 3 mm root crack (b) 5 mm over penetration

Figure 3.17: Two examples of D-scans from defects at the bottom of the weld.

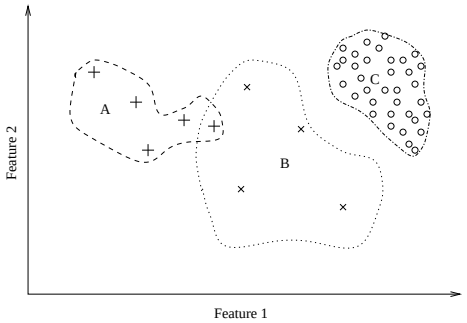


Figure 3.18: A fictitious two feature example.

labeled A (with dashed boundary), and one B (with dotted boundary) and finally class C (with dash-dotted boundary). There is also a number of examples from each class shown in the figure, where the $\times$:es are from class A, the $+$:es from class B, and the $\circ$:es are from class C. As one can see class A and B are overlapping and they also have a low number of examples which makes it difficult to design a classifier with a proper decision boundary based on those examples. Class C exemplifies the desired case with a sufficient number of examples and non-overlapping class boundaries. The two features used in this example are clearly not suitable for separating the A and B classes. However, for a different choice of features the classes may not be overlapping, and the problem can be solved.

In this section, some choices of feature extraction methods are discussed, as well as issues regarding pre-processing and region of interest selection.

Pre-processing

In this study, only the envelope of the acquired ultrasonic data is used. This was also the strategy the previous study [5, 6, 7]. The envelope is calculated by means of the Hilbert transform. The resulting data is also smoothed with a low-pass filter to reduce the measurement noise present in data.

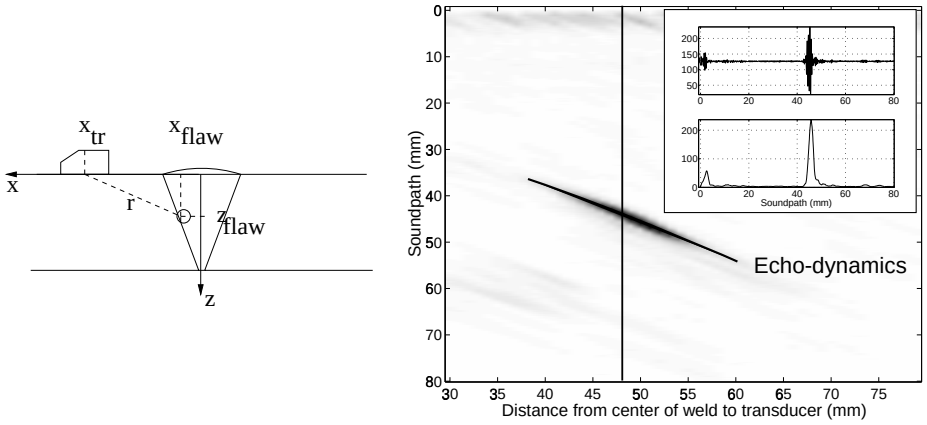
Extracting Defect Position

The flaw position is used for both region of interest (ROI) selection the depth normalization (which is required for the feature extraction performed later). It is therefore important that the position estimation is accurate and robust. The current method to find the flaw position is based on fitting an hyperbolic function to the flaw response in B-scan data, see [6]. This method is summarized below:

> The curves formed by a point scatterer in a B-scan, obtained with the contact measuring setup in shown Figure 3.19(a), are shaped as a part of an hyperbola given by the equation
>
> $$r = \sqrt{x_d^2 + z_{\text{flaw}}^2}, \tag{3.1}$$
>
> where $x_d = x_{\text{flaw}} - x_{\text{tr}}$ is the horizontal distance between transducer and the flaw, and z_{flaw} is the vertical position of the flaw. The position estimate $(\hat{x}_{\text{flaw}}, \hat{z}_{\text{flaw}})$ is found by minimizing the

summed squared error $\sum_i \parallel r_{max}^{(i)} - \hat{r}^{(i)} \parallel^2$, $i = 1, 2, \ldots, N$, for the N selected A-scans, where $r_{max}^{(i)}$ is the position of the max amplitude of the envelope of the ith A-scan.

Consider the A-scan 48 mm from the center of the weld, marked with a vertical line in Figure 3.19(b) (also included in the box in the same figure). The maximum response appears approximately at $r_{max}^{(i)} = 45$ mm.



(a) Contact measuring setup (b) B-scan

Figure 3.19: Illustration of the different distances used in Eq (3.1). *The* $r_{max}^{(i)}$ *positions in the A-scans appears along echo-dynamic curve marked in the B-scan.*

Depth Normalization

In an ultrasonic B-scan, a defect located close to the transducer will be seen in a fewer A-scans than a similar defect present further away from the probe, due to the lobe characteristics (cone-beam geometry) of the probe. By studying the echo-dynamics from two flaws at different depths, the echo-dynamics of the flaw closest to the transducer will have a narrower shape than the other flaw. A simple way to normalize is to re-sample the echo-dynamics (or wavelet coefficients) in some angle interval. That is, the feature vector (or matrix) is re-sampled in an angular scale instead of the original

linear scale. This is illustrated in Figure 3.20, where the two horizontal arrows indicates the distances where the flaws f_1, and f_2, are inside the ultrasonic beam. The depth normalization procedure consists of re-sampling

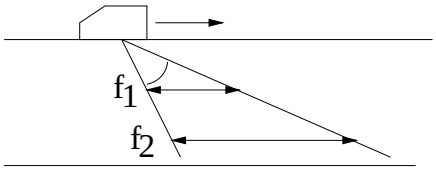
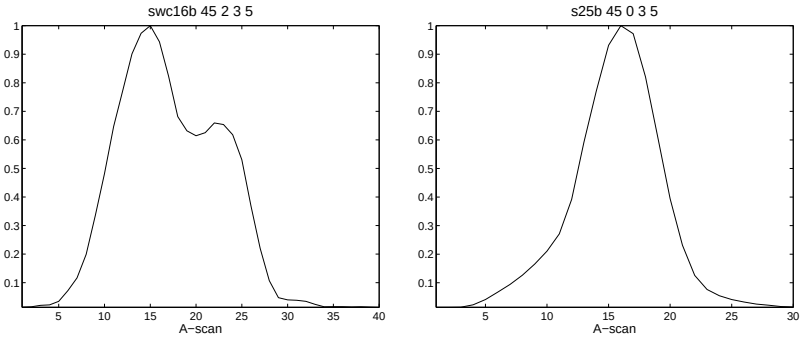


Figure 3.20: Illustration of the effect of the cone beam geometry for two defects at different depths. The defect "f_1" will be seen in fewer A-scans than defect "f_2", which must be compensated for, before the features can be used for classification.

the features for a suitable angular interval given the depth of the flaw. This implies interpolating features from flaws located close to the probe, and down sampling features for flaws that lie further away from the probe.

ROI Selection

Selection of ROI is an important issue since all further processing rely on it. The positioning of the analyzing window must be precise. If this is not the case, then the features fed to the classier will vary between different measurements, giving inconsistent results. However, it is not a trivial task to position an analyzing window accurately in the ultrasonic B-scan images encountered. Ideally an hyperbolic shaped analyzing window should be positioned around a flaw response in a B-scan, where the position of the window should be determined based on the (exact) position of the flaw. The problem is that the defect position is unknown and must be estimated. The procedure described above is, however, not accurate enough for the precise horizontal positioning required here. In the previous studies [6, 7] the echo-dynamics (max amplitude variation) of the flaw response was used for horizontal positioning. Figure 3.21 shows two examples of echo-dynamics. As one can see the echo-dynamics curves may be skewed, have more than one peak, etc. If a B-scan has two (or more) separate peaks (as in Figure 3.2(a)), then the situation becomes even more complicated. The approach used previously was to smooth the echo-dynamics, with low-pass filtering, which partially solves the problem. This method was suitable for the simulated and artificial defects. Experiments have been performed using center-of-



(a) Sidewall crack (b) Slag inclusion

Figure 3.21: Two examples of echo-dynamics (max amplitude variation between consecutive A-scans).

mass calculations in order to find a robust estimate for the center of the echo-dynamics. This approach was, however, too sensitive to long tails with high amplitude (energy) in the echo-dynamics. Therefore, the previously used algorithm is utilized here as well. The algorithm is summarized below:

1. Low-pass filter the echo-dynamics.

2. Find the A-scan of max amplitude of the filtered echo-dynamics.

3. Select a number N of A-scans centered around this position. The number N depends on the depth of the defect. Re-sampling will also be necessary to obtain features of equal length (as described above).

Classical Features

The perhaps most commonly used features for classification of defects during ultrasonic inspection is the rise time, pulse duration and fall time [16, 17]. These three features are calculated from the envelope of an A-scan as depicted in Figure 3.22. Typically one uses the 90% and 10% levels for the calculation. In the previous studies 90% and 35% where used [7], but these levels may be adjusted according to the level of noise and disturbances present in the measurements.

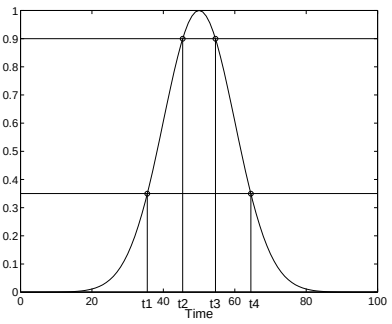


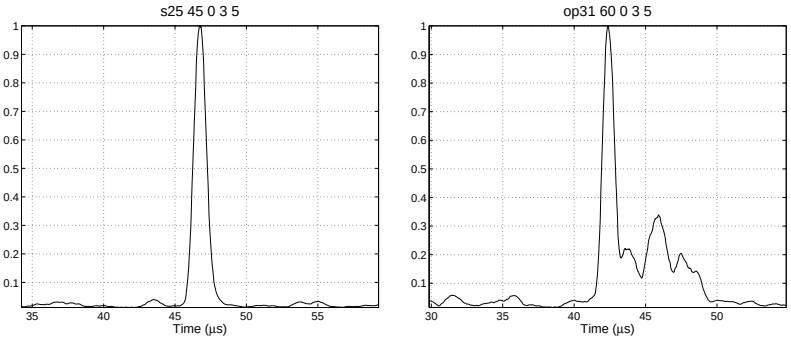
Figure 3.22: The four times used for calculating rise time, pulse duration and fall time.

When 2D data are available (B-scans) one commonly uses the echo-dynamics, discussed above, which gives a description of the amplitude variation between consecutive A-scans in a B-scan.

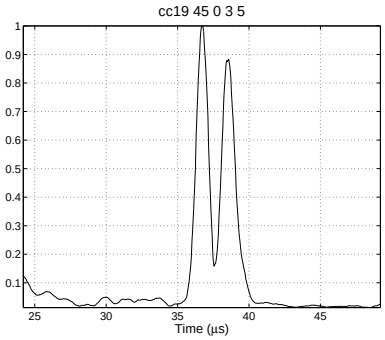
These basic features are reliable provided that the US pulses (echos) are well defined. However, realistic defects result in pulses with envelopes that are much different from the well defined bell in Figure 3.22, which considerably impairs the reliability of such features. Smoothing (low-pass filtering) partly alleviates these problems, but at the expense of some loss of information. Figure 3.23 shows the envelope of A-scans from three different types of defects. In spite of the very different shape of the waveforms, the rise time, pulse duration and fall time are rather similar for all of the signals in Figure 3.23. It is evident that more powerful features are needed if the classification should be feasible for this type of signals.

Feature Extraction using the Discrete Wavelet Transform

The *discrete wavelet transform* (DWT) described in Chapter 2 has several interesting features also for this application. The impulse-like nature and the locality of the basis functions in the DWT makes it suitable for modeling of ultrasonic signals. If, for example, an analyzing window is centered around an ultrasonic pulse, then it is possible to examine at which position and scale this pulse has significant energy, which is reflected in the wavelet coefficients given by the DWT. It should be noted that the envelopes of typical ultrasonic signals can be well described with only a few of the large scale components (wavelets), which result in a good data compression ability.



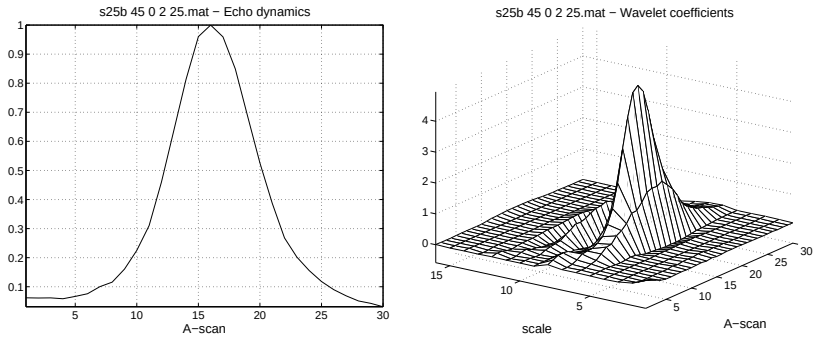
(a) Slag inclusion (b) Over penetration



(c) Center crack

Figure 3.23: Envelope of A-scans from three different flaws. The pulse duration and the raise- and fall times are similar despite the very different pulse shapes.

There are several different types of pre-defined mother wavelets available in common software packages, like the Wavelet toolbox for MATLAB [15]. The *Coiflet 2* mother wavelet, used in Chapter 2, is a fairly smooth wavelet suitable also for this application. The echo-dynamics, and the first 16 DWT coefficients, from the same (consecutive) A-scans are displayed in Figure 3.24. One can clearly see how the echo-dynamics is reflected in the



(a) (b)

Figure 3.24: (a) The echo-dynamics from a slag inclusion. (b) The first 16 wavelet coefficients from the same A-scans as in (a).

wavelet coefficients. Note also that there are significant energy (information) for more than one scale. Obviously, if only the echo-dynamics is used, a significant amount of information is lost, which then may impair classification performance.

3.4.2 Defect Classes

A conclusion, from the performed measurements, is that the characterization task is much more complex for the realistic defects than for the artificial, and simulated, counterpart. The variability of ultrasonic responses from the same type of defects appeared to be very large, which resulted in a considerable overlapping of flaw classes in feature space. This obviously causes problems with the flaw classification. A realistic goal is to categorize defects in sharp defects, like cracks and lack of fusion, and volumetric (or soft defects), like porosity and slag inclusions. Defects in the bottom of the weld are also easy to distinguish from other flaws since they all occur at the

same position. The flaw types are, therefore, divided in three main groups which are: sharp (crack-like) defects, volumetric defects, and defects at the bottom of the weld.

Sharp Defects and Volumetric Defects

The Figures 3.25–3.28 show envelopes of A-scans and echo-dynamics for crack-like defects and volumetric defects, respectively. At least two observations can be made from these images:

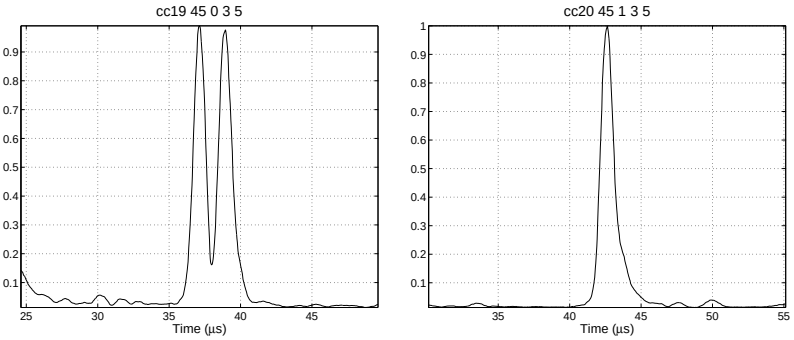
- The ultrasonic responses from crack type defects exhibit a large variation of features. This is especially clear for center cracks and lack of fusion defects.
- It is very difficult to distinguish sidewall cracks, and lack of fusion, from slag inclusions.

Analysis of Figure 3.25 to Figure 3.28 leads to the conclusion, which can be expressed in pattern recognition terms, as too large within-class variation for cracks and overlapping class regions between slag inclusion and LOF/SWC:s. Porosity is the type of defect that is easiest to separate from the other classes, due to the multiple echos which this type of defects produces. Another observation is that it is difficult to draw any general conclusion from the echo-dynamics (Figure 3.26 and Figure 3.28).

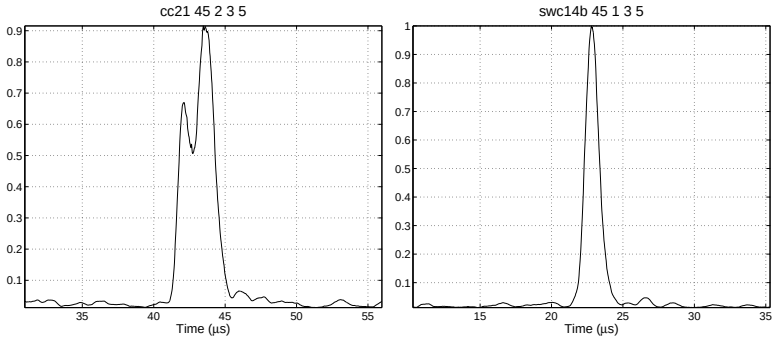
One conclusion from the measurements is that a larger number of examples from the sharp class of defects would be required, in order to see as many variations as would be needed to construct a fully automatic classifier based on training examples only. It is obvious that the use of the classical type of features is unsatisfactory in this case. This implies that more sophisticated tools are needed for feature extraction.

Defects at the Bottom of the Weld

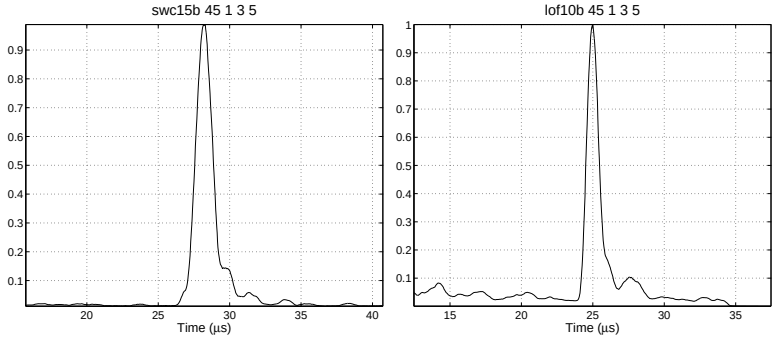
The defects located at the bottom of the weld include three types of flaws: over penetration, lack of penetration, and root cracks. Figure 3.29 shows one example of each type. The within-class variation seems much smaller for these defects than for the other defects. The class separation between the three different flaw types also appears larger than in the former case. Lack of penetration has a rather "clean" pulse shape, over penetration has typical



(a) Center crack (b) Center crack



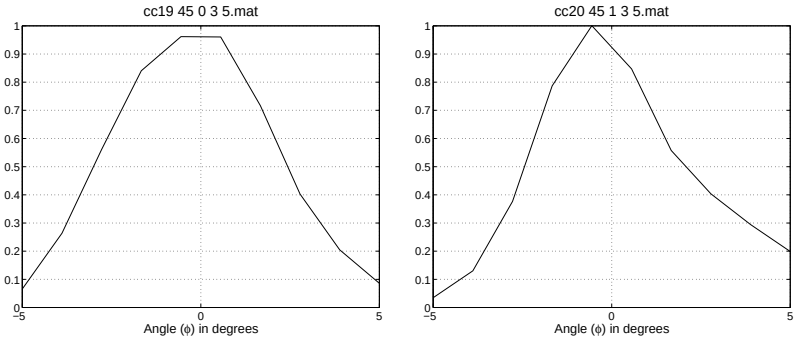
(c) Center crack (d) Sidewall crack



(f) Lack of fusion

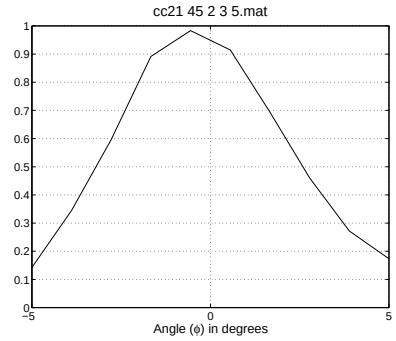
(f) Sidewall crack

Figure 3.25: Envelope of A-scans from sharp defects. The pulse shapes are well defined, but occasionally double echos are found.

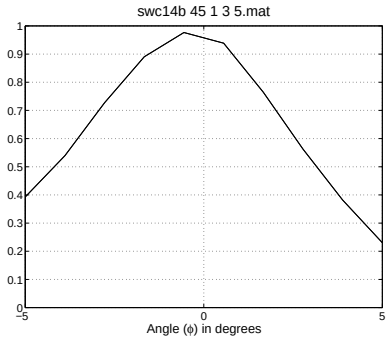


(a) Center crack

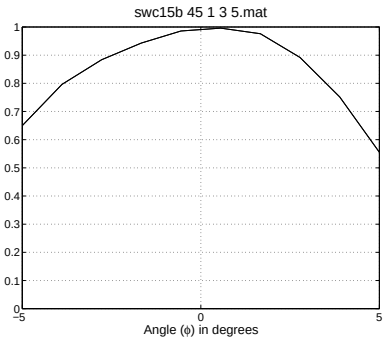
(b) Center crack



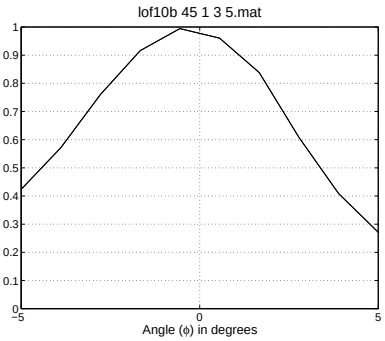
(c) Center crack



(d) Sidewall crack

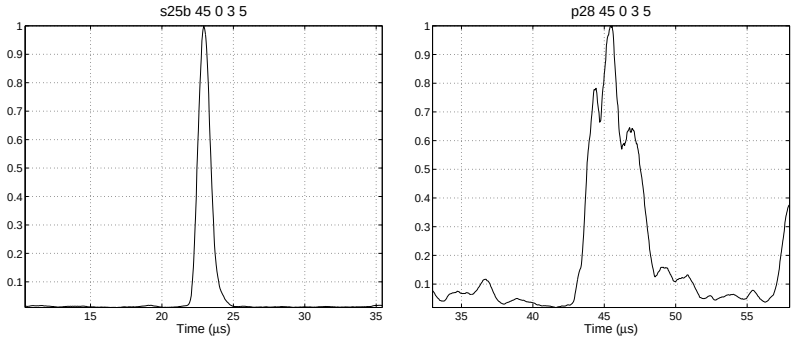


(e) Sidewall crack

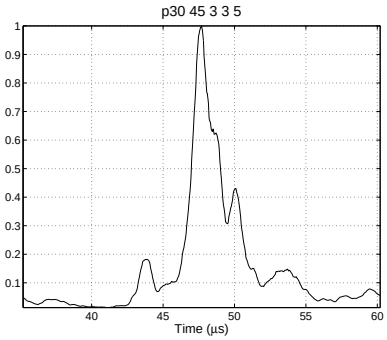


(f) Lack of fusion

Figure 3.26: Echo-dynamics in the angle range -5 to 5 degrees from crack-like defects.



(a) Slag inclusion (b) Porosity



(c) Porosity

Figure 3.27: Envelope of A-scans from volumetric defects. The pulse shape from the slag inclusions are rather well defined in contrast to porosity which exhibits many "ringings".

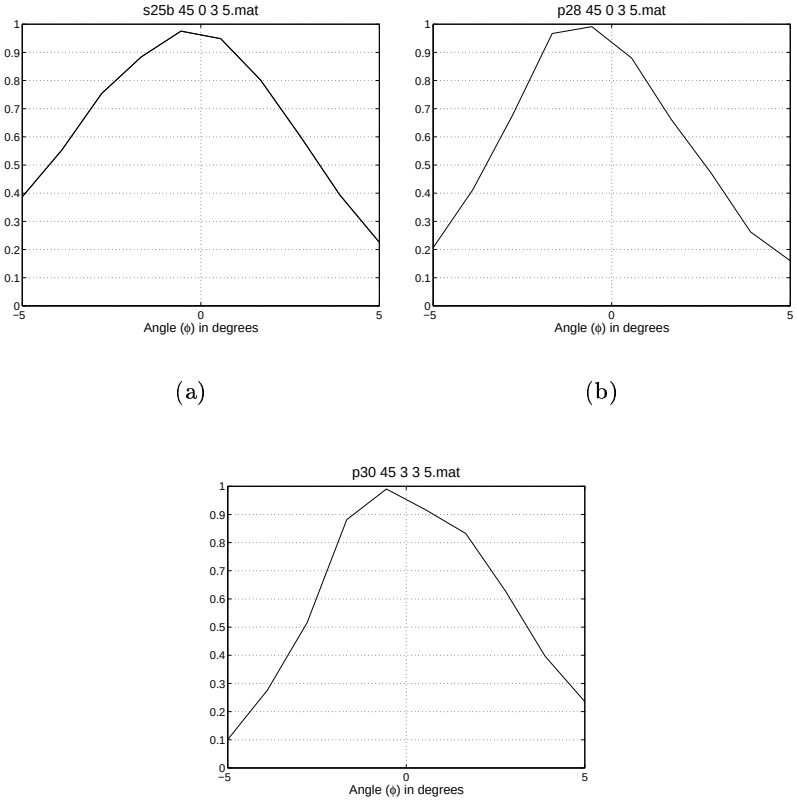
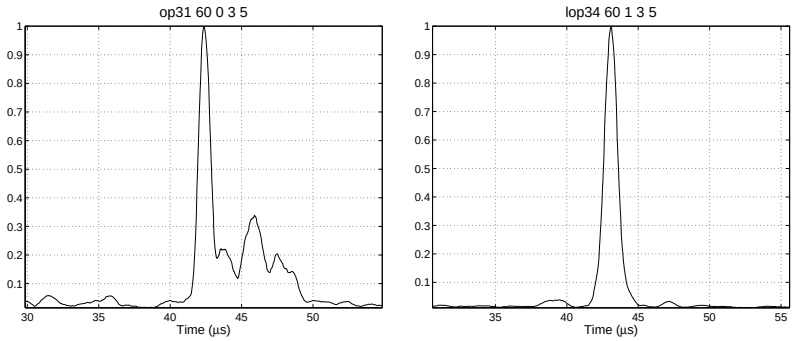
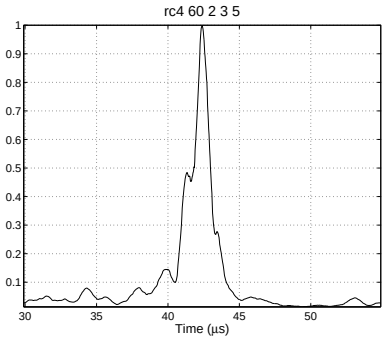


Figure 3.28: Echo-dynamics in the angle range -5 to 5 degrees from volumetric defects. (a) Slag inclusion (b)–(c) Porosity.



(a) Over penetration (b) Lack of penetration



(c) Root crack

Figure 3.29: Envelope of A-scans from defects at the bottom of the weld. The pulse shape from lack of penetration is well defined which makes it easy to distinguish from the overpentration, which exhibits large ringings after the main pulse, and from root cracks, which has a small pre-pulse before the main pulse.

ringings after the main pulse, and root cracks result often in a pulse which comes slightly before the main pulse, which can be seen in Figure 3.29.

The classical features may be sufficient for separation of the three flaw types, at least if the ringings of the over penetration, and the "pre-pulse" of the root cracks, are not separated too far from the main pulse. A large separation between the ringings (or pre-pulse) and the main pulse may result in that only the main pulse is used for feature extraction. In fact, the pre-pulse or ringings might not even have a pulse amplitude that is higher than lower limit (see Figure 3.22), which results in a total loss of these features.

However, if the DWT is used, more information is preserved about the pulse shape, and then there is no risk of loosing information of low amplitude pulses as long as they occur inside the analyzing window.

3.4.3 Natural Contra Artificial Defects

As mentioned earlier B-scan data was also acquired for the aluminum blocks with artificial defects used in the previous studies [5, 6, 7]. In these studies the classical features, described above, was shown to be sufficient. The purpose here for acquiring data from artificial defects was 1) to briefly verify some of the previous results and 2) to elucidate the difference to data acquired from realistic defects. Figure 3.30 shows a B-scan from block B1 with four artificial cracks (notches). The cracks have a depth of 2 mm, 4 mm, 4 mm and 8 mm (from left to right in the figure). One can see the diffraction echos, slightly above the main echos, whose location in the B-scan is in agreement with the size of the defects. Figure 3.31 shows one A-scan (and the envelope) from the same B-scan as in Figure 3.30. The pulse shape from the artificial defects are very "clean" compared to the ones in the steel blocks with real defects. There are no double echos or irregular pulse shapes present in the signals from the artificial cracks in the aluminum data.

The defect characterization (i.e. classification) task becomes much simpler since the within-class variation is much lower than for real flaws. The main implication of this is that the number of data needed "to span" the room of possible flaw signals is much lower for artificial defects than for the real defect counterpart. The features needed for classification are also simpler, echo-dynamics, rise time, pulse duration and fall time work well for artificial defects [7], contrary to the real flaw signals were more sophisticated features are needed.

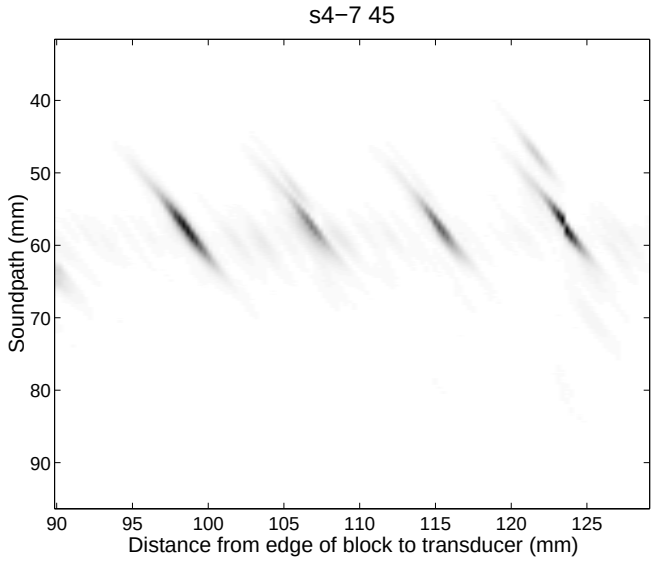
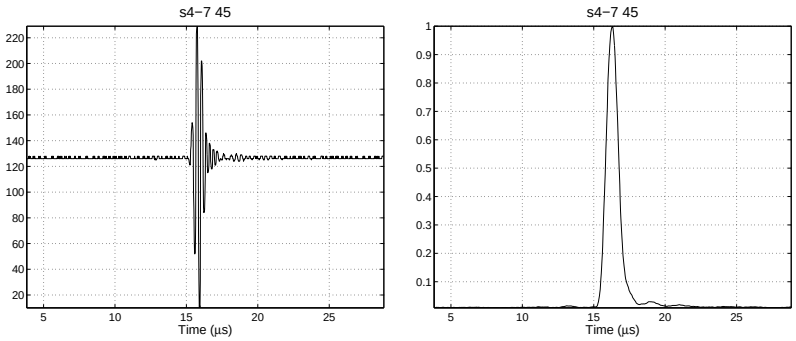


Figure 3.30: Four artificial cracks (notches) in the B1 aluminum block.



(a) (b)

Figure 3.31: One A-scan from a crack (notch) in aluminum block B1. (a) A-scan and (b) Envelope of the same A-scan.

3.5 Conclusions

During the evaluation of ultrasonic data acquired from the V-welded steel blocks it became evident that the characterization task is much more complex than for simulated and artificial flaw signals. The feature space of possible flaw signals is also considerably larger for the real defects than for the artificial counterparts, i.e., the variation of the ultrasonic signals within one type (class) of defects is much larger for real than for artificial defects.

Our goal was to separate soft (or volumetric) defects from the sharper ones (crack-like defects), but if one studies the echo-dynamics and the pulse shapes (i.e. the envelope) it becomes apparent that some sharp and soft defect types are very hard to separate. This implies that overlapping feature regions are encountered, especially when using classical features (fall/raise times, pulse duration and echo dynamics). To avoid overlapping class boundaries, more powerful feature extraction algorithms are needed to achieve a good classification performance.

High variation of the ultrasonic signals also has two further consequences: flaw position estimation (needed for feature extraction) may be poor and the amount of data needed to construct a reliable classifier is large. Below, a number of conditions that must be fulfilled in order to successfully build a classifier based on training examples is listed:

- Ensure that the measurements are good (informative) enough to distinguish between different types of defects. This is vital, because it can, of course not, be expected to be able to distinguish between different defects if the information needed is not present in the measurements.
- The features must be representative. The features that are fed to the classifier must preserve the information needed for successful classification.
- The number of training examples must be sufficient. As a rule of thumb one needs at least ten times as many examples as the parameters in the classifier, to avoid that the classifier learns the training examples and performs poor on unseen examples. Moreover, the examples must be enough representative to span the whole room of possible flaw signals for each defect class. If the last condition is not fulfilled the classifier will not able to classify all defect signals properly.

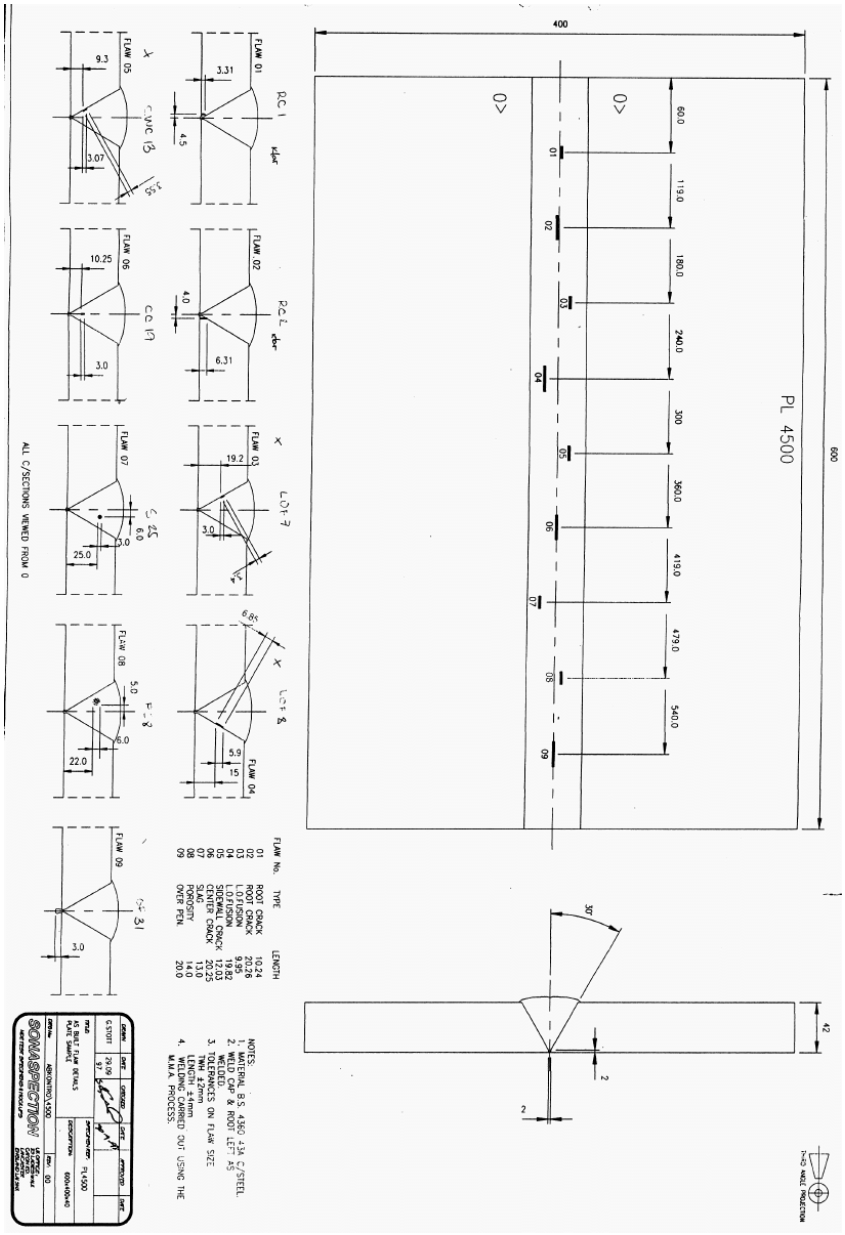
The second condition is clearly not fulfilled with classical features and more powerful methods are, therefore, needed. Note also that the first condition

may not be fulfilled using a single B-scan measurements only. A common practice in situations when it is difficult to categorize a measurement, is to combine measurements from several transducers (with different angles, center frequencies etc.) and TOFD measurements. This technique is usually known as *data fusion*.

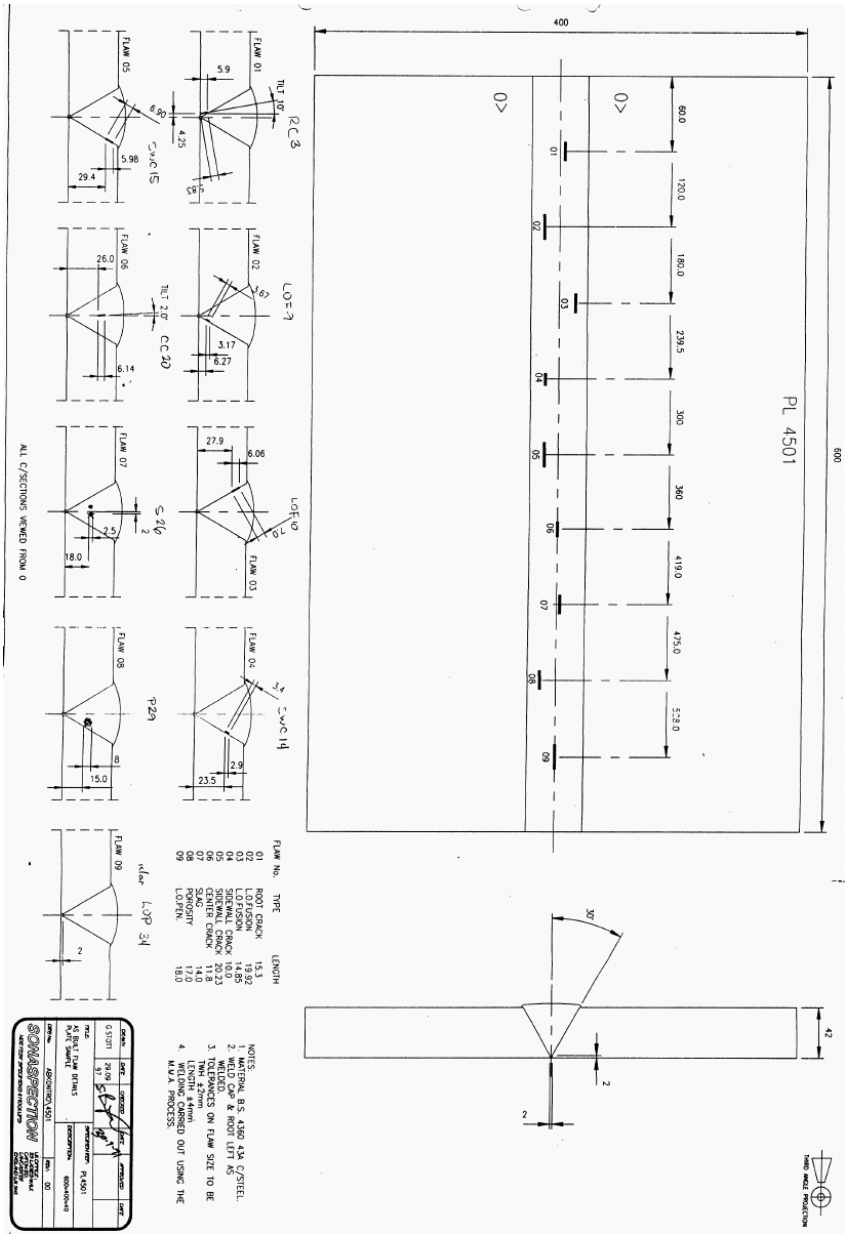
The last condition is more cumbersome. Clearly the number of training examples was not sufficient to span the room of all possible flaw signals, which is huge since one must account for different orientation, flaw size, crack roughness etc. Therefore, one could not expect to obtain a feasible classifier, based on the low amount of training data available here, using a standard pattern recognition approach. Hence, *a priori* knowledge *must* be incorporated to solve this difficult characterization problem. This knowledge can, for example, take the form of expert knowledge of experienced operators or the form of a advanced flaw modeling. A reasonable approach, for characterization of the type of defects encountered in this study, is to concentrate on improving flaw imaging and leave the classification tasks to experienced operators.

3.A Carbon Steel Block Drawings

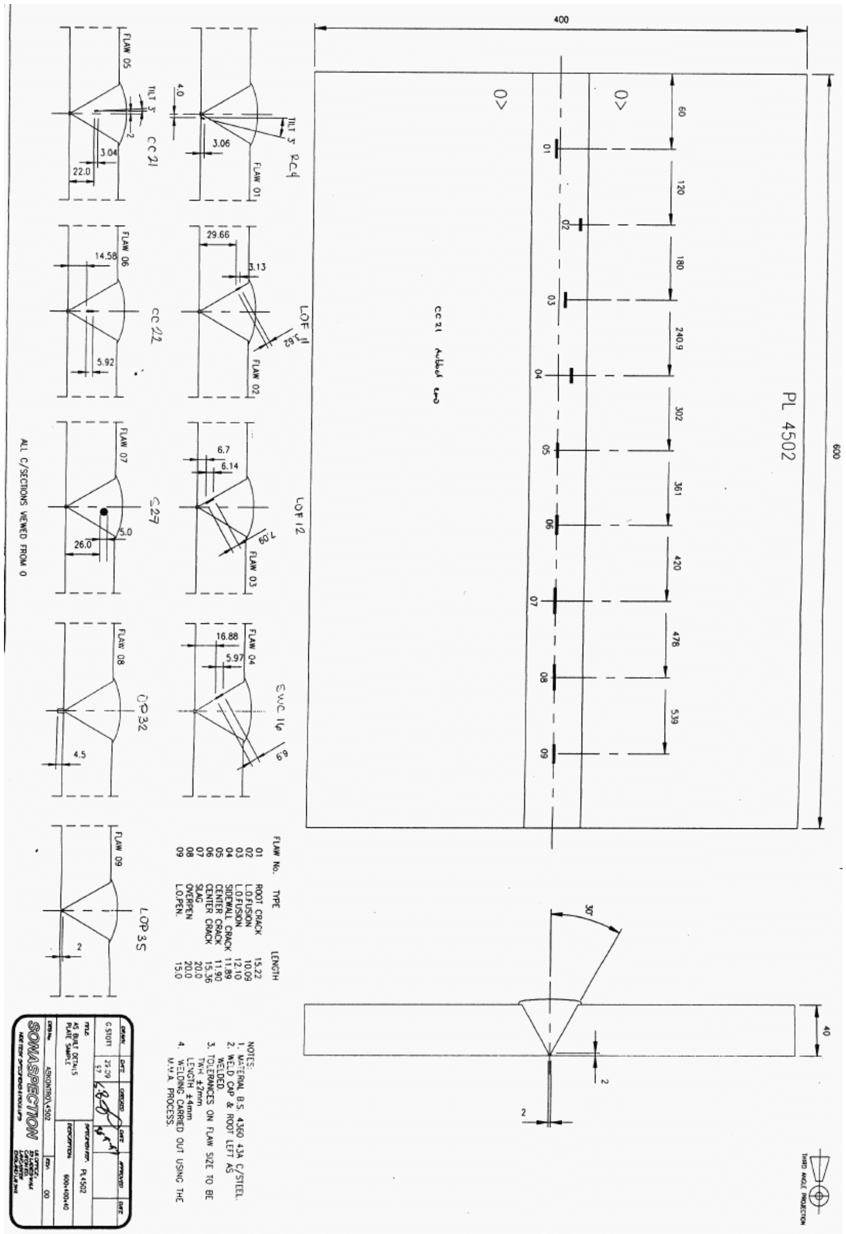
3.A.1 PL4500



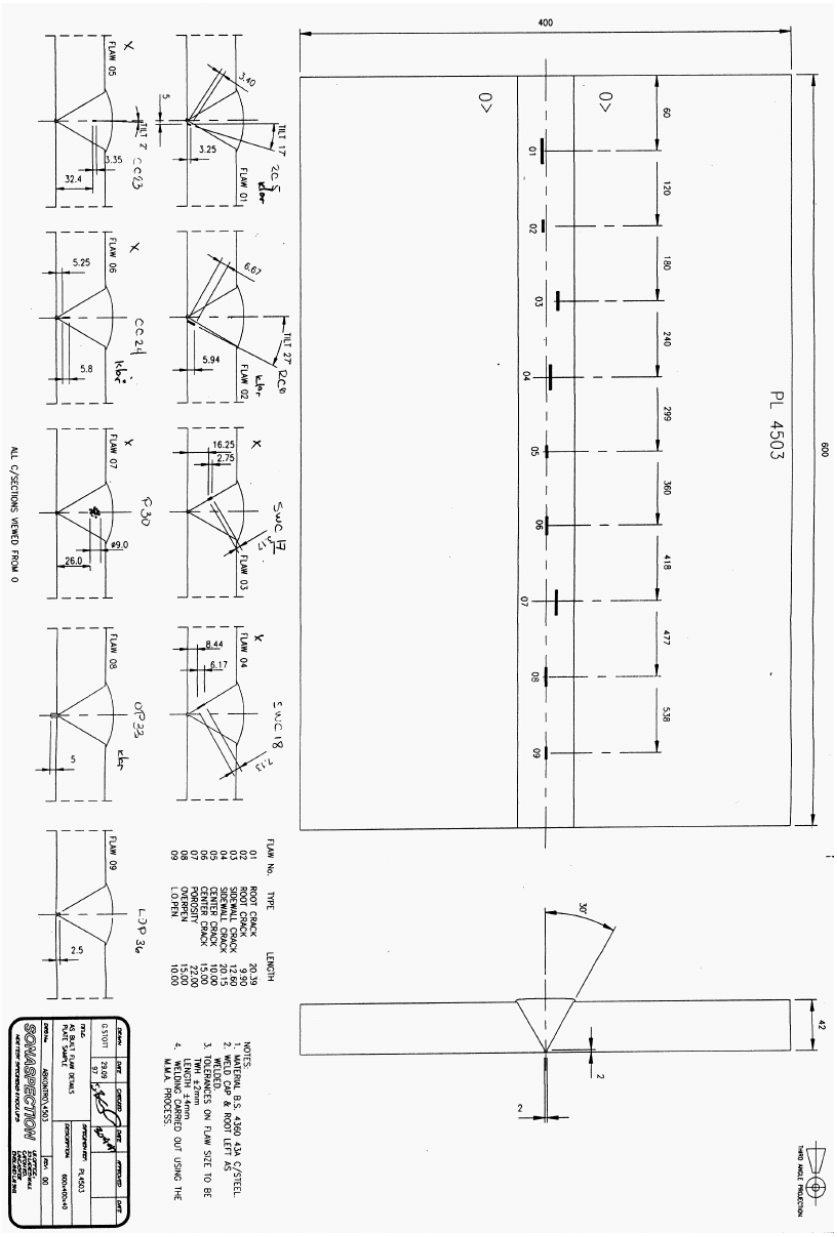
3.A.2 PL4501


PL 4501
400
600
60.0
120.0
180.0
239.5
300
360
419.0
475.0
538.0
01
02
03
04
05
06
07
08
09
FLAW 01
FLAW 02
FLAW 03
FLAW 04
FLAW 05
FLAW 06
FLAW 07
FLAW 08
FLAW 09
5.9
4.25
5.98
29.4
26.0
6.14
3.17
6.27
27.9
6.06
2.5
18.0
15.0
2.9
23.5
2
30°
42
FLAW No. TYPE LENGTH
01 ROOT CRACK 15.3
02 L.O.FUSION 19.92
03 L.O.FUSION 14.85
04 SIDEWALL CRACK 18.05
05 SIDEWALL CRACK 20.23
06 CENTER CRACK 11.8
07 SLAG 17.0
08 POROSITY 17.0
09 L.O.PEN 18.0
NOTES:
1. MATERIAL BS 4360 43A C/STEEL.
2. WELD CAP & ROOT LEFT AS WELDED.
3. TOLERANCES ON FLAW SIZE TO BE TN+ 4mm
4. LENGTH &4mm WELDING CARRIED OUT USING THE M.M.A. PROCESS.
ALL C/SECTIONS VIEWED FROM 0
THIRD ANGLE PROJECTION

3.A.3 PL4502


PL 4502
400
600
40
30°
ALL C/SECTIONS VIEWED FROM 0
FLAW No. TYPE LENGTH
01 ROOT CRACK 15.22
02 L.O.FUSION 10.09
03 L.O.FUSION 12.10
04 SIDEWALL CRACK 11.89
05 CENTER CRACK 11.90
06 CENTER CRACK 15.36
07 SLAG 20.0
08 OVERPEN 20.0
09 L.O.PEN. 15.0
NOTES:
1. MATERIAL B.S. 4360 43A C/STEEL.
2. WELD CAP & ROOT LEFT AS WELDED.
3. TOLERANCES ON FLAW SIZE TO BE TWH ±2mm LENGTH ±4mm
4. WELDING CARRIED OUT USING THE M.M.A. PROCESS.
SONASPECTION
THIRD ANGLE PROJECTION

3.A.4 PL4503


PL 4503
400
600
42
30
2
FLAW No. TYPE LENGTH
01 ROOT CRACK 20.39
02 ROOT CRACK 9.90
03 SIDEWALL CRACK 19.80
04 SIDEWALL CRACK 20.15
05 CLUSTER 10.00
06 CENTER CRACK 10.00
07 POROSITY 15.00
08 POROSITY 22.00
09 LOPEN 10.00
NOTES
1. MATERIAL B.S. 4360 43A C/STEEL.
2. WELD CAP & ROOT LEFT AS WELDED.
3. TOLERANCES ON FLAW SIZE TO BE TWH ±2mm LENGTH ±4mm
4. WELDING CARRIED OUT USING THE M.M.A. PROCESS.
ALL C/SECTIONS VIEWED FROM 0
CROSS SECTION

Chapter 4

Temperature Mapping using Ultrasonic Tomography

4.1 Introduction

Accurate control of gas temperature in closed spaces is an important issue for several applications such as power plant boilers and air-conditioning systems. The performance of a temperature control system depends on several factors, where the response time and the accuracy of the transducers are two important factors. A common method to obtain an estimate of the gas temperature is to insert one or several probes inside the volume of interest. The temperature distribution is then estimated from these measurements and then the heating system is controlled based on the estimate. There are, however, difficulties using this method to measure the temperature distribution. Typically the probes have a rather long response time and it may also be difficult to place the probes at suitable positions, due to hot environments (the probes may be damaged) etc. The application studied here is temperature mapping in air-conditioning systems. Here the probes must located somewhere on the walls or in the ceiling in the room, which is not optimal, since one usually is interested of the temperature inside the room and not on its boundaries.

An idea to improve the performance is to replace the probes, which usually are cheap thermocouples, with ultrasonic (US) transducers. The physical property that is used is the temperature dependence of the sound

velocity. Thus, by sending ultrasonic pulses along suitable paths, and measuring the time taken for each pulse, it should, in theory, be possible to reconstruct the temperature distribution in the room. Similar systems have been developed for other applications, such as, measurement of temperature distribution in industrial boilers, or in the outlet of the burner to measure the temperature of exhaust gases [18, 19, 20, 21].

The goal of this, so called *amenity sensor*, was to measure the temperature distribution in a room, using as few ultrasonic sensors as possible. The application was the next generation of office air-conditioning systems with improved comfort achieved by measuring the temperature distribution in a room and using this information in a temperature control system. In Figure 4.1, an example of a temperature distribution and a measurement setup using four (fan-beam) US paths, with a hot air-jet resulting in a temperature gradient in the center of the room is presented. The amenity sensor has a

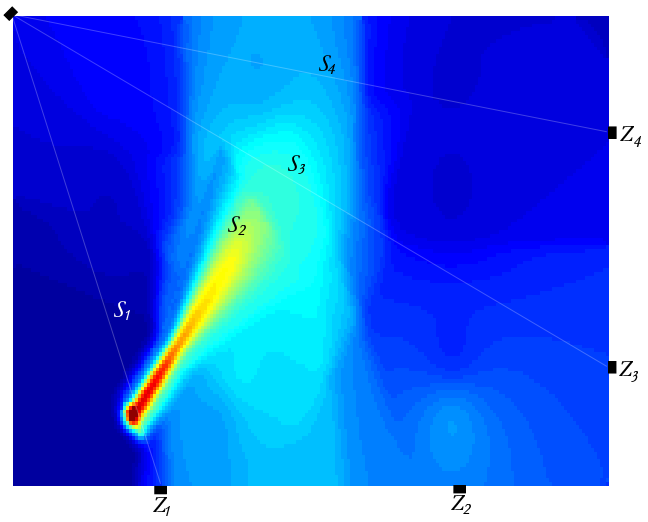


Figure 4.1: Example of a time of flight measurement setup.

number of advantages compared to an ordinary point measurement system using thermoelements. Firstly, the number of sensors required to achieve a reasonable mesh in the temperature map is relatively low. Secondly, the sensors can be placed along the walls, which facilitates their installation. Thirdly, the response time of the system should be much faster (cheap thermoelements have a response time in the range of 30 seconds and a TOF measurement will only take fractions of a second).

The development of the amenity sensor may be split in two separate tasks, the ultrasound measurement system, and the temperature reconstruction algorithm. Here we focus on the temperature reconstruction algorithms based on time of flight (TOF) measurements. The transducers contribute to a substantial part to the overall system cost, so it is important to reduce the number of transducers as much as possible without sacrificing the performance of the whole system too much. The objective here is to investigate feasibility of methods suitable for an accurate reconstruction of room temperature distribution based on a few TOF measurements. Our aim is to maximize the system accuracy while using as few sensors as possible. It is apparent that this problem is similar to computed tomography (CT) used in medicine but the fundamental difference is the number of measurements which is generally much larger in medical applications. This fact will, of course, influence the design/choice of reconstruction algorithm. We will show that number of sensors can be reduced if *a priori* knowledge on the temperature distribution is available.

The basic physical properties of the Amenity sensor are discussed briefly in Section 4.2. In Section 4.3 the problem is formulated in discrete space which results in a linear system of equations for solving the temperature mapping problem. The dimension of this equation system is high and it is in general also ill-conditioned. Since the number of equations also is smaller than the number of unknowns it is necessary to make certain prior assumptions in order to obtain reasonable performance. Strategies from CT, traditional regularization schemes, and a recent method by M. Gustafsson [22] are also discussed in this section, as well as aspects on dimension reduction and error performance. Since no real data were available for the simulations described in Section 4.4, a smooth phantom model was used to generate data for evaluation of the different algorithms. Finally, in Section 4.5 the conclusions are given.

4.2 Physical Model

The sound speed dependence of the temperature T can be described by

$$c_s = \sqrt{\frac{\gamma RT}{m}} = K\sqrt{T} \tag{4.1}$$

where $\gamma = C_p/C_v$ is the ratio between the heat capacitivities of the gas (air), R is the universal gas constant, and m is the molar weight of the gas. The

sound speed also depends of the humidity and the pressure etc, but these are second order effects [23] which are neglected here. If the temperature distribution is not constant the sound speed will be a function of the position in the room, $c_s(x, y)$, and hence the TOF will depend on the path traveled by the sound. For simplicity two dimensions are considered here, but the extension to three dimensions is straightforward.

Let the reciprocal of $c_s(x, y)$ be referred to as the *slowness function* $f(x, y)$. Then the reconstruction of the temperature distribution (i.e. the slowness function) can be formulated as the estimation of $f(x, y)$ from P parallel beam projections z_p, along straight lines defined by $s_p(l_p)$, or

$$z_p = \int_{s_p} f(x, y) dl_p \quad p = 1, 2, \ldots, P. \tag{4.2}$$

The equation of a line can be expressed as $x \cos\theta + y \sin\theta = t$.

For $f(x, y)$ to be uniquely determined, the TOF (z_p) must be known for all angles and rays [24], that is

$$z_\theta(t) = \iint_{\mathcal{R}^2} f(x, y)\delta(x \cos\theta + y \sin\theta - t) dx dy \tag{4.3}$$

must be known for all (continuous) θ and t, where $\delta(\cdot)$ is the Dirac delta function. Equation (4.3) is known as the *Radon* transform of $f(x, y)$ [25, 24, 26, 27]. Here we do not have access to all $z_\theta(t)$ and consequently, the estimate of $f(x, y)$ must be performed using the P available projections.

4.3 Reconstruction Techniques

There exist numerous approaches for solving the inverse problem of estimating $f(x, y)$ from the measurements z_p in Eq. (4.2). In traditional CT the approaches can roughly be divided in two groups, one transform based, and one which is based on finite series-expansion. The latter method can be divided further in subgroups depending on which basis functions that are used in the series-expansion, and which method that is used to solve the resulting linear equation system.

The transform based methods are based on the *Fourier slice theorem* which (mostly) leads to the *filtered back-projection algorithm* (FBA) [25]. The FBA requires a large number of projections to obtain good performance. In this application the number of projections is limited, mostly due to the cost of installing a large number of US sensors.

4.3.1 The Algebraic Approach

In the series-expansion approach one assumes that the function $f(x,y)$, which we want to estimate, can be expressed as a linear combination of the basis functions $b_j(x,y)$,

$$f(x,y) = \sum_j a_j b_j(x,y). \tag{4.4}$$

Thus, $f(x,y)$ is expressed as a weighted sum of a number of "base images". By using (4.2) and (4.4), the TOF measurements can be expressed as

$$z_p = \sum_j a_j \int_{s_p} b_j(x,y) dl_p + e_p \qquad p = 1,2,\ldots,P \tag{4.5}$$

where e_p is a noise term and a_j is the parameters we want to estimate.

The most commonly used basis in the series-expansion approach is the standard (or the natural) basis which consists of a unity value at the position corresponding to each pixel (m,n) of a (sampled) image and zero otherwise. Other examples are the Fourier basis, splines etc. [28, 18].

In a discrete formulation the integral in (4.5) will simply be replaced by a (weighted) sum

$$z_p = \sum_j a_j \sum_m \sum_n b_j(x_m, y_n) \delta(x_m \cos\theta_p + y_n \sin\theta_p - t_p) + e_p. \tag{4.6}$$

More specifically, let $f(x,y)$ and $b_j(x,y)$ be sampled on a rectangular $M \times N$ grid. Denote the sampled version of $b_j(x,y)$ as the $M \times N$ matrix $\mathbf{B}_j$ and the sampled $f(x,y)$ as $\mathbf{F}$ (also a $M \times N$ matrix). The discrete sum approximation of the line-integral in Eq. (4.6) can be written in matrix from as, $z_p = \sum_j a_j \boldsymbol{\phi}_p^T \mathrm{Col}(\mathbf{B}_j) + e_p = \sum_j a_j \boldsymbol{\phi}_p^T \mathbf{b}_j + e_p$, where $\mathrm{Col}(\cdot)$ is an operator which organizes a $M \times N$ matrix in lexicographic order (one single column vector of length MN). The vector $\boldsymbol{\phi}_p$ contain ones only at those positions (pixels) which are intersected by the projection s_p, and otherwise zeros.

Let $\mathbf{B} = [\mathbf{b}_1 \mathbf{b}_2 \ldots \mathbf{b}_{MN}]$, $\mathbf{f} = \mathrm{Col}(\mathbf{F})$, $\mathbf{a} = [a_1 \; a_2 \ldots \; a_{MN}]^T$, $\mathbf{z} = [z_1 \; z_2 \; \ldots \; z_P]^T$, $\mathbf{e} = [e_1 \; e_2 \; \ldots \; e_P]^T$, and $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \boldsymbol{\phi}_2 \ldots \boldsymbol{\phi}_P]$. Then Eq. (4.4) can be approximated as

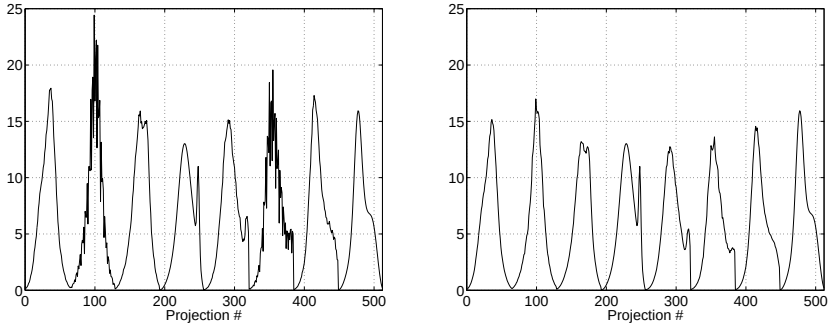
$$\mathbf{f} = \mathbf{B}\mathbf{a} \tag{4.7}$$

and hence, (4.6) can be written as

$$\mathbf{z} = \mathbf{\Phi}^T\mathbf{f} + \mathbf{e} = \mathbf{\Phi}^T\mathbf{B}\mathbf{a} + \mathbf{e} = \mathbf{W}\mathbf{a} + \mathbf{e}. \tag{4.8}$$

If the standard basis is considered, $\mathbf{B}$ reduces to $\mathbf{B} = \mathbf{I}$, and hence $\mathbf{a} = \mathbf{f}$.

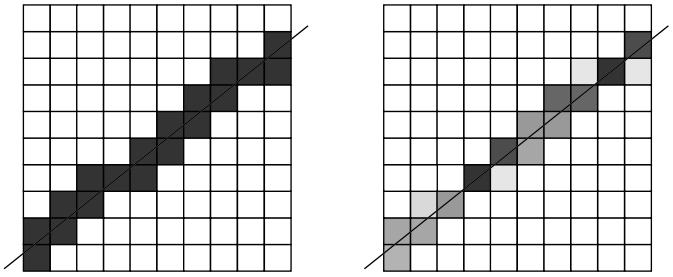
The projection matrix $\mathbf{\Phi}$ will be sparse—containing only ones and zeros. This crude binary method will however introduce model errors in the projections. The errors can be reduced by replacing the binary weight values in the projection matrix with the relative length of the ray inside each pixel intersected. This method is known as the *conventional line integral* method (CLI) [29]. More accurate approximations of (4.2) exists, known as *strip projections* [25, 29], but the CLI method is sufficient for our purposes. Figure 4.2 shows a comparison of the the binary and the CLI projection method, using parallel beam projections, for a smooth phantom (which is used in the simulations later in this chapter, see Figure 4.6(a)). Each peak



(a) Binary method (b) CLI method

Figure 4.2: Projections from a smooth phantom (shown in Figure 4.6(a)) using 64 parallel rays and 8 angles.

in the figure corresponds to 64 parallel projections, and there is one peak for each angle, which is 8 equally spaced angles between 0 and π, giving 512 projections in total. Clearly one can see that the binary method results in artifacts in the projections, especially for the angles $\pi/4$ and $3\pi/4$ (the second and the sixth peak in 4.2(a)). There is some evidence that some iterative algoritms may even diverge if the crude binary version of calculating projections is used [30], and therefore, the CLI method is adopted here. Figure 4.3 illustrates how the different elements in the projection matrix are



(a) Binary method (b) CLI method

Figure 4.3: An illustration of the two projection methods where the weight values in the projection matrix is proportional to is level of the shade in the corresponding pixels.

assigned in the binary and the CLI method—the gray level in each pixel is proportional to the element weight value.

4.3.2 Closed Form Solutions

The model in Eq. (4.8) can be seen as a system of equations which can be solved for $\mathbf{a}$ and consequently for $\mathbf{f}$. However, there will only be a unique solution to (4.8) if there is no noise, and if $P = MN$, that is, the number of measurements equals the number of basis functions used. If $P \neq MN$ or $\mathbf{e} \neq \mathbf{0}$, then the least-squares solution can be used,

$$\hat{\mathbf{a}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{z} \tag{4.9}$$

which is the solution to the optimization problem

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{z} - \mathbf{W}\mathbf{a}\|^2. \tag{4.10}$$

This solution will not exist if the matrix $\mathbf{W}^T\mathbf{W}$ is singular, which for example happens if $\dim(\mathbf{z}) < \dim(\mathbf{a})$. If $\mathbf{z}$ has a lower dimension than $\mathbf{a}$, then there will be infinitely many $\mathbf{a}$ that gives the same $\mathbf{z}$. A unique solution can be found by means of the generalized, or pseudo, inverse solution [31, 32]

$$\hat{\mathbf{a}} = \mathbf{W}^+\mathbf{z}. \tag{4.11}$$

The generalized inverse is calculated by means of a *singular value decomposition* (SVD), that is

$$\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{4.12}$$

where $\mathbf{U}$ is a $P \times P$ matrix, $\mathbf{V}$ is $MN \times MN$ matrix, which both are orthogonal, and $\mathbf{D}$ is a $P \times MN$ matrix of the block form

$$\mathbf{D} = \begin{bmatrix} \mathbf{\Sigma} \\ \mathbf{0} \end{bmatrix}. \tag{4.13}$$

The diagonal matrix $\mathbf{\Sigma}$ contains the square root of the r nonzero eigenvalues of both $\mathbf{W}^T\mathbf{W}$ and $\mathbf{W}\mathbf{W}^T$, where r is the rank of $\mathbf{W}$ [33]. The pseudo inverse, or the Moore-Penrose inverse, of $\mathbf{W}$ is then

$$\mathbf{W}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^T \tag{4.14}$$

where $\mathbf{D}^+ = [\mathbf{\Sigma}^{-1}\mathbf{0}]$. If some of the singular values are close to zero $\mathbf{\Sigma}$ becomes ill-conditioned which implies strong noise amplification. The solution is simply to truncate the SVD (TSVD)—only the "sufficiently" large singular values are used. This is then an *approximate* generalized matrix inverse. If the equation system (4.8) is inconsistent ($P \neq MN$ or $\mathbf{e} \neq \mathbf{0}$), then the estimate $\hat{\mathbf{a}}$ in Eq (4.11) is the *minimum norm* least squares solution. That is, if $\mathbf{a}$ is split in one part which is in the row space of $\mathbf{W}$, $\mathbf{a}_r$, and one part which is the null-space, $\mathbf{a}_n$, then the pseudo inverse gives the $\mathbf{a}$ which minimizes the norm $\|\mathbf{a}\|^2 = \|\mathbf{a}_r\|^2 + \|\mathbf{a}_n\|^2$ subject to $\mathbf{z} = \mathbf{W}\mathbf{a}$. Note that all $\mathbf{W}\mathbf{a}_n = \mathbf{0}$ since $\mathbf{a}_n$ is in the null-space of $\mathbf{W}$. The pseudo inverse solution simply sets $\hat{\mathbf{a}} = \mathbf{a}_r$.

4.3.3 Regularization Techniques

The remedy for solving the ill-posedness by truncating the SVD above is a form of *regularization*. Regularization is the standard approach for solving these types of ill-conditioned inverse problems. The basic feature of regularization is the trade off between the belief in the measurements and the belief in the *a priori* knowledge. Following Demoment [31], a regularized solution to (4.8) can be formulated as the solution to the following optimization problem

$$\begin{aligned} \hat{\mathbf{a}}(\mathbf{z}, \mu) &= \arg\min_{\mathbf{a}} \{J_1(\mathbf{a}, \hat{\mathbf{a}}_0) + \mu J_2(\mathbf{a}, \hat{\mathbf{a}}_\infty)\} \\ &= \arg\min_{\mathbf{a}} \{V(\mathbf{a})\} \end{aligned} \tag{4.15}$$

where J_1 is a measure of the fit to the measurements, J_2 is a measure of the fit to the *a priori* knowledge, $\hat{\mathbf{a}}_0$ is the least-squares solution (4.9), and $\hat{\mathbf{a}}_\infty$ corresponds to an *a priori* (smooth) distribution. Thus, the factor μ controls the belief in the measurements $\hat{\mathbf{a}}_0$ vs the belief in the prior knowledge $\hat{\mathbf{a}}_\infty$. A common choice for J_1 is the weighted quadratic distance

$$J_1(\mathbf{a}, \hat{\mathbf{a}}_0) = (\mathbf{a} - \hat{\mathbf{a}}_0)^T \mathbf{S}_1 (\mathbf{a} - \hat{\mathbf{a}}_0) \tag{4.16}$$

where $\mathbf{S}_1$ is a diagonal matrix. If the measurement noise is Gaussian, the matrix $\mathbf{S}_1$ is usually chosen as

$$\mathbf{S}_1 = \mathbf{W}^T \mathbf{\Lambda}^{-1} \mathbf{W} \tag{4.17}$$

where $\mathbf{\Lambda}$ is diagonal matrix containing the eigenvalues of the noise covariance matrix $\mathbf{C}_{ee} = E\{\mathbf{e}\mathbf{e}^T\}$. A common choice for J_2 is the *Kullback* distance [31, 34, 35]

$$J_2(\mathbf{a}, \hat{\mathbf{a}}_\infty) = \sum_{p=1}^{P} a_p \log(\frac{a_p}{\hat{a}_{\infty p}}) \tag{4.18}$$

which is the negative of the relative entropy of distribution a_p relative to the prior distribution $\hat{a}_{\infty p}$.[1] Another choice for J_2 is

$$J_2(\mathbf{a}) = \mathbf{a}^T \mathbf{S}_2^T \mathbf{S}_2 \mathbf{a} \tag{4.19}$$

where $\mathbf{S}_2$ can be a finite difference operator, which punishes high gradients in $\mathbf{a}$ giving more smooth reconstructions. If $\mathbf{S}_2 = \mathbf{I}$ in (4.19), then only the variance in $\hat{\mathbf{a}}$ will be punished.

This technique has, for example, been used by Bramantini et. al. [18], to measure temperature distributions in power plant boilers. A Fourier series basis was used with $\mathbf{S}_1 = \mathbf{I}$ in (4.16), and a gradient operator as $\mathbf{S}_2$, yielding the solution

$$\hat{\mathbf{a}} = (\mathbf{W}^T \mathbf{W} + \mu \mathbf{S}_2^T \mathbf{S}_2)^{-1} \mathbf{W}^T \mathbf{z}. \tag{4.20}$$

The tuning parameters in J_1 and J_2 above ($\mu, \mathbf{S}_1, \mathbf{S}_2$ etc.) are known as the *hyperparameters*, and they determine the behavior of the reconstruction.

[1] Note that the (relative) entropy used here is not really the same as the entropy used in the statistical sense. Here we assume that $f(x, y) > 0$ and $\iint f(x, x) dx dy = 1$, which results in a strong resemblance of $f(x, y)$ with a probability density function (PDF).

It is generally not an easy task to estimate these parameters, some examples are given in [31] and in the next subsection.

Note that the methods discussed in this section results in linear optimization problems (if the methods based on entropy are excluded), and any linear method can be formulated as $\hat{\mathbf{f}} = \mathbf{H}\mathbf{f}$, where $\mathbf{H}$ is the *degradation* matrix or the *point spread function* matrix of the reconstruction system. The columns in $\mathbf{h}_i$ in $\mathbf{H}$ tells how the corresponding elements f_i in $\mathbf{f}$ will spread to neighboring elements. Thus, by inspecting $\mathbf{H}$, the resolution at different parts of $\mathbf{f}$ can be determined (see Appendix 4.A).

Note also the high dimensionality involved in this kind of image reconstruction. For example, if the standard basis (or any other complete basis) is used, computation of a (regularizied) least squares solution to (4.8) involves inversion of an $MN \times MN$ matrix, which is huge already for relatively small N and M. Consider for example $N = M = 64$, then an inversion of a 4096×4096 matrix is required. To avoid the large matrix inversion, Eq. (4.8) can be solved in an iterative manner. Iterative methods is also the only option if the maximum entropy criteria above is used since the solution to the optimization problem (4.15) can not be expressed in a closed form.

4.3.4 Perturbation Analysis

As described above, the hyperparameters control the trade-off between the belief in *a priori* knowledge and the belief in the measurements. Another way of expressing this is that the reconstructions should be consistent with the measurements, that is, the goal is high *fidelity*, but at the same time $\hat{\mathbf{f}}$ should not be too sensitive to fluctuations in the measurements $\mathbf{z}$. In other words, *stability* with respect to measurements errors is also important. Let $\Delta\mathbf{z}$ be a small perturbation in $\mathbf{z}$, $\Delta\hat{\mathbf{f}}$ be the corresponding perturbation in $\hat{\mathbf{f}}$. Then assume a linear estimation method of the form $\hat{\mathbf{f}} = \mathbf{Q}\mathbf{z}$. If for example, Eq (4.16) and (4.19) are used, then $\mathbf{Q}$ becomes $\mathbf{Q} = \mathbf{B}(\mathbf{W}^T\mathbf{S}_1\mathbf{W} + \mu\mathbf{S}_2^T\mathbf{S}_2)^{-1}\mathbf{W}^T$. Then the error $\Delta\hat{\mathbf{f}}$ is bounded according to

$$\|\Delta\hat{\mathbf{f}}\| \leq \|\mathbf{Q}\| \cdot \|\Delta\mathbf{z}\| \tag{4.21}$$

where the Euclidean norm has been used [36].[2] The Euclidean norm of $\mathbf{Q}$ is equal to the largest singular value of $\mathbf{Q}$, σ_{max}. The singular vectors corresponding to the largest singular values will tell how the perturbation is "spread" over $\hat{\mathbf{f}}$. Marechél et. al. call these the *critical modes* of the

[2] If Eq (4.18) is used Eq (4.21) will be a first order approximation.

reconstruction [36]. Note that the error will be maximal if $\Delta\mathbf{z}$ is equal to the singular vector corresponding to σ_{max}. Thus, the maximum amplification of the error is $\sigma_{max}\frac{\|\Delta\mathbf{z}\|}{\|\Delta\hat{\mathbf{f}}\|}$ which we denote $\sigma(\mu)$ (a function of μ). Now the different regularization algorithms of the form (4.15) can be compared for a desired level of fidelity. The best algorithm will be the one that gives the lowest amplification $\sigma(\mu)$ of the relative error, defined as

$$\frac{\|\Delta\hat{\mathbf{f}}\|}{\|\hat{\mathbf{f}}\|} \leq \sigma(\mu)\frac{\|\Delta\mathbf{z}\|}{\|\mathbf{z}\|}. \tag{4.22}$$

Figure 4.4 illustrates one example of the trade-off between fidelity and stability. A large μ will give a low σ_{max} at the expense of a large mean quadratic error. The matrix $\mathbf{W}^T\mathbf{W}$ in (4.20) is singular in this example since the number of projections is fewer than the number of parameters to estimate, which results in a large SSE, defined as

$$V_{\text{SSE}} = \frac{1}{MN}\sum_i (f_i - \hat{f}_i)^2 \tag{4.23}$$

for small values of μ. Adding noise shifts the optimal value of μ to higher values. The optimum value of μ in the noiseless case, for this example, is approximately 0.8.

4.3.5 Iterative Algebraic Reconstruction Algorithms

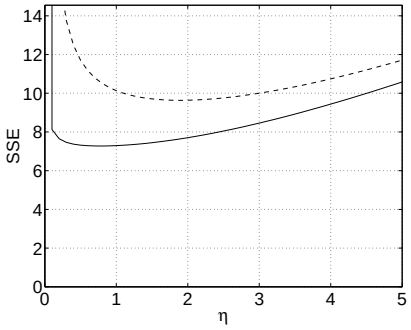
The parameter vector in Eq. (4.8) can be estimated using recursive algorithms that are not explicitly based on a particular optimization criteria. The iterative, or the recursive, reconstruction techniques can roughly be divided in two main groups, additive algebraic (ART), and multiplicative algebraic reconstruction techniques (MART) [25, 37, 28, 38]. As their name implies the correction term is additive for the first type and multiplicative for the latter. The updating formula is

$$\hat{a}_j(n+1) = \hat{a}_j(n) + k_j^{(a)}(n) \tag{4.24}$$

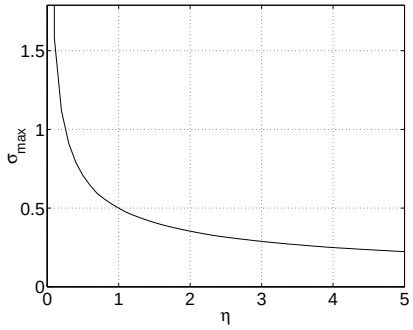
for additive ARTs, and for MARTs

$$\hat{a}_j(n+1) = \hat{a}_j(n) \times k_j^{(m)}(n). \tag{4.25}$$

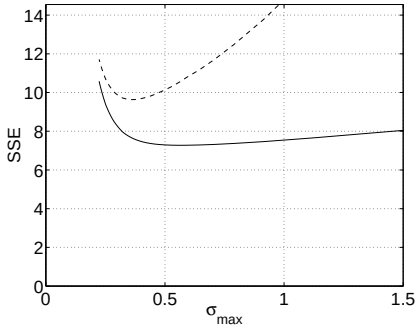
The sub-script j indicates that the jth element in $\hat{a}$ is updated, n is the iteration number, and $k_j(n)$ is the correction factor. The super-scripts (a)



(a) Average SSE vs. μ



(b) σ_{max} vs. μ



(c) Average SSE vs. σ_{max}

Figure 4.4: Plot of the dependence of the sum squared error and σ_{max} of the hyperparamter μ for 15 rays and 10 angles using a wavelet basis (256 basis functions). The average V_{SSE} is calculated from 1000 examples of a the smooth phantom (used later Section 4.4), and $\mathbf{S}_2 = \mathbf{I}$ in (4.20). Solid lines—Noise-free, dashed lines—White Gaussian noise added.

and (m), meaning additive and multiplicative, respectively. The most common correction factor $k_j^{(a)}(n)$ in (4.24) is based on Karczmarz "method of projections" for solving algebraic equations [25],

$$k_j^{(a)}(n) = \frac{z_p - \mathbf{w}_p^T \hat{\mathbf{a}}(n)}{\mathbf{w}_p^T \mathbf{w}_p} w_{p,j} \tag{4.26}$$

where $\mathbf{w}_p^T$ is the pth row in $\mathbf{W}$ (by convention all vectors are expressed in a column format here) and w_{pj} is the jth element in $\mathbf{w}_p$. The updating is performed using one projection per iteration, that is $p = (n \mod P) + 1$.

The correction factor for the MART algorithm is

$$k_j^{(m)}(n) = \left(\frac{z_p}{\boldsymbol{\phi}_p^T \hat{\mathbf{f}}(n)}\right)^{\eta \Phi_{p,j}} \tag{4.27}$$

where η is a relaxation factor which takes values between 0–1 [28, 37]. It can be shown that the MART algorithm maximizes the entropy under the constraints of the measurements, and that $\mathbf{f}$ is positive (see [35] for a proof). This implies that MART can not be applied to other basis directly because $\mathbf{a}$ will not be positive. In order for MART to handle other bases than the standard one, Eq. (4.25) must be modified. This can be accomplished by, first expressing (4.25) in matrix form

$$\hat{\mathbf{f}}(n+1) = \mathbf{K}(n)\hat{\mathbf{f}}(n) \tag{4.28}$$

where $\mathbf{K}(n)$ is a diagonal matrix $\mathbf{K}(n) = \text{diag}\{k_1^{(m)}(n)\, k_2^{(m)}(n)\, \cdots\, k_{MN}^{(m)}(n)\}$, and use (4.7) to express $\mathbf{f}(n)$, and finally multiply (4.28) from left with $\mathbf{B}^T$

$$\hat{\mathbf{a}}(n+1) = \mathbf{B}^T \mathbf{K}(n) \mathbf{B} \hat{\mathbf{a}}(n). \tag{4.29}$$

This becomes, however, computationally in-efficient since two matrix multiplications has to be performed for every TOF measurement z_p in (4.27). Hence, the modified MART algorithm (4.29) is not used here.

There exist several other ways to compute the correction factors for the (M)ART algorithms [28, 38, 38, 37]. For example, a similar relaxation factor as in (4.27) can be added to the ART algorithm as well. Note that if $w_{ij} = 0$ there is no updating of $\hat{a}_j$, that is, $k_j^{(m)}(n) = 1$ for MART, and $k_j^{(a)}(n) = 0$ for ART. Using the standard basis in (4.8) implies that only elements in $\mathbf{f}$ that are intercepted by a projection is updated.

The convergence rate of the algorithms is dependent on the order which the projections are treated, and on the relaxation parameter η. Herman [39]

has shown that a careful choice of updating order can result in a significant improvement in convergence speed for the ART algorithm.

Note also that when (4.8) is over-determined, $P > MN$, the ART solution will oscillate around the intersections of the hyperplanes that each row in (4.8) defines. If (4.8) is underdetermined there will be infinitely many solutions, but the ART algorithm will converge to the solution $\hat{\mathbf{a}}$ which minimizes $\|\mathbf{a}(0) - \hat{\mathbf{a}}\|^2$, where $\mathbf{a}(0)$ is the initial guess [25]. That is, for underdetermined problems, ART finds a solution that will depend on the initial guess, namely the $\hat{\mathbf{a}}$ which is closest to the initial guess $\mathbf{a}(0)$.

4.3.6 Simultaneous Updating in Algebraic Algorithms

In the previous section, the updating of $\hat{a}_j(n)$ in (4.24) and (4.25) was performed projection by projection. Another idea is to compute the average update from all projections and then update $\hat{a}_j(n)$. Using this method to update $\hat{\mathbf{a}}$ for the ART algorithm is known as the *simultaneous iterative reconstruction technique* (SIRT) [25]. Equation (4.24) then becomes

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) + \begin{bmatrix} \sum_i \frac{z_i - \mathbf{w}_i^T \hat{\mathbf{a}}(n)}{\mathbf{w}_i^T \mathbf{w}_i} w_{i,1} \\ \sum_i \frac{z_i - \mathbf{w}_i^T \hat{\mathbf{a}}(n)}{\mathbf{w}_i^T \mathbf{w}_i} w_{i,2} \\ \vdots \\ \sum_i \frac{z_i - \mathbf{w}_i^T \hat{\mathbf{a}}(n)}{\mathbf{w}_i^T \mathbf{w}_i} w_{i,MN} \end{bmatrix}. \tag{4.30}$$

The sums in (4.30) can be rewritten as

$$\begin{aligned} \sum_p \frac{z_p - \mathbf{w}_p^T \hat{\mathbf{a}}(n)}{\mathbf{w}_p^T \mathbf{w}_p} w_{p,j} &= \frac{1}{\sum_p \mathbf{w}_p^T \mathbf{w}_p} [w_{1,j}\ w_{2,j} \cdots w_{P,j}](\mathbf{z} - \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_P^T \end{bmatrix} \hat{\mathbf{a}}(n)) \\ &= \frac{1}{\sum_p \mathbf{w}_p^T \mathbf{w}_p} \tilde{\mathbf{w}}_j^T (\mathbf{z} - \mathbf{W}\hat{\mathbf{a}}(n)) \\ &= \eta \tilde{\mathbf{w}}_j^T (\mathbf{z} - \hat{\mathbf{z}}(n)) \end{aligned} \tag{4.31}$$

where $\tilde{\mathbf{w}}_j$ is j:th column vector of $\mathbf{W}$, $\eta = \frac{1}{\sum_p \mathbf{w}_p^T \mathbf{w}_p}$, and $\hat{\mathbf{z}}(n) = \mathbf{W}\hat{\mathbf{a}}(n)$. Eq (4.30) can then be written

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) + \eta \mathbf{W}^T (\mathbf{z} - \hat{\mathbf{z}}(n)). \tag{4.32}$$

Note that $\mathbf{W}^T(\mathbf{z} - \hat{\mathbf{z}}(n)) = \frac{\partial V(\mathbf{a}(n))}{\partial \mathbf{a}(n)}$ with $J_1 = \|\mathbf{z} - \hat{\mathbf{z}}(n)\|^2$ and $\mu = 0$ in (4.15). That is, Eq (4.32) is a steepest descent algorithm for solving the (non-regularized) LS problem in (4.9), also known as the Landweber algorithm [40, 41].

The two different orders of updating $\hat{\mathbf{a}}(n)$, used in ART and SIRT respectively, are known in the neural network community as *pattern* and *batch* learning.[3] The updating of weights in a neural network can be accomplished after presenting each example (pattern learning) or after presenting all examples (batch learning).

Batch versions of the MART algorithm can similarly be found by calculating a correction factor for all projections, as in the *simultaneous* MART (SMART) algorithm by Byrne [42] where

$$k_j^{(m)}(n) = \Pi_{p=1}^{P} \left(\frac{z_p}{\boldsymbol{\phi}_p^T \hat{\mathbf{f}}(n)}\right)^{\eta \Phi_{p,j}} \tag{4.33}$$

or as in the *expectation maximization maximum likelihood method* [43] where

$$k_j^{(m)}(n) = \sum_{p=1}^{P} \left(\frac{z_p}{\boldsymbol{\phi}_p^T \hat{\mathbf{f}}(n)}\right) \Phi_{p,j}. \tag{4.34}$$

Here is the modified version (4.29) more attractive since the (costly) matrix multiplications is only computed after each batch (all projections), and not after every single projection as in the previous section. Two other benefits of the simultaneous approach are also that; 1) the noise on the projections is averaged—simultaneous algorithms are known to produce much smoother images than the sequential (pattern) versions [30], and 2) they are easy to parallelize for multiprocessor machines.

Stability of the Steepest-descent Algorithm

If η is not carefully chosen the recursive algorithms above can easily become unstable and the solution may diverge. Small values of η will give a very slow convergence and large values may give an unstable algorithm. For the steepest-descent algorithm (4.32), the necessary and sufficient conditions for stability is that $\eta < \frac{2}{\lambda_{max}}$, where λ_{max} is the largest eigenvalue of $\mathbf{W}\mathbf{W}^T$ [43,

[3] The objective for neural nets (see also Chapter 2) is, for example, to train a net from a number of known examples and then classify the unknown patterns.

44]. This can be shown by noting that, in order for the error to decrease at each iteration, the following property must hold

$$V(n) - V(n+1) > 0, \tag{4.35}$$

where $V(n) = \|\mathbf{z} - \mathbf{W}\hat{\mathbf{a}}(n)\|^2$, and the updating in Eq. (4.32) has been used to calculate $V(n+1)$. Then $V(n) - V(n+1) = (\mathbf{z} - \mathbf{W}\mathbf{a}(n))^T \eta \mathbf{W}\mathbf{W}^T (2\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T)(\mathbf{z} - \mathbf{W}\mathbf{a}(n))$, where $\mathbf{I}$ is the identity matrix. Thus if (4.35) should hold the matrix $(2\mathbf{I} - \eta\mathbf{W}\mathbf{W}^T)$ must be positive-definite, that is, $|\frac{2}{\eta}| > |\mathbf{W}\mathbf{W}^T|$ which is guaranteed if $\eta < \frac{2}{\lambda_{max}}$.

The convergence of (4.32) will be slow if the spread between the eigenvalues of $\mathbf{W}\mathbf{W}^T$ is large—a large eigenvalue spread will give "narrow valleys" in the error surface which will give slow convergence if the starting vector is not chosen carefully [40].

4.3.7 Adaptive Learning Rate

One way to circumvent the problem of choosing an explicit η, also known as the *learning rate*, is to make the learning rate adaptive. A strategy for updating η can, for example, be:

> The learning rate is increased by a factor lr_{inc} if the present sum squared error, $V_{\text{SSE}}(n)$ (defined in (4.23)), is smaller than the last one, $V_{\text{SSE}}(n-1)$. If the present error exceeds the last one by a factor of lr_{ratio}, there is no parameter updating and the present learning rate, η is decreased by a factor lr_{dec}.

The factors lr_{inc} and lr_{ratio} should be chosen slightly larger than one and the factor lr_{dec} should be smaller than one.

In the simulations performed later in this chapter the learning rate was adjusted after each pass. That is, all projections were considered before η was updated according to the algorithm above.

4.3.8 Choice of Basis

The (regularized) LS solution involves a matrix inversion which is very computationally demanding due to the high dimensionality of the problem. One way to alleviate this problem is to use a truncated basis instead of the full

standard basis. This is accomplished by approximating $\mathbf{f}$ using

$$\mathbf{f} = \sum_{j=1}^{N_b} a_j \mathbf{b}_j, \tag{4.36}$$

where $N_b < MN$. Vaguely speaking, the dimension N_b of the vector $\mathbf{a}$ should be chosen according to the amount of information contained in the measurements $\mathbf{z}$. If standard regularization using the full standard basis is used, and only a few measurements are available, one has to use a large μ in Eq (4.20) which will smooth the estimate $\hat{\mathbf{f}}$ heavily.[4] However, if the number of variables is reduced—using a truncated basis—there are fewer parameters to estimate and the variance of the estimates should be lower. The choice of basis, and the number of parameters to use will depend on the available amount of prior knowledge. In our application it is reasonable to assume that the temperature distributions are rather smooth with local (smooth) gradients where heat sources are located—we will not have the high (contrast) gradients found in medicine, for example. Interesting families of basis are, for example, the wavelet family [11, 15, 45], comprised of local functions (functions with compact support). Examples of 1D wavelets can be found in Chapter 2. We use the 2D version of the same smooth wavelet (Coiflet 2) that was used in Chapter 2. Four examples of the 2D Coiflet 2 wavelet are shown in Figure 4.5.

The choice of basis can be seen as a user variable which depends on the application at hand. The optimal basis, in a sense that it minimizes $E\{||\mathbf{f} - \mathbf{Ba}||^2\}$ for a given N_b, consists of the eigenvectors of the covariance matrix, $\mathbf{C}_{ff} = E\{(\mathbf{f} - \mathbf{m}_f)(\mathbf{f} - \mathbf{m}_f)^T\}$ of $\mathbf{f}$, where $E\{\cdot\}$ is the expectation operator and $\mathbf{m}_f$ the mean of $\mathbf{f}$ [4]. The basis comprised of the eigen vectors is denoted as the *principal component* (PC) basis.

In order to use the PC basis, a number of representative examples of $\mathbf{f}$ is needed for estimation of the covariance matrix $\mathbf{C}_{ff}$. This is usually accomplished using the standard formula

$$\hat{\mathbf{C}} = \frac{1}{J-1} \sum_{j=1}^{J} (\mathbf{f}_j - \hat{\mathbf{m}})(\mathbf{f}_j - \hat{\mathbf{m}})^T, \tag{4.37}$$

where $\hat{\mathbf{m}}$ is the estimated mean and J is the number of available examples. Suitable temperature distribution examples can be found by means of simulations, real measurements, or a combination of both.

[4] If $P < MN$ then the matrix $\mathbf{W}^T\mathbf{W} = \mathbf{\Phi}\mathbf{\Phi}^T$ in (4.20) will not have full rank, and hence will not be invertible without regularization.

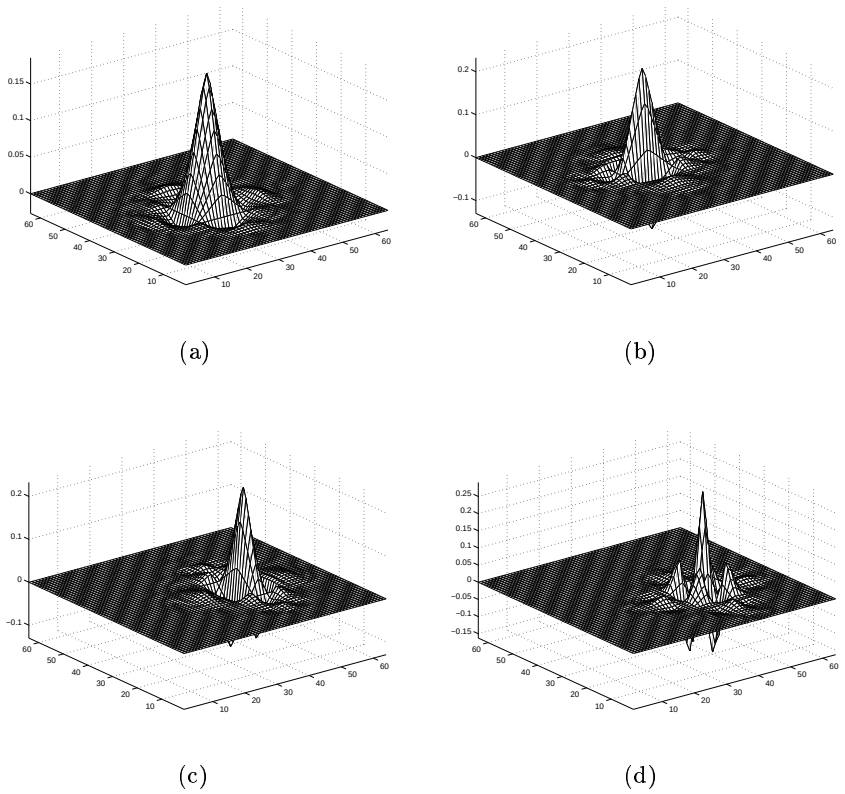


Figure 4.5: Four examples of the 2D Coiflet 2 wavelet.

If the prior knowledge is too weak to define a useful adapted basis, then one must resort to other more general bases, like the wavelet basis described above. Another common choice is the Fourier basis which is comprised of complex exponentials of the form $\exp[-i2\pi(ux/M + vy/N)]$, where u and v are the spatial frequencies. This basis can be motivated by the fact that temperature distributions must be smooth and can, therefore, be approximated by truncated Fourier series. Note that this basis is also used (implicitly) in the FBA algorithm. The FBA uses the full basis, but usually a Hamming window is applied to filter out high frequencies which otherwise are amplified due to the high-pass *wedge* filter which results from the polar to rectangular co-ordinate transformation used in the FBA algorithm [25].

4.3.9 Minimizing Reconstruction Errors (MRE)

All algorithms described above focus on estimating a parameter vector $\mathbf{a}$ which minimizes the *measurement* error $||\mathbf{z} - \mathbf{\Phi}^T\mathbf{B}\mathbf{a}||^2 = ||\mathbf{z} - \hat{\mathbf{z}}||^2$. Another more natural approach is to minimize the *reconstruction* error $||\mathbf{f} - \hat{\mathbf{f}}||^2$. Following [22], the idea is the choose the parameter vector which minimizes the *expected reconstruction error* $V_{\text{MRE}} = E\{||\mathbf{f} - \hat{\mathbf{f}}||^2\}$. Instead of first estimating $\mathbf{a}$ this can be accomplished in a single step as $\hat{\mathbf{f}} = \mathbf{Q}^T\mathbf{z} + \mathbf{d}$. Strait forward calculations (see [22]) give

$$\mathbf{Q} = \mathbf{C}_{zz}^{-1}\mathbf{\Phi}^T\mathbf{C}_{ff}, \tag{4.38}$$

where $\mathbf{C}_{zz} = E(\mathbf{z} - \mathbf{m}_z)(\mathbf{z} - \mathbf{m}_z)^T$, and $\mathbf{m}_z$ is the mean vector of $\mathbf{z}$. The optimal estimate of $\mathbf{d}$ is

$$\mathbf{d} = (\mathbf{I} - \mathbf{Q}^T\mathbf{\Phi}^T)\mathbf{m}_f. \tag{4.39}$$

Note that, if one studies the expression (4.38) above for the MRE method, it has strong resemblance with methods using the PC basis. The optimal MRE method also uses the eigenvectors of the covariance matrix of the temperature distributions. However, the MRE method also takes the covariance of the measurements into account which should result in superior performance.

4.4 Simulations

In this section the performance of the algorithms described above are compared for simulated data. At the time of printing no useful physical model was available for simulation of temperature distributions. It is, however,

	x_1	x_2	y_1	y_2	α
min	0.2	0.1	-0.2	-0.35	0.6
max	0.45	0.35	0.05	-0.1	1.4

Table 4.1: Parameter limits for the phantom.

reasonable to assume that the distributions should be rather smooth. There will, of course, be gradients due to sun light, air jets, heat radiators, humans, computers etc, in a real measurement situation. This temperature change will, though, not be unbounded. Here an artificial model based on a smooth phantom is used, also found in [22, 28]. This model suffice for the purpose here, which is to measure reconstruction performance from a low number of projections. The phantom is given by

$$\begin{aligned}\mathrm{ph}(x,y) = 1.09\alpha(0.3\cos(x,y) + 0.8\exp[-81(x-x_1)^2 - 36(y-y_1)^2] \\ + \exp[-64(x-x_2)^2 - 36(y-y_2)^2])\end{aligned} \quad (4.40)$$

where $\cos(x,y)$ is given by

$$\cos(x,y) = 0.2(1-\cos(2\pi(x+0.5)^{4/5})) \times (1-\cos(2\pi(y+0.5)^{2/3})) \quad (4.41)$$

for $|x| < 0.5$, $|y| < 0.5$, otherwise zero. The parameters in (4.40) and (4.41) were drawn from uniform distributions on intervals given in Table 4.1. The phantom were sampled in $[-0.5,-0.5] \times [0.5,0.5]$ on a 64×64 grid giving 4096 measurements. Figure 4.6 shows four realizations of the phantom. For simplicity parallel beam projections have been used, where the the number of projections and the angles of the projections have been varied. The performance has been measured with the sum squared error (Eq. (4.23)), for each method.

4.4.1 Iterative Algorithms vs. The Filtered Back-projection Algorithm

The iterative algorithms ART and MART, where first compared with the filtered back-projection algorithm (FBA) [25]. It is known that FBA requires a large set of projections in order to produce good reconstructions. In Figure 4.7 the FBA is compared to ART and MART (using the standard basis) with 91×8, and 91×64 projections respectively, for the example shown in Figure 4.6(a). They all perform well if 91×64 projections are used, but when

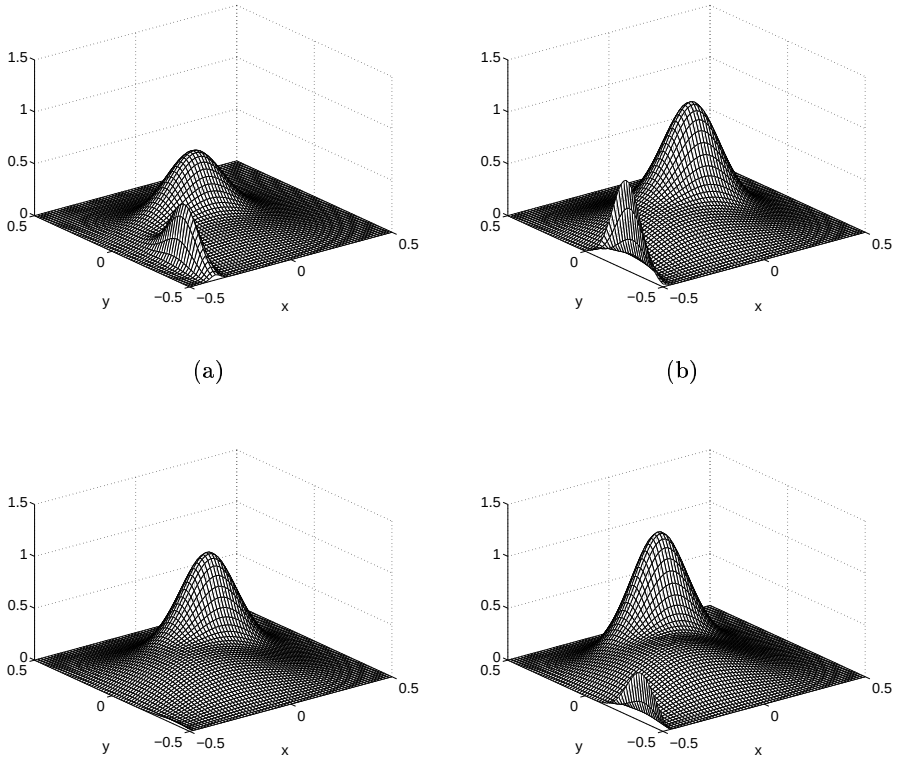
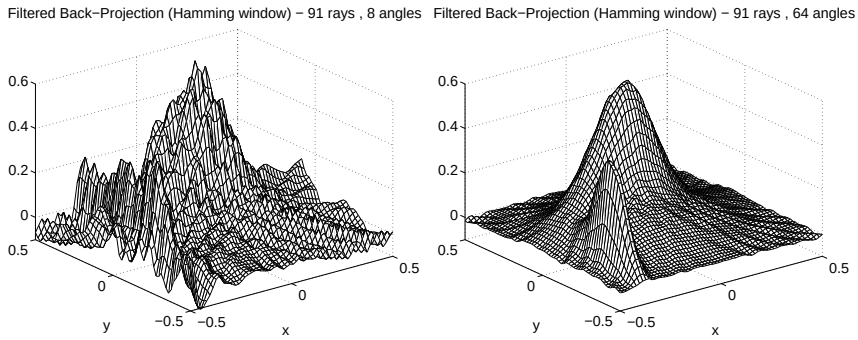
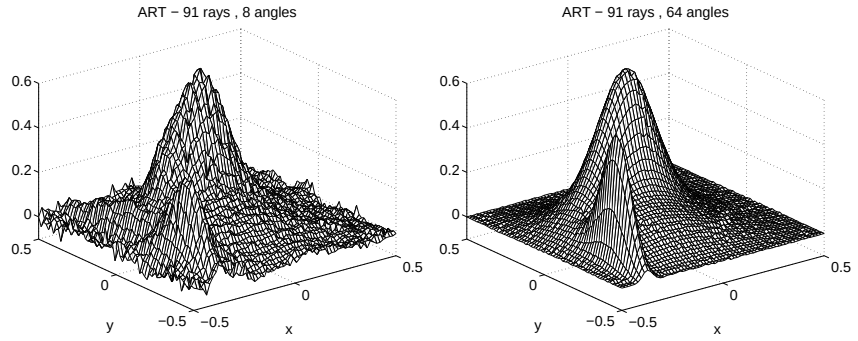


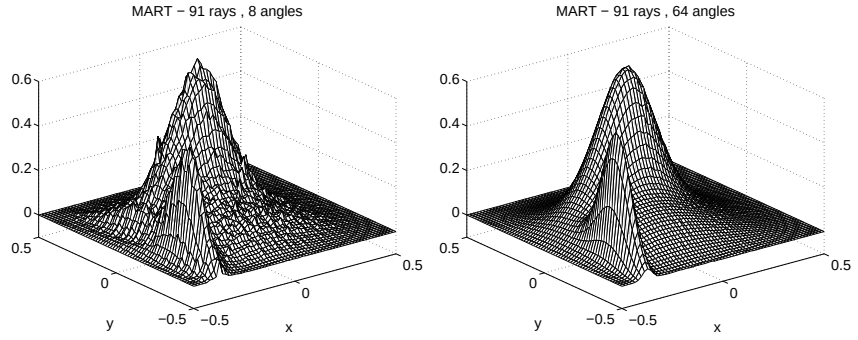
Figure 4.6: Four realizations of the phantom.



(a) FBA 91×8 (b) FBA 91×64



(c) ART 91×8 (d) ART 91×64



(e) MART 91×8 (f) MART 91×64

Figure 4.7: Reconstructions of the phantom shown in Figure 4.6(a) using 91 rays and, 8 and 64 angles respectively (the standard basis has been used in the ART and MART algorithms). As can bee seen the FBA do not perform well, compared to ART and MART, for the case when 8 angles (and 91 rays) has been used. Using 8 angles and 91 rays gives a total of 728 measurements which is too large for this application, indicating that the FBA is not suitable algorithm for reconstruction from sparse measurements.

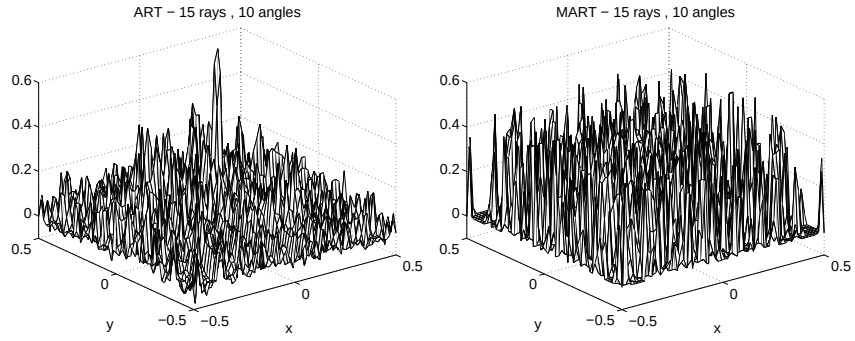
using 8 angles the FBA method clearly has the worst performance. In the temperature problem considered in this work $91 \times 8 = 728$ projections is still too large. One should also that, when using a low number of projections, some pixels will not be updated if the standard basis is used, since they are not intersected by a projection. Figure 4.8 shows reconstructions where the number of projections has been substantially reduced to 150. Some pixels in Figure 4.8(a) and (b) are not updated (≈ 600 out of 4096) but the number of parameters to estimate is still much higher then the number of measurements. This results in the high variance in the estimates, (a) and (b), which clearly requires some form of regularization. Using a reduced basis, like the wavelet basis in (c) and (d), reduces the variance substantially (at the expense of increased bias).

The conclusion from this experiment is that the FBA is not suitable method for reconstruction from sparse measurements. The experiments also showed the necessity to substantially reduce the dimension of the parameter vector (using a suitable reduced basis) for the ART algorithm.

4.4.2 Simulation Results

In this section the performance of a selection of the algorithms discussed previously in this chapter are compared. The propose is to evaluate the performance when the number of measurements is very low (in the order of 10 projections). One consequence of this is that it is not feasible to use the standard basis, which implies that some algorithms are unsuitable (see last subsection). Here the ART, SIRT, modified SMART (M-SMART), regularized least squares (R-LS), the truncated singular value decomposition method (TSVD), and the MRE method have been compared. The methods where compared by calculating the average sum squared error (SSE) from 100 examples, using a PC and a wavelet basis. Table 4.2 and Table 4.3 show the results using a PC basis (with 25 eigenvectors) and a wavelet basis (using the 64 largest scale wavelets) respectively.

All simulations with the iterative algorithms where performed using the adaptive learning rate method. For the R-LS method the η which yielded the lowest (average) SSE on the test set was chosen and the number of singular values used in the TSVD method was chosen in the same way. For the MRE method the regularity was assured by noting that the inverse of a symmetric matrix $\mathbf{C}$ can be expressed using its eigenvalues and eigenvectors



(a) ART—Standard basis

(b) MART—Standard basis

(c) ART—Wavelet 64

(d) ART—Wavelet 256

Figure 4.8: Reconstructions of the phantom shown in Figure 4.6(a) using the iterative ART and MART algorithms, with 15 rays and 10 angles for the Standard basis and the Coiflet 2 wavelets (with 64 and 256 basis functions respectively). The performance is clearly unsatisfactory for the standard basis when this relatively low amount of measurements (150 projections) has been used. Using the (reduced) wavelet basis improves the performance significantly.

Rays	Angles	Average SSE					
		ART	SIRT	M-SMART	R-LS	TSVD	MRE
3	3	58.0	49.3	54.0	48.8	48.8	12.5
5	5	25.2	16.1	10.9	11.9	14.9	2.9
15	10	2.0	1.25	1.33	1.23	1.20	0.46

Table 4.2: Average SSE using a 25 component PC basis (for 100 examples).

Rays	Angles	Average SSE					
		ART	SIRT	M-SMART	R-LS	TSVD	MRE
3	3	125.7	124.1	214.0	123.9	124.1	12.5
5	5	52.7	52.3	49.1	50.0	51.7	2.9
15	10	10.5	10.4	9.5	9.8	9.6	0.46

Table 4.3: Average SSE using a 64 component wavelet basis (for 100 examples).

with the formula

$$\mathbf{C}^{-1} = \sum_{j=1}^{r} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j^T \tag{4.42}$$

where r is the rank of $\mathbf{C}$, $\mathbf{v}_j$ are the eigenvectors and λ_j the corresponding eigenvalues. If $\mathbf{C}$ is ill-conditioned, only the first (sufficiently large) eigenvalues λ_j (and corresponding eigenvectors) are used for calculating the inverse. Note that the rank of $\mathbf{C}_{zz}$ will depend on both the (number of) measurements and the distribution of $\mathbf{f}$. Typically, the eigenvalues of $\mathbf{C}_{zz}$ drops off faster than the corresponding ones for $\mathbf{C}_{ff}$, indicating that $\mathbf{z}$ does not contain all information about $\mathbf{f}$ (see also [22]).

By observing the SSE one can conclude that the MRE method and methods using the PC basis have superior performance compared with methods using the wavelet basis. The particular choice of algorithm (if the MRE method is excluded) seems to be of less importance. The important issue is the choice of basis. The PC basis clearly has superior performance compared to the wavelet basis, and the MRE method has superior performance compared to all the other algorithms studied here. However, the SSE does not give a complete picture of the behavior of the methods. In Figure 4.9 and Figure 4.10 three reconstruction examples are shown which give a clearer

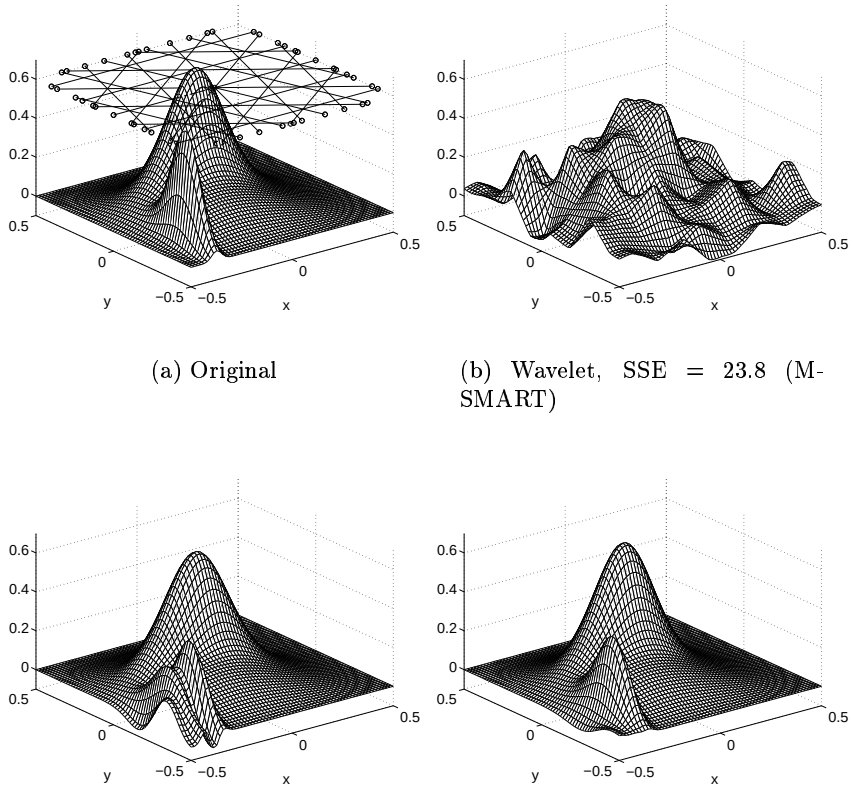
view of the reconstruction performance. Note that the reconstruction error is dependent on both the distribution of $\mathbf{f}$ and the location of the measurements $\boldsymbol{\phi}_i$. Using a low number of measurements (projections) that are located at positions outside areas where the main temperature variation occur will result in poor performance. If one has knowledge of the distribution of $\mathbf{f}$ one should, of course, choose the projections $\boldsymbol{\phi}_i$ based on that prior knowledge to obtain a low reconstruction error. There are however cases where it is reasonable to tolerate larger errors at some parts of $f(x, y)$. An example is an office space where the temperature in the center of the room probably is of more importance than the sidewall temperature. The US paths should then be chosen accordingly.

For this application, and in general when the number of measurements is much lower than the parameters to estimate, the importance of prior knowledge is vital for the reconstruction/estimation performance. This can be illustrated with the following example. Let us assume that the image (object) to be reconstructed is a cylinder with fixed radius but with a varying height. If no prior knowledge is available, one has to assume that the spatial frequencies of the image are unlimited and as a consequence of the Nyquist condition, one must sample infinitely dense in order to avoid aliasing. However, if it is known that the reconstructed object is a cylinder, only measurement is needed (projection) to estimate its height. Thus, the performance of the reconstruction process, in particular the choice of basis functions is to a large extend dependent on the available prior knowledge.

The simulations performed here clearly shows that that, using basis functions learned from example data (which can be regarded as prior knowledge) outperforms methods using more general bases, like the wavelet basis.

4.5 Conclusions

A short review of reconstruction algorithms appropriate for reconstruction of temperature distributions, which are based on a time of flight measurements performed by ultrasonic transducers, has been presented. The feasibility of the algorithms have been analyzed for this temperature mapping application. This has been accomplished by performing simulations for a standard soft phantom model. The temperature distribution was expressed using a linear model taking the form of a linear combination of basis functions. Two main issues, the choice of basis functions and the estimation algorithm was considered.



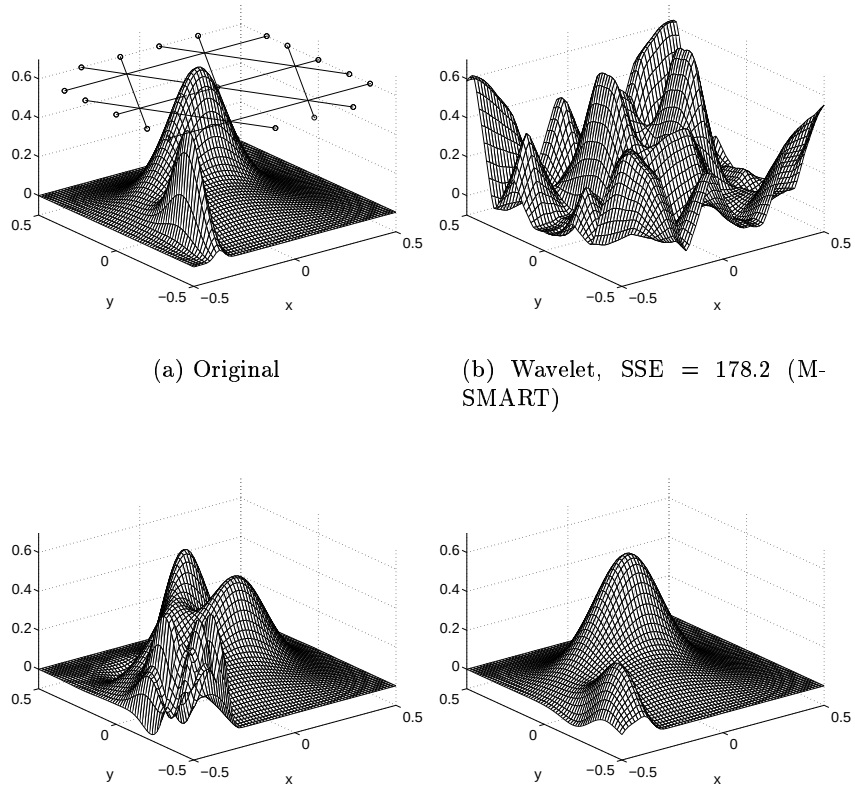
(a) Original

(b) Wavelet, SSE = 23.8 (M-SMART)

(c) PC, SSE = 4.5 (M-SMART)

(d) MRE, SSE = 0.73

Figure 4.9: Reconstructions using 5 angles and 5 rays for, the Coiflet 2 wavelet basis (using 64 wavelets), the PC basis (using 25 eigenvectors), and the MRE method. The US sensors (marked with circles) and the paths are overlaid in subfigure (a). The reconstruction using the wavelet basis exhibits rather large deviations from the original in areas which not are intersected by any projections. This behavior is not seen when the PC basis, and the MRE method, has been used.



(a) Original

(b) Wavelet, SSE = 178.2 (M-SMART)

(c) PC, SSE = 46.2 (M-SMART)

(d) MRE, SSE = 3.82

Figure 4.10: Reconstructions using 3 angles and 3 rays for, the Coiflet 2 wavelet basis (using 64 wavelets), the PC basis (using 25 eigenvectors), and the MRE method. The US transducers (marked with circles) and the paths are overlaid in subfigure (a). The deviations from the original for the wavelet reconstruction is even more pronounced than in Figure 4.10, and the PC reconstruction starts to have a similar behavior. The MRE method do not have these artifacts and the performance is clearly superior to the other methods.

Generally, the choice of reconstruction algorithm depends on several factors, such as, number of measurements, amount of prior knowledge, required resolution, available computational capacity and real time requirements, cost, etc. If the number of measurements is low, prior information about the reconstructed image must be incorporated to achieve good performance. We have shown that for this application it is possible to, based on simulations or real measurements, find particular basis functions adapted to the problem that clearly improve the reconstruction performance. For the iterative algorithms studied here, the estimation speed will depends on a convergence rate of the chosen algorithm for the particular basis functions. The slowest iterative algorithm tested was the M-SMART algorithm. However, the M-SMART algorithm has the advantage that it generates smoother reconstructions than both the ART and the SIRT algorithms. If the ART and the SIRT algorithms are compared, the SIRT algorithm generates smoother reconstructions than the ART ditto. This is probably due to the noise averaging performed in the simultaneous updating when all projections are considered before each update. Note that the SSE (sum squared error) measure does not reveal whether the estimates are smooth or not a large over-shoot and small oscillations may have the same SSE.

The general conclusion that can be drawn from all simulations is that the most important factor, determining the reconstruction performance, is the proper choice of basis. If the MRE method is excluded, all the other methods gave an SSE in the same order for the bases examined here. The MRE method was, however, superior to all other methods. The optimal MRE method also uses the eigenvectors of the covariance matrix of the temperature distributions, but the MRE method also takes the covariance of the measurements into account which results in superior performance.

It should also be noted that if a basis that is very well adapted to the problem at hand is used, poor performance can expected for new data that does not fit this basis. This can be both an advantage and a disadvantage depending on the application, since it makes the system less sensitive to outliers but also less general. In this application rather strong assumptions regarding the temperature distributions must be made, due to the low number of available measurements. Thus, the clearly preferred reconstruction algorithm is the MRE method.

4.A Point Spread Function Interpretation

The reconstruction algorithms, described previously in this chapter, all give estimates $\hat{\mathbf{f}}$ which to a varying degree are degraded compared to the original $\mathbf{f}$. This is due to the ill-posedness of the problem and the limited number of measurements. If linear algorithms are considered, this degradation can be written $\hat{\mathbf{f}} = \mathbf{H}\mathbf{f}$n where the matrix $\mathbf{H}$ is the *degradation* matrix. Now it is easy to see that each column $\mathbf{h}_i$ in $\mathbf{H}$ describes how pixel i in $\mathbf{f}$ "spreads" to neighboring pixels. The $\mathbf{h}_i$s are therefore known as the *point spread functions* (PSF) of the system. Note that the system in general is not translational invariant resulting in different PSFs for all pixels.

Let us begin by re-formulating the problem. First, assume for simplicity that $\mathbf{f}$ is expressed using an ON-basis

$$\mathbf{f} = \mathbf{B}\mathbf{a}, \tag{4.43}$$

and that the measurements are projections of the form

$$\mathbf{z} = \mathbf{\Phi}^T\mathbf{B}\mathbf{a} \tag{4.44}$$

$$= \mathbf{W}\mathbf{a}. \tag{4.45}$$

Then, perform a singular value decomposition of $\mathbf{W}$

$$\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{4.46}$$

and let $\mathbf{W}^+$ be the generalized inverse (which can be truncated as described in Section 4.3)

$$\mathbf{W}^+ = \mathbf{U}\mathbf{D}^+\mathbf{V}^T. \tag{4.47}$$

Then the estimate, $\hat{\mathbf{a}}$, can be expressed as

$$\hat{\mathbf{a}} = \mathbf{W}^+\mathbf{W}\mathbf{a}. \tag{4.48}$$

Inserting (4.46) and (4.47) in (4.48), gives

$$\hat{\mathbf{a}} = \mathbf{V}\mathbf{D}^+\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{a} \tag{4.49}$$

$$= \mathbf{V}\mathbf{D}^+\mathbf{D}\mathbf{V}^T\mathbf{a} \tag{4.50}$$

since $\mathbf{U}$ is orthonormal. Then, by using (4.43), $\hat{\mathbf{f}}$ can be expressed as

$$\hat{\mathbf{f}} = \mathbf{B}\mathbf{V}\mathbf{D}^+\mathbf{D}\mathbf{V}^T\mathbf{B}^T\mathbf{f} \tag{4.51}$$

$$= \mathbf{H}\mathbf{f} \tag{4.52}$$

$$= [\mathbf{h}_1\mathbf{h}_2\ldots\mathbf{h}_{MN}]\mathbf{f}. \tag{4.53}$$

The vector $\mathbf{h}_i$ is now the PSF for element i in $\mathbf{f}$. If $\mathbf{W}$ is invertible (regular), then $\hat{\mathbf{a}} = \mathbf{W}^{-1}\mathbf{W}\mathbf{a} = \mathbf{I}\mathbf{a} = \mathbf{a}$, and if $\mathbf{B}$ is a complete ON-basis, then $\mathbf{H}$ will be the identity matrix and the reconstruction will be perfect. Since $\mathbf{W}$ generally is ill-conditioned, and hence no exact inverse exists ($\mathbf{H} \neq \mathbf{I}$), the $\mathbf{h}_i$:s will have some spread around element i.

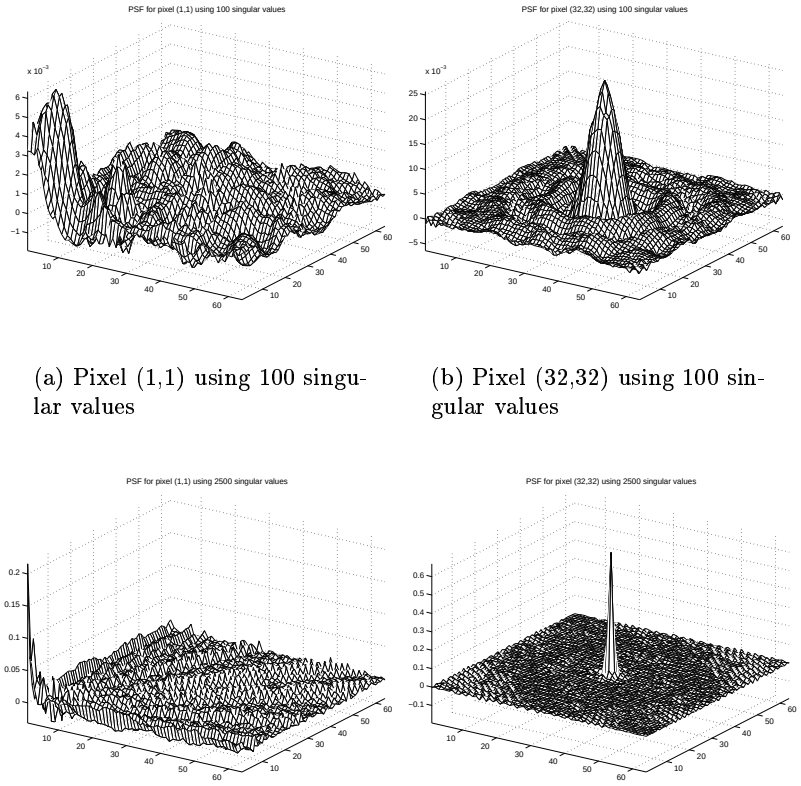
This method gives us a powerful method of examining the resolution in different locations in the image $\mathbf{f}$, for the particular reconstruction method that is used. This technique has been used by Smith et. al. [32] in reconstruction of SPECT images, using truncated generalized inverses. In general if $\mathbf{a}$ is linearly estimated as

$$\begin{aligned} \hat{\mathbf{a}} &= \mathbf{Q}\mathbf{z} && (4.54)\\ &= \mathbf{Q}\mathbf{W}\mathbf{a} && (4.55)\\ &= \mathbf{Q}\Phi^T\mathbf{B}\mathbf{a} && (4.56)\\ &= \mathbf{Q}\mathbf{\Phi}^T\mathbf{f} && (4.57) \end{aligned}$$

then $\hat{\mathbf{f}}$ will be

$$\begin{aligned} \hat{\mathbf{f}} &= \mathbf{B}\hat{\mathbf{a}} && (4.58)\\ &= \mathbf{B}\mathbf{Q}\mathbf{\Phi}^T\mathbf{f} && (4.59)\\ &= \mathbf{H}\mathbf{f} && (4.60) \end{aligned}$$

that is, $\mathbf{H}$ can easily be determined from $\mathbf{Q}$, and we have a method to examine the expected resolution for *any* linear estimation method. This can be very useful since $\mathbf{H}$ can be designed so that good resolution is obtained in important areas in $\mathbf{f}$. The choice of $\mathbf{H}$ then involves choosing suitable basis functions, a suitable estimation method (choice of $\mathbf{Q}$) and, of course, suitable projections $\mathbf{\Phi}$, for the problem at hand. Figure 4.11 shows the PSFs using 100 and 2500 singular values respectively for a corner pixel and for the center pixel using the standard basis. Clearly the resolution is poorer (larger spread) in the corner than in the center. This depends on the way the projections were performed. Here 64 parallel projections at 64 angles were used, giving 4096 projections. The projections were made in such a way that for the projection angle $\theta = 0$, there was precisely one projection passing each pixel. For an angle $\theta = \pi/4$ there will be no projections passing through some of the pixels—the ones in two of the "opposite corners" of the image relative to the projections. This was done deliberately just to show the difference between PSFs at different pixel positions. The resolution using 2500 singular values is much better both in the corners and the center compared to using only 100 singular values.



(a) Pixel (1,1) using 100 singular values

(b) Pixel (32,32) using 100 singular values

(c) Pixel (1,1) using 2500 singular values

(d) Pixel (32,32) using 2500 singular values

Figure 4.11: Point spread functions using 100 and 2500 singular values in the approximate generalized inverse.

Note that if the number of projections used is low some pixels will have PSF:s that are zero everywhere—because no projections is passing these particular pixels. Thus, neighboring pixels must "spread" information to these pixels as well. This implies that if a low number of projections is used, then it is desirable to have some spread in the PSF:s so that information "average out" to neighboring pixels, which otherwise would be zero. Another implication of this is that if high resolution is required one must measure densely at regions of interest.

Bibliography

[1] C.G. Windsor. Can we train a computer to be a skilled inspector? *Insight*, 37(1):36–49, January 1995.

[2] A. MacNab and I. Dunlop. A review of artificial intelligence applied to ultrasonic defect evaluation. *Insight*, 37(1):11–16, January 1995.

[3] Christopher M. Bishop. Theoretical foundations of neural networks. Technical report, Neural Computing Research Group Dept. of Computer Science & Applied Mathematics, 1996.

[4] Keinosuke Fukanaga. *Introduction to Statistical Pattern Recognition.* Academic Press, second edition, 1990.

[5] B. Eriksson and T. Stepinski. Ultrasonic characterization of defects, part 1. literature review. Technical report, Swedish Neuclear Power Inspectorate, 1994. SKI Report 94:11.

[6] B. Eriksson and T. Stepinski. Ultrasonic characterization of defects, part 2. theoretical studies. Technical report, Swedish Neuclear Power Inspectorate, 1995. SKI Report 95:21.

[7] Bo Eriksson Tadeusz Stepinski and Bengt Vagnhammar. Ultrasonic characterization of defects, part 3. experimental verification. Technical report, Swedish Neuclear Power Inspectorate, 1996. SKI Report 96:75.

[8] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing.* Addison-Wesley, 1992.

[9] L. Udpa and S.S. Udpa. Eddy current defect characterization using neural networks. *Materials Evalution*, 48, March 1990.

[10] S.R. Satish and W. Lord. A parametric approach to eddy current ndt signal processing. *Nondestructive Testing Communications*, 1:65–78, 1983.

[11] Gilbert Strang and Truong Nguyen. *Wavelets and Filter Banks.* Wellesly - Cambridge Press, 1996.

[12] Andrew G. Bruce, David L. Donoho, Houng-Ye Gao, and R. Douglas Martin. Denoising and robust non-linear wavelet analysis. In *Proceedings of the SPIE The International Society for Optical Engineering*, volume 2242, pages 325–36, 1994.

[13] Howard Demuth and Mark Beale. *Neural Network Toolbox for use with MATLAB, User's Guide V 3.0.*

[14] Yvette Mallet, Danny Coomans, Jerry Kautsky, and Olivier De Vel. Classification using adaptive wavelets for feature extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1058–1066, October 1997.

[15] Michel Misiti, Yves Misiti, Georges Oppenheim, and Jean-Michel Poggi. *MATLAB Wavelet Toolbox User's Guide.*

[16] J.L. rose Y.H. Jeong E. Allway and C.T. Cooper. A methology for reflector classification in complex geometric weleded structures. *Materials Evaluation*, Januari 1984.

[17] J. Rose J. Nestleroth O. Ganglbauer J. Ausserwoeger and F. Wallner. Flaw classification in welded plates employing a multidimensional feature-based decision process. *Materials Evaluation*, April 1984.

[18] Mauro Bramanti Emanuele A. Salerno Anna Tonazzini Sauro Pasini and Antonio Gray. An acoustic pyrometer system for tomographic thermal imaging in power boilers. *IEEE Transactions on Instrumentation and Measurement*, 1996.

[19] A.P. Ryzhov A.I. Shandro and V.G. Meshcheryakov. Determining the gas temperature in furnace of a type p-67 boiler by an acoustic method. *Thermal Engineering*, 41(11):902–906, 1994.

[20] Rudolph H. Nichols. An acoustic technique for rapid temperature distribution measurement. *Journal Acoustic Society of America*, 77(2):759–764, February 1985.

[21] John A. Kleppe. High-temp gas measurement using acoustic pyrometry. *Sensors*, 13(1):17–22, January 1996.

[22] M. Gustafsson. Solving inverse problems in ultrasonics using principal component analysis. In *Proc. of the Fifth International Symposium on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland, Aug. 25-29 1998*, 1998.

[23] Rymantas Kazys. Ultrasonic amenity sensor—feasibility study. Technical report, IMRA, 1996.

[24] Martin D. Altschuler Yair Censor Gabor T. Herman Arnold Lent Robert M. Lewitt Sargur N. Srihari Heang Tyd and Jararam K. Udupa. Mathematical aspects of image reconstructions from projections. In L.N. Kanal and A. Rosenfeld, editors, *Progress in Patter Recognition*, volume 1, pages 323–375. North-Holland, 1981.

[25] A.C. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging*. IEEE Press, 1988.

[26] Ronald N. Bracewell. *Two-Dimensional Imaging*. Prentice Hall, 1995.

[27] Harish P. Hiriyannaiah. X-ray computed tomography for medical imaging. *IEEE Signal Processing Magazine*, pages 42–59, March 1997.

[28] D. Verhoeven. Multiplicative algebraic computed tomographic algorithms for the reconstruction of multidirectional interferometric data. *Optical Engineering*, 32(2), Feb 1993.

[29] Shih-Chung B. Lo. Strip and line path integrals with square pixel matrix: A unified theory for computational ct projections. *IEEE Transactions on Medical Imaging*, 7(4):355–363, December 1988.

[30] Roger S. Fager Kumar V. Peddanarappagari and Gopal N. Kumar. Pixel-based reconstruction (pbr) promising simultaneous techniques for ct reconstructions. *IEEE Transactions on Medical Imaging*, 12(1):4–9, March 1993.

[31] Guy Demoment. Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37(12):2024–2036, December 1989.

[32] Mark F. Smith Carey E. Floyd Jr. Ronald J. Jaszczak and R. Edvard Coleman. Reconstruction of spect images using generalized matrix inverses. *IEEE Transactions on Medical Imaging*, 11(2):165–175, June 1992.

[33] Gilbert Strang. *Linear Algebra and its Applications*. Harcourt Brace & Company, 1988.

[34] Gerlarld Minerbo. Ment: A maximum entropy algorithm for reconstructing a source from projection data. *Computer Graphics and Image Processing*, 10:48–68, 1979.

[35] Maria Luiza Reis and Nilson Costa Roberty. Maximum entropy algorithms for image reconstruction from projections. *Inverse Problems*, 8(623–644), 1992.

[36] P. Maréchal D. Togane A. Celler and J.M Borwein. Assessment of the performance of reconstruction processes for computed tomography. In *IEEE Nuclear Science Symposium*, volume 2, pages 1353–1357, 1998.

[37] Yair Censor. Finite series-expansion reconstruction methods. *Proceedings of The IEEE*, 71(3):409–419, March 1983.

[38] P.M.V. Subbarao P. Munshi and K. Muralidir. Performance of iterative tomographic algorithms applied to non-destructive evaluation with limited data. *NDT&E International*, 30(6):259–370, 1997.

[39] Gabor T. Herman. Algebraic reconstruction techniques can be made computationally efficient. *IEEE Transactions on Medical Imaging*, 12(3):600–609, September 1993.

[40] Tin-Su Pan and Andrew E. Yagle. Acceleration of landweber-type algorithms by suppression of projection on the maximum singular vector. *IEEE Transactions on Medical Imaging*, 11(4):479–487, December 1992.

[41] Tin-Su Pan Benjamin M. Tsui and Charles L. Byrne. Choice of initial conditions in the ml reconstruction of fan-beam transmission with truncated projection data. *IEEE Transactions on Medical Imaging*, 16(4):426–438, 1997.

[42] Charles L. Byrne. Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Transactions on Image Processing*, 2(1), January 1993.

[43] Charles L. Byrne. Accelerating the emml algorithm and related iterative algorithms by rescaled block-iterative methods. *IEEE Transactions on Image Processing*, 7(1):100–109, January 1998.

[44] Simon Haykin. *Adaptive Filter Theory.* Prentice Hall, 1991.

[45] Olivier Rioul and Martin Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 1991.