NANYANG TECHNOLOGICAL UNIVERSITY

AUTOMATED DETECTION OF WELDING DEFECTS IN RADIOGRAPHIC IMAGES

WANG XIN

SCHOOL OF MECHANICAL AND AEROSPACE ENGINEERING

2007

Automated Detection of Welding Defects in Radiographic Images

Wang Xin

School of Mechanical and Aerospace Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2007

Abstract

Non-destructive testing (NDT) is widely used in many fields. Weld line defect detection is very popular in NDT. Most weld lines should be tested before use, especially in ships, pipes, aircraft etc. In the field of welding defects inspection, the X-ray technique is one of the most popular methods. Currently all the radiographs have to be stored as films and interpreted manually by an experienced interpreter. Human interpretation of weld quality based on film radiography is very subjective, inconsistent, labor intensive and sometimes biased. Also after some time, it may be necessary to check the film again. Hence it becomes another problem to find a film from the large number stored. After a long period, the quality of the films will also drop.

With the developments of image processing algorithms and computer technology, it is possible to let a computer do these jobs. However, the weld line is not flat. Hence, the backgrounds of the images are not uniform. Also the defects are always very small. Due to the degraded quality and the small size of the defects, interpretation of radiographic images is a difficult task.

In this thesis, an automatic detection system to detect welding defects in radiographic images is presented. After obtaining the digital radiographs, image preprocessing is applied to improve the quality of the radiographic image. In this thesis, different methods for improving the quality of radiographic images are investigated. Through comparative analysis, morphological enhancement and adaptive wavelet thresholding can enhance the quality of radiographic images.

A key step in the automated interpretation process is the segmentation of indications from the background. In this study, two new segmentation algorithms are proposed. One method is multiscale edge detection based on wavelet transform (MEWT). According to the wavelet multi-scale character, the coefficients of the wavelet transforms are integrated on a series of scales to look for the best scale where the edges are well discriminated from noise to extract edge features. The other algorithm is multi-level thresholding based on fuzzy entropy and genetic algorithm (MTFEGA). The radiographic image is segmented using multi-level thresholding based on maximum fuzzy entropy. The procedure to find the optimal thresholds is implemented by a genetic algorithm, which can overcome the computational complexity problem. These two algorithms can succeed in segmenting welding defects present in radiography, contrary to conventional methods. The experiment results show that the MTFEGA algorithm can extract smaller defects with lower contrast in comparison with MEWT. So, the MTFEGA algorithm is applied to extract potential defects.

Afterwards two group features: texture features and morphological features are extracted, a feature selection and classification system based on a support vector machine (SVM) is proposed to recognize defects. The top 16 best features are selected based on SVM criteria and receiver operating characteristic curves(ROC) are used as inputs to a designed SVM classifier. With this method, 97.99% of the existing flaws are detected with 14.81% being false alarms. The behavior of the proposed classification method is compared with various other classification techniques: k-means, linear discriminant, k-nearest neighbor classifiers and feed forward neural network, which have been used in the past. The results show the proposed system based on the support vector machine has the best result. These results show the efficiency of the automatic detection system in defect detection.

Acknowledgments

I wish to express my sincere gratitude to all the people who have assisted me during the years of my postgraduate study at Nanyang Technological University. I am most grateful to my supervisor, Associate Professor Brian Stephen Wong, who has given me guidance in my study and research during the past three years. His patient instruction and constructive discussions always help me find out the right way in my research.

I am grateful to Dr Weimin Bai, who have made valuable comments and suggestions that improve my research considerably. I would also like to thank all my friends in Robotics Research Centre, who have offered me a hand in one way or another while I was pursuing my research, and also created a happy and warm environment. In addition, I wish to convey my appreciation to all the staff in the Robotics Research Centre for their support and assistance.

Finally, I must thank my dear parents for their love, infinite patience, continuous encouragement and support throughout these years.

Contents

List of Figures

List of Tables

List of Abbreviation

NDT	non-Destructive test
UT	ultrasonic testing
RT	radiographic testing
PMT	photo multiplier
lp/mm	lines pairs per millimeter
MRF	Markov Random Field
DFT	discrete fourier transform
FFT	fast fourier transform
LoG	Laplacian of Gaussican
SNR	signal-to-noise ratio
DWT	discrete wavelet transform
CWT	continuous wavelet transform
GA	genetic algorithm
NGTDM	neighborhood gray tone difference matrix
GLDM	gray level difference matrix
RF	random field
SVM	support vector machine
SRM	structural risks minimization
ROC	receiver operating characteristic
KNN	k-nearest neighbors
ANN	artificial neural network
FDA	flexible discriminant analysis
PDA	penalized discriminant analysis
MDA	mixture discriminant analysis

Chapter 1

Introduction

1.1 Background

Most structural failures such as building collapses, foundation settlements, aircraft crashes and ships sinking are preceded by some kind of warning – often too subtle to be seen or heard unaided. Non-destructive testing (NDT) gives us the tools to read these warnings. NDT methods are useful at all stages of a structure's life from new construction quality control to verification of as-built conditions through health monitoring to residual lifetime prediction and monitoring of a demolition. NDT procedures are related in all disciplines relying on measuring certain physical properties of materials, or structures and from these, inferring or deducing related properties to arrive at the information that is required. The key to each of the methods is the data reduction and analysis.

NDT has no clearly defined boundaries. A simple technique such as visual inspection is a form of non-destructive testing, as also might be the measurement of an obscure physical property. It used to be considered that there were five major methods – radiographic, ultrasonic, magnetic, electrical and penetrate – but all these can be subdivided [2]. Many NDT methods have reached the stage of development where they can be used by a skilled operator following detailed procedural instructions.

NDT is required both to detect defects with high reliability and to provide accurate defect size information. It can be used on many types of materials and structures e.g. welds. Because defects in weld metal affect the strength of the welded joint, welded structures often have to be tested non-destructively, particularly for critical applications where weld failure can be catastrophic, such as in pressure vessels, load-bearing structures, power plants

and pipelines. Inspection of welded structures is essential to ensure that the quality of welds meets the requirements of the design and operation, thus assuring safety and reliability. A variety of NDT methods are available for the inspection of welding defects [3]. X-ray or γ-ray radiography, together with ultrasonic and magnetic particle inspection, are the mainstays of weld inspection. Visual inspection is the primary evaluation method of many quality control programmes. It can be easily carried out, is inexpensive, and usually does not require special equipment other than magnifying glasses, boroscopes, or television camera systems. It is used most effectively for the inspection of welds where quick detection and correction of flaws or process related problems can result in significant cost savings. For more critical welded structures such as high-pressure vessels, the nature, location, and magnitude of the flaws must be mapped in order to determine their acceptability by further fracture mechanics analysis. To this end, more sophisticated NDT methods such as ultrasonic testing (UT) and radiographic testing (RT) are needed. Ultrasonic inspection uses sound waves of short wavelength and high frequency to detect flaws. Usually pulsed beams of high frequency ultrasound are used via a hand-held transducer which is placed on the specimen. Any sound from that pulse that returns to the transducer like an echo is shown on a screen which gives the amplitude of the pulse and the time taken to return to the transducer. Defects anywhere through the specimen thickness reflect the sound, back to the transducer. Flaw size, distance and reflectivity can be interpreted. RT is the other commonly used NDT method for detecting internal welding flaws. Until the advent of ultrasonic inspection, radiography was the only available method for finding buried defects in welds. It is based on the ability of X-rays or γ-rays to pass through metal and other materials opaque to ordinary light, and produce photographic records by the transmitted radiant energy [4]. Because different materials absorb either X-ray or γ-rays to different extent, penetrated rays show variations in intensity on the receiving films. That provides a means to examine the internal structure of a weld. Current acceptance codes for welds have evolved principally from knowledge of the inherent advantages and limitations of radiographic testing. The underlying physical principles of radiographic inspection have been known for nearly a century and, not surprisingly, radiography has evolved during this time into a mature technology [5]. The pipeline images are usually ordinary laser camera images and not X-rays, but the problem involves many similar considerations. And natural gas pipelines in the North Sea and on land in continental Europe are aging and identification of signs of corrosion and decay is important.

The interpretation of a radiograph is currently being done manually by experienced interpreters whose task consist of detecting, recognizing and quantifying the defects images.

However, the radiograph quality, the welding over-thickness, the bad contrast, the noise and the weak sizes of defects make their job difficult. The interpretation of a radiograph is much more than just looking at the film. Although radiographs are interpreted by comparing with standards, to interpret correctly and analyze the results of any radiographic examination, the interpreter must first judge the quality of the radiograph. The quality of the radiograph is determined by the following techniques: density, penetrameter selection and sensitivity, identification of the film, coverage of part, and artifacts. Secondly, the interpreter must also be able to identify rejectable discontinuities and judge them as true flaws or artifacts attributed to the radiographic process. An interpreter will be able to make sound judgments if he has the knowledge of the component or part configuration and manufacturing process [6] [7]. Defect quantification is subject to human judgement and subjective considerations, such as capabilities and experiences of the interpreter because it takes time to train a film interpreter. Also, recently it has been difficult to employ skilled interpreters. Moreover, for the identification process, not only the interpreters' skill influences the testing result, but also it is difficult for skilled interpreters to assess small flaws within a short time. The whole area of NDT of welded structures is currently undergoing a period of rapid change brought about by a combination of technological revolutions, especially after the computer was invented. Computer visual image processing systems have some good characteristics, allowing objective assessment, high-speed judgment, non-human error etc. Therefore, an image processing system would allow weld defects to be detected using X-ray radiography in the presence of background noise.

The computer has developed very quickly in terms of both hardware and software. It is possible to use computer vision to detect the defects instead of using a human being. Different persons will give different results to a single film. But with the help of a computer we can set up a unique criterion. Also, when the image is saved in digital files, it is not necessary to keep many films and use lots of time and money to maintain them. When it is necessary to find some image in the future, it becomes an easy procedure.

Because X-rays are harmful to human beings, especially when the specimen is thick and there is a need to use high energy X-rays to penetrate the metal, they must be far away from the test place and develop the films later. So the result will be obtained later. To obtain a real time, online system for radiographic evaluation, some techniques have been developed to display the image on a screen. If a fluorescent screen, which converts X-rays to light, is put behind a specimen, an image of the specimen can be seen on this screen. The image is usually so faint that fine detail cannot be discerned and flaw sensitivities are

usually poor. A suitably sensitive closed-circuit television (CCTV) camera can be focused on the fluorescent screen and the amplification circuits associated with the camera are used to produce a bright image on a television monitor screen. Since the image is presented on a television monitor, it can be remote from the X-ray equipment and all radiation hazards are eliminated [2].

However, the interpretation is difficult due to the complex situation. The most important part of the radiographic inspection is the "reading" of the radiograph. The importance of appropriate film viewing conditions has already been emphasized along with desirable screen luminance, masking conditions etc. Ideal automated image recognition systems, using computer programs for pattern recognition, are still quite difficult, so that film interpretation still depends on the skill and experience of the inspector or radiologist [2].

Since to detect very small defects, films are commonly considered as a reference for all image systems and radiographic testing with film is an expensive and time-consuming technique (exposure time and development of the film), some attempts have been made to fully automate the radiographic inspection cycle. However, to date there are no satisfactory results that allow the detection of the weld discontinuities with optimal trade off the error types, false alarms and missed defects and keeping both small.

The specifications concerning defect type, minimum size of defect and acceptability criteria are supplied more or less by the industry, but since no control is actually performed on the weld, the knowledge about acceptability is vague. The only piece of information that can be regarded as absolute concerns cracks, which are unacceptable regardless of their size.

There are many kinds of defects in a weld-line. Table 1.1 lists the most common types of the defects in a weld-line as well as a brief description of them.

1.2 Objective and Scope

Figure 1.1 is an image scanned from a radiograph of a weld line with defects in specimen. Usually the defect characteristics in the image are:

1. The defects are all quite small.
2. The defects' positions are random. It is impossible to predict where the defects can

Table 1.1: Description of welding defects

Types of Defects	**Description**
Porosity	It occurs as voids cased by gas trapped in the weld metal. The voids may occur as spherical, elongated, or ”worm hole” shapes and in patterns that are random, clustered, or linear. On the radiograph the spherical voids have the appearance of rounded dark area while the nonspherical voids have an elongated dark area with a smooth outline.
Slag Inclusions	They are particles of slag entrapped in the weld metal or along the fusion planes. The particles appear darker than the surrounding area and may be irregular in shape or elongated in the direction of the deposited weld bead
Lack of Fusion	It is a discontinuity cased by molten weld metal which has failed to bond the base metal or to a previously deposited weld bead. On the radiograph it appears as a dark indication usually elongated and varying in width.
Cracks	It is a rupture of solidifield metal. Cracks associated with welding may be longitudinal, transverse, or radially oriented and may occur in the weld metal, base metal or through both. On the radiograph, it appears irregular, intermittent or continuous lines.
Incomplete Penetration	It is a discontinuity that occurs at the root of welds designed for through penetration where full penetration has not been achieved. The discontinuity appears on a radiograph as a straight dark line that may be either continuous or intermittent.
Burn Through	It is a melting of the metal from the root of the weld or through the backing strip. It appears on the radiograph as an individual darkened area of elongated or rounded contour which may be surrounded by a lighter ring.

appear.

3. The contrast is very low. The gray level between defects and the background is similar.

4. The background is not uniform. The defects are not the only dark regions in the image. Some parts of the image may have a darker gray level.

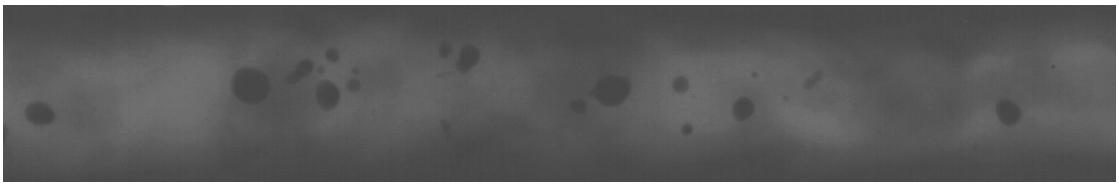


Figure 1.1: Example of a radiographic image

The defects cannot be detected only by the gray level difference. A single method cannot result in an ideal consequence. In order to find the defects in this kind of image, we should imitate the decision of human vision, observing the whole image and comparing the defects not only with the background in its adjacent neighborhood, but also with the information from a larger range.

The objective of this project is to develop a system to automatically detect welding defects using the images produced by analog radiographs.

Figure 1.2 is the flow chart of the system. The main idea is: after obtaining digitized radiographs, firstly, image preprocessing methods are applied to improve the quality of radiographic images. Then the potential defects are extracted using a segmentation method. After extracting features of potential defects, a feature selection and classification system is used to classify the potential defects as the defect or non-defect. The whole system is not only automatic, but also an "intelligent" procedure that will assist greatly in examining the weld. In order to realize this high automation, only a few parameters are needed to be set, so that the human factor can be reduced.

In this study, we do research as follows:

- To analyze and compare the preprocessing methods of digital radiographic images.
- To develop the segmentation method to extract potential defects from the background.
- To investigate the features of defects.
- To develop a defect classification method.

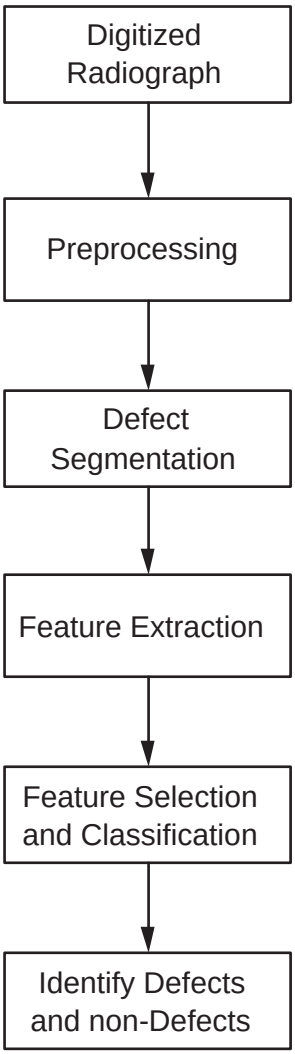


Figure 1.2: Flow diagram of the system

Chapter 2

Literature Review

2.1 Characteristics of X-rays

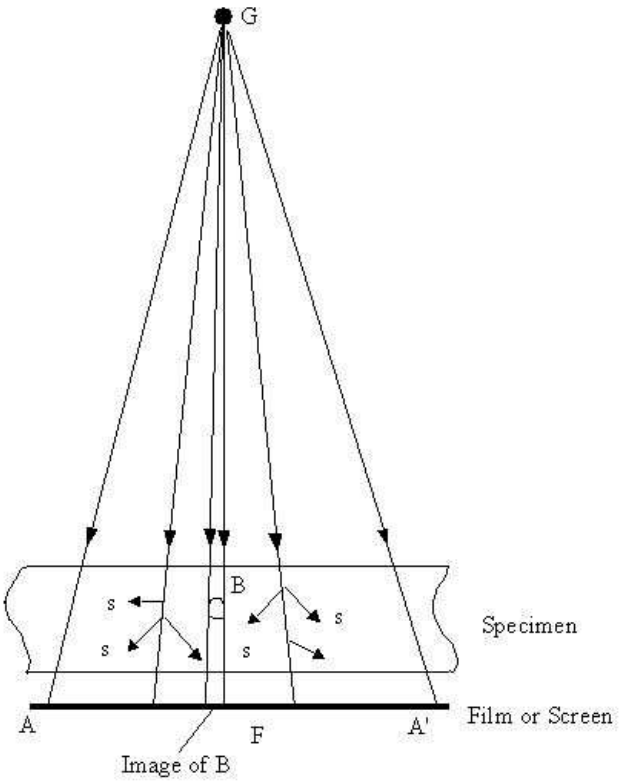


Figure 2.1: Principles of film radiography, where AA' is the film plane, B is a cavity in the specimen, imaged at F, G is the radiation source, and S is the scattered radiation generated inside the specimen

Since the discovery of X-rays in 1895, it has been realized that they can be used for non-

destructive testing of materials [2] [8]. X-rays are a form of electromagnetic radiation, of the same physical nature as visible light, radio-waves, etc., but which have a wavelength which allows them to penetrate all materials with partial absorption during transmission. They include a fairly wide waveband of radiation from about 10nm, which will usefully penetrate only very small thicknesses of solid material, to about 10^{-4}nm, which will penetrate up to about 500mm steel.

X-rays travel in straight lines outwards from a source: for all practical purposes they cannot be focused, so the usual set-up for producing a radiograph is as shown in Figure 2.1, using a small diameter source, G, and a sheet of photographic film as a detector. A beam of penetrating ionising radiation passes through a specimen to expose the films.

2.1.1 General Principles

In Figure 2.1, the X-rays travel in straight lines from the source to the film, so that if there is a cavity in the specimen, as shown at B, which causes a lower absorption along the path GBF, more radiation reaches the film at point F, compared with other points. Consequently an X-ray 'image' of the cavity is produced which will be the projection of the cavity of very nearly natural size. (Because G is quite far from the specimen compared with the film. The magnification is small.) Thus, a two-dimensional image of a three-dimensional cavity is formed.

To produce a radiograph, the X-rays are allowed to reach the film for an appropriate exposure-time, which depends on the intensity of the X-rays, the thickness of the specimen, and the characteristics of the film. The film is then processed (developed, fixed, washed and dried), so that the X-ray image can be seen as different levels of gray (film density). The film is then placed on an illuminated screen so that the image can be examined and interpreted.

When X-rays are absorbed in a material, some X-ray energy is re-emitted as scattered radiation, which under some conditions can travel in a different direction to the primary beam. Thus at point F on the film, some radiation will travel directly along the line GBF, and this forms the image of the cavity B, however scattered radiation, S, can also reach F,

and this is non-image forming. The ratio:

$$\frac{Image \quad forming \quad radiation}{Non-image \quad forming \quad radiation} = \frac{Direct \quad radiation}{Scattered \quad radiation} = \frac{I_D}{I_s}$$

where I_D and I_S are the intensities of the radiation, is an important parameter in industrial radiography. The total radiation reaching point F in unit time is $(I_D + I_S)$ and the ratio

$$\frac{I_D + I_S}{I_D} = 1 + \frac{I_S}{I_D} \tag{2.1}$$

is known as the 'build-up factor'.

The basic law of X-ray absorption is given by

$$I_X = I_0 \exp(-\mu x) \tag{2.2}$$

where x is the thickness of the material, I_0 is the incident intensity of radiation, I_X is the transmitted intensity, and μ is a constant, known as the 'linear absorption coefficient' with dimension cm^{-1}, and its value depends on the material and the X-ray wavelength.

2.1.2 Unsharpness of the Image

Geometric Unsharpness

One important physical factor should be mentioned at this stage. The source of X-rays or gamma-rays is usually a small area, a few millimeters in diameter: it is never a true point source. Consequently, for any defect in the specimen which is not close to the film, there is a blurring of the image, known as the penumbra, or geometric unsharpness, U_g (Figure2.2). In this figure, the size of the source, s, has been exaggerated, for clarity, and the 'defect' has been taken to be a small, physically sharp step on the surface, at O. Then an element of the source at M will image the step on the film at M' and an element of the source at P will image at P'. Thus the image of a sharp step is blurred on the film, over a distance $d_{P'N'M'}$. By simple geometry

$$\frac{d_{P'M'}}{d_{MP}} = \frac{d_{N'O}}{d_{ON}}$$

and if for simplicity $d_{NO} = a \; and \; d_{N'O} = b$,

$$d_{P'M'} = \frac{sb}{a} = \; geometric \quad unsharpness \; , U_g \tag{2.3}$$

and this relationship holds for the image of the edge of any specimen detail.

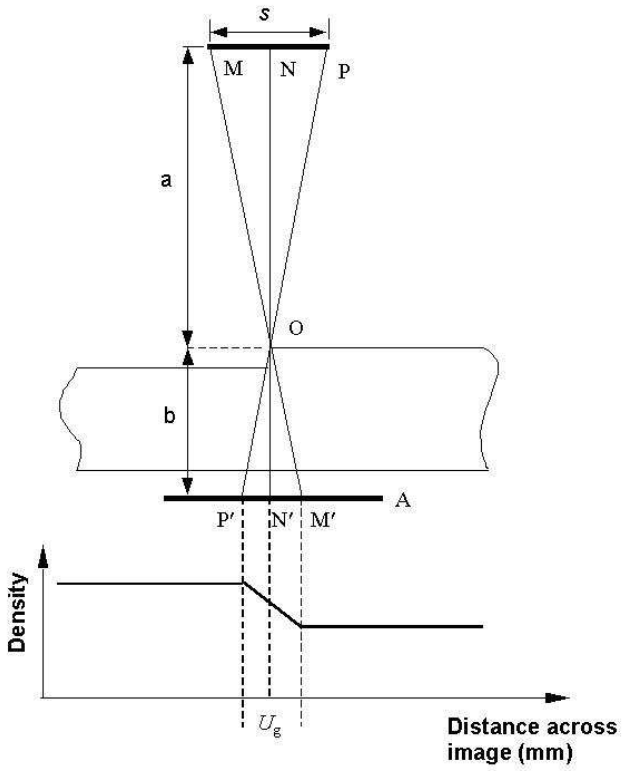


Figure 2.2: Geometric unsharpness on specimen detail at O, with a curve of the density distribution across the image (source diameter exaggerated in size)

Film Unsharpness

When an X-ray quantum is absorbed in the film emulsion, it sensitizes a silver halide crystal; furthermore, it may have sufficient energy to release electrons, which can reach and sensitize adjacent silver halide crystals. One X-ray quantum can therefore sensitize a small volume of silver halide crystals, thus producing a small disc instead of a point image and this is equivalent to producing an inherent or film unsharpness, U_f. If, in Figure 2.3 a sharp metal edge is laid on the film, and the image of this edge on the processed film is examined, it will be found that an X-ray quantum absorbed in a silver halide crystal just to the ledge of O has sensitized silver halide grains to the right of O, and the density distribution across the image follows curve in practice. Also high-energy quanta will produce a greater effect.

The geometric unsharpness is the main disadvantage of radiography. It is present because radiation from an X-ray tube is not produced from a point source.

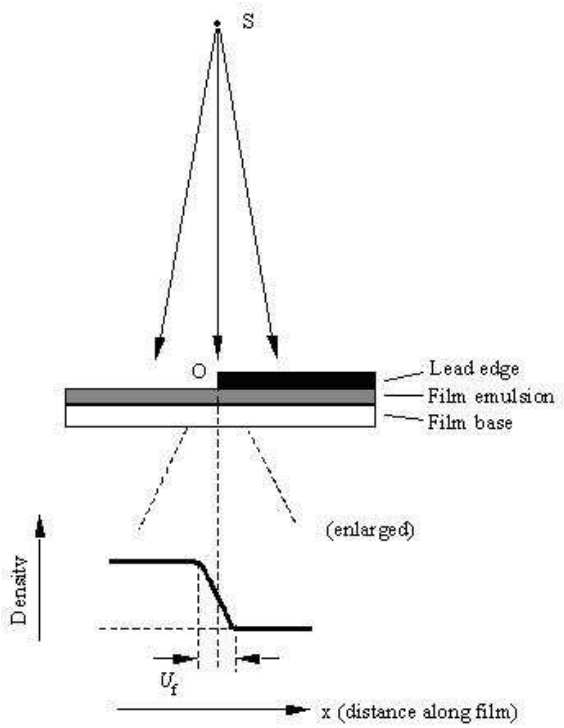


Figure 2.3: Film unsharpness, U_f: the curve shows the density distribution on the film

2.2 Digital Radiographs

Radiography using photographic film is a robust technique for welding defects inspection. It has been used for over a century. In fact, one of the first discoveries about X-ray energy was its ability to affect photographic film.

Radiography with film is a time consuming technique; the technician must perform many operations to produce and interpret an image. Among these steps is handling of the film in a darkroom both prior to and after the X-ray exposure (which itself may require up to several minutes). The end result is a "snapshot" of the object of interest that must typically be viewed on a high intensity light box under subdued general lighting conditions. Post processing of the image is generally limited to adjusting the intensity of the viewing light. Long term storage, duplication or distribution of a film based image are also problematic, time consuming and expensive. Finally, a film-based system must include a budget for consumable items, film and chemistry, and provide for the disposal of hazardous waste.

Today, newer technologies have enabled to produce digital radiographs. Film digitization systems can digitize the film to radiographic image. Electronic sensors are available to

produce radiographic images.

Digital radiography allows automated weld inspection using automated interpretation systems. We can use image processing methods to adjust both the brightness and contrast of the image, apply filters to reduce noise or "snow", enhance edge detail, and quickly and precisely measure any details of interest in the image. Storage, duplication, and distribution of these images are simple, quick, and inexpensive. Mass storage of images for archival purposes is also simpler and cheaper.

2.2.1 Film Digitization Systems

The existing film digitization systems to non-destructive testing (NDT) covered are laser scanners, CCD line scanners and CCD array cameras. Based on the physical properties of the X-ray film the parameters are defined, which have to be fulfilled by the digitization system [9].

Laser Scanner

The principle of laser scanner is point by point digitization as shown in Figure 2.4. The film is moved in front of a collection tube. A laser beam with a fixed diameter passes the film. The diffuse transmitted light through the film is integrated by the collection tube an registered by a photo multiplier (PMT) on top of the collection tube. During the scan the folding mirror moves the laser beam along a horizontal line on the film. The film is moved with a speed of 75 lines/sec. The resulting voltage at the photo multiplier is proportional to the light intensity behind the film. After logarithmic amplification a digitization with 12 bit yields grey values that are proportional to the optical density of the film.

The laser scanner has two main advantages: 1. The quantity of collected light is considerably higher. The laser focuses the whole light intensity on the point to measure. The radiation passing the film is collected over a spatial angle as great as possible in the detector. 2. Scattered light from regions with low optical density passing to regions of high optical density, thereby distorting the measurement, is nearly avoided by pointwise illumination.

The essential difference to other scanners (CCD line scanner, CCD camera) is the reversed optical way. The laser scanner illuminates with focused light and measures the diffuse

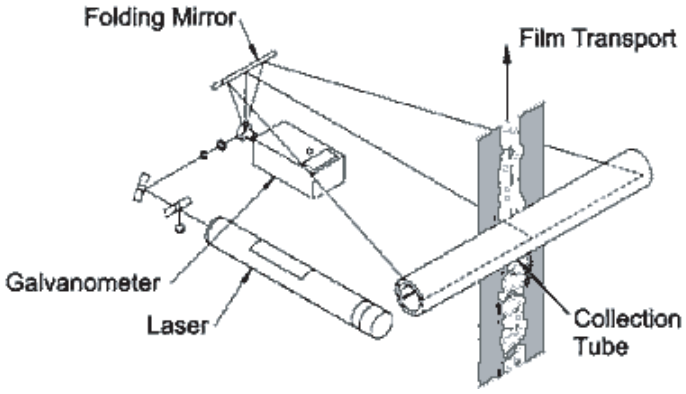


Figure 2.4: Principle of the laser scanner

light intensity behind the film. Other methods illuminate with diffuse light (the film is illuminated with a diffuser) and measure the light intensity that passes the film in one direction (camera objective or human eye in classical film inspection).

CCD Line Scanner

The principle of a CCD line scanner is line by line digitization as shown in Figure 2.5. The film is illuminated by a light bar to which the light of a projector lamp is passed by a light guide. A Teflon plate ensures a diffuse film illumination. The illuminated line is projected by an objective on a CCD line detector. The film is moved under the light bar to scan the whole film area. The combination of light bar, light guide, and projector lamp is the illumination of highest power.

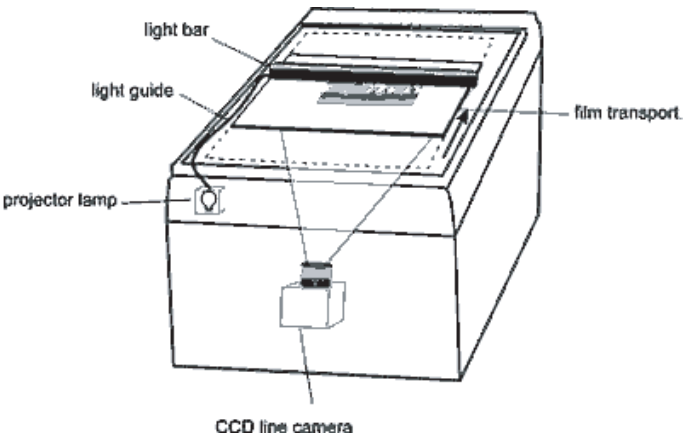


Figure 2.5: Principle of the CCD line scanner

Line by line digitization is a good compromise between speed and influence of scattered radiation at high dynamical ranges (great differences between optical densities) that are typical for NDT films. Furthermore, modern CCD line arrays have operation modes for

anti-blooming (avoid to transfer charges from saturated pixels to unsaturated neighbors) and integration time control. Thus, the scanner can be adjusted optimally to the optical density range of the film to digitize.

CCD Camera

The principle of CCD camera is array digitization as shown in Figure 2.6. This procedure resembles classical film inspection, only the human eye has been replaced by a camera.

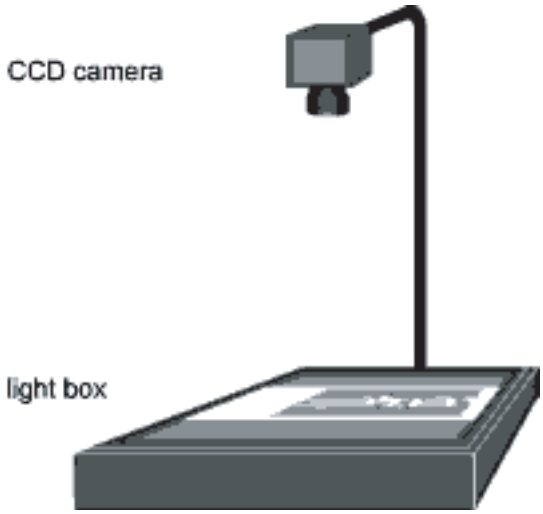


Figure 2.6: Principle of the CCD array camera

The advantages of CCD camera are: simple structure, cheapest realization, no moving parts during digitization. The two disadvantages of CCD camera are: 1. The whole information of the image is projected onto the camera array. Due to scattered light only a small dynamic range of the film can be digitized. Depending on the quality of objective and camera array an optical density range between 2 and 3 (at best) can be achieved. 2. Since the number of pixel for each CCD array is limited, only a small region of the film can be digitized with sufficient spatial resolution. These regions can measure from 5×5 cm^2 to 10×10 cm^2 in size. For this reason CCD cameras are not suited for archiving of complete films which are typically of a size of 8×24 cm^2 and larger.

2.2.2 Digital Detectors

Digital radiography uses electronic sensors to convert the X-ray energy into a video (television) or directly to a digital (computer) format. The image is presented on a TV or

computer monitor almost instantaneously. Figure 2.7 shows an imaging sensor system. The imaging sensor system produces digital images, with film life detail, and allows for easy viewing on a computer screen and electronic transmittal to other location. The amorphous silicon flat panel detectors as the most flexible digital imaging system currently are available [10].

Figure 2.7: Amorphous Silicon Imaging Sensor Plate

The Amorphous silicon (a-Si) imaging technology, which was developed by medical equipment manufacturers for digital radiography, has been boosted by the recent breakthroughs in thin film transistor arrays similar to those found in notebook computer screens. As a result, the a-Si detectors, the latest generation of which can generate images in a digital 16-bit format yielding over 65,000 gray scales for analysis, have become efficient enough in achieving the resolution needed for a wide range of industrial applications. The detectors are manufactured in various sizes with resolutions up to 5 lp/mm. Industrial applications for a-Si X-ray detectors, ranging from in-line inspection of printed circuit board (PCB) assemblies to searching for cracks in aircraft fuselage, also include non-destructive testing of pipelines, welds and nuclear waste, neutron radiography and X-ray tomography. The technology is based on a two-dimensional, solid-state, amorphous (non-crystalline) silicon 'imaging array' that contains hydrogen. The arrays, which can be fabricated up to an area of $12^{'}16$ square inches, contain about one million sensors. Combined with a cesium iodine (CsI), the sensor presents an ideal solution for high-resolution X-ray imaging applications. A scintillator is deposited directly onto the surface of the arrays. X-ray photons striking the phosphor are converted to visible light, which is absorbed and converted to an electric charge by the photodiodes. The charge is integrated on each photodiode so that each pixel collects a signal proportional to the local flux of the X-ray beam. When the array

circuitry scans the diodes, the charge is converted into a video signal, which reproduces the X-ray image. The signal is read out in real time as a digital electronic image using thin film transistors made of the same amorphous silicon material. The image is then manipulated. It can either be read out and displayed continuously at 5 to 30 frames per second, or integrated over many frames to be displayed at a frame every few seconds, to improve sensitivity. In both cases the feedback to the operator is immediate.

In this study, films are digitized using a laser digitizer with photomultipliers, manufactured by Computerised Information Technology Ltd, UK. The scanner provides an optical density range from 0 to 4.1. The films are digitized with a spatial resolution of $100\mu m$ and a gray level resolution of 8 bit per pixel.

2.3 Review of Automated Welding Defects Inspection

Radiographic testing is one of the most important nondestructive testing techniques for weld inspection. Traditionally, experienced interpreters evaluate the weld quality based on radiography. It is time and manpower consuming work. In addition, human interpretation of weld quality based on film radiography is very subjective, inconsistent and sometimes biased. Therefore, it is desirable to develop a automated welding defects detection system to increase the objectivity, accuracy and efficiency of radiographic inspection. Computer vision is a key factor in the implementation of an automated radiographic inspection system. Because of the problems associated with visual detection, currently there is a great deal of work and research on the development of automated systems for inspection and analysis of radiographs. Some of the important achievements in this area are presented below. To date, there are no satisfactory results that allow the detection of the weld discontinuities without false alarms.

2.3.1 Method of Gayer et al.

Gayer et al. [11] proposed a two-step process for the automatic recognition of welding defects through radiography. This method tried to imitate the way a human inspector inspects radiographs: first, a general glance with coarse resolution, followed by fine focusing on defective regions.

The first step was a fast search for potential defects in the X-ray image. Based on the

principle that a defect was characterized by relatively irregular behavior of the gray levels in the X-ray image, two different algorithms were developed for the fast search procedure: calculation of the relative contribution of high frequencies and calculation of a derivative function. The spectrum of the X-ray image was determined with the help of a fast Fourier transformation, which was calculated either row by row or column by column in little 32×32 windows. When the sum of the higher frequencies of a window was greater than a given threshold value, the entire window was marked as potentially defective. Another possibility was suggested by the authors as part of this task: a window was selected as potentially defective when the sum of the first derivative of the rows and columns of a window was large enough. The second step was identification and location of true defects. Only those regions which were previously classified as being potentially defective were studied here. It could be accomplished either by comparing defects with known defects' templates, or by thresholding the image. The first lead to a matching between the potential defect and typical defects which were stored in a library as templates. Whenever a large resemblance between the potential defect and a template was found, the potential defect was classified as a true defect. The second algorithm estimated a defect-free X-ray image of the test piece by modeling every line of an interpolated spline function without special consideration for the potentially defective region. Following this, the original and the defect-free images were compared. True defects were identified when a large difference occurs compared to the original input image.

Gayer's method only works well for the defects types which are included in the template set but would fail when presented with defects type or shape not represented.

2.3.2 Methods Based on Neural Networks

Lawson and Parker [12] proposed to apply neural networks to automate detection of defects in radiographic images. The method generated a binary image from the test image. The authors used a multi-layer perceptron network trained to segment the image. The backpropagation technique was used to train the weights within the network. The methodology adopted comprised two stages. The first stage was the segmentation of the weld from the background content of the radiographic image. The training of the network was achieved with a weld template. A three layer perceptron was used for the network architecture. The input to the network comprised 6 elements: horizontal and vertical image position, local mean, median, maximum pixel value and actual pixel value. The network

was subjected to a maximum of 40,000 training samples from the three images. The second stage was the segmentation of defects from the weld region. A network with two hidden layers was trained on a test set of image data previously segmented by a conventional adaptive threshold method. The optimum network performance was achieved using 50,000 random sets of inputs generated from the training image. The experiment results on five X-ray images showed that detection using neural networks was superior to the segmentation method using adapted thresholds.

Nafaâ [13] proposed the application of neural network in the edge detection of X-ray images containing defects of welding. They designed a network window classifier which classified the central pixel of a relatively small area in the image. To extract the edges, the window must slide, pixel by pixel, over the entire image. Then, the input layer of the network corresponded to the portion of image covered by the window (3x3). Therefore, the neural network had 9 neurons receiving the gray-scale values of pixels composing the square window. The output layer contained one neuron whom the state identified the edges from the background and created the segmented image.

Zscherpel [14] [15] [16] applied a neural network to detect the cracks in welds. They extracted features using morphological, derivative-of-Gaussian, Gaussian weighted image moment vector, FFT and wavelet based filters, and subsequent classified using backpropagation neural networks to detect the crack.

The obtained results showed the effectiveness of using neural networks to detect defects in X-ray images. However, the inconvenience of the method based-on neural networks is that there are no accurate rules for the option of the hidden layers number and the neurons number in each layer. Also it needs big training sets. Also, they do not provide much insight in comparision with other nonlinear repression technique.

2.3.3 Method of Zhao

Zhao [17] presented a method for defects detection in a weld line. The entire algorithm consisted of three steps: image enhancement, defects location and defects detection. The proposed low-level image processing method included image enhancement, edge detection and region growing. The defects' were located by detecting the change of the gray level. But due to the disadvantage of edge detection used in the noise image, the edge cannot present the exact feature of the image. Hence, a region growing method was followed

by using gradient map to ensure the growing direction. After the process, the result was not ideal enough, so a fine detection was required. The proposed fine detection was a combination of k-means and MRF. Since the background of the image changed greatly from one place to the others, the means of the defects and the background had to be decided in each area. The k-means provided the mean values of the defect and the background in the small area. Furthermore, the MRF could be used.

Zhao's method is only based on gray level. When the size and contrast of a defect are both large enough, the computers can find it quite correctly. But when the contrast and the size are both small, the defect will be missed by the computer system.

2.3.4 Background Subtraction Methods

Hyatt et al. [18] presented a multiscale method for segmenting flaw indications from the background radiographic images. The method was designed to remove the overall background structure while reserving the defect details.

Liao and Li [19] proposed a welding flaw detection approach based on curve fitting. The key idea of this work was to simulate a 2D background of a normal welding bead characterized by low spatial frequencies in comparison with the high spatial frequencies of the image of the defects. Thus, a 2D background was estimated by fitting each vertical line of the weld to a polynomial function. Then, the obtained image was subtracted from the original image. The defects were detected where the difference was considerable. The whole process consisted of four parts: preprocessing, curve fitting, profile-anomaly detection and postprocessing. The preprocessing was used to remove the background by thresholding and normalizing the line images. The threshold value was chosen by observing the histograms of the scaled images. The curve fitting was used to smooth the profile by filtering out the local variations. The profile-anomaly detection detected and identified the anomalies in the tested profiles. The profile anomalies caused by welding defects was classified into peak-anomaly, trough-anomaly and slant-concave-anomaly. Then the results obtained from the processed line profiles were put together to generate 2-D flaw-maps. The postprocessing removed isolated anomalies from the previous step and updated the flaw-map.

Wang and Liao [20] proposed a fuzzy k-nearest neighbor and multi-layer perceptron neural network to classify different types welding defects. The whole system consisted five

parts: noise reduction, contrast enhancement, defect segmentation, feature extraction and pattern classification. The median filter was used to reduce the radiographic image noise. The histogram equalization algorithm was applied to enhance the contrast. Defect segmentation was applied to extract defects from the background using the background subtraction method and the histogram thresholding method. Then features used for classification, distance from center, circularizes, compactness, major axis, width and length, elongation, Heywood diameter and average intensity and standard deviation of intensity, were extracted. Finally the k-nearest neighbor and the three layer perceptron neural network classifiers were used for classification. Two approaches were compared using the statistical bootstrap method. The results indicate that the multiplayer perceptron neural network is superior to the k-nearest neighbor methods.

Subsequently, Liao [7] developed a fuzzy expert system for the classification of welding discontinuities. First, three features, width, the mean square error between the object and its Gaussian and the peak intensity, were extracted for each object in the line image. Then the fuzzy rules were extracted from feature data based on a modified fuzzy c-means algorithm. Finally, the performance of the fuzzy expert system was compared with the multiplayer perceptron neural network. If appropriately designed, the fuzzy c-means algorithm yields better performance than a multiplayer perceptron neural network.

Background subtraction method only works well when contrast between defects and background is large enough. But when the contrast is low, the defect can not be extracted using this method.

2.3.5 Defect Recognition using Statistical Classifiers

A method for automated recognition of welding defects was proposed by Silva et al. [21]. In the first step, the median filter and contrast enhancement technique were used to improve the quality of X-ray image. Then the potential defects were segmented in the X-ray image. Geometric and gray value features (contrast, position, aspect ratio, width area ratio, length area ratio and roundness) were extracted. The most relevant features were used as input data on a linear discriminant classifier.

Sofia and Redouane [22] presented a method for automated recognition of welding defects using a k-nearest neighbor classifier. The detection followed a pattern recognition methodology: i) Segmentation: regions of pixels were found, and isolated from the rest

of the X-ray image using a watershed algorithm and morphological operations (erosion and dilation). ii) Feature extraction: the regions were measured and shape characteristics (diameter variation and main direction of inertia based on invariant moments) were quantified. iii) Classification: the extracted features of each region were analyzed and classified using a k-nearest neighbor classifier. According to the authors, the method achieved a good detection rate.

Mery [23] proposed automatic detection of defects using texture features and statistical classifiers. The proposed method followed three steps: segmentation of potential flaws, feature extraction and classification. The first step was segmentation. The laplacian of gaussian (LoG) edge detector was used. This LoG filter searched for changes in the gray values of the image, thus identifying zones delimited by edges that indicate flaws. However, this filter was noise sensitive. After the edges were detected the closed regions were considered as potential flaws. The second step was feature extraction. Two group texture features, features based on the cooccurrence matrix and features based on Gabor functions were analyzed. Features based on the cooccurrence matrix gave a measurement of how often one gray value would appear in a specified spatial relationship to another gray value on the image. Features based on 2D Gabor functions presented an appropriate choice for tasks requiring simultaneous measurement in both space and frequency domains. The last step was classification. The most relevant features were analyzed and classified using statistical classifiers. The polynomial, Mahalanobis and nearest neighbor classifier were analyzed.

Statistical classifiers can recognize defects well after defects are segmented from background successfully. However, most defects in the specimen used by these researchers are a bit larger than that could be described as fine defects which is of most concern.

In order to compare these methods better, a quantitative comparison quoting reported sensitivity and specificity values for each method as well as database sizes are summarized in Table 2.1. However, the direct comparison of performance levels is not possible because of the different databases used.

2.4 Texture

Texture is one of the most commonly used features used to analyze and interpret images. It is a measure of the variation of the intensity of a surface, quantifying properties such

Table 2.1: Comparison of different classifier

Method	Detection Rate	Minimum Width of defect	Database Size
Method of Gayer et al.	100%	5	4 images with 4 defects
Method based-on neural networks	94%	/	2314 regions of interest cut out of several images
Method of Zhao	83%	10	8 images with 52 defects
Background subtraction method	93%	/	24 images with175 defects
Defect recognition using statistical classifiers	91%	15	1 image with 198 defects

as smoothness, coarseness, and regularity. The texture feature is a value, computed from the image of an object, which quantifies some characteristics of the gray level variation within the object. It gives information about the distribution of the gray values in the image. Normally, the texture feature is independent of the object's position, orientation, size, shape and average gray level. It is difficult to use texture information for any practical application in image processing unless we can somehow quantify it. There are four general approaches to texture quantification, the use of each of which may have advantages and disadvantages for particular applications. Statistical, structural, spectral and model-based methods are more or less suitable for individual problems and their suitability originates from the nature of the problem. In this section, a brief description of these major texture evaluation methods is given.

2.4.1 Statistical Methods

A very popular class of methods for texture description and analysis uses measures that arise from the statistical processing of the pixels' gray-scale values. We can distinguish two different categories of statistical methods, which differ in the preprocessing applied to

the pixel values prior to the extraction of features.

Direct Measurement

This group of methods directly uses the statistics of the gray level values of the image pixels in order to derive textural features. Probably the most popular tool for the computation of statistical textural features is the co-occurrence matrix (or spatial gray level dependence matrix SGLDM), introduced by Haralick in 1973 [24], which will be presented in detail in section 5.2.1.

The neighborhood gray tone difference matrix (NGTDM) was introduced in [25]. The matrix is in fact a vector S with size equal to the number of gray level values in the image. If $\bar{A}$ is the average gray level intensity in the neighborhood of a pixel with gray level value of i, the this pixel contributes the amount $|i - \bar{A}|$ to the element $S(i)$. The values in the NGTDM are a measure of the deviation of the gray level of image pixels from the average gray level around them. In fine textures, the matrix will have high values, while in coarse textures will have low values. From the NGTDM, it was suggested that five features may be calculated: coarseness, contrast, "business", complexity and strength [25].

A similar concept is the basis for the gray level difference matrix(GLDM), which is a vector R defined for various pixel distances d and angles θ. If a pixel with gray level value i is at distance d and θ from a pixel with gray level value j, its contribution to GLDM matrix $R_{d,\theta}$ is to increment by the value of $R_{d,\theta}(|i - j|)$ [26].

Other statistical features derived directly from the image under examination include those related to image edges and the length of primitives (run length) [27]. One can examine the gradient image produced via application of an edge detector to the original and extract a variety of features related to the magnitude, frequency, directives and other properties of the edges [28]. Run length measurements [27] [28] may also generate a number of features useful for texture-based classification. The run length is measured as the number of pixels along a particular direction which lies within a gray level margin around the value of the starting pixel. Obviously, the run length is direction-dependent and has high values in coarse textures and low values in fine textures.

In [26], it was shown that the co-occurrence matrix method provides a more robust description of an image texture than do the GLDM, the run length method and the power spectral density method [29], which examines the Fourier power content of image regions

and belongs to another category of texture description methods, the spectral methods discussed in the next section.

Measurement Following Linear Filtering

The purpose of these methods is to enhance some characteristics of the texture, before actually measuring them. To this end, the image is first convolved with a number of small rectangular filters. The size of the filters determines how ”local” the result of each convolution may be used as a feature itself, while secondary features may be extracted as well, such as the variance of values in a pixel neighborhood.

The main problem that needs to be addressed when local linear filtering is used is the choice of the filters. It has been shown [30] that the filters that perform optimally are those implementing the Karhunen-Loève transform. However, in order to obtain optimality these filters must differ from texture to texture. This renders them unpractical, especially in cases where there is no prior knowledge of the examined textures. Instead, other substitutes have been proposed which approximate the Karhunen-Loève transform and have been shown to work well with a variety of textures. Among these are the discrete cosine and sine transforms and the discrete hadamard transform [30]. Their implementation is quite simple and various features may be produced from them.

A more empirical approach to the same problem was presented by Laws in [31]. Even though the reasoning followed by Laws is very different to the one leading to the discrete transform mentioned above, the similarity between the resulting masks is obvious.

2.4.2 Structural Methods

This category of texture quantification methods assumes the existence of texture primitives, which are combined with each other using placement rules to produce higher level texture structures. The quantification methods attempt to detect the primitives as well as the associated rules [32].

2.4.3 Spectral Methods

This category method is associated with the measurement of the spatial frequency content of the image. These approaches are somewhat similar to structural methods in the sense that a texture generated by a set of textural primitives will have high spectral content at spatial frequencies inversely proportional to the size of the primitives.

A common spectral method in texture analysis is the use of the autocorrelation of an image or image region [33]. When a coarse texture is examined, the autocorrelation value drops slowly with distance, and vice versa for fine textures. In general, the variation of the autocorrelation function indicates the extent of periodicity as well as the order of the size of any primitives in an image. The autocorrelation function and the power spectral density of an image form a Fourier transform pair. This leads to the use of the 2-D Fourier transform, either in the form of the discrete fourier transform (DFT), or the fast fourier transform (FFT) for the determination of textural features. Such processing may provide the principal directions of the primitives and the main frequencies contained within them [34].

The 2-D Fourier transform may also be applied in a selected window of an image. Some researchers have used frequency-related techniques in windows inside the images to localize the frequency content in space and locally characterize texture. These techniques are known as discrete image transform techniques. Among these efforts, we mention the use of the 2-D Gabor filters [35], which are meant to simulate the physiological function of human vision and 2-D wavelet transform. The 2-D Gabor filters will be presented in detail in section 5.2.1. There is a close relation between such frequency-related textures analysis methods and the statistical methods using linear filtering. Indeed, the former may be considered as a subset of the latter if they are regarded at a particular scale.

2.4.4 Model-based Methods

Model-based methods assume that texture is the result of underlying processes and they aim at identifying these processes and their parameters. We briefly review some model-based methods in this section.

First, there are the random field (RF) models, which consider the gray-scale values of an image pixel $I(x)$ as a linear combination of the gray-scale of its neighboring pixels $I(x')$, plus a noise component:

$$I(x) = \sum_{x' \in N} h(x')I(x') + n(x) \tag{2.4}$$

where N is the neighborhood of x and $n(x)$ is the noise component. For a given image and for a given class of an RF model, what needs to be done is to estimate the model coefficients $h(x')$, as well as other parameters of the model, which is usually done using the least squares method, even though other methods exist. Examples of classes of RF models that have attracted attention in the past are Gaussian Markov random fields [36], autoregressive [37] and long correlation models [38].

Another group of model-based methods are those relying on fractal analysis of images. This kind of analysis is suitable for textures which are at some degree self similar in different scales, meaning that their primitives can be decomposed to a set of their own scaled-down copies.

2.5 Statistical Learning Theory

The statistical learning theory, was introduced in the late 1960's. Until the 1990's it was a purely theoretical analysis of the problem of function estimation from a given collection of data [39]. It views a supervised classification problem as an input-output relationship [40] [41] [42]. In the case of a two-class classification problem, an algorithm learns from a given set of k training samples, $(\mathbf{x}_1, y_1), ..., (x_k, y_k)$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$, which are drawn from a fixed but unknown cumulative (probability) distribution function $P(\mathbf{x}, y)$, where $\mathbf{x}$ is an N-dimensional observed data vector and y_i is a class label assigned to each data vector. $\mathbb{R}$ is the set of all real numbers. A decision rule or classification rule is represented by $\{f_\alpha(\mathbf{x}) : \alpha \in \Lambda\}$, $f_\alpha : \mathbb{R}^n \rightarrow \{-1, +1\}$, where Λ is the set of parameters used in the decision rule [43]. For example, in a multilayer neural network, Λ is st of weights of the network.

The aim of classification is to assign the class label y, based on the training samples $\mathbf{x}$ and a decision rule f_α, that provides the smallest possible error over the independent (or unknown) data samples or the smallest possible expected risk defined as,

$$R(a) = \int L(y, f_\alpha(\mathbf{x}))dP(\mathbf{x}, y) \tag{2.5}$$

The function f_α is called the hypothesis. The set $\{f_\alpha \mathbf{x}) : \alpha \in \Lambda\}$ called the hypothesis space [43], and $L(y, f_\alpha(\mathbf{x}))$ is the loss or discrepancy between the response y of the supervisor or teacher to given input $\mathbf{x}$ and the response $f_\alpha(\mathbf{x})$ provided by the learning machine. In other words, the expected risk is a measure of the performance of a decision rule that assigns the class label y to an input data vector $\mathbf{x}$. However, evaluation of the expected risk is difficult, since the cumulative distribution function $P(\mathbf{x}, y)$ is unknown and thus one may not be able to evaluate the integral in Equation (2.5). The only known information is contained in the training samples. Therefore, a stochastic approximation of the integral in Equation 2.5 is desired that can be computed empirically by a finite sum given by,

$$R_{emp}(\alpha) = \frac{1}{k} \sum_{i=1}^{k} L(y_i, f_\alpha(\mathbf{x}_i)) \tag{2.6}$$

This sum is known as the empirical risk. The value $R_{emp}(\alpha)$ is a fixed number for a given α and a particular training data set.

2.5.1 Empirical Risk Minimization

Based on the law of large numbers [44], the empirical mean of a random variable converges to its expected value if the size of the training samples is infinitely large. This remark justifies the use of empirical risk $R_{emp}(\alpha)$ instead of the risk function $R(\alpha)$. However, convergence of the empirical mean of the random variable to its expected value does not imply that the value α that minimizes the empirical risk will also minimize the risk function $R(\alpha)$. If convergence of the minimum of the empirical risk to the minimum of the expected risk does not occur, this principle of empirical risk minimization is said to be inconsistent. In this case, even though the empirical risk is minimized, the expected risk may be high. In other words, a small error rate of a learning machine on the training samples does not necessarily guarantee high generalization ability (i.e. the ability to work well on unseen data). This situation is commonly referred to as overfitting. Vapnik and Chervonenkis [45] [46] have shown that consistency occurs if and only if convergence in probability of the empirical risk to the expected risk is substituted by uniform convergence in probability. Note that convergence in probability of $R(\alpha)$ means that for any $\varepsilon > 0$ and for and $\eta > 0$, there exists a number $k_0 = k_0(\varepsilon, \eta)$ such that for any $k > K_0$, the inequality $R(\alpha_k) - R(\alpha_0) < \varepsilon$ holds true with a probability of at least $1 - \eta$ [39]. ε is a small number

and η is referred to as the level of significance-similar to the α value in statistics. Uniform convergence in probability is defined as

$$\lim_{k\to\infty} \Pr(\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)| > \varepsilon) \to 0, \quad \forall\varepsilon \tag{2.7}$$

where "sup A" is the supremum over a nonempty set A and is defined as the smallest scalar x such that $x \geq y$ for all $y \in A$. Uniform convergence is necessary and sufficient condition for the consistency of the principle of empirical risk minimization.

Vapnik and Chervonenkis [45] have also shown that necessary and sufficient condition for the consistency amounts to the fitness of the VC-dimension in the hypothesis space. The VC-dimension is a measure of the capacity of a set of classification functions or the complexity of the hypothesis space [47]. The VC-dimension, generally denoted by h, is an integer that represents the largest number of data points that can be separated by a set of functions f_α in all possible ways. For example, for a binary classification problem, the VC-dimension is the maximum number of points which can be separated into two classes without error in all possible 2^k ways. The proof of consistency of the empirical risk minimization (ERM) can be found in Vapnik [40] [41].

The theory of uniform convergence in probability also provides a bound on the deviation of empirical risk from the expected risk given by

$$R(\alpha) \leq R_{emp}(\alpha) + \varphi(\frac{h}{k}, \frac{\log(\eta)}{k}) \tag{2.8}$$

where k is the number of training samples, and the confidence term φ is defined by

$$\varphi(\frac{h}{k}, \frac{log(\eta)}{k}) = \sqrt{\frac{h(log\frac{2k}{h} + 1) - log(\eta/4)}{k}} \tag{2.9}$$

The parameter h is the VC-dimension of a set of classifiers, and the bound in Equation **??**Raemp) holds for any $\alpha \in \Lambda$, and $k > h$ with a probability of at least $1 - \eta$ such that $0 \leq \eta \leq 1$. From the bound in Equation **??**Raemp), a good generalization performance (i.e. the smallest expected risk $R(\alpha)$) can be obtained when the empirical risk as well as the ration between the VC-dimension and the number of training samples is small. With a fixed number of training samples, the empirical risk is usually a decreasing

function of the VC-dimension while the confidence term is an increasing function. This means that there exists an optimal value of the VC-dimension that can give the smallest expected risk. Therefore, to obtain accurate classification, the choice of an appropriate VC-dimension is also crucial.

2.5.2 Structural Risk Minimization

For the selection of an appropriate VC-dimension for a given set of functions, Vapnik [42] proposed the principle of structural risk minimization (SRM) that is based on the fact that the minimization of the expected risk is possible by simultaneous minimization of the two terms in Equation (2.8). The first is the empirical risk term and second is the confidence term, which depends on the VC-dimension of the set of functions. SRM minimizes the expected risk function with respect to both the empirical risk and the VC-dimension. To achieve this aim, a nested structure of hypothesis space is introduced by dividing the entire class of functions into nested subsets

$$H_1 \subset H_2 \subset \cdots \subset H_n \subset \cdots \tag{2.10}$$

Each hypothesis space has the property that $h(n) \leq h(n+1)$ where $h(n)$ is the VC-dimension of the set H_n. This implies that the VC-dimension of each hypothesis space is finite. The principle of SRM can be mathematically represented as

$$\min_{H_n}(R_{emp}(\alpha) + \varphi(\frac{h}{k}, \frac{log(\eta)}{k})) \tag{2.11}$$

Even though the SRM is mathematically defined, its implementation may not be easy due to the difficulty in computing the VC-dimension for each H_n. Only a few models are known for the computation of the VC-dimension [48]. Thus, the SRM procedure is to minimize the empirical risk for each hypothesis space so as to identify the hypothesis space with the smallest expected risk. The hypothesis with the smallest expected risk is the best compromise between the empirical risk (i.e. approximation of the training data) and the confidence term (i.e. measure of the complexity of the approximating function). The SVM algorithm is an implementation that is able to minimize the empirical risk as well as a bound on the VC-dimension [43].

Chapter 3

Image Preprocessing for Defects Detection

3.1 Introduction

The gray level of the area in an X-ray image depends on the distance the X-rays pass through. The longer the X-rays pass through in the metal, the more absorption there will be. Therefore fewer X-rays reach the film. In this area a lighter part will be obtained on the film. From Figure 3.1 it can be seen, although the weld-seam is not flat steel, it does not change sharply.

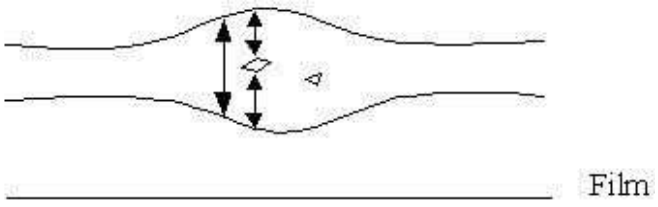
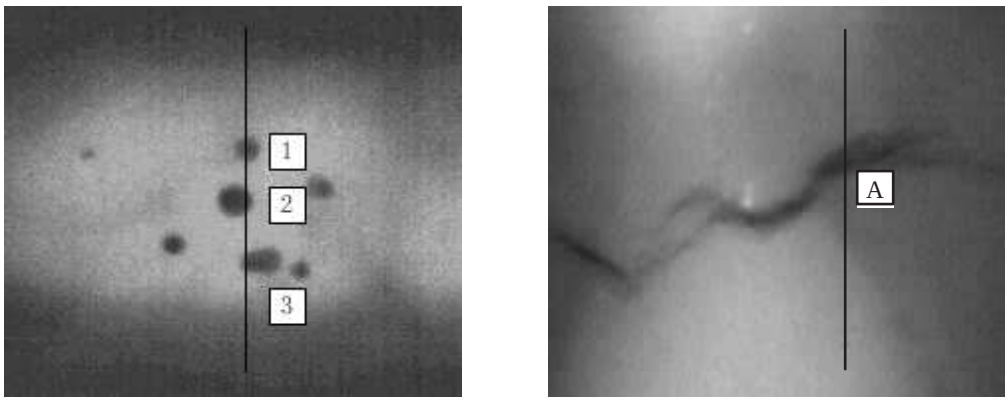


Figure 3.1: Illustration of the darkness on the film depending on the thickness the X-rays pass through

Figure 3.2 is part of a whole image. Figure 3.3 shows the pixels' gray level profile of the image shown in Figure 3.2. From the curve, it can be seen:

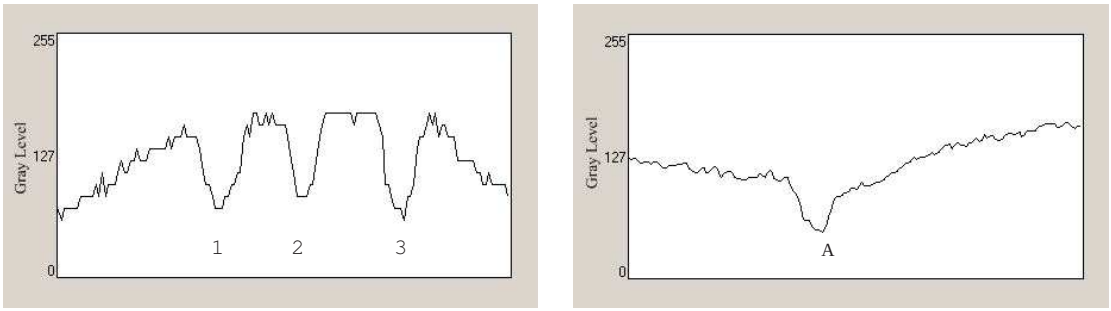
1. The gray level is only in a small range, so although we use 8 bit (256 levels) to record the image most of the levels are wasted.
2. There are depressions at the defects. But the image also contains significant noise. The sizes of the noise are small, but the change rates are usually large.



(a) Part of image with porosity defect

(b) Part of image with crack defect

Figure 3.2: Part of image with defect



(a) Gray levels along the line shown in Figure 3.2(a)

(b) Gray levels along the line shown in Figure 3.2(b)

Figure 3.3: Gray level profile

Because of these characteristics of radiographic images, in order to find the defects effectively, some pre-adjustments should be processed. In this study, morphological enhancement method and wavelet thresholding are applied to preprocess the radiographic images and the results are compared with other commonly used methods.

3.2 Contrast Enhancement

Radiographic images usually have poor contrast. The aim of contrast enhancement is to improve the quality of radiographic images. In the original radiographic images, the distribution of gray levels is highly skewed towards the darker side. Therefore, it is desirable to stretch the histogram distribution to a rectangular shape instead of a skewed one. In this

section, three enhancement methods, morphological enhancement, histogram equalization and contrast limited adaptive histogram equalization, are investigated.

3.2.1 Morphological Enhancement

Morphological contrast enhancement is based on mathematical morphology theory. Mathematical morphology is a topological and geometrical based approach to image analysis. Mathematical morphology can be defined as the theory for analysis of spatial structures. It aims to analyze the shape and form of objects using the language of set theory [49].

Morphological filtering is typically defined as grouping different pixels in the images based on their color, spatial frequency and intensity. Objects in the morphologically processed image are usually well identified by a group of pixels that represent the objects shape. The main morphological filters used in shape reconstruction (grouping different objects) are erosion, dilation, opening and closing [50] [51].

Morphological operators often take a binary image and a structuring element as input and combine them using a set of operator. The concepts of erosion and dilation can be extended to gray-scale images suing the means of threshold decomposition, which uniquely represents a gray-scale image as a collection of cross-section, or binary image [34] [52]. A binary image is just a special case of gray-scale image (that is, the number of gray levels is two).

The gray-scale erosion of an image $f(x, y)$ by a structuring function $b(x, y)$, which is also a gray-scale image, is denoted by $f \ominus b$ and is defined as

$$(f \ominus b)(s,t) = min\{f(s+t, t+y) - b(x,y) | (s+x), (t+y) \in D_f; (x,y) \in D_b \quad (3.1)$$

where D_f and D_b are the domains of f and b, respectively.

The gray-scale dilation of an image $f(x, y)$ by a structuring function $b(x, y)$, is denoted by $f \oplus b$ and is defined as

$$(f \oplus b)(s,t) = min\{f(s-t, t-y) + b(x,y) | (s-x), (t-y) \in D_f; (x,y) \in D_b \quad (3.2)$$

The expressions for opening and closing of gray-scale images have the same form as their binary counterparts. The gray-scale opening of an image $f(x, y)$ by a structuring function $b(x, y)$, is denoted by $f \circ b$ and is defined as

$$f \circ b = (f \ominus b) \oplus b \tag{3.3}$$

Similarly, the closing of $f(x, y)$ by b, denoted $f \bullet b$, is

$$f \bullet b = (f \oplus b) \ominus b \tag{3.4}$$

The morphological top-hat transform of an image, th, is defined as the difference between the original image and the opened image.

$$th = f - (f \circ b) \tag{3.5}$$

Similarly the bottom-hat transform of an image, bh, is defined as the difference between the closed image and the original image.

$$bh = (f \bullet b) - f \tag{3.6}$$

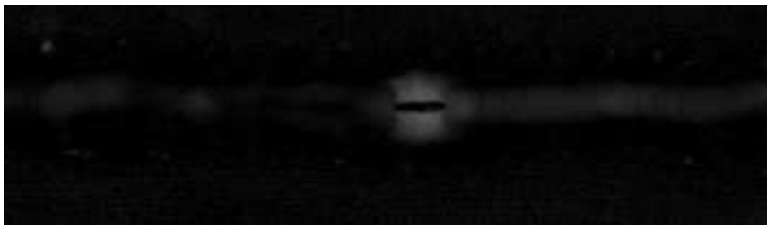
Morphological contrast enhancement is based on the notion of morphological top-hat and bottom-hat transform. Top-hat and bottom-hat filters can be used to extract light objects (or, conversely, dark ones) from a dark (or light) but slowly changing background [53].

We use both the top-hat and the bottom-hat filters on the original image, and combine the results by adding to the original image the result of the top-hat filter, and subtracting the result of the bottom-hat. The enhancement process is shown in Figure 3.4.

3.2.2 Histogram Equalization

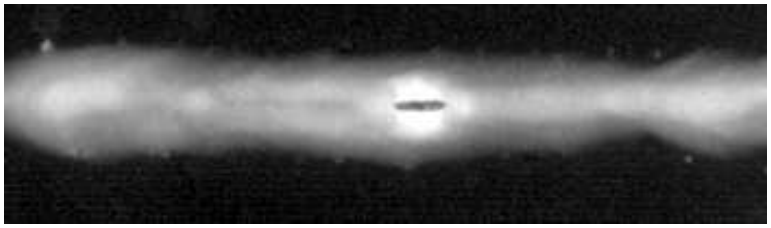
Currently, the most frequently used technique to enhance the contrast of radiographic images is histogram equalization (HE). It is based on the assumption that a good gray-level assignment scheme should have equally distributed brightness levels over the whole

(a) Original image



(b) Image after using top hat filter

(c) Image after using bottom hat filter



(d) Adding to the original image the result of the top-hat filter

(e) Subtracting the result of the bottom-hat

Figure 3.4: Morphological enhancement process

brightness scale. Individual pixels retain their brightness order. However, the values are shifted so that they are equally distributed over the brightness scale. The result of the brightness transformation should be that the cumulative histogram becomes a straight line.

If G' is the transformed gray value corresponding to the original gray value G of any pixel, the principle of histogram equalization postulates that

$$G' = F(G) = G'_{min} + \Delta G' \cdot H(G)/N \tag{3.7}$$

where $\triangle G' = G'_{max} - G'_{min}$, G'_{max} and G'_{min} represent the upper and lower limits of the transformed gray values respectively, $H(G)$ represents the cumulative histogram of the gray values of the original images, and represents the number of pixels over which the histogram is taken.

3.2.3 Contrast Limited Adaptive Histogram Equalization

HE transforms image pixels based on overall image statistics. Adaptive histogram equalization (AHE) involves selecting a local neighborhood centered around each pixel, calculating and equalizing the histogram of the neighborhood, and then mapping the centered pixel based on the new equalized local histogram [54]. For example, at each point in an input image we could consider a 8×8 window around that point. The 64-element histogram could then be used to determine a mapping function to histogram equalize that point based on the neighborhood. Since each point is based on its own neighborhood, the mapping function can vary over the image.

Contrast limited adaptive histogram equalization (CLAHE) seeks to reduce the noise produced in homogeneous areas by basic adaptive histogram equalization, and was originally developed for medical imaging. It has been successful for the enhancement of portal images [55]. The homogeneous areas can be characterized by a high peak in the histogram associated with the contextual regions since many pixels fall inside the same gray range. With AHE, a local histogram is calculated and used to obtain the final value. High peaks in the histogram lead to large values in the final image because of integration. This problem can be corrected by limiting the amount of contrast enhancement at every pixel, which is achieved by clipping the original histogram to a limit. CLAHE is an improved version of AHE. It can overcome the limitations of standard histogram equalization and AHE.

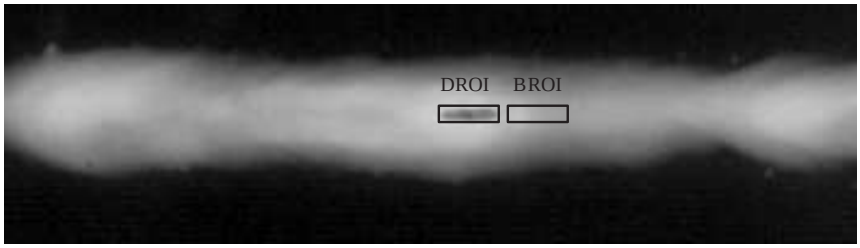
3.2.4 Performance Evaluation

The performance of the enhancement is characterized through visual inspection, quantitative measures such as those based on background and detail region variance and comparative measures such as image profiles. An experiment comparison is made among HE, CLAHE and morphological enhancement.

We enhanced the radiographic images by HE, CLAHE and morphological enhancement. The result of histogram equalization on a typical radiography image can be seen in Figure 3.5(b). As a digital radiographic image has only a finite number of gray scales, an ideal equalization is not possible. It causes some pixels with initially different brightness values to be assigned the same value, and other values to be missing altogether. From Figure 3.5(b) and Figure 3.6(b), we can see that the histogram equalization enhances the contrast for brightness values close to maxima in the histogram and decreases contrast near the minima. That is, it improves the contrast in the image in areas of poor contrast at the expense of those areas where there is already good contrast. Figure 3.5(b) shows that histogram equalization in its basic form can give a result that is even worse than the original image. Large peaks in the histogram can also be caused by large areas of similar brightness. Frequently these correspond to areas of background, and are essentially uninteresting. The effect of histogram equalization on these areas is to enhance the visibility of noise. The feature of interest in the radiographic images such as defects need enhancement locally. However, the technique does not also adapt to local contrast requirements; minor contrast differences can be entirely missed when the number of pixels falling in a particular gray range is small. Applying contrast limited adaptive histogram equalization on the image in Figure 3.5(a) results in image that can be found in Figure 3.5(c). The local contrast is largely improved. And the minor contrast differences of defects with background can be kept. However, the most striking feature of the image is the nonuniform background that has become more visible. Figures 3.5(d) shows morphological enhancement with a disk-shaped structuring element. The defect contrast is improved. At the same time, the background nonuniform is greatly reduced.

The estimate is performed separately in defect regions (defect variance, DV) and in background regions (background variance, BV) of the radiographic image. We expect reasonably high values of DV in the enhanced images, while the BV value should remain low in order to indicate limited noise amplification.

The defect region of interest (DROI) and background region of interest (BROI) used for



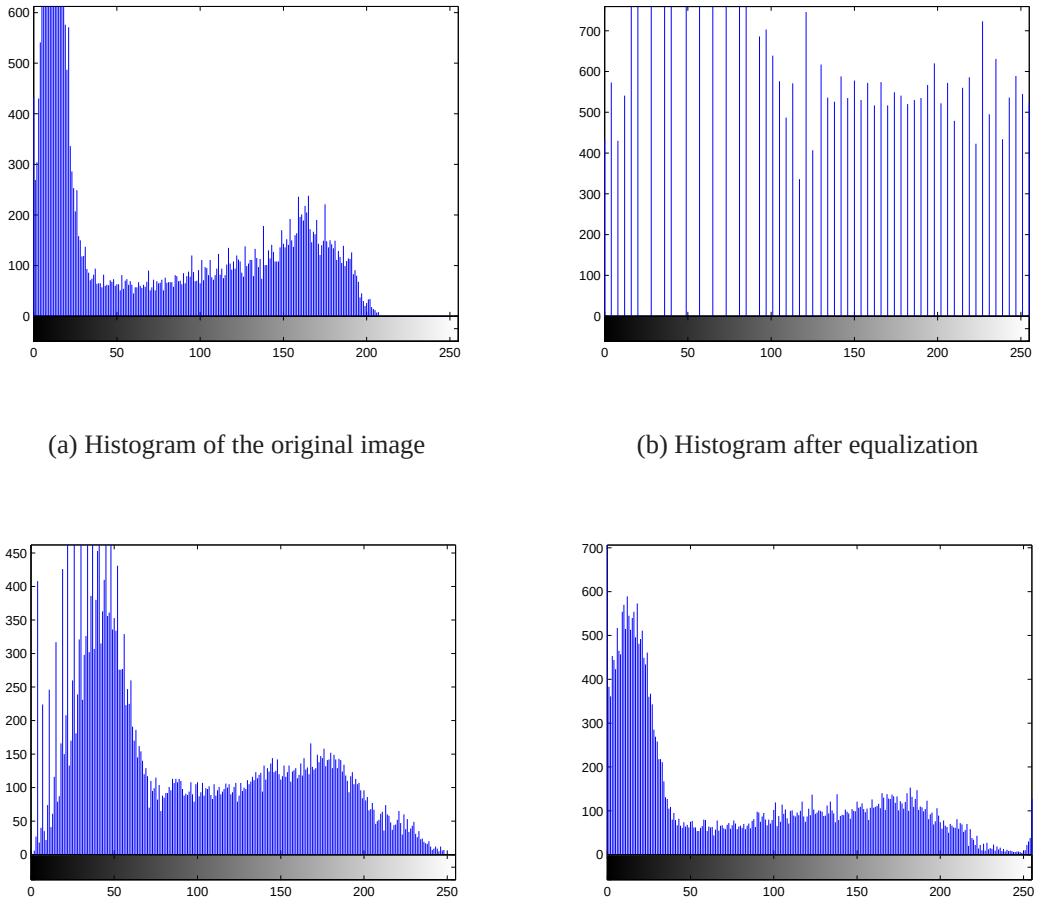
(a) Original image

(b) Image after histogram equalization

(c) Image after CLAHE

(d) Image after morphological enhancement

Figure 3.5: Contrast enhancement of a radiographic image



(a) Histogram of the original image

(b) Histogram after equalization

(c) Histogram after CLAHE

(d) Histogram after morphological enhancement

Figure 3.6: Histogram of a radiographic image

calculating the DV and BV are highlighted. The contrast to noise ratio (CNR) is also calculated for quality measurement. It is defined as [56]

$$CNR = \frac{|\mu_d - \mu_b|}{\sqrt{0.5(\sigma_d^2 + \sigma_b^2)}} \tag{3.8}$$

where μ_d and σ_d are the mean and the standard deviation computed in the DROI, μ_b and σ_b are the mean and the standard deviation computed in the BROI.

The data in Table 3.1 summarizes the quantitative measures obtained for the radiographic image shown in Figure 3.5. The morphological enhancement algorithm shows a moderate improvement the DV over the HE and CLAHE algorithms. HE maintains the highest noise amplification. The morphological contrast enhancement is able to enhance the fine details

of radiographic image without noise emphasis or the over-accentuation of background tissue. So, we apply this method for image enhancement.

Table 3.1: Performance measures for enhancement of the radiographic image

Enhancement algorithm	DV	BV	CNR
Original	737.007	149.6076	27.8985
HE	739.1787	175.3282	26.3685
CLAHE	3027.8	351.5897	65.1050
Morphological enhancement	5338.1	449.0186	90.8885

3.3 Noise Reduction

The common noise sources that corrupt images can be divided into three categories.

1. Firstly, images originally recorded on photographic film are subject to degradation by film grain noise.
2. Secondly, the conversion of an image from optical to electrical form is a statistical process since, in reality, each picture element receives a finite number of photons.
3. Finally, electronic amplifiers that process the signal introduce thermal noise.

Considerable efforts have been devoted to modeling noise from these these sources. The most common source of noise is counting statistics in the image detector due to a small number of incident particles. The simplest form of spatial averaging is simply to add together the pixel brightness values in each small region of the image, divided by the number of pixels in the neighborhood, and use the resulting value to construct a new image. Since the noise in this image is random due to the counting statistics of the small number of photons, the improvement in image quality or signal-to-noise ratio is just the square root of the number of the pixels.

Noisy images may also occur due to instability in the light source or detector during the time required for scanning or digitizing an image. The pattern of this noise may be quite

different from the essential Gaussian noise due to counting statistics, but it still shows up as a variation in brightness in uniform regions of the scene.

In this section, wavelet thresholding is applied to reduce the radiographic image noise and compared with the median filter.

3.3.1 Discrete Wavelet Transforms

"Wavelet" is a relatively new but rapidly growing area. Basically, it refers to a class of transforms with an added advantage in its ability in scaling and translating. Just like "Fourier Transform", it can act as a filter, which will be shown mathematically later. It is this behavior that leads to its widely accepted use in vastly different areas. Some of the areas are image, communication and noise suppression. Given the abundant literature on basic wavelet theories, one can easily find extensive presentations on the topic of interest. For example, [57] [58] [59] have now become classics, [60] [61] [62] [63] are popular books.

In most areas, "Wavelet transform" has taken the place of the early "Fourier transform" due to its flexibility both in terms of scaling and translation.

In practice, we can compute a wavelet transform only over finitely many scales. This is because the observed data is limited between a non-zero small (fine) scale and a finite large (coarse) scale. According to Mallat [64], one can normalize the observable finest scale to $2^0 = 1$, and the coarsest scale to 2^J where J is dependent on the sample size of the data.

In order to model this scale limitation, we introduce a scaling function $\phi(x)$ and its Fourier transform $\Phi(\omega)$ that satisfies almost everywhere:

$$|\Phi(\omega)|^2 = \sum_{j=1}^{+\infty} |\Psi(2^j\omega)|^2 \tag{3.9}$$

similar to the dilation of a wavelet, the dilated scaling function and its Fourier transform are given by:

$$\phi_{2^j}(x) = 2^{-j}\phi(2^{-j}x) \tag{3.10}$$

$$\Phi_{2^j}(\omega) = \Phi(2^j\omega) \tag{3.11}$$

The approximation to a function $f(x)$ at scale 2^j and the Fourier transform of the approx-

imation are given by:

$$S_{2^j} f(x) = \phi_{2^j} * f(x) \tag{3.12}$$

$$S_{2^j} F(\omega) = \Phi(2^j \omega) F(\omega) \tag{3.13}$$

Consider a finite-scale representation of (3.9),

$$|\Phi(\omega)|^2 = \sum_{j=1}^{J} |\Psi(2^j \omega)|^2 + |\Phi(2^J \omega)|^2 \tag{3.14}$$

By Parseval's theorem it can be shown that:

$$\|S_{2^0} f(x)\|^2 = \sum_{j=1}^{J} \|W_{2^J} f(x)\|^2 + \|S_{2^J} f(x)\|^2 \tag{3.15}$$

(3.15) shows that a finite-scale wavelet transform represented by $\{(W_{2^j} f(x))_{j=1,2,\ldots,J}, S_{2^J} f(x)\}$ is an isometry. It is interpreted as a multiresolution representation. The information lost in smoothing $S_{2^j} f$ to $S_{2^{j+1}} f$ is stored in $W_{2^j} f$, called the "detail signal" [59]. The original signal $S_{2^0} f$ can be recovered from $S_{2^J} f$ by adding details recursively at each scale.

Now consider a sequence $\{S_n\}$ of finite energy, i.e.

$$\sum_{n=-\infty}^{+\infty} |s_n|^2 < \infty \tag{3.16}$$

Mallat has shown that if there exist two constants, $C_2 > C_1 > 0$, such that,

$$\forall \omega \in R, \quad C_1 \leq \sum_{k=-\infty}^{+\infty} |\Phi(\omega + 2k\pi|^2 \leq C_2 \tag{3.17}$$

there exists a function $f(x) \in \boldsymbol{L}^2(\boldsymbol{R})$ such that,

$$\forall n \in \boldsymbol{Z}, \quad S_{2^0} f(n) = s_n \tag{3.18}$$

(3.18) means that a sequence $\{s_n\}$ can be viewed as uniform sampling of $S_{2^0} f(x)$, the approximation of some function $f(x)$ at scale $2^0 = 1$. To generalize this connection, the uniform sampling sequences of a finite-scale wavelet transform of $f(x)$,

$$\{(W_{2^j} f(n + \upsilon))_{j=1,2\ldots,J}, \;\; S_{2^J} f(n + \tau)\} \tag{3.19}$$

are called a discrete wavelet transform (DWT) of the sequence $\{s_n\}$. Here υ and τ are sampling shifts depending on the particular scaling function and wavelet.

The computation of the DWT of a sequence relies on two discrete filters, H and G. They are related to scaling function and wavelets by

$$\Phi(\omega) = H(\frac{\omega}{2})\Phi(\frac{\omega}{2}) \tag{3.20}$$

$$\Psi(\omega) = G(\frac{\omega}{2})\Phi(\frac{\omega}{2}) \tag{3.21}$$

where $H(\omega)$ and $G(\omega)$ are 2π-periodic and satisfy,

$$H(0) = 1 \tag{3.22}$$

$$|H(\omega)|^2 + |H(\omega + \pi)|^2 \leq 1 \tag{3.23}$$

$$|H(\omega)|^2 + |G(\omega)|^2 = 1 \tag{3.24}$$

3.3.2 Denoising Using Wavelet Thresholding

Wavelet thresholding (first proposed by Donoho [65] [66] [67]) is a signal estimation technique that exploits the capabilities of wavelet transform for signal denoising and has recently received extensive research attention. This method has great efficiency in preserving the true underlying data while suppressing noise.

Wavelet denoising attempts to remove the noise present in the signal while preserving the signal characteristics, regardless of its frequency content. It involves three steps:

1. Calculate the wavelet transform of the image.
2. Threshold the wavelet coefficients by discarding (setting to zero) the coefficients with relatively small or insignificant magnitudes.
3. Compute the inverse wavelet transform to get the denoised image.

The radiographic image without noise is represented as a two-dimensional matrix $\mathbf{g} = \{g_{ij}\}$. The noisy radiographic image $\mathbf{f} = \{f_{x,j}\}$ is modeled as

$$f_{i,j} = g_{i,j} + n_{i,j} \qquad i,j = 1,...,N \tag{3.25}$$

where $\{n_{i,j}\}$ is *iid* as $N(0,\sigma^2)$.

Let $\mathbf{Y} = \mathcal{W}\mathbf{f}$ denote the matrix of wavelet coefficients of $\mathbf{f}$, where $\mathcal{W}$ is the two-dimensional wavelet transform operator. The wavelet decomposition of the radiographic image is done as shown in Figure 3.7. In the first level of decomposition, the image is split into 4 subbands, namely the HH_1, HL_1, LH_1 and LL_1 subband. The HH_1 subband gives the diagonal details of the image. The HL_1 subband gives the horizontal features while the LL_1 subband not shown is the low resolution residual consisting of low frequency components and it is this subband that is further split at second level of decomposition into HH_2, HL_2, LH_2 and LL_2 subbands.

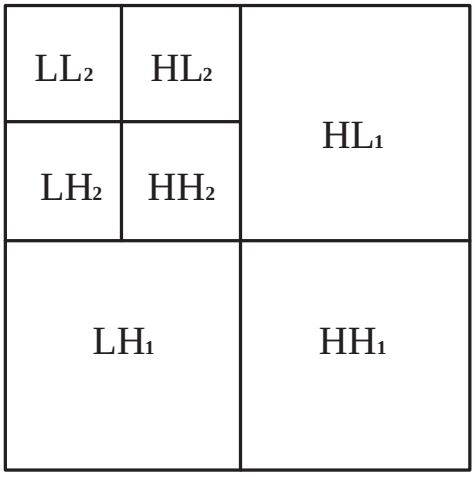


Figure 3.7: Sub-bands of the 2-D wavelet transform

The wavelet thresholding denoising method filters each coefficient $Y_{i,j}$ from the detail subband with a threshold function to obtain $\hat{X}_{i,j}$. The denoised estimate is then $\hat{\mathbf{g}} = \mathcal{W}^{-1}\hat{\mathbf{X}}$, where $\mathcal{W}^{-1}$ is the inverse wavelet transform operator.

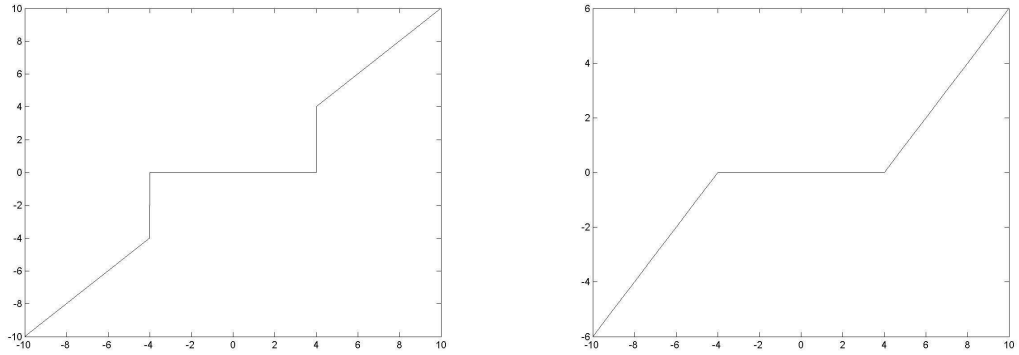
Hard thresholding and soft thresholding are two thresholding methods frequently used.

In case of hard thresholding (refer to Figure 3.8(a)),

$$D(Y,\lambda) \equiv \begin{cases} Y & if\ |Y| > \lambda \\ 0 & otherwise \end{cases} \tag{3.26}$$

In case of soft thresholding (refer to Figure 3.8(b)), or wavelet shrinkage,

$$D(Y,\lambda) = sgn(Y)max(0,|Y| - \lambda) \tag{3.27}$$



(a) Hard thresholding ($\lambda = 4$)

(b) Soft thresholding ($\lambda = 4$)

Figure 3.8: Hard thresholding and soft thresholding

The soft thresholding rule is chosen over hard thresholding due to the following reasons: soft thresholding has been shown to achieve near minimax rate over a large number of Besov spaces [68]. Moreover, it is also found to yield visually more pleasing images. Hard thresholding is found to introduce artifacts in the recovered images. So, in what follows, soft thresholding will be the primary focus. We now study two soft thresholding techniques: VisuShrink and BayesShrink.

VisuShrink

The threshold $\lambda_{UNIV} = \sqrt{2 \ln N}\sigma$ (N being the signal length, σ being the noise variance) is well known in wavelet literature as the universal threshold. It is the optimal threshold in the asymptotic sense and minimizes the cost function of the difference between the function and the soft thresholded version of the same in the L_2 norm sense (i.e. it minimizes $E\|Y_{Thresh} - Y_{Orig.}\|$).

VisuShrink is thresholding by applying the universal threshold proposed by Donoho and Johnstone [66]. This threshold is given by $\sigma\sqrt{2 \log M}$ where σ is the noise variance and M is the number of pixels in the image. It is proved in [66] that the maximum of any M values i.i.d as $N(0, \sigma^2)$ will be smaller than the universal threshold with high probability, with the probability approaching 1 as M increases. Thus, with high probability, a pure noise signal is estimated as being identically zero.

BayesShrink

BayesShrink is a subband adaptive threshold computed for each detail subband. In BayesShrink [68] we determine the threshold for each subband assuming a generalized gaussian distribution (GGD). The GGD is given by

$$GG_{\sigma_X,\beta}(x) = C(\sigma_X, \beta) \exp -[\alpha(\sigma_X, \beta)|x|]^{\beta} \tag{3.28}$$

$-\infty < x < \infty, \beta > 0$, where

$$\alpha(\sigma_X, \beta) = \sigma_X^{-1} \left[\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}\right]^{1/2}$$

and

$$C(\sigma_X, \beta) = \frac{\beta \cdot \alpha(\sigma_X, \beta)}{2\Gamma\left(\frac{1}{\beta}\right)}$$

and $\Gamma(t) = \int_0^{\infty} e^{-u} u^{t-1} du$.

The parameter σ_X is the standard deviation and β is the shape parameter. It has been observed [68] that with a shape parameter β ranging from 0.5 to 1, we can describe the distribution of coefficients in a subband for a large set of natural images. Assuming such a distribution for the wavelet coefficients, we empirically estimate β and σ_X for each subband and try to find the threshold T which minimizes the Bayesian risk, i.e. the expected value of mean square error.

$$\tau(T) = E(\hat{X} - X)^2 = E_X E_{Y|X}(\hat{X} - X)^2 \tag{3.29}$$

where $\hat{X} = \eta T(Y), Y|X \sim N(x, \sigma^2)$ and $X \sim GG_{X,\beta}$. The optimal threshold T^* is then given by

$$T^*(\sigma_x, \beta) = \arg\min_T \tau(T) \tag{3.30}$$

This is a function of the parameters σ_x and β. Since there is no closed form solution for T^*, numerical calculation is used to find its value.

It is observed that the threshold value set by

$$T_B(\sigma_X) = \frac{\sigma^2}{\sigma_X} \tag{3.31}$$

is very close to T^*.

The estimated threshold $T_B = \sigma^2/\sigma_X$ is not only nearly optimal but also has a intuitive appeal. The normalized threshold, T_B/σ is inversely proportional to σ , the standard

deviation of X , and proportional to σ_X , the noise standard deviation. When $\sigma/\sigma_X \ll 1$, the signal is much stronger than the noise, T_B/σ is chosen to be small in order to preserve most of the signal and remove some of the noise; when $\sigma/\sigma X \gg 1$, the noise dominates and the normalized threshold is chosen to be large to remove the noise which has overwhelmed the signal. Thus, this threshold choice adapts to both the signal and the noise characteristics as reflected in the parameters σ and σ_X.

The GGD parameters, σ_X and β, need to be estimated to compute $T_B(\sigma_X)$. The noise variance σ^2 is estimated from the subband HH_1 by the robust median estimator [68],

$$\hat{\sigma} = \frac{Median(|Y_{ij}|)}{0.6745}, Y_{ij} \subseteq subbandHH_1 \tag{3.32}$$

The parameter β does not explicitly enter into the expression of $T_B(\sigma_X)$. Therefore it suffices to estimate directly the signal standard deviation σ_X. The observation model is $Y = X + V$, with signal X and noise V independent of each other, hence

$$\sigma_Y^2 = \sigma_X^2 + \sigma^2 \tag{3.33}$$

where σ_Y^2 is the variance of Y. Since Y is modeled as zero-mean, σ_Y^2 can be found empirically by

$$\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i,j=1}^{n} Y_{ij}^2 \tag{3.34}$$

where $n \times n$ is the size of the subband under consideration. Thus

$$\hat{T}_B(\hat{\sigma}_X) = \frac{\hat{\sigma}^2}{\hat{\sigma}_X} \tag{3.35}$$

where

$$\hat{\sigma}_X = \sqrt{max(\hat{\sigma}_Y^2 - \hat{\sigma}^2, 0)} \tag{3.36}$$

In the case that $\hat{\sigma}^2 \geq \hat{\sigma}_Y^2$, $\hat{\sigma}_X$ is taken to be zero, i.e, $\hat{T}_B(\hat{\sigma}_X) = \max(|Y_{ij}|)$, and all coefficients are set to zero.

To summarize, BayesShrink performs soft thrsholding, with the data-driven, subband-dependent threshold.

$$\hat{T}_B(\hat{\sigma}_X) = \frac{\hat{\sigma}^2}{\hat{\sigma}_X}$$

3.3.3 Median Filter

Median filtering is a frequently used technology to remove the noise from radiographic images. A median filter deploys a small mask template, which is usually 3×3 or 5×5. The template operation may be calculated by either correlation or convolution operators. The median filter replaces a pixel's gray level with the median value of its neighborhood.

$$G'(x,y) = median\{G(x_1,y_1)|(x_1,y_1) \quad is \quad in \quad N(x.y)\} \tag{3.37}$$

where $N(x,y)$ is the immediate neighbors of the pixel (x,y).

The use of weighting kernels to average together pixels in a neighborhood is a convolution operation, which has a direct counterpart in frequency space image processing. It is a linear operation in which no information is lost from the original image. There are other processing operations that can be performed in neighborhoods in the spatial domain that also provide noise smoothing. But, these are not linear and do not utilize or preserve all of the original data.

The most widely used of these methods is based on ranking of the pixels in a neighborhood according to brightness. Then, for example, the median value in this ordered list can be used as the brightness value for the central pixel. As in the case of the kernel operation, it is used to produce a new image and only the original pixel values are used in the ranking for the neighborhood around each pixel.

The so-called median filter is an excellent rejecter of certain kinds of noise, for instance "shot" noise in which individual pixels are corrupted or missing from the image. If a pixel is accidentally changed to an extreme value, it will be eliminated from the image and replaced by a "reasonable" value, the median value in the neighborhood. A median filter is able to remove the noise and replace the bad pixels with reasonable values while causing a minimal distortion of degradation of the image. Of course, the computational effort required rises quickly with the number of values to be sorted, even using specialized methods which keep partial sets of the pixels ranked separately so that as the neighborhood is moved across the image, only a few additional pixel comparisons are needed.

Application of a median filter can also be used to reduce the type of random noise shown before in the context of averaging. There are two principal advantages to the median

filter as compared with multiplication by weight. First, the method does not reduce the brightness difference across steps, because the values available are only those present in the neighborhood region, not an average between those values. Second, median filtering does not shift boundaries as averaging may, depending on the relative magnitude of values present in the neighborhood. Overcoming these problems makes the median filter preferred both for visual examination and for measurement of images.

3.3.4 Comparison of Median Filter and Wavelet Thresholding

Because gray level changing is the main feature of the defects, we need choose a method that can remove the noise as well as least affect the sharpness of the edges.

Considering the character of each method and the X-ray image character, we compare following three denoising methods: Median filter; VisuShrink with universal threshold and BayesShrink with subband threshold.

First, we compare different methods based on the signal-to-noise ratio (SNR) of the denoised image. SNR is defined as

$$SNR = 10 \cdot \log_{10} \frac{var(Y)}{mse(Y, \hat{Y})} \tag{3.38}$$

where Y is the original clean image and $\hat{Y}$ is the denoised image.

The three denoised methods are applied to an noised image (refer to Figure 3.9). The performance of these methods is shown in Figure 3.10. Table 3.2 and Figure 3.10 show that the SNR of VisuShrink and the median filter is similar. BayesShrink improves the SNR much more than the other methods.

What is more, gray level changing is the main feature of the defects, the method which can remove the noise as well as least affect the sharpness of the edges is better. BayesShrink is more visually appealing and adapts to discontinuities in images better than other methods.

In order to compare these three methods more carefully, the gray level of the pixels along the line (refer to Figure 3.11) is shown in Figure 3.12. The gradient (first order derivative) of the gray level along the line (refer to Figure 3.11) is shown in Figure 3.13.

The median filter and VisuShrink can produce smoother images (from curve in Figure 3.12(a), Figure 3.12(b)), and change rate of the gray level and the gradient (the first deriva-

(a) 256×256 image of 'Lena'

(b) Noisy version of Lena ($SNR = 5.63db$)

Figure 3.9: Image with noise (Original image of 'Lena' from the Internet)

(a) Denoised using Median Filter ($SNR = 10.32db$)

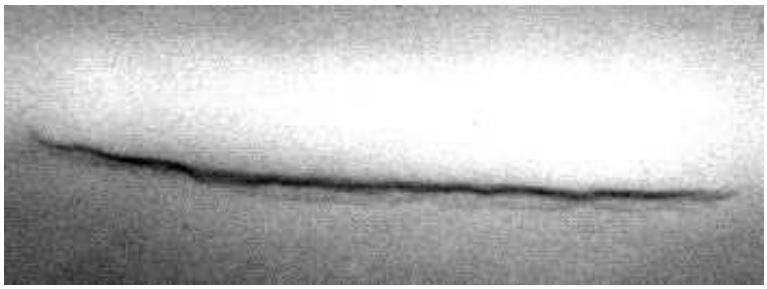
(b) Denoised using Visu-Shrink (SNR =10.33db)

(c) Denoised using Bayes-Shrink (SNR= 11.31db)

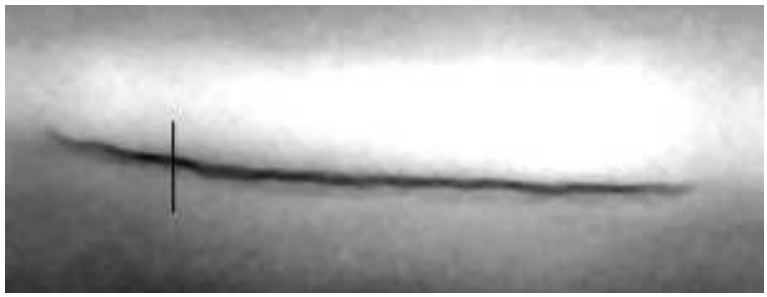
Figure 3.10: Performance of the various image enhancement methods on lena

Table 3.2: Signal-to-Noise Ratio of different methods (ref. Figure 3.10)

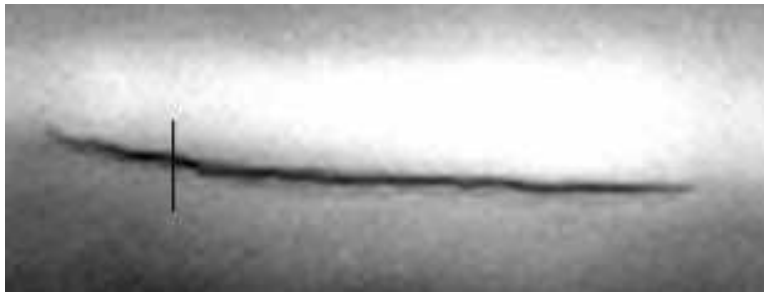
Method	SNR
Noise Lena	5.63db
Median Filter	10.32db
Universal thresholding (VisuShrink)	10.33db
Subband adaptive thresholding (BayesShrink)	11.31db



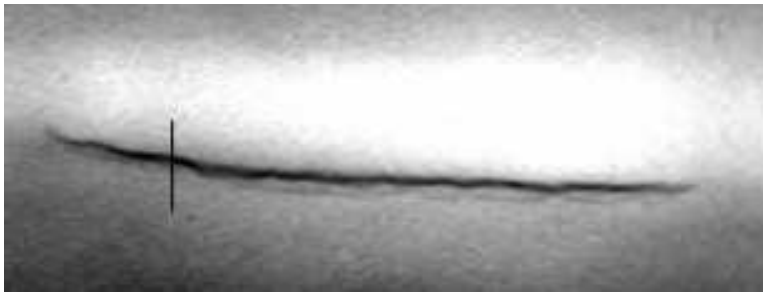
(a) X-ray image which includes crack



(b) Denoised using Median Filter

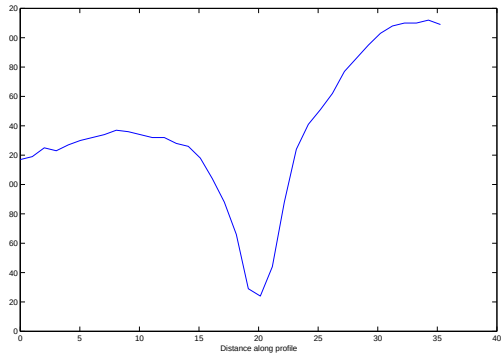


(c) Denoised using VisuShrink

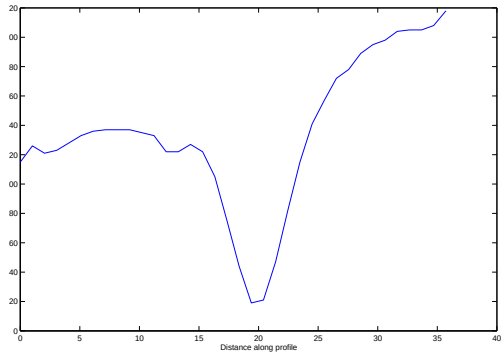


(d) Denoised using BayesShrink

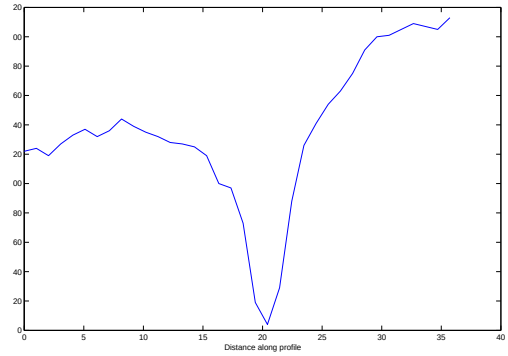
Figure 3.11: Performance of the various methods on an X-ray image with crack



(a) Median Filter

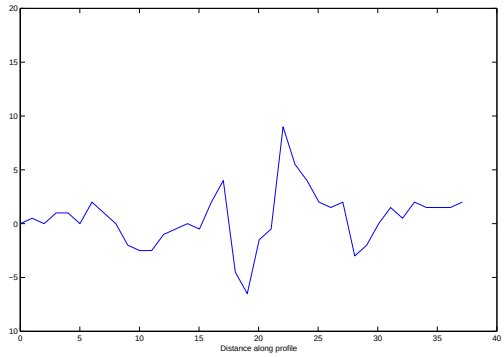


(b) VisuShrink

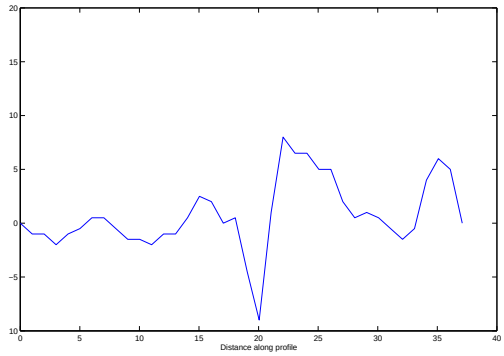


(c) BayesShrink

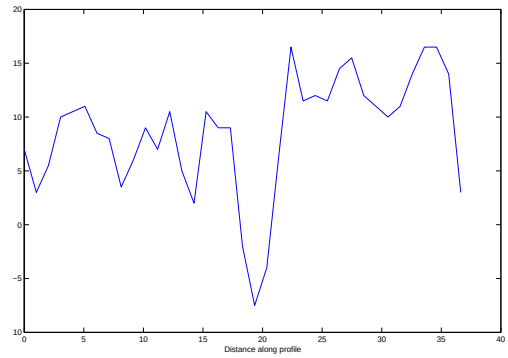
Figure 3.12: Gray level along the line shown in Figure 3.11



(a) Median Filter



(b) VisuShrink



(c) BayesShrink

Figure 3.13: Gradient of the gray level along the line shown in Figure 3.11

Table 3.3: The effect comparison of different methods (ref. Figure 3.12, Figure 3.13)

Method	Gray Level Change		Gradient Change	
	Change Region	Change Rate	Change Region	Change Rate
Median Filter	$25 \sim 215$	190	$-6.5 \sim +9$	15.5
VisuShrink	$25 \sim 215$	190	$-9 \sim +8$	17
BayesShrink	$5 \sim 215$	205	$-7 \sim +17$	24

tive) (from the curve in Figure 3.13(a), Figure 3.13(b)) is lower. Whereas, the BayesShrink acquires a little rugged image (from curve in Figure 3.12(c)), but the change rate of the gray level and the gradient (the first derivative) (from the curve in Figure 3.13(c)) is higher. Table 3.3 shows these in detail. In image processing, at the edge, the larger the change rate the better. In Table 3.3, it is obvious that the output of the BayesShrink has a larger change rate at the edge than other methods. Therefore, subband adaptive thresholding (BayesShrink) is used in this project.

3.4 Summary

In this study, morphological enhancement and adaptive wavelet thresholding are proposed to improve the quality of radiographic images. The comparative analysis between the proposed methods and currently frequently used methods has showed the effectiveness of these methods. They show promising results on radiographic images. The morphological enhancement can not only improve the local contrast of the radiographic images but also reduce the noise produced in homogeneous areas. The adaptive wavelet thresholding technique can remove the image noise while keeping the sharpness of defects' edges well. Therefore, morphological enhancement and adaptive wavelet thresholding can greatly enhance radiographic image and they will be helpful for defect recognition.

Chapter 4

Segmentation of Radiographic Images

4.1 Introduction

Segmentation is a fundamental stage in the automatic welding defects detection system. Its application can extract potential defects which are subsequently classified. In general, automated segmentation is often considered as the most difficult task in image processing [69].

Image segmentation is a process in which a region or an object of interest is extracted from the rest of the image. It provides the most fundamental way to extract useful features for further image analysis and scene interpretation. Mathematically, by using the definitions in Equation 4.1, segmentation can be expressed in the following way:

$$\begin{aligned} &\cup_{l=1}^{N} R_l = I \\ &R_l \cap R_m = \emptyset \quad \forall l \neq m \\ &P(R_l) = True \quad for\ l = 1, 2, \ldots, N \\ &P(R_l \cup R_m) = False \quad \forall l \neq m \\ &\cup : Union \\ &\cap : Intersection \end{aligned} \tag{4.1}$$

Where, I is the set of all image pixels, $P(.)$ is a uniformity predicate, N is the number of segmented regions and $\{R_1, R_2, R_3, \ldots, R_N\}$ represents the segmented regions.

Segmentation, separation of defects from the background, is important for defect recognition. In the radiographic image, the defects are all quite small and their positions are

random. Some defects are not as dark as some part of the background. So, it is difficult to segment the radiographic image using conventional segmentation methods. In this study, two new segmentation methods are proposed to segment radiographic images. One is multiscale edge detection based on wavelet transform. The other is multi-level thresholding based on maximum fuzzy entropy and genetic algorithm.

4.2 Conventional Segmentation Methods

In this section, we investigate defect segmentation using conventional image segmentation methods. The segmentation problem can be approached with different methods, which generally can be classified into three main methods [70] [71] [72] [73], namely,

1. Threshold techniques
2. Edge-based methods
3. Region-based methods

4.2.1 Thresholding techniques

Thresholding is a particularly useful region-approach for scenes containing solid objects resting upon a contrasting background [69]. When using a threshold rule for image segmentation, one assigns all pixels at or above the threshold gray level to the object. All pixels with gray level below the threshold fall outside the object. The boundary is then that set of interior points, each of which has at least one neighbor outside the object.

$$f_T(x,y) = \begin{cases} 1, & \text{if} \quad f(x,y) \geq T \\ 0, & \text{if} \quad f(x,y) < T \end{cases} \tag{4.2}$$

where T is the threshold.

Thresholding works well if the objects of interest have uniform interior gray level and rest upon a background of different, but uniform, gray level [69].

Thresholds are either global or local. The global thresholding uses a fixed threshold for all pixels in the image and therefore works only if the intensity histogram of the input image

contains neatly separated peaks corresponding to the desired subject(s) and background(s). Figure 4.1 shows some typical histograms along with suitable choices of threshold. If distinct peak does not exist, then it is unlikely that simple thresholding will produce a good segmentation. In this case, adaptive thresholding may be a better choice. Local thresholding, on the other hand, selects an individual threshold for each pixel based on the range of intensity values in its local neighborhood. This allows for thresholding of an image whose global intensity histogram does not contain distinctive peaks. Adaptive thresholding typically takes a greyscale or color image as input and, in the simplest implementation, outputs a binary image representing the segmentation. For each pixel in the image, a threshold has to be calculated. If the pixel value is below the threshold it is set to be the background value, otherwise it assumes the foreground value.

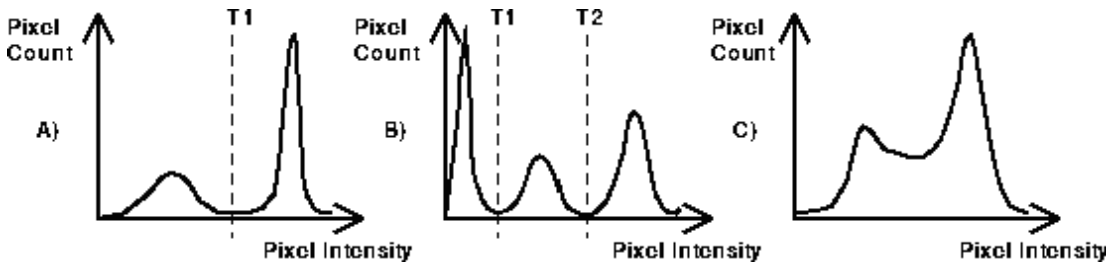
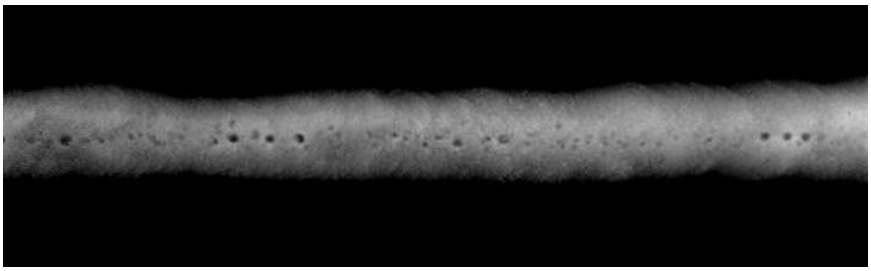
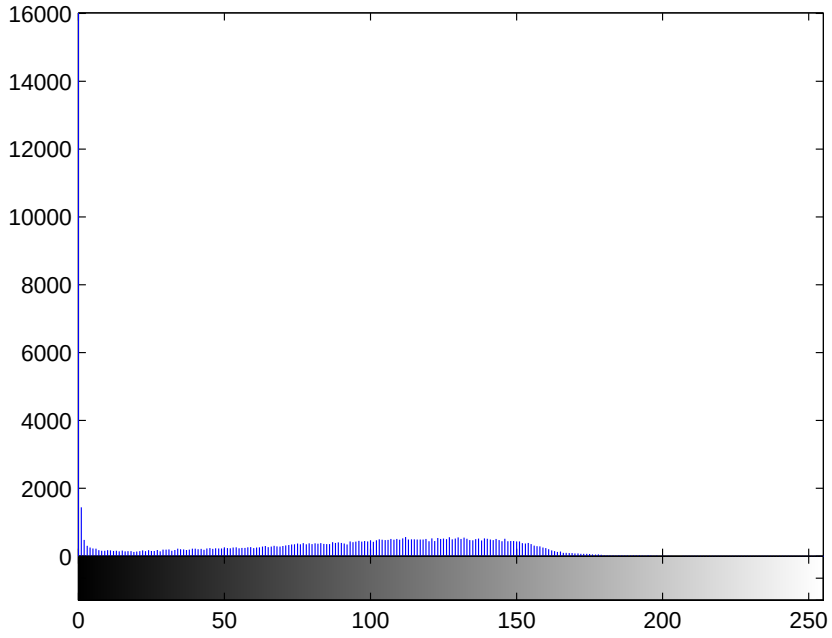


Figure 4.1: A) shows a classic bi-modal intensity distribution. This image can be successfully segmented using a single threshold T1. B) is slightly more complicated. Here we suppose the central peak represents the objects we are interested in and so threshold segmentation requires two thresholds: T1 and T2. In C), the two peaks of a bi-modal distribution have run together and so it is almost certainly not possible to successfully segment this image using a single global threshold

The histogram can indicate an appropriate threshold for the segmentation of the desired object. For images with distinct objects and background, the histogram will be bimodal. Thus, the global threshold level can be chosen as the gray level that corresponds to the valley of the histogram. From the Figure 4.2 we can see that it is difficult to choose a global threshold to extract the defects from the background for the radiographic image. We need to find an optimal gray level threshold to separate defects from the background. Optimal thresholding methods depend on the maximization or minimization of a merit or performance function. If assumptions are made about the shape of the histogram, a model can be comprised as the sum of several component distributions. However, some defects still can not be segmented using an optimal threshold as shown in Figure 4.3.

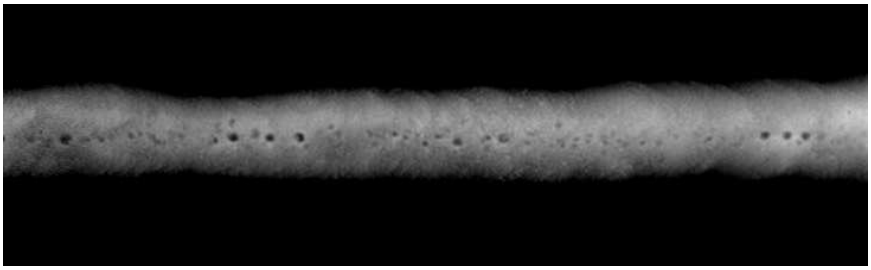


(a) Image after preprocessing



(b) Histogram of the image

Figure 4.2: Histogram of a radiographic image with porosity defects



(a) Image after preprocessing

(b) segmentation result using a optimal threshold

Figure 4.3: Segment the radiographic image using the optimal threshold

4.2.2 Edge-based methods

Edge detection is one of the most commonly used operations in image analysis, and there are probably more algorithms in the literature for detecting edges than for any other tasks. The reason for this is that edges form the outline of an object. An edge is the boundary between an object and the background, and it also indicates the boundary between overlapping objects. This means that if the edges in an image can be identified accurately, all of the objects can be located and basic properties such as area, perimeter, and shape can be measured.

Edge-based methods are based on the postulate that the pixel values change rapidly at the edge between two regions. The detection of the defects is based on gray level changing. So the edges contain the most information. There are many ways to perform edge detection. The most common ways may be grouped into two categories: Laplacian and gradient.

The Laplacian method searches for zero crossings in the second derivative of the image to find the edges. The Laplacian is a scalar second-derivative operator for functions of two dimensions. It is defined as

$$\nabla^2 f(x,y) = \frac{\partial^2}{\partial x^2} f(x,y) + \frac{\partial^2}{\partial y^2} f(x,y) \tag{4.3}$$

It is commonly approximated digitally by either of the convolution kernels shown in Figure 4.4.

$$\begin{array}{ccc} 0 & -1 & 0 \\ -1 & +4 & -1 \\ 0 & -1 & 0 \end{array} \qquad \begin{array}{ccc} -1 & -1 & -1 \\ -1 & +8 & -1 \\ -1 & -1 & -1 \end{array}$$

Figure 4.4: Laplacian convolution kernels

Since it is a second derivative, the Laplacian will produce an abrupt zero- crossing at an edge. The Laplacian is a linear, shift-invariant operator, and its transfer function is zero at the origin of frequency space. Thus, a Laplacian-filtered image will have zero mean gray level.

If a noise-free image has sharp edges, the Laplacian can find them. The binary image that results from thresholding a Laplacian-filtered image at zero gray level will produce closed, connected contours when interior points are eliminated. The presence of noise, however, imposes a requirement for low pass filtering prior to using the Laplacian.

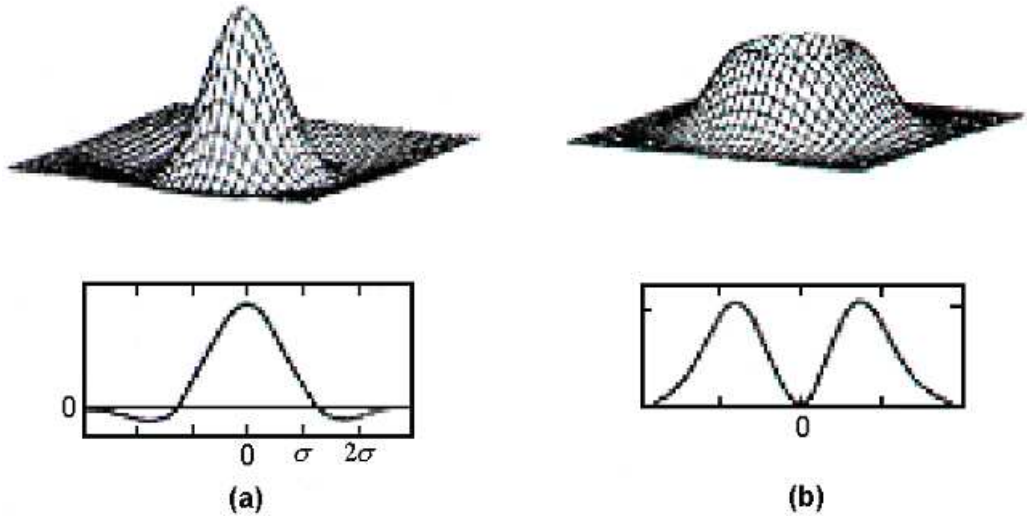


Figure 4.5: The Laplacian-of-Gaussian filter(a) Impulse response; (b) Transfer function

A Gaussian low pass filter is a good choice for this pre-smoothing. Since convolution is associative, the Laplacian and Gaussian impulse responses can be combined into a single Laplacian of Gaussian (LoG) kernel:

$$-\nabla^2 \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\pi\sigma^2}} = \frac{1}{\pi\sigma^4}(1 - \frac{x^2+y^2}{2\pi\sigma^2}) e^{-\frac{x^2+y^2}{2\pi\sigma^2}} \tag{4.4}$$

This impulse response is separable in x and y and thus can be implemented efficiently. It has the shape of the impulse response of a general band pass filter, namely a positive peak in a negative dish (Figure 4.5). The parameter σ controls the width of the central peak and, thus, the amount of smoothing.

The gradient method detects the edges by looking for the maximum and minimum in the first derivative of the image, including Robert, Prewitt and Sobel edge operators.

Robert Edge Operator

One local differential operator for finding edges is the Robert edge detector [74]. It consists of two derivatives at right angles, estimated digitally as the difference between pairs of pixels. These two differences are the first derivative of brightness in two perpendicular directions and each orients at 45 degrees to the principal grid of points. Being first derivatives, they are not directly suitable for use as an edge detector. But if they are combined as $\sqrt{D_1^2 + D_2^2}$, then it is an even power, and also a single value that combines both directions so as to give a uniform response to an edge in any direction.

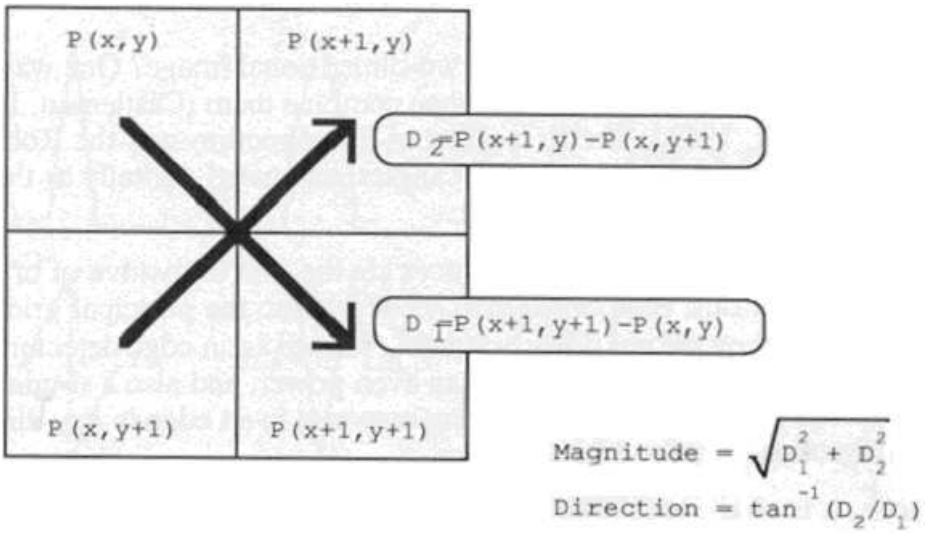


Figure 4.6: Robert edge operator

Furthermore, it is also possible to determine the orientation of the local edge. The direction of maximum gradient in brightness is given by

$$\tan^{-1}(D_2/D_1) - \pi/4 \tag{4.5}$$

and the edge is at right angles to this direction.

This method suffers from several practical drawbacks. First, the square root function is a demanding one for the computer (recall that this operation must be performed for every pixel in the image). This is sometimes overcome by using various approximations or shortcuts. For instance, it is not really essential to use the square root to determine the function's maxima, since the sum of squares will do as well. An even more drastic simplification is to use the maximum value of the two derivatives, which avoids all of the arithmetic. However, this makes the detector significantly more sensitive to edges in the 45 degree directions than to those aligned with the pixel grid.

Prewitt Edge Operator

The two convolution kernels shown in Figure 4.7 form the Prewitt edge operator [34]. Each point in the image is convolved with both kernels, and the maximum determines the output. The Prewitt operator likewise produces an edge magnitude image.

$$\begin{matrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ +1 & +1 & +1 \end{matrix} \qquad \begin{matrix} +1 & 0 & -1 \\ +1 & 0 & -1 \\ +1 & 0 & -1 \end{matrix}$$

Figure 4.7: Prewitt edge operator

Sobel Edge Operator

Sobel's algorithm is a nonlinear edge detection method [72]. It is an example of an algorithm with complex mathematical underpinnings that is rather simple to implement. The two convolution kernels shown in Figure 4.8 form the Sobel edge operator. As with the Prewitt edge operator, each point in the image is convolved with both kernels. One kernel responds maximally to a generally vertical edge and the other to a horizontal edge. The maximum value of the two convolutions is taken as the output value for that pixel. The resultant image is an edge magnitude image.

$$\begin{matrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{matrix} \qquad \begin{matrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{matrix}$$

Figure 4.8: Sobel edge operator

Canny Edge Detector

The Canny edge detector [75] is an edge detection method that is optimal for step edges corrupted by white noise. Canny used three criteria to design his edge detector: 1. Low error rate. Error includes misdetection and false alarm. 2. The edge points are well localized. 3. Single response to an edge. Unfortunately, there is always a trade-off between detection and localization for conventional gradient-based edge detectors, that is, the larger the mask, the more correct edges are detected, while larger edge location deviation results. To balance these two criteria, Canny designed a numerical optimization model to calculate the parameters of the detector. While in practice such optimization is hard to achieve, he approximated the optimal detector by the first derivative of a Gaussian, which provides a significant improvement of computation efficiency.

To quantify these criteria, the following functions are defined:

$$SNR(f) = \frac{A\left|\int_{-W}^{0} f(x)dx\right|}{n_0\sqrt{\int_{-W}^{+W} f^2(x)dx}} \tag{4.6}$$

$$Localization(f) = \frac{A\,|f'(0)|}{n_0\sqrt{\int_{-W}^{+W} f'^2(x)dx}} \tag{4.7}$$

where A is the amplitude of the signal, n_0^2 is the variance of noise and $f(x)$ is the filter for edge detection. SNR defines the signal-to-noise ratio and $Localization$ defines the localization of the filter $f(x)$.

Suppose forming a spatially scaled filter f_s from f, where $f_s(x) = f(x/s)$. we get the following "uncertainty principle":

$$SNR(f_s) = \sqrt{s}SNR(f) \tag{4.8}$$

$$Localization(f_s) = \frac{1}{\sqrt{s}} Localization(f) \tag{4.9}$$

That is, increasing the filter size increases the signal-to-noise ratio but also decreases the localization by the same factor. This suggests maximizing the product of the two. So the object function is defined as:

$$J(f) = \frac{\left|\int_{-W}^{0} f(x)dx\right|}{\sqrt{\int_{-W}^{+W} f^2(x)dx}} \frac{|f'(0)|}{\sqrt{\int_{-W}^{+W} f'^2(x)dx}} \tag{4.10}$$

The optimal filter that is derived from these requirements can be approximated with the first derivative of the Gaussian filter,

$$f(x) = -\frac{x}{\sigma^2} e^{-\frac{x}{2\sigma^2}} \tag{4.11}$$

The choice of the standard deviation for the Gaussian filter, σ, depends on the size, or scale, of the objects contained in the image. For images with multiple size objects, or unknown size one approach is to use Canny detectors with different σ values. The outputs of the different Canny filters are combined to form the final edge image.

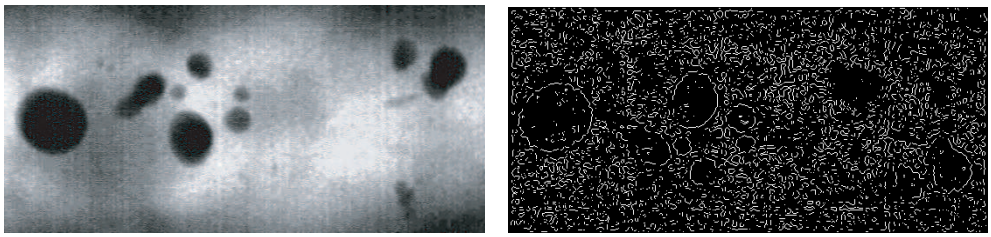
The algorithm of the Canny detector can be described by three steps:

1. Calculate a mask of the first derivative of a Gaussian as the approximation of the optimal mask, filter the image with this mask.

2. Non-maximum suppression in a direction perpendicular to the edge is applied, to retain maxima in the image gradient.

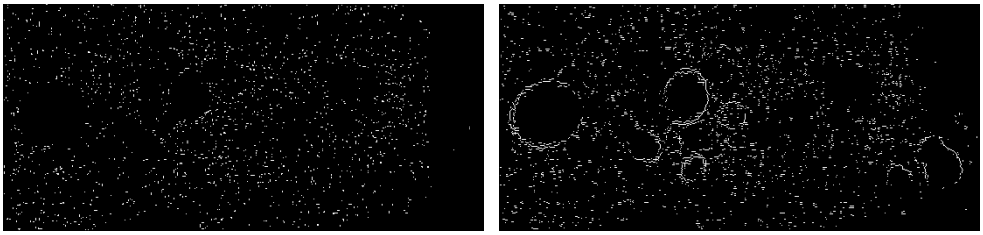
3. Weak edges are removed using thresholding. The thresholding is applied with hysteresis, or double threshold. The high threshold is used to find 'seeds' for strong edges. These seeds are grown into an edge in both directions as long as possible, so long as you can do this without the edge strength falling below the low threshold. This reduces streaking in the output edges.

The performance of these edge-based methods is shown in Figure 4.16. It is clear from Figure 4.16(b) that LoG misdetects some edges and is noise sensitive. The reason is that



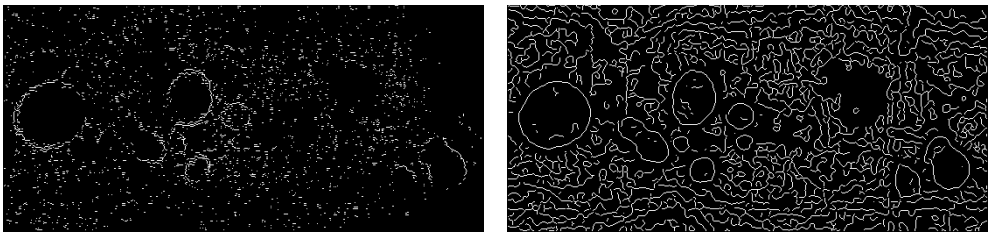
(a) Image after preprocessing

(b) LoG edge operator



(c) Roberts edge operator

(d) Perwitt edge operator



(e) Sobel edge operator

(f) Canny edge operator

Figure 4.9: Performance of conventional edge detection methods

the Laplacian is a good high-pass filter, but not a particularly good tool for demarcating edges [76]. In most cases, boundaries or edges of features or regions appear at least locally as a step in brightness, sometimes spread over several pixels. The Laplacian gives a larger response to a line than to a step, and to a point than to a line. In an image that contains noise, typically present as points varying in brightness due to counting statistics, detector characteristics, etc., the Laplacian will show such points more strongly than the edges or boundaries that are of interest. So, LoG is not a appropriate edge detector for radiographic image.

The result shows that the Robert edge operator can not detect most of the edges of an X-ray image. In fact, the Robert edge operator calculates the edge parameter between pixels rather than aligns with the pixel grid, and by using the difference between pairs of pixels. So, it is rather sensitive to any noise present in the image. The Perwitt and Sobel edge

operators have better results than the Robert edge operator, however, they lost some defect edges and are sensitive to the noise. So, the gradient detectors are not good choice for radiographic images.

The canny edge detector is a fast robust detector for detection edges in X-ray image, it can detect most all of the edges of defects. However, it can not distinguish between noise and edges well. Adjusting the high and low thresholds, the mask size and the Gaussian parameter is also troublesome.

4.2.3 Region-based methods

Region-based segmentation looks for uniformity within a sub-region, based on a desired property, e.g. intensity, color, and texture. Clustering techniques encountered in pattern classification literature have similar objectives and can be applied for image segmentation [77].

Region-based methods rely on the postulate that neighboring pixels within the one region give similar valleys. This leads to the class of algorithms known as region growing of which the ”split and merge” technique [78] is probably the best known. The general procedure is to compare one pixel to its neighbors. If a criterion of homogeneity is satisfied, the pixel is said to belong to the same class as one or more of its neighbors. Various homogeneity criteria have been investigated for region growing.

One of the most current regions growing methods is single linkage. It considers pixels as vertices in a graph. Neighboring pixels having similar properties are joined by an arc. The image segments are maximal sets of pixels belonging to the same connected component. The simplest single linkage scheme is similarly defined in the following way. Pixels p and p' are considered as related (we write $p\ R\ p'$) if their gray levels $f(p)$ and $f(p')$ are not too distant. More precisely,

$$p\ R\ p' \Leftrightarrow |f\ (p) - f\ (p')| \leq \alpha \tag{4.12}$$

where α is a fixed threshold.

This technique is attractive due to its simplicity, but the choice of α necessitates knowledge about the gray levels of the processed image to obtain good results.

On its own, R is not an equivalence relation. But the following relation is obtained by transitivity saturation:

$$p\,R\,p' \Longleftrightarrow there\ exists\ a\ path\ p = p_1, p_2, \ldots, p_n = p'\ with\ \forall k\ p_k\ R\ p_{k+1} \tag{4.13}$$

is an equivalence relation, giving a classification which does not depend on the seed's position in the region. But there is a drawback: $p\,R\,p'$ does not mean $|f(p) - f(p')| \leq \alpha$. Thus, this method can lead to a disastrous chaining effect especially for images with low contrast shape boundaries or images with a lighting shift.

In [79], an improvement in simple region growing had been introduced. They replaced the fixed threshold α by a moving threshold $\alpha(p)$ generated by the knowledge of the gray levels in a neighborhood $V(p)$ of pixel p in the following way:

$$\alpha(p) = \underset{q \in V(p)}{Min} \{|f(p) - f(q)|\} \tag{4.14}$$

Then, two pixels belong to the same region if

$$|f(p) - f(p')| \leq Max\{\alpha(p), \alpha(p')\} \tag{4.15}$$

In this way, the equivalence relation properties are preserved but the chaining effect can still occur. In order to prevent overstepping, they consider different solutions, but with the common drawback that the obtained regions are generally too small.

In global (as opposed to single) linkage region growing, pairs of neighboring pixels are not compared on the basis of similarity. This method is based on the comparison of the value of the current pixel with the mean value of an already existing neighboring region. According to its value and to the mean value of the existing region, the pixel is merged into one of the regions whose mean is updated. In global linkage region growing, pixels are assigned to a neighboring region until this region can no longer grow.

Actually, the choice of the homogeneity criterion is critical for even moderate success [80], [81], and in all instances the results are upset by noise.

In [82], a method known as "seeded region growing" is presented, which is based on the conventional region growing postulate of similarity of pixels within region, but whose mechanism is closer to that of the watershed method [69]. Instead of tuning homogeneity

parameters as in conventional region growing, seeded region growing is controlled by choosing a (usually small) number of pixels, known as seeds.

Another common approach in region-based segmentation is characterizing statistical uniformity of sub-regions using parametric models, so called "statistical estimation". With this approach, two sub-regions are considered to be uniform, and consequently merged, if they can be represented by a single instance of the model, i.e. if they have common parameter values within a threshold. In practice, the parameters of a sub-region cannot be observed directly but can only be inferred from the observed data and the knowledge of the imaging process. In statistical approaches, this inference is often made using Bayess rule [83] and the conditional PDF $p(I(x,y)|\theta_m)$, which presents the conditional probability that certain data $I(x,y)$ (or statistics derived from the data) will be observed, given that sub-region m has the parameter values of θ_m. In typical statistical region merging algorithms [84], stochastic estimates in the parameter space are obtained for different sub-regions, and merging decisions are based on the similarity of these parameters.

A limitation of most estimation-based segmentation methods is that they do not explicitly represent the uncertainty in the estimated parameter values and, therefore, are prone to error when parameter estimates are poor. A Bayesian probability of homogeneity directly exploits all of the information contained in the statistical image models, instead of estimating parameter values [85]. The probability of homogeneity is based on the ability to formulate a prior probability density on the parameter space, and measures homogeneity by taking the expectation of the data likelihood over a posterior parameter space.

Image segmentation is often approached by "edge-preserving" smoothing operations as well as the partitioning operation. Edge-preserving smoothing techniques can be classified roughly two approaches [86]: Markov random field (MRF) including energy-based methods [87] and diffusion-based methods [88]. Both approaches show similar restoration characteristics because the diffusion-based methods can be viewed as an energy-based method that uses only the prior energy term at a given temperature [83]. Snyder et al. [89] proposed an edge-preserving smoothing method for image segmentation based on the technology called mean field annealing (MFA) [83]. MFA is an energy-based method for finding the minimum of complex functions which typically have many minima [90]. For the image segmentation problem, a proper energy function is defined intending to keep the edges and to smooth the rest of areas in the image. The segmentation is performed by minimizing the energy function using MFA.

4.2.4 Watershed Transform

The watershed transform could be considered as a region based segmentation approach. It is a popular segmentation method coming from the field of mathematical morphology. In mathematical morphology, an image is usually interpreted as a topographical surface, and its gray level is considered an altitude. Figure 4.10 illustrates the watershed transformation performing on a gray-level image. If a drop of water falls on the contour, then it will flow along a descending path to a local minimum. A collection of pixels on the contour is defined as the watershed line. The two regions separated by the watershed line are called the catchment basins. Each catchment basin is associated with a local minimum. In an image, the contour lines appear in the places where the gray level changes sharply in comparison with the other pixels of the neighborhood. To ensure the watershed lines will follow the contour lines in the image, the watershed algorithm is usually applied to the gradient image. This gradient-based watershed transformation achieves a useful result for image segmentation. The watershed transformation has several advantages such as closed contours, non-intersected regions, each region containing a single local extrema (usually minimum), and the union of all regions and watersheds being the original whole surface.

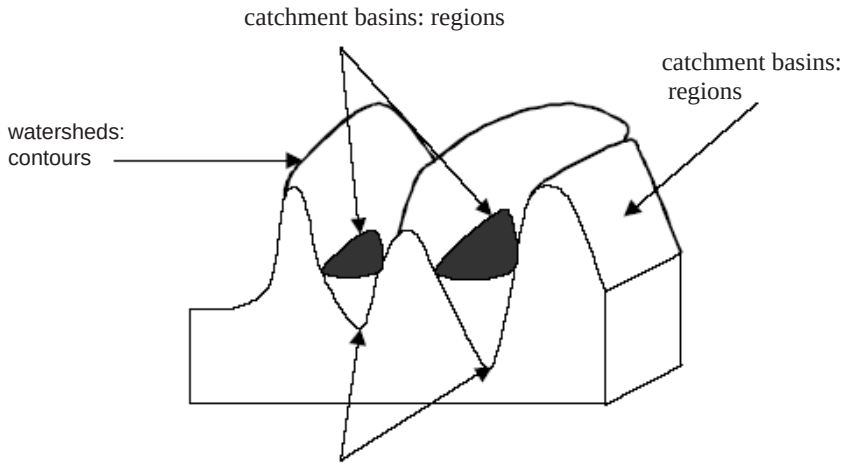
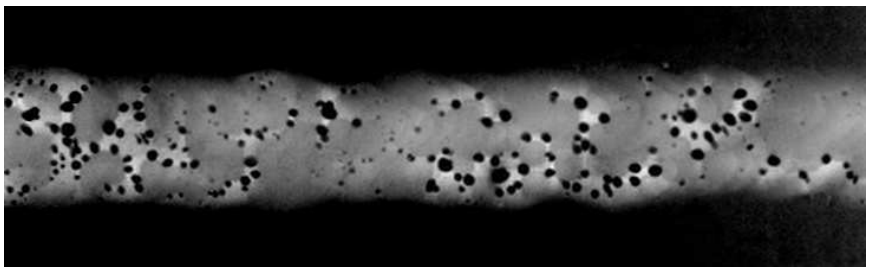


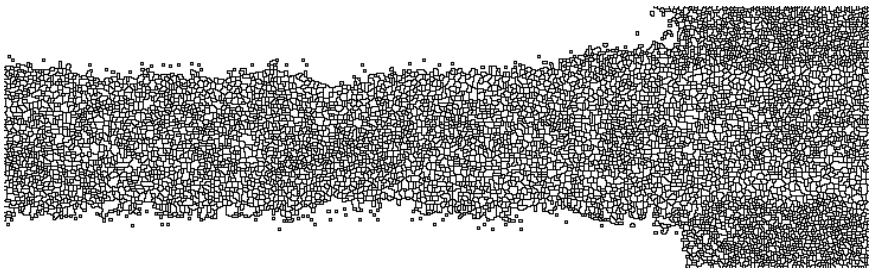
Figure 4.10: Watershed transformation on gray-level image. Each catchment basin contains single local minima. The bold line indicates the watershed line that corresponds to the contour of the image.

The watershed transformation is applied on the gradient of the gray level image. The gradient image is interpreted as a topological surface. By using the watershed transformation on the gradient image, the image is segmented into homogeneous regions, which

are low in contrast and with low gradient. This method can provide closed contours. However, the watershed algorithm is highly sensitive to gradient noise, it results in the over-segmentation with too many regions in the image such as Figure 4.11.



(a) Image after preprocessing



(b) Image after applying watershed transform

Figure 4.11: Over-segmentation

An approach used to control over-segmentation is based on the concept of markers. A marker is a connected component belonging to an image. The regularization of the gradient is accomplished by imposing markers as the regional minima of the gradient and then suppressing all the other minima by way of a morphological reconstruction operation. Hence, one must provide internal markers, one for each calcification and also a background or external marker. The markers are obtained in a semi-automatic way. The internal markers are single pixels placed manually inside each calculation with a mouse. The external markers are obtained from the internal markers as a result of the watershed transformation of the inverse of the input mammographic section using the internal markers as the marker set. Unfortunately, the markers selection and extraction are not so easy for welding defects detection. The defects to be detected are so complex and so varied in shape, gray level and size that it is very hard to find reliable algorithms enabling their extraction.

4.2.5 Summary

Each of these methods has its advantages and disadvantages for welding defects segmentation. If an optimal threshold, which can extract defects from background, is found, thresholding method is more appropriate for defects segmentation. The edge-based segmentation techniques are computationally fast and do not require a priori information about the image content. A common problem of edge-based segmentation is that often the edges do not enclose the object completely. Region-based approaches are generally less sensitive to noise, and usually produce more reasonable segmentation results as they rely on global properties rather than local properties, but their implementation complexity and computational cost can be often quite large. The watershed transform can provide closed contours. However, the watershed algorithm is highly sensitive to gradient noise, it results in the over-segmentation. Marker-controlled watershed segmentation can control over-segmentation. However, the selection and extraction of markers are not so easy for welding defects detection. The defects to be detected are so complex and so varied in shape, gray level and size that it is very hard to find reliable algorithms enabling their extraction. For images which contain large areas of uniformity, region-growing processing is probably a better choice. For a nonuniform radiographic image, it is not a good choice. Since it is difficult to segment defects using these conventional methods, we propose two segmentation methods: a multiscale edge detection algorithm based on the wavelet transform and a multi-level thresholding algorithm based on maximum fuzzy entropy and genetic algorithm.

4.3 Multiscale Edge Detection Algorithm Based On Wavelet Transform

Since it is difficult to detect edges using conventional edge detectors, we propose a new multiscale edge detection algorithm based on wavelet transform (MEWT). According to wavelet multiscale character, after obtaining edge information at different scales we integrate the coefficients of the wavelet transforms on a series of scales to look for the best scale where the edges are well discriminated from noises to extract edge features.

4.3.1 Continuous Wavelet Transforms

The continuous wavelet transform (CWT) was first developed by Grossmann and Morlet [57]. Let $\psi(x) \in \boldsymbol{L}^2(\boldsymbol{R})$ be an admissible mother wavelet that satisfies

$$\int_{-\infty}^{+\infty} \psi(x)dx = 0 \tag{4.16}$$

Let $\psi_s(x)$ be a factor $s \in \boldsymbol{R}^+$,

$$\psi_s(x) = \frac{1}{s}\psi(\frac{x}{s}) \tag{4.17}$$

The CWT of a function $f(x)$ at scale s and position x in scale-space plane is defined by the convolution product

$$\boldsymbol{W}: \quad W_s f(x) = \psi_s * f(x) \tag{4.18}$$

The CWT is an isometry, i.e.,

$$\|f(x)\|^2 = \int_0^{+\infty} \|W_s f(x)\|^2 \frac{ds}{s} \tag{4.19}$$

The inverse CWT can then obtained by:

$$\boldsymbol{W}^{-1}: \quad f(x) = \int_0^{+\infty} \tilde{\psi}_s * W_s f(x) \frac{ds}{s} \tag{4.20}$$

where $\tilde{\psi}(x) = \psi(-x)$.

An arbitrary two variable function $g(s,x) \in \boldsymbol{L}^2(\boldsymbol{R}^2)$ is not necessarily the CWT of some function. For $g(s,x)$ to be in the range of $\boldsymbol{W}$, it must satisfy the reproducing equation,

$$g(s,x) = \int_0^{+\infty} ds' \int_{-\infty}^{+\infty} K(s,s';x,x')g(s',x')dx' \tag{4.21}$$

where

$$K(s,s';x,x') = \int_{-\infty}^{+\infty} \tilde{\psi}_s(u-x)\psi_{s'}(x'-u)du \tag{4.22}$$

is the wavelet reproducing kernel (WRK).

For practical applications, the scale-space must be discretized. The scale parameter s is often discretized to a dyadic sequence, $\{2^j\}_{j\in Z}$. A wavelet is a function $\psi(x)$ whose average is zero. We denote by $\psi_{2^j}(x)$ the dilation of $\psi(x)$ by a factor 2^j

$$\psi_{2^j}(x) = \frac{1}{2^j}\psi(\frac{x}{2^j}) \tag{4.23}$$

The wavelet transform of $f(x)$ at scale 2^j and at the position x is defined by the convolution product

$$\boldsymbol{W}: \quad W_{2^j} f(x) = f * \psi_{2^j}(x) \tag{4.24}$$

The Fourier transform of $W_{2^j} f(x)$ is

$$FW_{2^j} f(\omega) = F(\omega)\Psi(2^j\omega) \tag{4.25}$$

Assume that there exist two strictly positive constants A_1 and B_1 such that

$$\forall\omega \in \boldsymbol{R}, A_1 \leq \sum_{j=-\infty}^{+\infty} \left|\Psi(2^j\omega)\right|^2 \leq B_1 \tag{4.26}$$

When it is ensured the whole frequency axis is covered by dilations of $\Psi(\omega)$ by $(2^j)_{j\in Z}$, then $F(\omega)$, and thus, $f(x)$ can be recovered from the dyadic wavelet transform defined by (4.23). The reconstructing wavelet $\chi(x)$ is any function whose Fourier transform satisfies

$$\sum_{j=-\infty}^{+\infty} \Psi(2^j\omega)\Xi(2^j\omega) = 1 \tag{4.27}$$

If property (4.26) is valid, there exist an infinite number of functions $\Xi(x)$ that satisfy (4.27). The function $f(x)$ is recovered from its dyadic wavelet transform with summation

$$\boldsymbol{W}^{-1} \quad f(x) = \sum_{j=-\infty}^{+\infty} W_{2^j} f * \chi_{2^j}(x) \tag{4.28}$$

Similar to the continuous case, for a sequence $g_j(x)_{j\in Z}$ to be the dyadic wavelet transform of some function, $g_j(x)$ must satisfy the reproducing equation:

$$\forall_j \in \boldsymbol{Z}, g_j(x) = \sum_{j=-\infty}^{+\infty} g_j * K_{i,j}(x) \tag{4.29}$$

with the WRK defined by,

$$K_{i,j}(x) = \tilde{\psi}_{2^j} * \psi_{2^j}(x) \tag{4.30}$$

4.3.2 Multiscale Edge Detection Based on Wavelets

In order to detect the contours of small structures as well as the boundaries of large objects, several researcher in computer vision have introduced the concept of multiscale edge detection [91], [92], [93]. The wavelet transform is closely related to multiscale edge detection and can provide a deeper understanding of these algorithms [64].

A remarkable property of the wavelet transform is its ability to characterize the local regularity of functions. For an image $f(x, y)$, its edges correspond to singularities of $f(x, y)$, and thus are related to the local maxima of the wavelet transform modulus. Therefore, the wavelet transform is an effective method for edge detection.

We assume that the smoothing function $\theta(x)$ is twice differentiable and define, respectively, $\psi^a(x)$ and $\psi^b(x)$ as the first- and second-order derivatives of $\theta(x)$

$$\psi^a(x) = \frac{d\theta(x)}{dx} \ \ and \ \ \psi^b(x) = \frac{d^2\theta(x)}{dx} \tag{4.31}$$

By the definition in (4.31), the functions $\psi^a(x)$ and $\psi^b(x)$ can be considered to be wavelets because their integral is equal to 0.

$$\int_{-\infty}^{\infty} \psi^a(x)dx = 0 \ \ and \ \ \int_{-\infty}^{\infty} \psi^b(x)dx = 0 \tag{4.32}$$

A wavelet transform is computed by convolving the signal with a dilated wavelet. The wavelet transform of $f(x)$ at the scale s and position x, computed with respect to the wavelet $\psi^a(x)$, is defined by

$$W_s^a f(x) = f * \psi^a(x) \tag{4.33}$$

The wavelet transform of $f(x)$ with respect to $\psi^b(x)$ is

$$W_s^b f(x) = f * \psi^b(x) \tag{4.34}$$

It follows that

$$W_s^a f(x) = f * (s\frac{d\theta_s}{dx})(x) = s\frac{d}{dx}(f * \theta_s)(x), \ and \tag{4.35}$$

$$W_s^b f(x) = f * (s^2\frac{d\theta_s}{dx^2})(x) = s^2\frac{d^2}{dx^2}(f * \theta_s)(x) \tag{4.36}$$

Therefore, the wavelet transforms $W_s^a f(x)$ and $W_s^b f(x)$ are, respectively, the first and second derivative of the signal smoothed at the scale s. The local extrema of $W_s^a f(x)$ thus correspond to the zeros of $W_s^b f(x)$ and to the inflection points of $f * \theta_s(x)$.

Let $\psi^1(x,y) = \frac{\partial\theta(x,y)}{\partial x}$ and $\psi^2(x,y) = \frac{\partial\theta(x,y)}{\partial y}$. So, $\psi_s^1(x,y) = \frac{1}{s^2}\psi^1(\frac{x}{s}, \frac{y}{s})$ and $\psi_s^2(x,y) = \frac{1}{s^2}\psi^2(\frac{x}{s}, \frac{y}{s})$. Let $f(x,y) \in \mathbf{L^2(R)}$. The wavelet transform of an image $f(x,y)$ at the scale s has two components defined by

$$W_s^1 f(x,y) = f * \psi_s^1(x,y) \;\; and \;\; W_s^2 f(x,y) = f * \psi_s^2(x,y) \tag{4.37}$$

It is straight forward to show that

$$\begin{pmatrix} W_s^1 f(x,y) \\ W_s^2 f(x,y) \end{pmatrix} = s \begin{pmatrix} \frac{\partial}{\partial x}(f * \theta_s)(x,y) \\ \frac{\partial}{\partial y}(f * \theta_s)(x,y) \end{pmatrix} = s\overrightarrow{\nabla}(f * \theta_s)(x,y) \tag{4.38}$$

Hence, edge points can be located from the two components, $W_s^1 f(x,y)$ and $W_s^2 f(x,y)$ of the wavelet transform.

Here, we choose Gaussian as the smoothing function for wavelet. i.e.

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.39}$$

Thus, the corresponding wavelet $\psi^1(x,y)$ and $\psi^2(x,y)$

$$\psi^1(x,y) = \frac{\partial g}{\partial x} = \frac{x}{2\pi\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.40}$$

$$\psi^2(x,y) = \frac{\partial g}{\partial y} = \frac{y}{2\pi\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.41}$$

Mallat and Zhong [64] constructed dyadic wavelets and calculated the local maxima of dyadic wavelet transform at each scale and formed a multiscale edge representation of the image. Furthermore, Mallat [64] proved that Canny edge detection is equivalent to finding the local maxima of a wavelet transform. The function (4.40), (4.41) is also an explicit form of the wavelet model for Canny edge detection. From the point of view of wavelet transforms, we can use a more effective algorithm to adjust the scale of the filters.

The Gaussian functions g_t have a semigroup property, that is,

$$g_{t_1+t_2} = g_{t_1} * g_{t_2} \tag{4.42}$$

So, we can find the cascade algorithm to calculate the wavelet transform of an image $f(x, y)$ on a series of scales as follows.

$$\begin{aligned} W^1 f(ns, x, y) &= \psi^1_{ns} * f(x, y) \qquad (4.43) \\ &= \frac{\partial g_{ns}}{\partial x} * f(x, y) \\ &= \frac{\partial (g_{(n-1)s} * g_s)}{\partial x} * f(x, y) \\ &= \frac{\partial (g_{(n-1)s} * f)}{\partial x} * g_s(x, y) \qquad (4.44) \\ &= W^1 f((n-1)s, x, y) * g_s(x, y) \end{aligned}$$

Thus, for any integer $n > 1$

$$W^1 f(ns, x, y) = \underbrace{g_s * ... * g_s}_{n-1} * W^1 f(s, x, y) \tag{4.45}$$

Similar,

$$W^2 f(ns, x, y) = \underbrace{g_s * ... * g_s}_{n-1} * W^2 f(s, x, y) \tag{4.46}$$

In an image, all edges are not created equal. Some are more significant than others, and some are blurred and insignificant. The edges of more significance are usually more important and more likely to be kept intact by wavelet transform. The insignificant edges are sometimes introduced by noise and preferably removed wavelet transform. In mathematics, the sharpness of an edge can be described by a Lipschitz exponent. Mallat and Hwang [94] showed that Lipschitz exponents can be measured by wavelet transform.

We first introduce the definition of Lipschitz exponent in 1-D [94].

Definition 1: Let $0 \leq \alpha \leq 1$. A function $f(x)$ is uniformly Lipschitz α over an interval $[a, b]$ if and only if there exists a constant K such that for any $(x_0, x_1) \in [a, b]^2$

$$|f(x_0) - f(x_1)| \leq K|x_0 - x_1|^\alpha \tag{4.47}$$

Theorem 1: Let $0 < \alpha < 1$. A function $f(x)$ is uniformly Lpschitz α over $[a, b]$ if and only if there exists a constant $K > 0$ such that for all $x \in [a, b]$, the wavelet transform satisfies

$$|W_{2^j} f(x)| \leq K(2^j)^\alpha \tag{4.48}$$

Theorem 1 proves that the Lipschitz exponent of a function can be measured from the evolution across scale of the absolute value of the wavelet transform.

From (4.48), we derive that

$$\log_2 |W_{2^j} f(x)| \leq \log_2(K) + \alpha j \tag{4.49}$$

The definition is extended to 2-D.

Definition 2: Let $0 < \alpha < 1$. A function $f(x, y)$ is uniformly Lipschitz α over an open set Ω of $\boldsymbol{R}^2$ if and only if there exists a constant K such that for all (x_0, y_0) and (x_1, y_1) in Ω

$$|f(x_0, y_0) - f(x_1, y_1)| \leq K|(x_0 - x_1)^2 + (y_0 - y_1)^2|^{\alpha/2} \tag{4.50}$$

The Lipschitz regularity of $f(x, y)$ over Ω is the superior bound of all α such that $f(x, y)$ is uniformly Lipschitz α.

Theorem 2: Let $0 < \alpha < 1$. A function $f(x, y)$ is uniformly Lipschitz α over an open set of $\boldsymbol{R}^2$ if and only if there exists a constant K such that for all points (x, y) of this open set

$$M_{2^j} f(x, y) \leq K(2^j)^\alpha \tag{4.51}$$

Theorem 2 is the 2-D extension of Theorem 1.

From (4.51), we derive that

$$\log_2(M_{2^j} f(x, y)) \leq \log_2(K) + \alpha j \tag{4.52}$$

The resolution of an image is directly related to the proper scale for edge detection. High resolution and small scales will result in noisy and discontinuous edges; low resolution and large scales will result in undetected edges. The scale is not adjustable with classical edge detectors, but with a wavelet transform, we can construct our own edge detectors with proper scales.

From Theorem 2, we can use the coefficients of the wavelet transform across scales to measure the local Lipschitz regularity. That is, when the scale increases, the coefficients of the wavelet transform are likely to increase where the Lipschitz regularity is positive, but they are likely to decrease where the Lipschitz regularity is negative. The locations with lower Lipschitz regularity are more likely to be details and noise.

Wavelet filters of large scales are more effective for removing noise, but at the same time increase the uncertainty of location of edges. Wavelet filters of small scales preserve the exact location of edges, but cannot distinguish between noise and real edges. We can use the coefficients of the wavelet transform across scales to measure the local Lipschitz regularity. We can use a large-scale wavelet at positions where the wavelet transform decreases rapidly across scales to remove the effect of noise, while using a smaller-scale wavelet at positions where the wavelet transform decreases slowly across scale to preserve the precise position of the edges. Using the cascade algorithm to observe the change of wavelet transform coefficient between each adjacent scale, we can distinguish different types of edges.

The diagram of the method is shown in Figure 4.12.

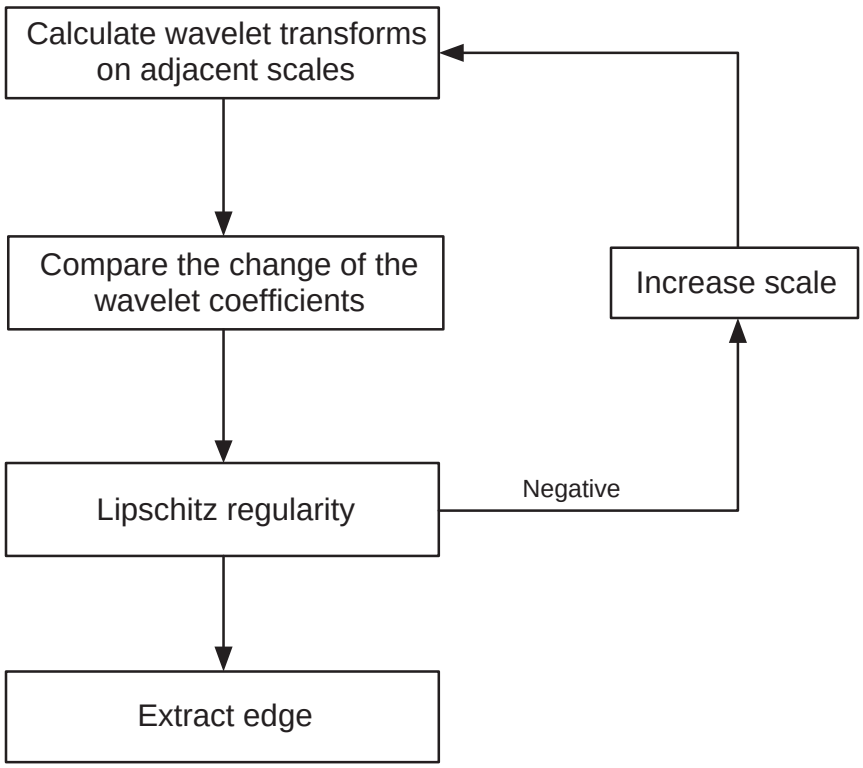
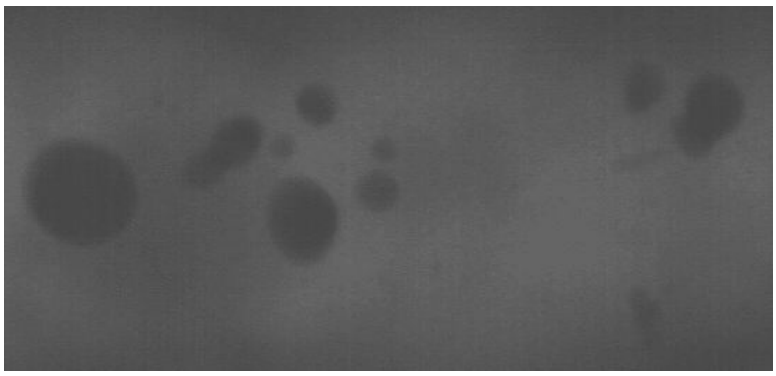


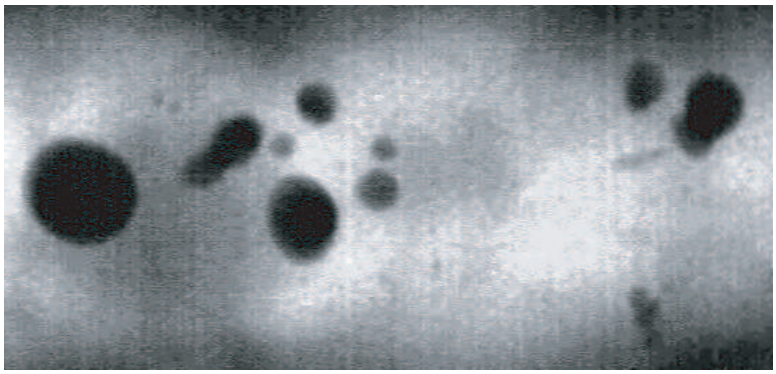
Figure 4.12: The diagram of the proposed multiscale edge detection algorithm

4.3.3 Experimental Results

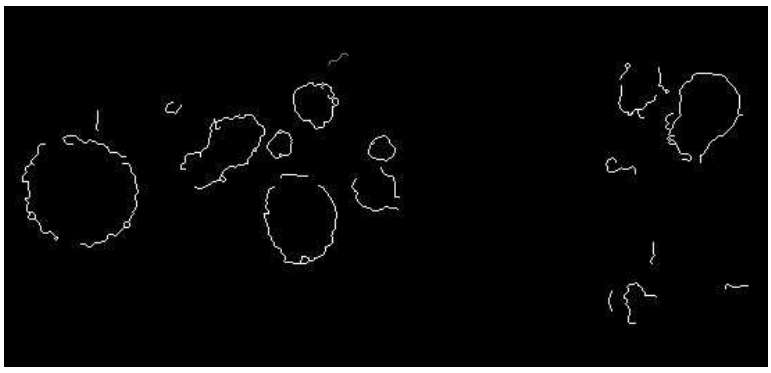
The proposed multiscale edge detection algorithm is implemented in the MATLAB language. In the experiments, dozens of gray-level (8 bit intensity) images have been used to test the performance of the proposed method.



(a) Original image

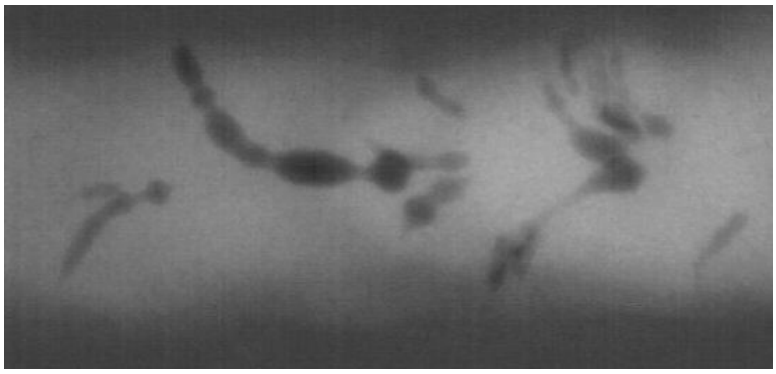


(b) Result of preprocessing

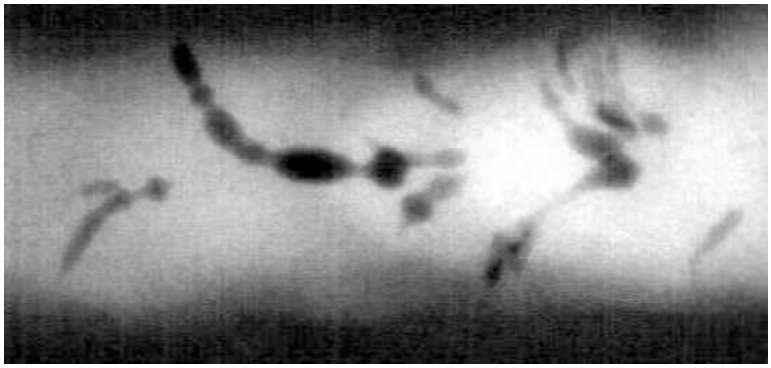


(c) Edge of defect

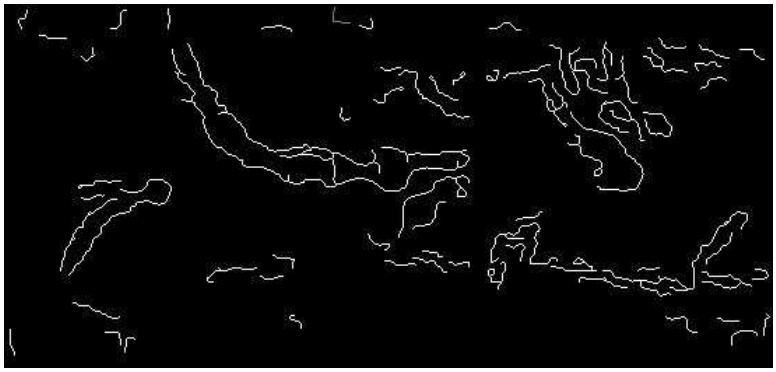
Figure 4.13: Scattered porosity



(a) Original image

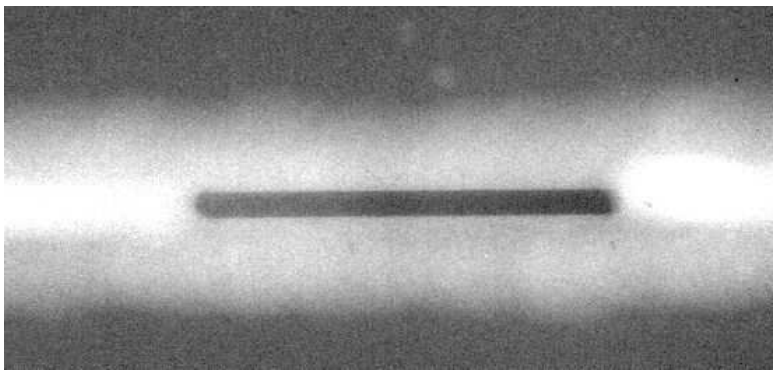


(b) Result of preprocessing

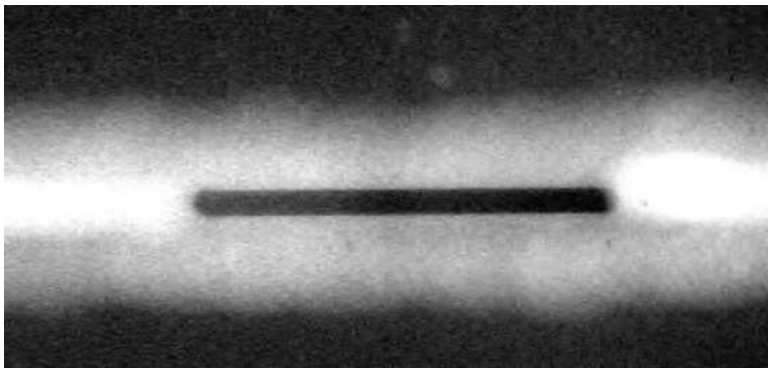


(c) Edge of defect

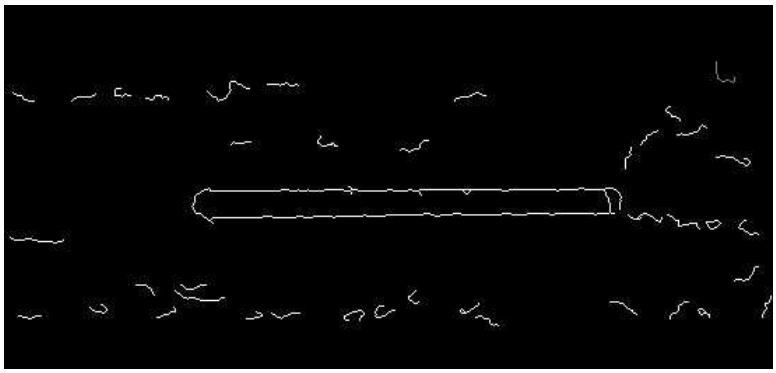
Figure 4.14: Elongated porosity



(a) Original image



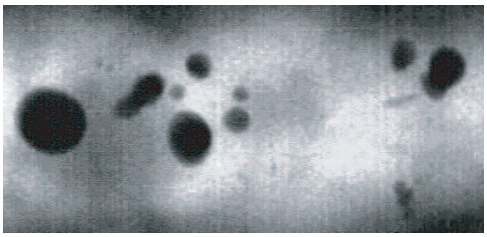
(b) Result of preprocessing



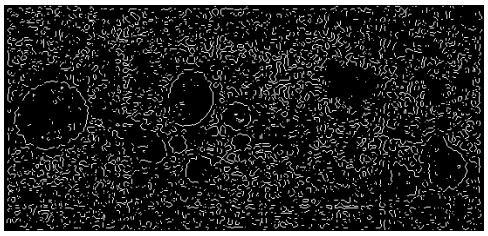
(c) Edge of defect

Figure 4.15: Incomplete root penetration

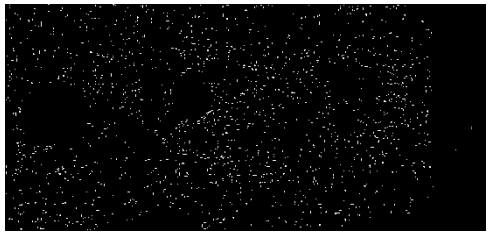
The main steps of the method are shown in Figure 4.13, Figure 4.14 and Figure 4.15. From (a) to (c), (a) is the original image. (b) is the image after preprocessing. (c) is the edge of defect using multiscale edge detection based on wavelet transform. The obtained results show that most of the edges of defects can be found, and the algorithm is appropriate not only to linear defects such as penetration but also volumetric defects such as porosity.



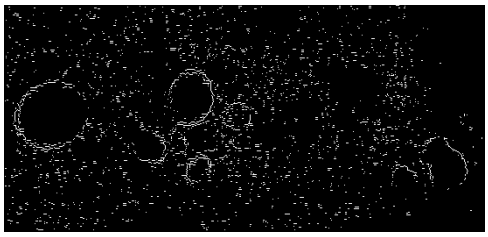
(a) Image after preprocessing



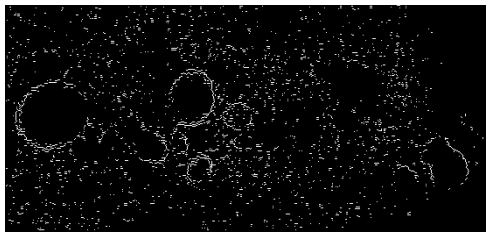
(b) LoG edge operator



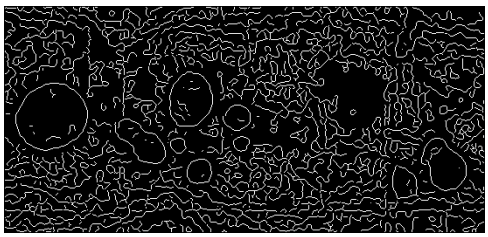
(c) Roberts edge operator



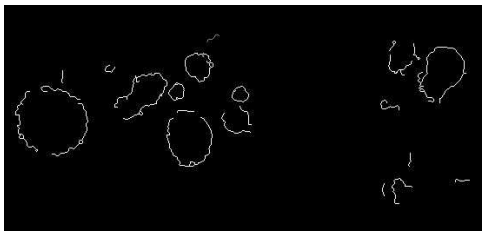
(d) Perwitt edge operator



(e) Sobel edge operator



(f) Canny edge operator



(g) Multiscale edge detection algorithm based on wavelet

Figure 4.16: Different edge-based methods comparison for the radiographic image

In Figure 4.16, the proposed algorithm is compared with other classical edge detection methods. This algorithm is not so noise sensitive as other methods. It can distinguish between noise and a defect's edge well for radiographic images. However, it is quite sensitive to the texture noise, such as thatshown in Figure 4.13, Figure 4.14 and Figure 4.15. The texture noise is introduced by the frequency error produced by the scanner's mechanical error and the bad quality or damage of the film. It can be seen from the image that the texture-noise is like some of the lines in the image. This kind of noise can make continuous edge and some false edges are found in edge detection. So the final result contains not only the defect we are looking for, but also this kind of noise. Sometimes the edges are so blur that the gradients are small. Under less than ideal conditions, edge image will have big gaps.

4.4 Multi-level Thresholding Algorithm based on Maximum Fuzzy Entropy and Genetic Algorithm

Since it is difficult to extract all the defects from background using two-level thresholding methods, We propose another method for the accurate segmentation of defect using multi-level thresholding based on fuzzy entropy and genetic algorithm (MTFEGA) . Firstly, we compute standard deviation (STD) of the radiographic image, which is used to estimate the underlying brightness probability distribution of the image. Then we apply two-level or three-level or four-level threhsolding based on the standard deviation. When $STD \leq 50$, two-level thresholding is applied; when $50 < STD \leq 70$, three-level thresholding is applied; when $STD > 70$, four-level thresholding is applied. The diagram of the method is shown in Figure 4.17.

4.4.1 Fuzzy Set Theory

The fuzzy set theory, first proposed by Zadeh [95] [96], can be regarded as the extension of classical set theory. It is an implementation of classes or groupings of data with boundaries that are not sharply defined (i.e., fuzzy). In classical set theory or crisp theory, data sets are represented by "crisp" definitions. Any member can either "belong" or "do not belong" to a data set. In a fuzzy set any member can belong to the data set with a certain value or membership. The benefit of extending crisp theory and analysis methods to fuzzy techniques is a strength in solving real-world problems. It is inevitable that

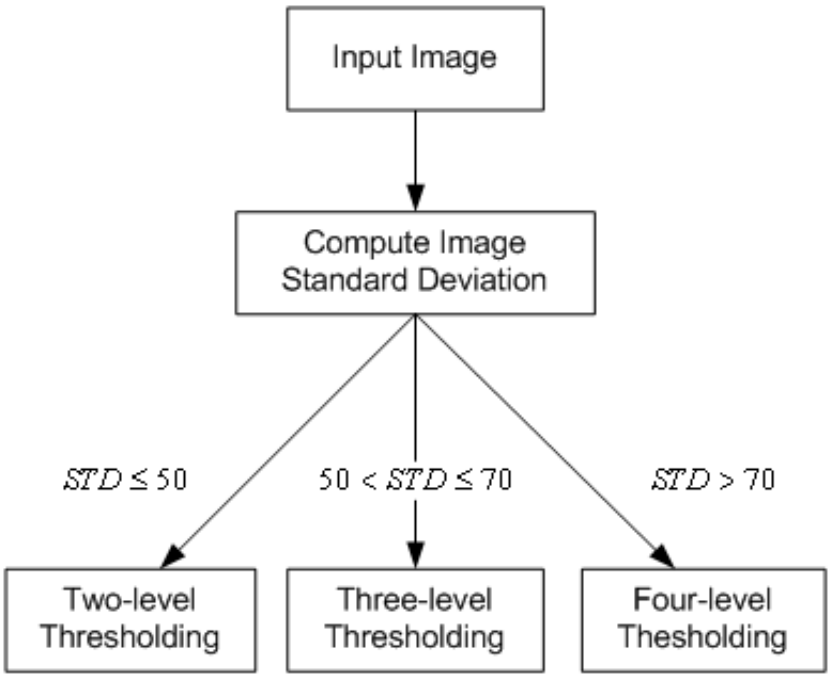


Figure 4.17: The diagram of the proposed multi-level thresholding algorithm

these real-world problems have some degree of imprecision and noise in the variables and parameters measured and processed for the application.

A classical set A is normally defined as a collection of element. Each element x in the universe either belongs to a set or not. Therefore the membership $\mu_A(x)$ is either 1 for those element in the set ($x \in A$) or 0 for those out of the set ($x \notin A$). A fuzzy set is an extension of a classical set in which an element may partially belong to a set. A fuzzy set A in the observed space X is characterized by a membership function $\mu_A(x)$ that associates each element x of X with a real number in the interval $[0, 1]$. The interval $[0, 1]$ is called the fuzzy domain. The value of $\mu_A(x)$ is the grade of x belonging to A. Generally, a fuzzy set A where $A \subset X$ is defined as

$$A = \{x, \mu_A(x) | x \in X\} \tag{4.53}$$

4.4.2 Genetic Algorithm

Since their formal introduction in 1975 by Holland [97], genetic algorithms have been applied to a variety of fields-from medicine and engineering to business-to optimize functions which do not lend themselves to optimization by traditional methods. Other applications of GAs include automatic programming and simulation of natural systems. More recently, the study and practical development of the GA by Goldberg [98] has resulted in great growth in the application of GAs to optimization problems. As succinctly stated by Goldberg, GAs are ”search procedures based on the mechanics of natural selection and natural genetics.” Random choice is used as a tool to guide a global search in the space of potential solutions.

A genetic algorithm (GA) [98] [99] is a randomized search and optimization techniques guided by the principles of evolution and natural genetics, having a large amount of implicit parallelism. GAs differ from traditional optimization and search methods in several respects. Rather than focusing on a single candidate solution (point in design space), genetic algorithms operate on populations of candidate solutions, and the search process favors the reproduction of individuals with better fitness values than those of previous generations (optimal individuals). Whereas calculus-based and gradient (hillclimbing) methods of solution are local in the scope of their search and depend on well-defined gradients in the search space, GAs are useful for dealing with many practical problems containing noisy or discontinuous fitness values. Enumerative searches are also inappropriate for many practical problems. Because they exhaustively examine the entire search space for solutions, they are only efficient for small search spaces, while the global scope of the GA makes it suitable for problems with large search spaces. Thus, GAs not only differ in approach from traditional optimization methods but also offer an alternative method for cases in which traditional methods are inappropriate.

GAs have been applied to continuous optimization problems, but it is rarely as effective as continuous optimization methods. Evolutionary programming is appropriate for continuous problems; GAs are not, being inherently discrete. The genetic algorithm as a discrete optimization process is distinct from more conventional optimization techniques in four ways:

1. GAs encode designs (feasible points) in a string, and it is this encoding that the GA works with: each individual in a population is an encoding of a possible solution to

the discrete optimization problem being analyzed.

2. GAs work simultaneously with a population of designs, not a single design or candidate solution.

3. GAs use only an objective function to evaluate candidate solutions, not derivatives or other auxiliary information.

4. GAs use random change in their search, not (solely) deterministic rules.

The process used by genetic algorithms to evolve solutions to optimization problems is analogous to the natural process of evolution by natural selection. Evolution as a natural process allows complex, highly adapted organisms to develop and thrive in an environment through the processes of genetic change and natural selection. Sexual reproduction (sexual in the sense of occurring between two parent individuals as opposed to one) provides the preservation of existing genetic information and the creation of new genetic information, and individuals in a population survive based on their fitness in their environment. Fitness is a quality measure of an individual's viability with respect to such criteria in the natural environment as food supply, competition for food and mates, and predation. The genetic information carried by more fit individuals is more likely to be passed on to ensure generations simply because more fit individuals are more likely to survive to reproduce—Darwinian survival of the fittest.

GAs apply the natural evolutionary processes of evaluation and selection to string representations of the arguments of the function being optimized. Structures (individuals in natural systems) are encoded into one or more strings (chromosomes). These individuals reproduce, and fit individuals persist from generation to generation, yielding improved designs.

The structure is analogous to the phenotype in natural systems and corresponds to a candidate solution to the optimization problem or a point in the design space, while the string encoding of the arguments to the function being optimized is analogous to the genotype. A decoding from the string representation to the structure is made for the purpose of fitness analysis by the objective function. The objective function yields a quantitative measure of an individual's utility or goodness, to be used as a selection criterion.

4.4.3 Two-level Thresholding

For two-level thresholding, Ostu's method [100] is applied to select the threshold. It can be regarded as the simplest and most standard method for an automated threshold selection. It only uses the zero-order and first-order cumulative moments of the gray level histogram. The two classes of pixels are separated by the between-class variance, σ_b:

$$\sigma_b^2(T) = P_1(\mu_1 - \mu)^2 + P_2(\mu_2 - \mu)^2 = P_1 P_2(\mu_1 - \mu_2)^2 \tag{4.54}$$

where P and μ are the corresponding probability and mean values of class 1 and 2, respectively.

The optimal threshold is the one that maximizes the between-class variance (or, conversely, minimizes the within-class variance). The optimal threshold T^* is given by

$$\sigma_b^2(T^*) = max\ \sigma_b^2(T) \qquad 1 \leq T < L \tag{4.55}$$

where L is the highest gray level of the pixels for a given image(e.g., gray level 255 for the 8 bit image).

4.4.4 Multi-level Thresholding Based on Maximum Fuzzy Entropy

For multi-level thresholding, the problem is how to determine the best thresholds. The relationship between a probability partition and a fuzzy c-partition (FP) in thresholding are explored by Zhao [101]. Based on this relationship and entropy approach, a technique derived to get best fuzzy c-partition can be used to find the optimum thresholds.

Let $D \equiv \{(i,j) : i = 0, 1, \ldots, M-1; j = 0, 1, \ldots, N-1\}$, $G = \{0, 1, \ldots, l-1\}$, where M, N and l are three positive integers. Then a digitized image is considered a mapping $I : D \rightarrow G$. $I(x,y)$ is the gray level value of the image at the pixel (x,y).

$$\begin{aligned} I(x,y) \in G \qquad & \forall (x,y) \in D \\ D_k = \{(x,y) : I(x,y) = k, (x,y) \in D\} \qquad & \\ k = 0, 1, ..., l-1 & \end{aligned} \tag{4.56}$$

$$h_k = \frac{n_k}{N * M} \qquad k = 0, 1, ..., l-1 \tag{4.57}$$

where n_k denotes the number of pixels in D_k.

The following conclusions can be formed:
$\cup_{k=0}^{l-1} D_k = D$ and $D_j \cap D_k = \phi\ (j \neq k)$. Thus $0 \leq h_k \leq 1$, $\sum_{k=0}^{l-1} h_k = 1$, $k = 0, 1, ..., l-1$. $H = \{h_0, h_1, \ldots, h_{l-1}\}$ is the histogram of the image. $\prod_l = \{D_0, D_1, \ldots, D_{l-1}\}$ is a probability partition (PP) of D with a probabilistic distribution.

$$p_k = p(D_k) = h_k \qquad k = 0, 1, ..., l-1 \tag{4.58}$$

The radiographic images are stored with 8 bits, so the gray levels l is 256. So, an 255 level digitized radiographic image is characterized by the PP, $\prod_l$, of its domain derived from equations (4.57) and (4.58).

A fuzzy set is an extension of a classical set in which an element may partially belong to a set. A fuzzy set A in the observed space X is characterized by a membership function $\mu_A(x)$ that associates each element x of X with a real number in the interval $[0, 1]$. The interval $[0, 1]$ is called the fuzzy domain. The value of $\mu_A(x)$ is the grade of x belonging to A. Generally, a fuzzy set A where $A \subset X$ is defined as

$$A = \{(x, \mu_A(x))\,|x \in X\} \tag{4.59}$$

where $0 \leq \mu_A \leq 1$.

One of the interesting point about transforming an image from the intensity domain into the fuzzy domain is how much information it can keep. According to the information theory [34], Shannon entropy can be defined as

$$H(A) = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{4.60}$$

where $\sum_{i=1}^{N} p(x_i) = 1$ and x_i, $i = 1, ..., N$ are the possible outputs from source A with the probability $p(x_i)$. The larger entropy $H(A)$ is, the more information A has. The width and attribute of the fuzzy region is decided by maximum fuzzy entropy, in turn the thresholds can be decided by fuzzy parameters.

Three-level Thresholding

For a considered radiographic image, the domain D of the original image is classified into three parts: E_d, E_m,E_b. E_d is composed of pixels with low level pixels, E_m is composed of pixels with middle gray levels, and E_b of high level pixels. $\prod_3 = \{E_d, E_m, E_b\}$ is an unknown probabilistic partition of D, whose probability distribution is: $p_d = P(E_d)$, $p_m = P(E_m)$, $p_b = P(E_b)$.

For three-level thresholding, we use the simplest function that is monotonic to approximate the memberships of bright μ_b, medium μ_m and dark μ_d. Where $\mu_d(k) = p_{d|k}$, $\mu_m(k) = p_{m|k}$ and $\mu_b(k) = p_{b|k}$. The three membership functions are shown in Figure 4.18. The membership functions have four parameters a_1, b_1, a_2 and b_2 (see Figure 4.18). In other words, two thresholds t_1, t_2, for three-level thresholding are depend on a_1, b_1, a_2, b_2.

Let

$$t_1 = \frac{1}{2}(a_1 + b_1)$$
$$t_2 = \frac{1}{2}(a_2 + b_2) \tag{4.61}$$

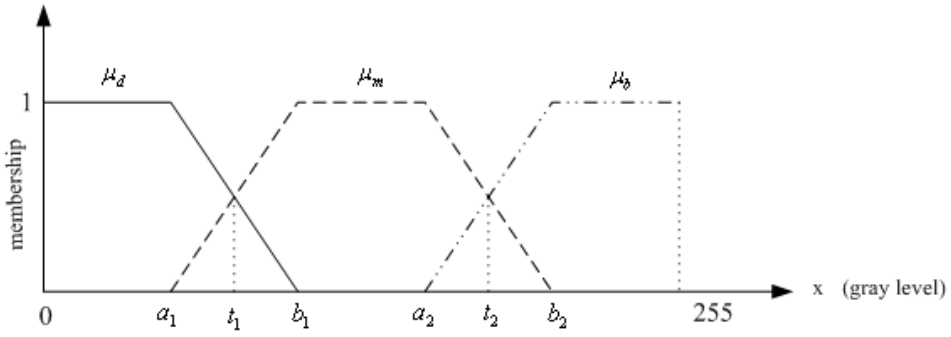


Figure 4.18: Fuzzy 3-partition

$$\mu_d(k) = \begin{cases} 1 & k \leq a_1 \\ \frac{k-b_1}{a_1-b_1} & a_1 < k \leq b_1 \\ 0 & k > b_1 \end{cases} \tag{4.62}$$

$$\mu_m(k) = \begin{cases} 0 & k \leq a_1 \\ \frac{k-a_1}{b_1-a_1} & a_1 < k \leq b_1 \\ 1 & b_1 < k \leq a_2 \\ \frac{k-b_2}{a_2-b_2} & a_2 < k \leq b_2 \\ 0 & k > b_2 \end{cases} \tag{4.63}$$

$$\mu_b(k) = \begin{cases} 0 & k \leq a_2 \\ \frac{k-a_2}{b_2-a_2} & a_2 < k \leq b_2 \\ 1 & k > b_2 \end{cases} \tag{4.64}$$

And the following conditions are satisfied $0 < a_1 \leq b_1 \leq a_2 \leq b_2 < 255$, for each $k = 0, 1, ..., 255$, let

$$D_{kd} = \{(x, y) : I(x, y) \leq t_1, (x, y) \in D_k\}$$
$$D_{km} = \{(x, y) : t_1 < I(x, y) \leq t_2, (x, y) \in D_k\}$$
$$D_{kb} = \{I(x, y) > t_2, (x, y) \in D_k\}$$

Then $p_{kd} = P(D_{kd}) = p_k * p_{d|k}$, $p_{km} = P(D_{km}) = p_k * p_{m|k}$ and $p_{kb} = P(D_{kb}) = p_k * p_{b|k}$. Note that $p_{d|k} + p_{m|k} + p_{b|k} = 1$ for $k = 0, 1, ..., 255$. When the pixel belongs to D_k, it is evident that $p_{d|k}$, $p_{m|k}$ and $p_{b|k}$ are the conditional probability of a pixel when it is classified into the class "d" (dark), "m" (medium) and "b" (bright) respectivley. So,

$$p_d = P(E_d) = \sum_{k=0}^{255} P(D_{kd}) = \sum_{k=0}^{255} p_k * p_{d|k} = \sum_{k=0}^{255} p_k * \mu_d(k) \tag{4.65}$$

$$p_m = P(E_m) = \sum_{k=0}^{255} P(D_{km}) = \sum_{k=0}^{255} p_k * p_{m|k} = \sum_{k=0}^{255} p_k * \mu_m(k) \tag{4.66}$$

$$p_b = P(E_b) = \sum_{k=0}^{255} P(D_{kb}) = \sum_{k=0}^{255} p_k * p_{b|k} = \sum_{k=0}^{255} p_k * \mu_b(k) \tag{4.67}$$

Then the following conclusion can be gained:

$$\begin{aligned} p_d &= \sum_{k=0}^{255} p_k * \mu_d(k) \\ p_m &= \sum_{k=0}^{255} p_k * \mu_m(k) \\ p_b &= \sum_{k=0}^{255} p_k * \mu_b(k) \end{aligned} \tag{4.68}$$

In this thesis, an entropy function is used to justify if the information of an image is mostly retained after thresholding, i.e. to measure the compatibility between the PP of D and the FP of G. The fuzzy entropy function is given below.

$$H(a_1, a_2, b_1, b_2) = -p_d \log p_d - p_m \log p_m - p_b \log p_b \tag{4.69}$$

where p_d, p_m and p_b are given in equation (4.68). The value of $H(a_1, a_2, b_1, b_2)$ is used to measure the compatibility between the histogram $H = \{h_0, h_1, \ldots, h_{255}\}$ of a given X-ray image and the fuzzy 3-partition $\prod_3 = \{\mu_b, \mu_m, \mu_d\}$ of G. The larger $H(a_1, a_2, b_1, b_2)$ is, the more compatible H and $\prod_3$.

The fuzzy entropy varies along with four variables a_1, b_1, a_2, b_2. We can find an optimal combination of (a_1, b_1, a_2, b_2) so that the total fuzzy entropy $H(a_1, a_2, b_1, b_2)$ has the maximum value. Then the most appropriate two thresholds t_1 and t_2 can be computed.

Four-level Thresholding

For a considered radiographic image, the domain D of the original image is classified into four parts: E_{dd}, E_{md}, E_{mb} and E_{bb}. $\prod_4 = \{E_{dd}, E_{md}, E_{mb}, E_{bb}\}$ is an unknown

probabilistic partition of D, whose probability distribution is: $p_{dd} = P(E_{dd})$, $p_{mb} = P(E_{mb})$, $p_{bb} = P(E_{bb})$.

For four-level thresholding, three membership functions are considered: dark μ_{dd}, medium dark μ_{md}, medium bright μ_{mb} and bright μ_{bb}, where $\mu_{dd}(k) = p_{dd|k}$, $\mu_{md}(k) = p_{md|k}$, $\mu_{mb}(k) = p_{md|k}$, $\mu_{md}(k) = p_{mb|k}$ and $\mu_{bb}(k) = p_{bb|k}$. The four membership functions are shown in Figure 4.19.

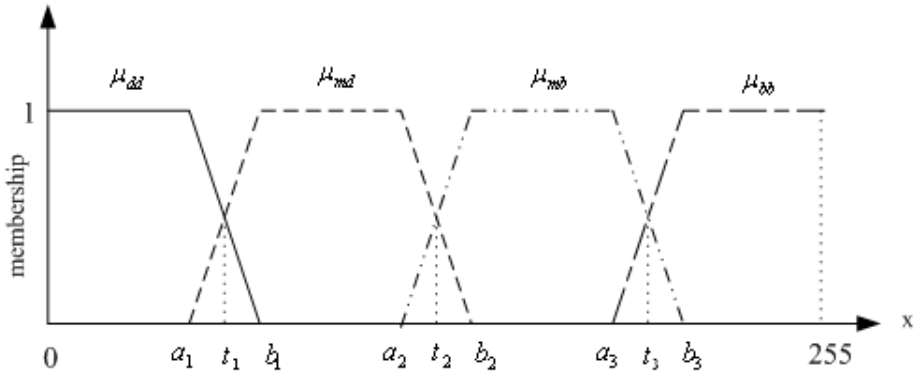


Figure 4.19: Fuzzy 4-partition

$$\mu_{dd}(k) = \begin{cases} 1 & k \le a_1 \\ \frac{k-b_1}{a_1-b_1} & a_1 < k \le b_1 \\ 0 & k > b_1 \end{cases} \tag{4.70}$$

$$\mu_{md}(k) = \begin{cases} 0 & k \le a_1 \\ \frac{k-a_1}{b_1-a_1} & a_1 < k \le b_1 \\ 1 & b_1 < k \le a_2 \\ \frac{k-b_2}{a_2-b_2} & a_2 < k \le b_2 \\ 0 & k > b_2 \end{cases} \tag{4.71}$$

$$\mu_{mb}(k) = \begin{cases} 0 & k \leq a_2 \\ \frac{k-a_2}{b_2-a_2} & a_2 < k \leq b_2 \\ 1 & b_2 < k \leq a_3 \\ \frac{k-b_3}{a_3-b_3} & a_3 < k \leq b_3 \\ 0 & k > b_3 \end{cases} \tag{4.72}$$

$$\mu_{bb}(k) = \begin{cases} 0 & k \leq a_3 \\ \frac{k-a_3}{b_3-a_3} & a_3 < k \leq b_3 \\ 1 & k > b_3 \end{cases} \tag{4.73}$$

The membership functions have six parameters a_1, b_1, a_2, b_2, a_3, b_3. In other words, two thresholds t_1, t_2 and t_3, for four-level thresholding are depend on a_1, b_1, a_2, b_2, a_3, b_3. And the following conditions are satisfied $0 < a_1 \leq b_1 \leq a_2 \leq b_2 \leq a_3 \leq b_3 < 255$, for each $k = 0, 1, ..., 255$, let

$$D_{kdd} = \{(x, y) : I(x, y) \leq t_1, (x, y) \in D_k\}$$
$$D_{kmd} = \{(x, y) : t_1 < I(x, y) \leq t_2, (x, y) \in D_k\}$$
$$D_{kmb} = \{(x, y) : t_2 < I(x, y) \leq t_3, (x, y) \in D_k\}$$
$$D_{kbb} = \{I(x, y) > t_3, (x, y) \in D_k\}$$

Then $p_{kdd} = P(D_{kdd}) = p_k * p_{dd|k}$, $p_{kmd} = P(D_{kmd}) = p_k * p_{md|k}$, $p_{kmb} = P(D_{kmb}) = p_k * p_{mb|k}$ and $p_{kbb} = P(D_{kbb}) = p_k * p_{bb|k}$. Note that $p_{dd|k} + p_{md|k} + p_{mb|k} + p_{dd|k} = 1$ for $k = 0, 1, ..., 255$. When the pixel belongs to D_k, it is evident that $p_{dd|k}$, $p_{md|k}$, $p_{mb|k}$ and $p_{bb|k}$ are the conditional probability of a pixel when it is classified into the four classes. So,

$$p_{dd} = P(E_{dd}) = \sum_{k=0}^{255} P(D_{kdd}) = \sum_{k=0}^{255} p_k * p_{dd|k} = \sum_{k=0}^{255} p_k * \mu_{dd}(k) \tag{4.74}$$

$$p_{md} = P(E_{md}) = \sum_{k=0}^{255} P(D_{kmd}) = \sum_{k=0}^{255} p_k * p_{md|k} = \sum_{k=0}^{255} p_k * \mu_{md}(k) \tag{4.75}$$

$$p_{mb} = P(E_{mb}) = \sum_{k=0}^{255} P(D_{kmb}) = \sum_{k=0}^{255} p_k * p_{mb|k} = \sum_{k=0}^{255} p_k * \mu_{mb}(k) \tag{4.76}$$

$$p_{bb} = P(E_{bb}) = \sum_{k=0}^{255} P(D_{kbb}) = \sum_{k=0}^{255} p_k * p_{bb|k} = \sum_{k=0}^{255} p_k * \mu_{bb}(k) \tag{4.77}$$

The fuzzy entropy function is given below. Here, an entropy function is used to justify if the information of an image is mostly retained after thresholding.

$$H(a_1, a_2, b_1, b_2, a_3, b_3) = -p_{dd} \log p_{dd} - p_{md} \log p_{md} - p_{mb} \log p_{mb} - p_{bb} \log p_{bb} \tag{4.78}$$

The fuzzy entropy varies along with six variables a_1, b_1, a_2, b_2, a_3, b_3. We can find an optimal combination of $(a_1, b_1, a_2, b_2, a_3, b_3)$ so that the total fuzzy entropy $H(a_1, a_2, b_1, b_2, a_3, b_3)$ has the maximum value. Then the most appropriate three thresholds t_1, t_2 and t_3 can be computed as follows:

$$\begin{aligned} t_1 &= \frac{1}{2}(a_1 + b_1) \\ t_2 &= \frac{1}{2}(a_2 + b_2) \\ t_3 &= \frac{1}{2}(a_3 + b_3) \end{aligned} \tag{4.79}$$

4.4.5 Genetic Algorithm Implementation

In this study, a genetic algorithm (Figure 4.20) is implemented to find the optimal fuzzy parameters so that the total fuzzy entropy has the maximum value. It can provide a near-optimal solution for an objective or fitness function of an optimization problem. It can also overcome many defects in other optimization techniques such as calculus based techniques, exhaustive technique and knowledge based techniques.

After a population of individuals has been initialized, the fitness of the individuals is calculated according to the given fitness function. Parents are then selected with a probability proportional to their fitness. Once the parents have been selected, individuals for the next generation are formed using two main genetic operators, crossover and mutation. This process is repeated until an acceptable solution is found. The reproduction and crossover operators determine which parents will have offspring, and how genetic material is exchanged between the parents to create those offspring. Mutation allows for random alteration of genetic material. Reproduction and crossover operators tend to increase the quality of the populations and force convergence. Mutation opposes convergence and replaces genetic material lost during reproduction and crossover.

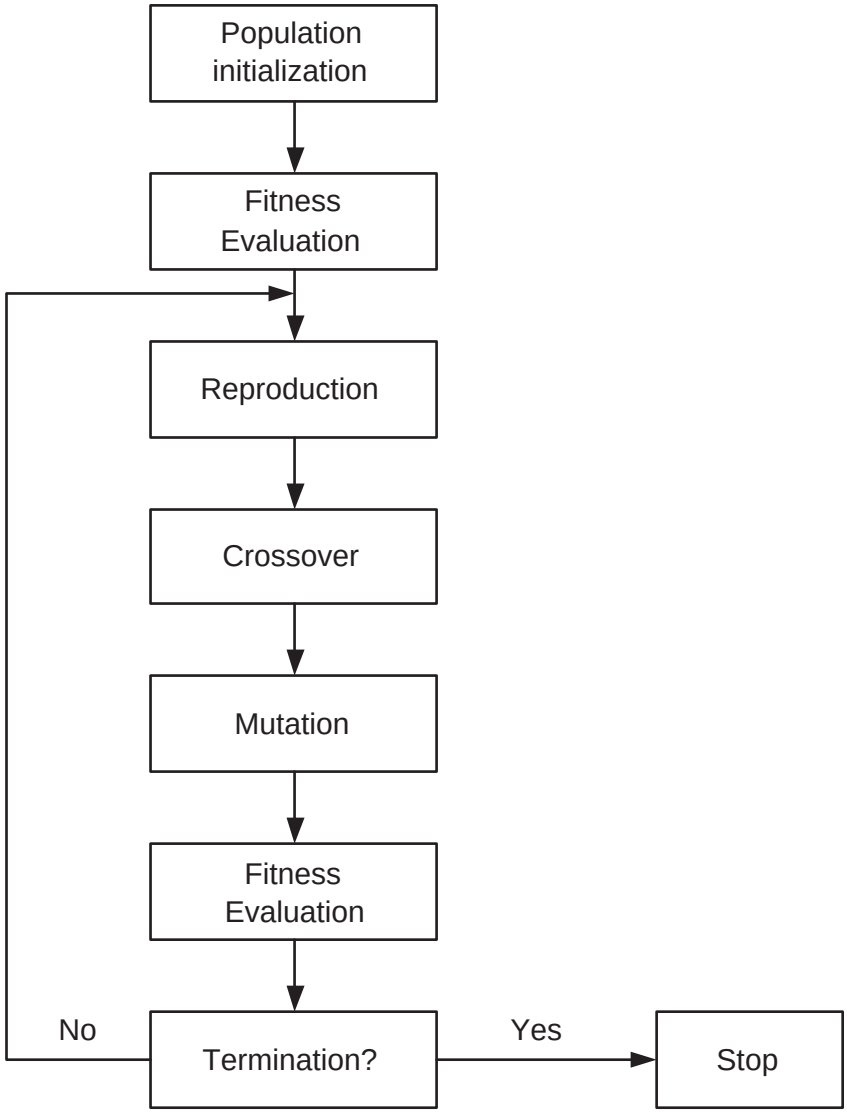


Figure 4.20: The GA algorithm

The algorithm is described in detail as follows.

Population Initialization

The first step is to encode the parameters into string. Binary representation is often used in GAs where each gene has a value of either 0 or 1. Other presentations have been proposed, for example, floating point representations [102], integer representations [103], gray- coded representations [104] and matrix representations [105]. Floating point representations are faster, more consistent and have higher precision than binary representations [102]. So, in this study, the floating point representation is used to encode the parameters. For three-level thresholding, each string is a sequence of floating point representing the four parameters a_1, b_1, a_2, b_2, and these four parameters follow the increasing order and the maximum value of b_2 is 255. For four-level thresholding, each string is a sequence of floating point representing the six parameters a_1, b_1, a_2, b_2, a_3, b_3, and these six parameters follow the increasing order and the maximum value of b_3 is 255. Then a population of N strings is randomly generated. In our experiments, N is equal to 300.

Fitness Computation

The fitness function is a particular type of objective function that quantifies the optimality of a solution (that is, a chromosome) in the genetic algorithm so that that particular chromosome may be ranked against all the other chromosomes. For three-level thresholding, we choose the entropy function 4.69 as the fitness function. For four-level thresholding, we choose the entropy function 4.78 as the fitness function.

Reproduction

A reproduction operator combines the relative fitness of generation chromosomes with some randomness in order to determine parents of the following generation. In this study, roulette wheel method which is one of the computationally simplest and most popular methods is used. This method calculates the ratio $f_i / \sum_i f_i$ for each chromosome i, which is considered its probability of survival into the next generation. As explained by Ansari and Hou [106], this approach gives strings with higher fitness values f_i a greater probability of survival. In addition, since the number of strings in a population is held constant over time, the reproduction operator will generate a new population of the same size. This implies that chromosomes with higher fitness values will eventually dominate the population.

Because the GA is blind in nature, it is possible for the offspring to be worse than their parents and some fitter chromosomes may be lost from the evolutionary process. To overcome this problem, in the study the elitist strategy is used in conjunction with the roulette wheel strategy. Elitism automatically forms a child in the new generation with the exact qualities of the parent with the highest merit. This approach guarantees that the highest fitness in a given generation will be at least as high as in the previous generation.

Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two children chromosomes. It is the main explorative operator in the GA. It is performed in two steps. First, the selected parents are mated at random. Second, a random position in the gene structure is selected for each pair of mates, and the remaining segments of the parents are swapped with crossover probability p_c.

In this study, convex crossover with crossover probability of p_c is used. If x, y are two parents, the parent x is replaced by $x' = \lambda x + (1-\lambda)y$ and the parent y is replaced by $y' = \lambda y + (1-\lambda)x$, where $0 < \lambda < 1$.

Mutation

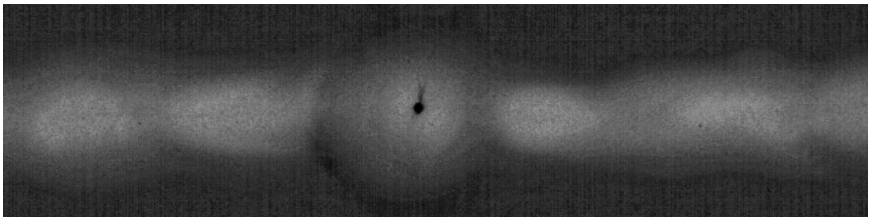
Mutation is a background operator, mainly used to explore new areas in the search space and to add diversity to the population of chromosomes in order to prevent being trapped in a local optimum. It is applied to the offspring chromosomes after crossover is performed.

Each chromosome undergoes mutation with probability p_m. For binary representation of chromosomes, a gene is mutated by simply flipping its values. Since we are considering real number representation in this thesis, we use nonuniform mutation. For a given parent x, if the gene x_k is selected for mutation, then the resulting offspring is $x' = [x_1 \cdots x'_k \cdots x_m]$ where x'_k is selected with equal probability from the two choices: $x'_k = x_k + \gamma(255 - x_k)(1 - \frac{t}{T})^b$ or $x'_k = x_k + \gamma x_k(1 - \frac{t}{T})^b$, where γ is a random number chosen uniform form $[0, 1]$, t is the current generation number, T is the maximum number of the generation and b is the parameter determining the degree of nonuniformity.

4.4.6 Experimental Results

The proposed algorithm is implemented in MATLAB language. In the experiments, dozens of gray-level (8 bit intensity) images have been used to test the performance of the proposed method.

One, two, or three optimal thresholds for each image based on the character of each radiographic image is obtained. The original images are partitioned into two parts, three parts or four parts as shown in Figure 4.21, Figure 4.22 and Figure 4.23. The obtained results show most of the defects can be extracted using this proposed method.



(a) Image after preprocessing

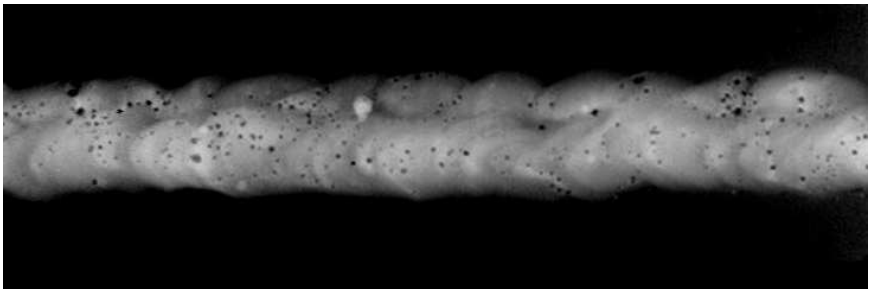
(b) Image after segmentation ($t = 83$)

Figure 4.21: A radiographic image with crack defect

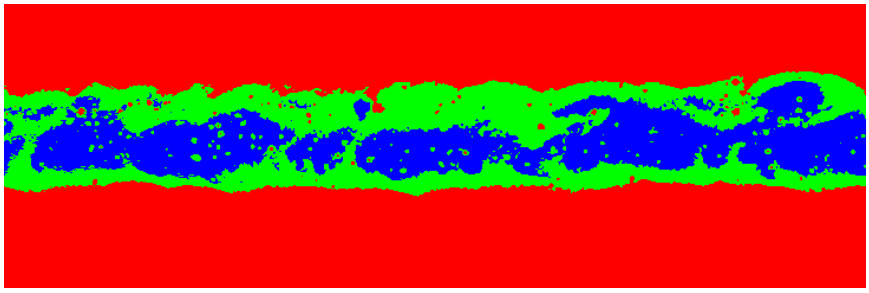
The contrast between the minimum defect that could be segmented and background is measured. The minimum contrast ratio (MCR) is defined to evaluate the segmentation performance:

$$MCR = \frac{contrast\ between\ minimum\ defect\ that\ could\ be\ segmented\ and\ background}{(255-0)\ gray\ levels} \tag{4.80}$$

25 radiographic images are measured. From the Table 4.1, we can see that MCR of the proposed algorithm is smaller than the two-level thresholding method. So, the proposed

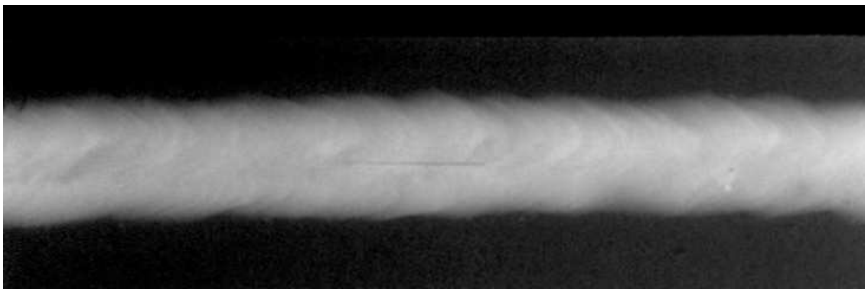


(a) Image after preprocessing

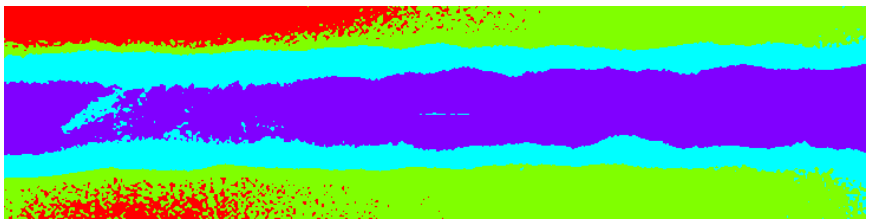


(b) Image after segmentation ($t_1 = 64$, $t_2 = 135$)

Figure 4.22: A radiographic image with porosity defect

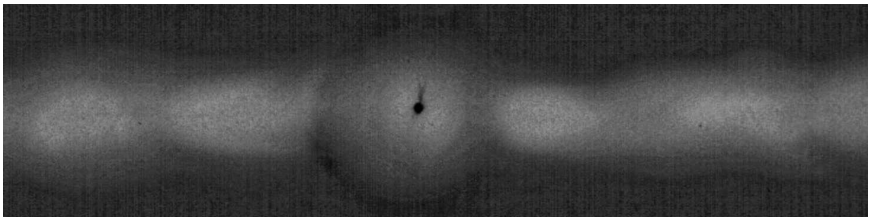


(a) Image after preprocessing



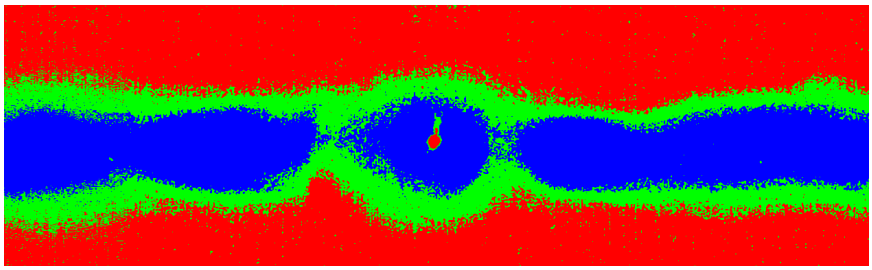
(b) Image after segmentation ($t_1 = 2$, $t_2 = 53$, $t_3 = 162$)

Figure 4.23: A radiographic image with incomplete penetration defect

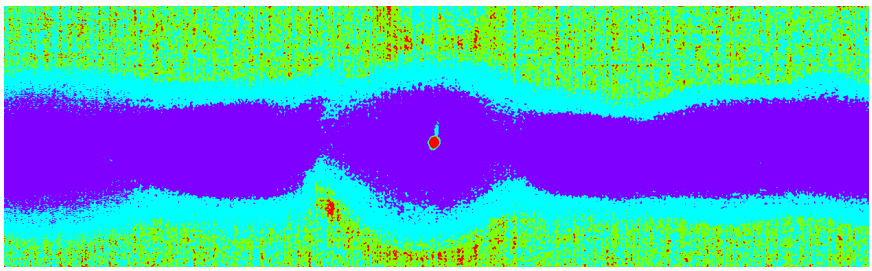


(a) Image after preprocessing

(b) Image after two-level thresholding

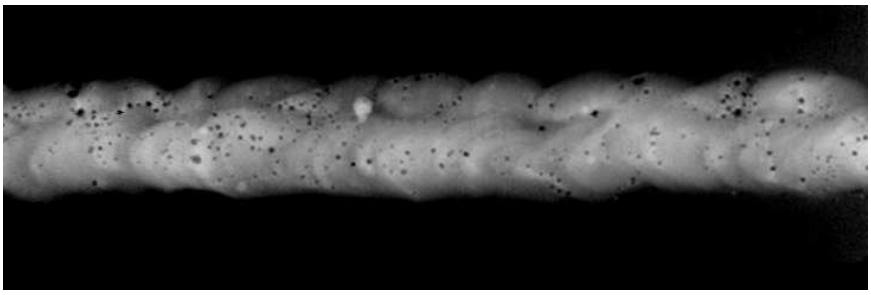


(c) Image after three-level thresholding



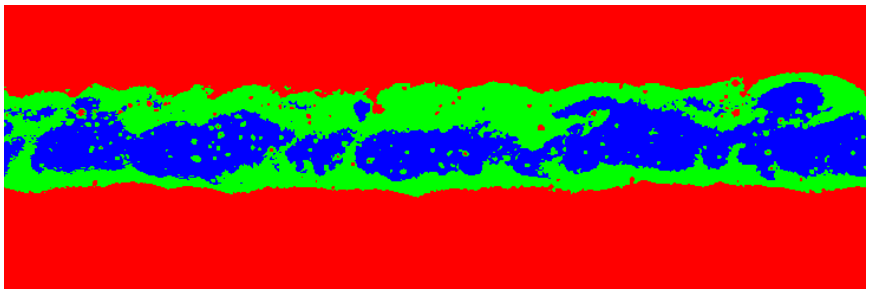
(d) Image after four-level thresholding

Figure 4.24: The result of multi-level thresholding ($STD \leq 50$)

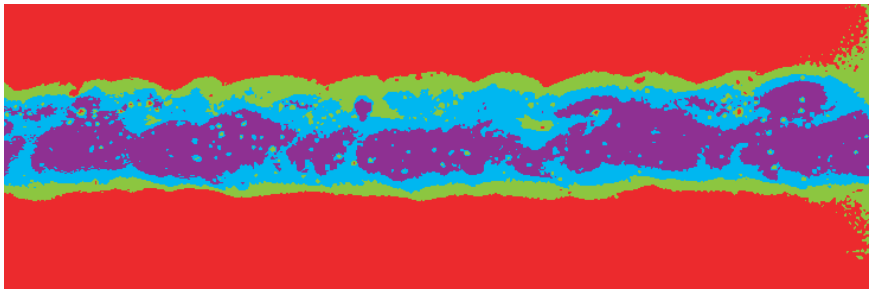


(a) Image after preprocessing

(b) Image after two-level thresholding

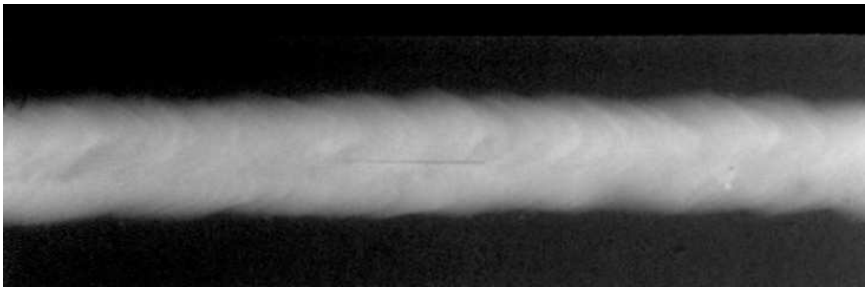


(c) Image after three-level thresholding



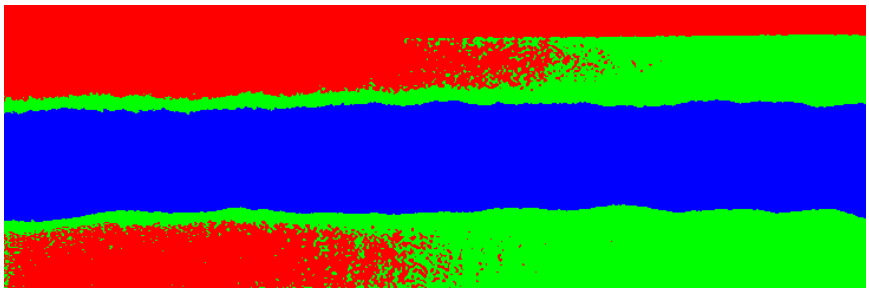
(d) Image after four-level thresholding

Figure 4.25: The result of multi-level thresholding ($50 < STD \leq 70$)

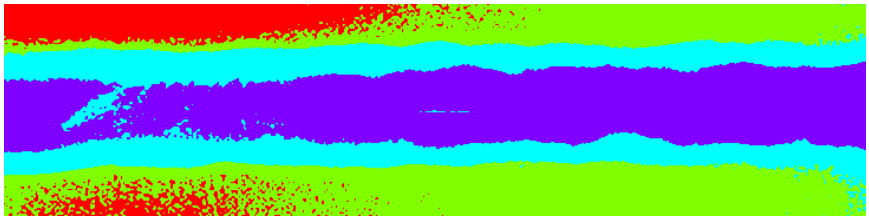


(a) Image after preprocessing

(b) Image after two-level thresholding



(c) Image after three-level thresholding



(d) Image after four-level thresholding

Figure 4.26: The result of multi-level thresholding ($STD > 70$)

algorithm can extract more defects.

Table 4.1: The minimum contrast ratio

Method	MCR
Proposed method	3.92%
Two-level thresholding	12.50%

The performance of multi-level thresholding is shown in Figure 4.24, Figure 4.25 and Figure 4.26. The results show the proposed method is a robust method. This method can reduce the over-segmentation comparing with only using three-level thresholding or four-level thresholding. When $STD \leq 50$, the two-level thresholding gives the best result. The three-level and four-level thresholding result in over-segmentation. When $50 < STD \leq 70$, the three-level thresholding gives the best results. Two-level thresholding misses some defects. Four-level thresholding leads to over-segmentation. When $STD > 70$, four-level thresholding can segment defects, while two-level and three-level thresholding can not extract defects. .

4.5 Segmentation Performance Evaluation

Two new segmentation algorithms developed for extracting welding flaws from digitized radiographic image have been presented. In this section, the proposed methods, MTFEGA and MEWT, are compared with each other. 25 radiographic images are segmented using these two methods. In 25 images, 544 defects are detected by human vision. The data in Table 4.2 summarizes the defect number segmented for each image. MTFEGA segments 473 defects and MEWT segments 374 defects.

The length and width of the defect are defined as the projection of the defect along the major axis and minor axis of the ellipse that has the same normalized second central moments as the potential region. Axes and orientation of the ellipse is shown in Figure 4.27.

The size of minimum defect that could be segmented and the contrast between the minimum defect that could be segmented and background for each image are measured. The results are summarized in Table 4.3, where the length, width and MCR are tabulated for these two methods. These results are illustrated in Figure 4.28.

Table 4.2: Segmented defect number

Defect Type	Image	Human Vision	MTFEGA	MEWT
Linear Porosity	1	27	25	20
	2	21	18	15
	3	12	10	7
Clustered Porosity	4	58	52	40
	5	49	43	28
	6	38	30	24
	7	20	11	8
Coarse Scattered Porosity	8	65	59	50
	9	47	45	38
	10	46	40	35
	11	26	23	17
Fine Scattered Porosity	12	38	33	26
	13	25	20	14
	14	18	13	10
	15	8	8	7
Lack of Fusion	16	18	16	12
	17	8	7	5
	18	3	3	3
Slag Inclusions	19	6	6	5
	20	5	5	4
	21	2	2	2
Crack	22	1	1	1
	23	1	1	1
	24	1	1	1
	25	1	1	1
	Total Number	544	473	374

The smallest defect segmented by MTFEGA method is 2.3 both in length and width. The smallest defect segmented by MEWT method is 5.7 in length and 3.8 in width. The size of

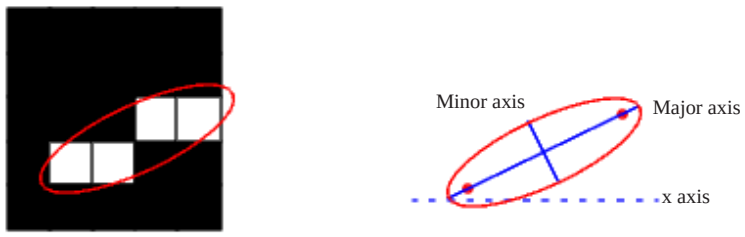


Figure 4.27: Axes and orientation of the ellipse

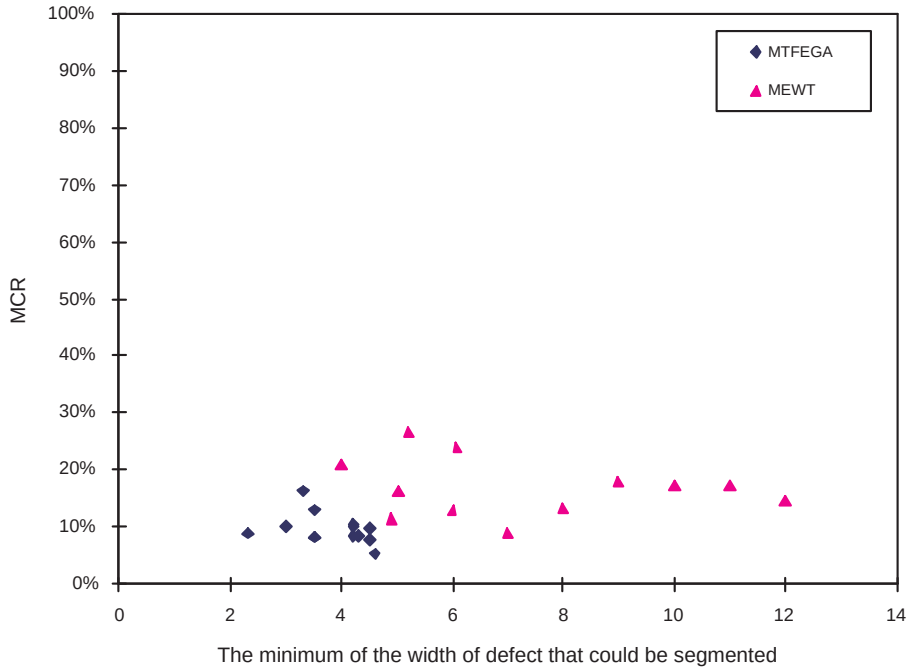


Figure 4.28: The segmented defects vs their sizes and contrast

minimum defect segmented by MTFEGA method is much smaller than the size by MEWT. The MCR of MTFEGA is 3.92% and the MCR of MEWT is 7.06%. The MTFEGA method can segment smaller defects with lower contrast compared to the MEWT method.

The drawback of these two methods is that the size of the defect segmented is a little different from the real size of the defect. From table 4.3, we can see that the size of the defect extracted by MTFEGA is smaller than real size of the defect and that the size of defect extracted by MEWT is bigger than the real size of the defect.

The experiments' results show the effectiveness of these two methods for image segmentation. The MTFEGA approach is a better choice for our dataset for segmenting low contrast

radiographic images with small size defects.

Table 4.3: Performance of MTFEGA and MEWT

Image	MTFEGA					MEWT				
	Real size		MTFEGA			Real size		MEWT		
	Length	Width	Length	Width	MCR	Length	Width	Length	Width	MCR
1	2.3	2.3	2.3	2.3	9.02%	6.3	5.2	8.3	7.2	26.67%
2	4.5	3.3	3.5	2.3	16.47%	8.0	6.1	8.6	7.5	23.92%
3	3.2	3.0	2.8	2.8	10.20%	7.8	4.9	9.2	6.4	11.37%
4	4.8	4.2	3.7	3.0	10.59%	7.8	5.7	7.3	6.4	16.47%
5	3.5	3.5	2.3	2.3	12.94%	6.6	5.5	7.9	7.2	12.94%
6	4.6	3.5	3.3	2.6	8.24%	5.2	5.2	8.6	6.8	9.02%
7	7.7	4.2	4.1	2.9	8.63%	7.3	6.4	8.5	7.4	13.33%
8	4.8	4.2	2.3	2.3	10.20%	6.7	6.7	7.7	6.4	17.25%
9	4.4	4.3	4.1	3.0	8.63%	7.6	5.3	7.7	7.3	17.25%
10	4.8	4.6	4.1	3.0	5.49%	7.3	5.7	7.3	6.9	14.51%
11	5.7	4.5	3.3	2.1	7.84%	5.8	4.6	8.0	7.2	17.65%
12	5.7	4.5	3.3	2.1	4.31%	6.9	6.1	7.9	7.6	7.06%
13	7.9	4.1	4.5	3.3	5.88%	7.4	6.5	8.9	7.4	19.61%
14	6.9	6.1	4.4	4.3	6.27%	10.1	7.2	8.3	7.6	6.27%
15	22.3	5.5	8.1	3.2	9.02%	26.6	7.0	25.1	5.9	20.78%
16	19.1	3.2	7.3	2.1	3.92%	27.1	3.6	15.7	2.3	10.20%
17	8.6	2.2	3.5	1.2	5.10%	16.0	4.8	16.9	7.3	12.55%
18	8.9	5.0	6.7	3.2	16.47%	8.9	5.0	10.1	7.0	16.47%
19	6.5	4.5	5.2	2.9	9.80%	6.8	5.4	7.1	6.3	20.78%
20	4.6	3.5	3.7	3.0	8.24%	5.6	5.2	8.2	7.8	18.04%
21	27.6	24.8	27.4	24.5	35.29%	27.6	24.8	35.0	30.5	35.29%
22	25.5	6.5	21.0	3.2	29.41%	25.5	6.5	27.2	8.0	29.41%
23	9.6	4.5	7.2	3.3	31.37%	9.6	4.5	10.1	6.4	31.37%
24	27.9	5.0	22.4	6.5	28.63%	27.9	5.0	21.3	3.3	28.63%
25	5.7	3.8	5.2	4.2	18.04%	5.7	3.8	10.7	9.4	18.04%

Chapter 5

SVM-based Feature Selection and Classification

5.1 Introduction

After extracting potential defects using the MTFEGA method, in order to perform an automated classification task, a set of measurements need to be taken from the potential discontinuity that must be classified. With the extraction of the set of features that will be used for the defect classification, the next step is the process of discriminating between defects and non-defects. In this study, a classification system based on statistical learning theory, called the support vector machine (SVM) is applied to the defect classification. Figure 5.1 shows the diagram of the system.

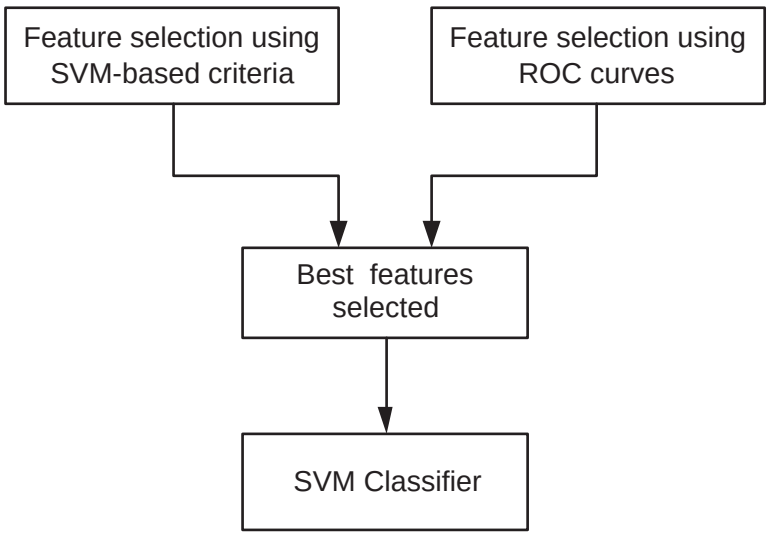


Figure 5.1: Classification system based on SVM

Support vector machines are a relatively new generation of techniques for classification and regression problems. The process of learning involves identification of a model based on the training data that are used to make predictions about unknown data sets.

Linear discrimination functions are the simplest form of discrimination functions. To overcome the limitation of only linear decision functions some attempts have been made to incorporate nonlinearity into the classical algorithm [107]. Hastie et al. [108] introduced the so called model of flexible discriminant analysis(FDA). FDA is reformulated in the framework of linear regression estimation and a generalization of this method is given by using nonlinear regression techniques. The proposed regression techniques implement the idea of using nonlinear mappings to transform the input data into a new space in which again a linear regression is performed.

In the case of too many predictors, such as the pixels of a digitized image, we do not want to expand the set: it is already too large. Breiman and Ihaka [109] proposed an idea to fit an linear discriminant analysis model, but penalize its coefficients to be smooth or otherwise coherent in the spatial domain, i.e. as an image. We call this procedure penalized discriminant analysis(PDA). With FDA itself, the expanded basis set is often so large that regularization is also required.

Hastie and Tibshirani [110] developed mixture discriminant analysis (MDA). MDA is to model each class by a mixture of two or more Gaussians with different centroids, but with every component Gaussian, both within and between classes, sharing the same covariance matrix. This allows for more complex decision boundaries, and allows for subspace reduction as in LDA.

A tree-structured method for classification and least-square regression was first proposed by Morgan and Sonquist [111] in order to handle interactions in survey data. Their idea was to recursively partition the data until a criterion is met, so that the partitions reduce prediction error. The computer program is called AID (Automatic Interaction Detection). Breiman et al. [109] proposed classification and regression tree (CART), which extended the basics of AID. Rather than stopping the partitions when a criterion is met. CART first grows a large tree, and then cuts it back to a smaller size. subsequently, a number of tree-structured methods based on CART have been developed. However, because AID and CART evaluate all possible partitions, they are computationally very expensive. Furthermore, such an exhaustive search may cause a bias in covariate selection because it is sensitive to outliers and prefers endcuts.

In real world applications these approaches have to deal with numerical problems due to the dimensional explosion resulting from nonlinear mappings. In the recent years approaches that avoid such explicit mappings by using kernel functions have some popular. The main idea is to construct algorithms that only afford dot products of pattern vectors which can be computed efficiently in high-dimensional spaces. Examples of this type of algorithms are the support vector machines (SVM).

SVM, first introduced by Boser et al. [112] and discussed in more detail in [40] [41], has its roots in statistical learning theory [39] which aims to create a mathematical framework for learning from input training samples with known identity and predict the outcome of data points with unknown identity. This results in two important theories. The first theory is called empirical risk minimization (ERM) where the aim is to minimize the learning or training error. The second theory is called structural risk minimization (SRM), which is aimed at minimizing the upper bound on the expected error over the whole dataset. SVMs are based on the SRM theory.

A SVM is a supervised learning technique based on the principle of optimal separation of classes. The goal is to find a linear separating hyperplane that separates the classes of interest provided the data is linearly separable. The optimum hyperplane provides the maximum margin between the two classes. For each class, the data vectors forming the boundary of classed are located on supporting hyperplanes - the term used in the theory of convex sets. Thus, these data vectors are called the “support vectors”. It is noteworthy that the data vectors located along the class boundary are the most significant ones form SVMs. Here is a simple example in 2D where a user is trying to separate data of two classes of samples (red circles and green squares) as shown in Figure 5.2.

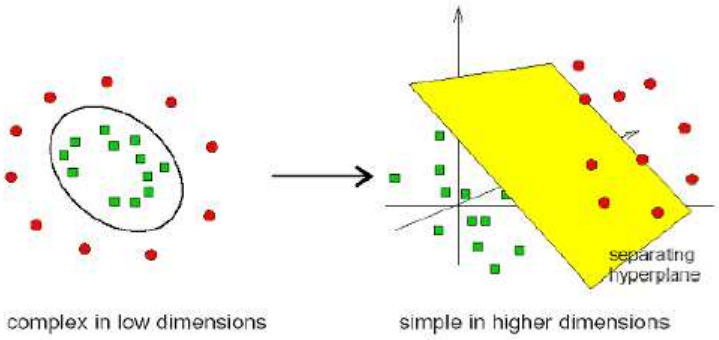


Figure 5.2: 2D example of separating data to two classes [1]

One of the main advantages of SVMs is that they are maximal margin classifiers. For

example, in Figure 5.3 on the following page, D is a better separator than A, B or C since it will be more likely to classify new samples that are close to the current decision boundary.

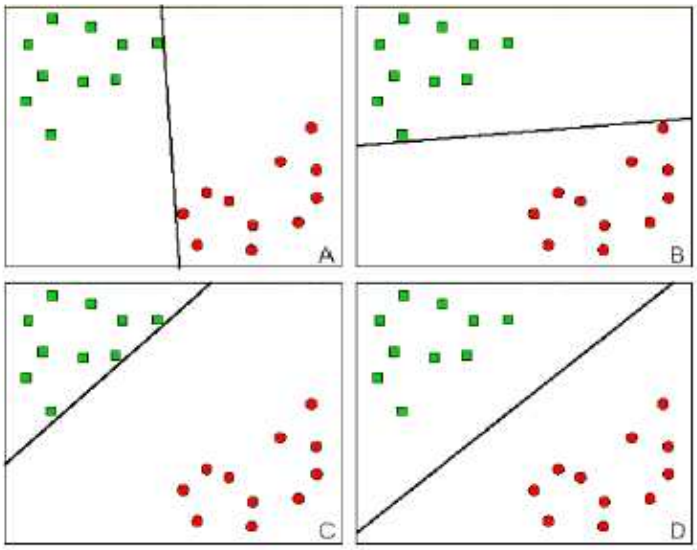


Figure 5.3: Maximal margin classifiers [1]

Two results make this approach successful:

1. The generalization ability of this learning machine depends on the VC dimension (a measure of the capacity of a learning algorithm) of the set of functions that the machine implements rather than on the dimensionality of the space. A function that describes the data well and belongs to a set with low VC dimension will generalize well regardless of the dimensionality of the space.

2. Construction of the classifier only needs to evaluate an inner product between two vectors of the training data. An explicit mapping into the high dimensional feature space is not necessary.

5.2 Defect Features

Most researchers extract the morphological features associated with the shape and size of the defects. Mery [23] proposed to extract texture features. In this study, we will extract these groups features: texture features and morphological features.

5.2.1 Texture-based Features

For our problem, the description of texture of the defect in the radiographic image, two texture features: co-occurrence matrix and Gabor filter are extracted.

Co-occurrence Matrix

There are several reasons to choose co-occurrence matrix as one method for our analysis:

1. The defects's positions are random. It is impossible to predict where the defects appear.
2. The background is uniform. The pattern of optical density is rather random. The defects are not the only dark region in the image. Some parts of the image may have a darker gray level.
3. The size of defects is small. The smallest defect we deal with has the length or width around 2-5 pixels.
4. Co-occurrence matrices have the significant advantage that they describe the pixels' spatial inter-relationships in an image in a way that is unaffected by monotonic gray level transformations.

A co-occurrence matrix is a square matrix with elements corresponding to the relative frequency of occurrence of pairs of gray level of pixels separated by a certain distance in a given direction. Formally, the elements of a G gray level co-occurrence matrix P_d for a displacement vector $d = (dx, dy)$ is defined as [69]:

$$P_{\mathrm{d}}(i,j) = |\{((r,s),(t,v)) : I(r,s) = i, I(t,v) = j\}| \tag{5.1}$$

where I denotes an image of size $N \times N$ with G gray levels, $(r,s),(t,v) \in N \times N$, $(t,v) = (r+dx, s+dy)$ and $|.|$ is the cardinality of a set. Figure 5.4 shows the four directions of the co-occurrence matrix. Figure 5.6 shows co-occurrence matrices for Figure 5.5.

Haralick [24] proposed 14 measures of textural features which are derived from the co-occurrence matrices, and each represents certain image properties such as coarseness, con-

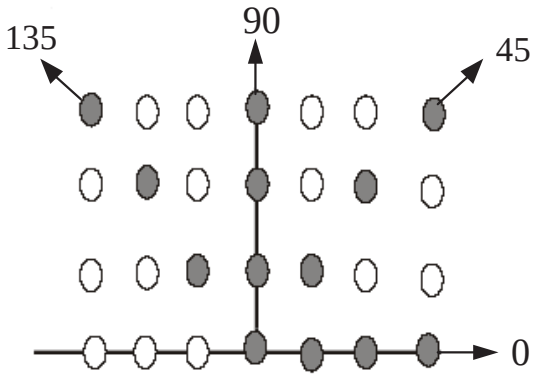


Figure 5.4: The four directions for the co-occurrence matrix

1	1	2	2
1	1	2	2
3	3	1	1
3	3	1	1

Figure 5.5: A 4×4 image with 3 gray-level values 1-3

Grey level	1	2	3
1	4	2	0
2	0	2	0
3	2	0	2

(a)

Grey level	1	2	3
1	2	2	0
2	0	1	0
3	2	1	1

(b)

Grey level	1	2	3
1	4	2	0
2	0	2	0
3	2	0	2

(c)

Grey level	1	2	3
1	3	1	1
2	1	1	0
3	1	0	1

(d)

Figure 5.6: Co-occurrence matrices for Figure 5.5: (a) 0° , (b) 45° , (c) 90° , (d) 135°

trast, homogeneity and texture complexity. Those used in this work for extracting features

in the defect detection of radiographic images are:

1) Shannon Entropy:

$$f_E = -\sum_i \sum_j p(i,j) \log p(i,j) \tag{5.2}$$

Entropy measures the complexity of the image. Complex textures tend to have higher entropy.

2)Contrast:

$$f_C = -\sum_i \sum_j (i-j)^2 p(i,j) \tag{5.3}$$

Contrast feature is a measure of the image contrast or the amount of local variations present in an image.

3)Angular Second Moment:

$$f_A = -\sum_i \sum_j \{p(i,j)\}^2 \tag{5.4}$$

The angular second moment, also called energy, is a measure of textural uniformity of an image. Energy reaches its highest value when gray level distribution has either a constant or a periodic form.

4)Inverse Difference Moment:

$$f_I = -\sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j) \tag{5.5}$$

The inverse difference moment measures image homogeneity. Hence it is suitable measure for detection of disorders in image. For homogeneous textures value of angular second moment turns out to be small compared to non-homogeneous ones.

In Equations (5.2)-(5.5), $p(i,j)$ refers to the normalized entry of the co-occurrence matrices. That is $p(i,j) = P_{\mathbf{d}}(i,j)/R$, where R is the total number of pixel pairs (i,j). For a displacement vector $\mathbf{d} = (dx, dy)$ and image of size $M \times N$, R is given by $(M-dx)(N-dy)$.

In this work, the texture features are extracted for four directions ($0°$, $45°$, $90°$ and $135°$) for distance $\mathbf{d} = 1$.

Gabor Filter

A common technique for implementing multiresolution analysis is the wavelet transforms. However, wavelet bases are shift invariant and, therefore, it is difficult to characterize a texture pattern from the wavelet coefficients since the wavelet descriptors depend on pattern location [113]. Gabor filters can also decompose the image into components corresponding to different scales and orientations. Gabor filters achieve optimal joint localization in spatial and spatial frequency domain [35].

In this work, measurements in both the spatial and the spatial frequency domains are considered. As mentioned above, Gabor filters have the ability to perform multi-resolution decomposition due to its localization both in spatial and spatial frequency domain. Filters with smaller bandwidths in the spatial-frequency domain are more desirable because they allow us to make finer distinctions among different textures. On the other hand, accurate localization of texture boundaries requires filters that are localized in the spatial domain. However, normally the effective width of a filter in the spatial domain and its bandwidth in the spatial-frequency domain are inversely related according the uncertainty principle. That is why Gabor filters are well suited for this kind of problem.

In the spatial domain, the Gabor function is a complex exponential modulated by a Gaussian function. The Gabor functions are a complete (but nonorthogonal) basis set given by

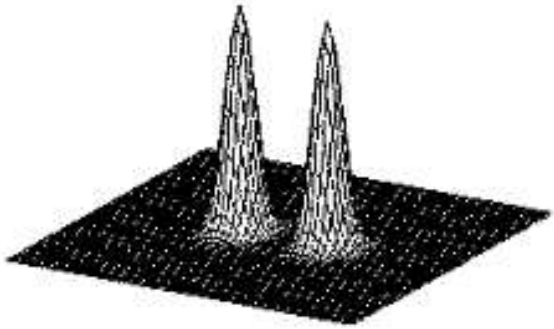
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)\right]\exp(2\pi j u_0 x) \tag{5.6}$$

where σ_x and σ_y denote the Gaussian envelope along the x and y axes, and u_0 defines the radial frequency of the Gabor function.

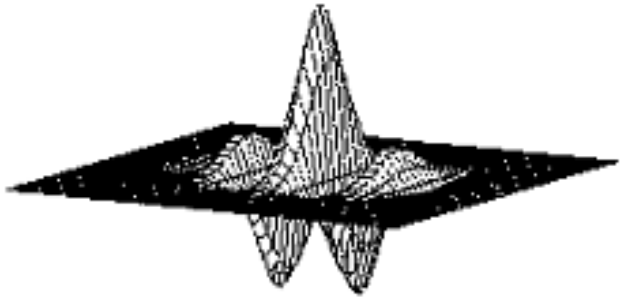
In the frequency domain, the Gabor function acts as a bandpass filter and the Fourier transform of $f(x,y)$ is given by

$$F(u,v) = \exp\left\{-\frac{1}{2}\left[\frac{(u-u_0)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right]\right\} \tag{5.7}$$

where $\sigma_u = \frac{1}{2\pi\sigma_x}$ and $\sigma_v = \frac{1}{2\pi\sigma_y}$. An example of a Gabor filter in spatial domains is given in Figure 5.7.



(a)



(b)

Figure 5.7: A directional Gabor filter in the frequency (a) and spatial(b) domains

Most Gabor filters have small response to regions of uniform luminance. This results in sensitivity to background luminance levels, which signifies a first order difference between regions. This can be corrected by subtracting the mean pixel value of each filter from each of pixel value. In the frequency domain this results in the subtraction of a sinc function with the same value as the filter at the origin. Unfortunately, the sinc function induces ripples in the filter's response, especially at low frequencies. However, if the amplitude of the sinc is sufficiently low then these ripples are not detectable in the spectral plots for the filters. A Gabor filter is built in the Fourier domain, and $F(u, v)$ is kept zero at $u = v = 0$. This ensures that the filters do not respond to regions with constant intensity.

In addition to radial frequency and orientation, frequency and orientation bandwidths are also very important. The filter parameters are decided such that the space frequency plane is covered nearly uniformly. In our implementation, we have used the following scheme to cover the frequency plane nearly uniformly [114], here m is scale n is orientation.

$$f_{mn}(x,y) = \alpha^{-m} f(x',y') \tag{5.8}$$

$$x' = \alpha^{-m}(x \cdot \cos\theta + y \cdot \sin\theta) \tag{5.9}$$

$$y' = \alpha^{-m}(-x \cdot \sin\theta + y \cdot \cos\theta) \tag{5.10}$$

where $\alpha > 1$, $m, n = interger$. $\theta = n\pi/K$ and K is the total number of orientations. The scale factor α^{-m} ensures that the energy is independent of m (scale)

$$E_{mn} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f_{mn}(x,y)|^2 dxdy \tag{5.11}$$

Let U_l and U_h denote the lower and upper center frequencies of interest. Let K be the number of orientations and S be the number of scale in the multiresolution decomposition. The the design strategy is to ensure that the half peak magnitude cross-sections of the filter responses in the frequency spectrum touch each other. This results in the following formulas for computing the filter parameters σ_u and σ_v.

$$\alpha = (U_h/U_l)^{1/(S-1)} \tag{5.12}$$

$$\sigma_u = \frac{(\alpha-1)U_h}{(\alpha+1)\sqrt{2ln2}} \tag{5.13}$$

$$\sigma_v = \tan(\frac{\pi}{2K})[U_h - 2\ln 2(\frac{\sigma_u^2}{U_h})][2\ln 2 - \frac{(2\ln 2)^2\sigma_u^2}{U_h^2}]^{-1/2} \tag{5.14}$$

where $m = 0, 1, ..., S-1$.

Each of the complex Gabor filters has the real and imaginary parts that are conveniently implemented as the spatial mask $L \times L$ sizes. In order to have a symmetric region of support, L should be an odd number. In our work, the filter parameters used are $\alpha = 2$,

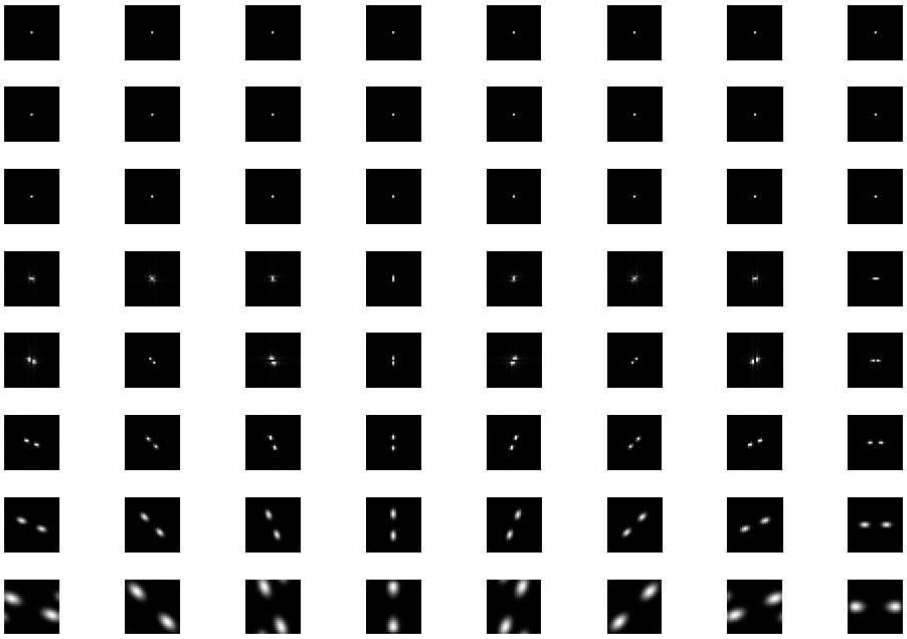


Figure 5.8: Gabor kernels at 8 orientations and 8 frequencies in the frequency domain

$U_h = 0.4$, $K = 8$, and $S = 8$. Figure 5.8. shows the Gabor kernels at 8 orientations and 8 frequencies in frequency domain.

The Gabor filters are applied to each segmented window W that contains the hypothetical defect and its surrounding. The filter windows G_{mn} are computed using the 2D convolution of the window W of the radiographic image with the Gabor mask as follows:

$$G_{mn} = \{[f_{mn}(x,y)_r * W(x,y)]^2 + [f_{mn}(x,y)_i * W(x,y)]^2\}^{1/2} \tag{5.15}$$

where $*$ denotes the 2D convolution operation, and $f_{mn}(x,y)_r$ and $f_{mn}(x,y)_i$ represent the real and imaginary parts of the Gabor filter separated from 5.8. The Gabor features, denoted by g_{mn}, are defined as the average output of G_{mn}, ie, it yields $K \times S$ features for each segmented window:

$$g_{mn} = \frac{1}{p_w q_w} \sum_{i=1}^{p_w} \sum_{j=1}^{q_w} G_{mn}(i,j) \tag{5.16}$$

where the size of the filtered windows G_{mn} is $p_w \times q_w$.

5.2.2 Morphological Features

This category features descriptors comprises area, length, shape factors of potential defects.

1. Area

Area is defined as the number of pixels interior to or on the potential defect boundary.

2. Length

Length is the projection of the potential defect along the major axis of the ellipse that has the same normalized second central moments as the potential region.

3. Width

Width is the projection of the potential defect along the minor axis of the ellipse that has the same normalized second central moments as the potential region.

4. Elongation

Equation (5.17) computes the ratio between the length and width of the potential defect, which takes a valued between 0 and 1.

$$elongation = \frac{width}{length} \tag{5.17}$$

5. Orientation

The orientation of the potential defect stretching in the weld is calculated as the angle (in degrees) between the horizontal line(x-axis) and the major axis.

6. Ratio of width to area (RWA)

Equation (5.18) computes the ratio between the width and the area of the potential defect.

$$RWA = \frac{width}{area} \tag{5.18}$$

7. Compactness

This feature measures the object shape that is calculated by

$$compactness = \frac{perimeter^2}{area} \tag{5.19}$$

where perimeter is the number of boundary points around the defect area. A circular object has a smaller compactness valued than a non-circular object.

5.2.3 Summary

The complete set of features is shown in Table 5.1. For each potential discontinuity, 87 features are extracted.

Table 5.1: Extracted features for defect detection

Feature No.	Feature description
From f_1 to f_{64}	Gabor features
From f_{65} to f_{80}	Co-occurrence matrix features
From f_{81} to f_{87}	Morphological features

5.3 Design of Support Vector Machines

In this section, the SVM will be designed for feature selection and classification. Based on statistical learning, Vapnik [45] [115] formulated the SVM. It is based on SRM to achieve the goal of minimizing the simultaneous bounds on the VC-dimension and the empirical risk. SVM claims to guarantee generalization, i.e., the decision rule reflects the regularities of the training data rather than the incapabilities of the learning machine. It also allows various other learning machines to be constructed under a unified framework, hence simplifying comparisons and promoting understanding.

Let $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{ni})^T$, $i = 1, ..., M$, be a sample of $\mathbf{x} \in \mathbb{R}^n$ and belong to Class I (defect) or Class II (non-defect). For linearly separable data, it is possible to determine a hyperplane that separates the data leaving one class on one side of the hyperplane, the other on the other side. This plane can be described by the equation:

$$f(x) = \mathbf{w}^T\mathbf{x} + b = \sum_{j=1}^{n} w_j x_j + b = 0 \tag{5.20}$$

where $\mathbf{w} \in \mathbb{R}^n$ is a weight vector and b is a scalar. The vector $\mathbf{w}$ and the scalar b determine the position of the separating hyperplane.

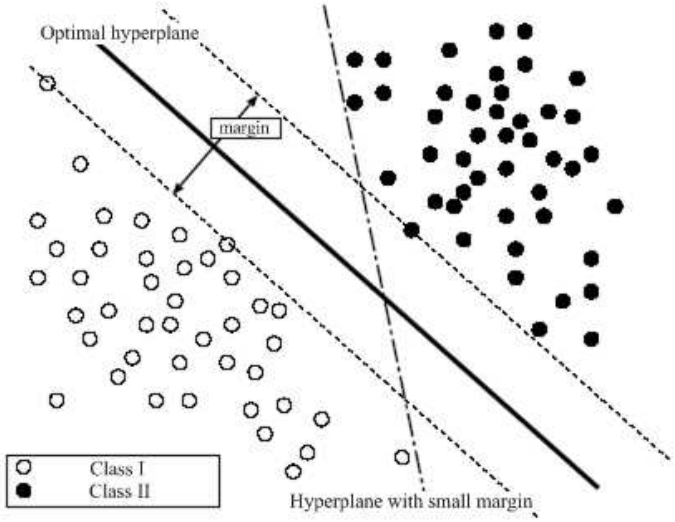


Figure 5.9: Overview of SVMs in linear separable case [1]

Figure 5.9 shows an overview of SVMs in the linear separable case. SVMs optimize parameters $\mathbf{w}$ and b based on maximum margin strategy. The margin is defined as the distance between two separate hyperplanes which is equal distant from and a parallel with the optimal hyperplane, like hyperplane in Figure 5.9. The strategy theoretically guarantees low generalization error for unknown example even in a high dimensional feature space.

The margin can be written as norm of $\frac{2}{\|\mathbf{w}\|}$. Thus, we can find an optimal hyperplane maximizing margin by solving the following problem:

$$\begin{aligned} &minimize && \frac{1}{2}\|\mathbf{w}\|^2 \\ &subject\ to && y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \end{aligned} \tag{5.21}$$

The solution of this problem is obtained using the Lagrangian theory:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i(y_i(\mathbf{w}\cdot\mathbf{x_i} + \mathbf{b}) - \mathbf{1}) \tag{5.22}$$

The corresponding dual is found by differentiating with respect to $\mathbf{w}$ and b,

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \tag{5.23}$$

$$\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{5.24}$$

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \tag{5.25}$$

However, in practice, data sets are not linearly separable due to noise and mixture of classes during the selection of training data. So, it is impossible in practice to create a linear separating hyperplane to separate classes of defect and non-defect without any misclassification error for a given training data set. This problem can be tacked by using a soft margin classifier [116] [117]. Soft margin classification relaxes the requirement that every data point belonging to the same class must be located on the same side of a linear separating hyperplane. It introduces slack variable ξ to take into account the noise or error in the dataset due to misclassification.

$$\begin{aligned} minimize \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \xi_i \\ subject\ to \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \end{aligned} \tag{5.26}$$

where $\xi_i \geq 0$ for all i. This new formulation trades off the two goals of finding a hyperplane with large margin (minimizing $\|\mathbf{w}\|$), and finding a hyperplane that separates the data well (minimizing the ξ_i). The parameter C controls this trade-off. Figure 5.10 illustrates the soft margin SVM.

Often, a linear separating hyperplane is not able to classify input data without some error. Under such circumstances, the data are transformed to a higher dimensional space using a non-linear transformation that spreads the data apart such that a linear separating hyperplane may be found. However, due to very large dimensionality of the feature space, it is not practical to compute the inner product of two transformed data vectors. Nevertheless, this may be achieved by using a kernel trick [118] instead of explicitly computing

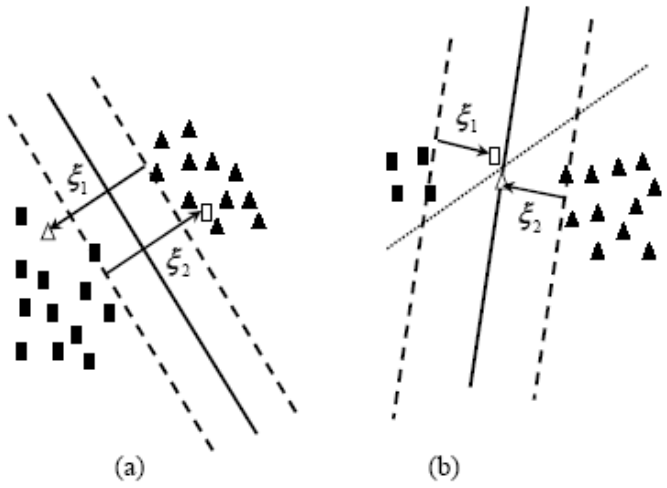


Figure 5.10: (a) The data points are not linearly separable. The solid black line is the SVM solution. The white triangle and the white rectangle are misclassified. The slack variables designate the distance of these points from the dashed lines for the corresponding classes. (b) The classes are separable. The dotted line is the solution when the tradeoff parameter C is very large (e.g., infinite), and this gives us the maximum margin classifier for the separable case. If the tradeoff parameter is small, then one allows errors (given by the two slack variables), but one gets a much larger margin.

transformations in the feature space. The kernel function is played by substituting a kernel function in place of the inner product of two transformed data vectors. The use of the kernel trick reduces the computational effort by significant amount. Using the kernel trick makes the maximum margin hyperplane fit in a feature space. The feature space is a non-linear map from the original input space, usually of much higher dimensionality than the original input space. In this way, non-linear SVMs can be created [119]. Support vector machines non-linearly map their n-dimensional input space into a high dimensional feature space. In this high dimensional feature space a linear classifier is constructed.

The basic idea of kernel trick is to nonlinearly map the data to a feature space of high or possibly infinite dimensions, $\mathbf{x} \rightarrow \phi(\mathbf{x})$, then compute a decision function of the form:

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \tag{5.27}$$

We then apply the linear SVM algorithm in this feature space. A linear separating hyperplane in the feature space corresponds to a nonlinear surface in the original space. We can now rewrite Equation 5.26 using the data points mapped into the feature space, and we obtain Equation 5.28.

$$minimize \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i \tag{5.28}$$

$\xi_i \geq 0$ for all i, where the vector $\mathbf{w}$ has the same dimensionality as the feature space and can be thought of as the normal of a hyperplane in the feature space.

The solution of this problem is obtained using the Lagrangian theory and one can prove that vector $\mathbf{w}$ is of the form:

$$\mathbf{w} = \sum_{i=1}^{l}\alpha_i^* y_i \phi(\mathbf{x_i}) \tag{5.29}$$

where α_i^* is the solution of the following quadratic optimization problem:

$$maximize \;\; W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C}\delta_{ij}$$

$$subject\ to \qquad \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{5.30}$$

where $\alpha_i \geq 0 \quad i = 1, ..., l$, δ_{ij} is the Kronecker symbol. $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{5.31}$$

Some of the commonly used kernels include Linear, Gaussian RBF (Radial Basis Functions), polynomial functions, and sigmoid polynomials as shown in Table 5.2.

The design of an SVM for a classification task consists of two tasks: choosing the kernel function and setting a value for the parameter C. The parameter C is also called an error penalty, because it deals with the trade-off between maximum margin and classification

Table 5.2: Some kernel functions

Kernel Function	Definition	Parameters
Linear	$\mathbf{x} \cdot \mathbf{x}_i$	
Gaussian RBF	$exp(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2})$	σ is a user defined value
Polynomial of degree d	$(\mathbf{x} \cdot \mathbf{x}_i + 1)^d$	d is a positive integer
Sigmoid	$\tanh(k(\mathbf{x} \cdot \mathbf{x}_i) - \theta$	k and θ are user defined values

error during training. A high error penalty will force the SVM training to avoid classification errors. It is clear that with high error penalty, the optimizer gives a boundary that classifies all the training points correctly. This, however, can give very irregular boundaries that may not lead to good performance of the classifier in the test set. In this study, we choose $C = 100$. Comparing with a lower error penalty $C = 1$ and a higher error penalty $C = 1000$, $C = 100$ can reach a good trade-off between maximum margin and the classification error. The selection of kernel function also has influence on the decision boundary. Usually an RBF kernel is favored, because they are not sensitive to outliers and do not require inputs to have equal variances. So, we choose the Gaussian RBF kernel function.

5.4 Feature Selection based on SVM and ROC

The purpose of feature selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generates the data. There are many potential benefits of variable and feature selection: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, and defying the curse of dimensionality to improve prediction performance.

In this study, an SVM-based feature selection algorithm is applied to select features. However, this algorithm relies on a backward feature selection, which is computationally tractable but not necessarily optimal. We improve the performance of the feature selection result by combining SVM-based feature selection with an ROC feature selection process.

5.4.1 Feature Selection Methods

Branch and Bound Methods

The simplest and most commonly used feature selection methods are known as branch and bound techniques [120] [121]. These techniques are guaranteed to converge to the optimal set of a given number of features if the measure used to compare different sets is a monotonic function. That means if, for a feature set F the suitability measure S has a value $S(F)$ then, for any other set F', with $F \subset F'$, $S(F')$ must satisfy

$$S(F') \leq S(F) \tag{5.32}$$

A consequence of this inequality is that the value of the measure S will always increase (the value of a corresponding error function will always decrease) by including one more feature in the proposed set of features and , therefore, this monotonicity does not allow the determination of the optimal size of the required feature space: the minimum value of the error function will only be reached if all the possible features have been included. However, the error function allows the comparison between features sets with the same number of features.

Given the monotonicity of the error function, some criterion must be used for the determination of the optimal set of features. Typically, this criterion is chosen as follows:

1. The asymptotic KNN error measure.
2. The Bayes' error.
3. The Bhattacharyya distance between classes [122].
4. The Kolmogoroff-Smirnoff distance between classes [120].

The branch and bound search technique works in the following way: let N be the total number of features available and N^* the optimal number of features. For one random set of N^* features, say $F_{m,0}$, the value of the suitability measure $S(F_{m,0})$ is calculated and it is set as a threshold. Assuming that the best feature set will be the one with the highest corresponding value of S, a search is initiated by forming all the possible feature sets of population $N - 1$. For each one of these, the value of E is calculated. If one or

more of these feature sets give rise to a value of S lower than the threshold, the all the subsets of these feature sets are excluded from any further search as they are guaranteed to present a suitability criterion value lower than $S(F_{m,0})$. From the other feature sets of population $N-1$, the one with the highest S is selected and all of its subsets of containing $N-2$ features are considered. The process is repeated until all the possible feature sets of between N^* and N features have been either examined or excluded. Every time a set with exactly N^* features is found with a higher than the threshold value of S, it becomes the currently selected feature set and its corresponding S valued the new threshold value.

This search algorithm is guaranteed to converge to the optimal feature set for the particular suitability criterion, offering significant savings compared to exhaustive search. However, when the number of features is large, even the branch and bound technique is computationally expensive. Also, the suitability criteria usually require the estimation of the class-conditional probability density functions which, for large numbers of features, is impractical.

Forward Feature Selection Method

This feature selection method starts by assuming an empty initial set of features. Then each feature in the data set is examined separately and the one which provides the best classification results on its own is selected as the first feature. In the second iteration all the possible combinations of the first feature with each one of the remaining features is examined and the pair of features giving the best classification results is kept. This process continues until such time that no further addition to the selected feature set results in a significant classification improvement.

Two processes need to be determined in a forward feature selection method: the evaluation of the classification results and the termination criterion. For the first process, an obvious choice is the fraction of samples correctly classified, or some other measure such as class separation, while the termination criterion is usually a statistical significance test.

The sub-optimality of the forward feature selection is due to the fact that a number of features, when considered together, may provide discriminatory properties, while when considered individually, they may perform very poorly [123]. Therefore, they may never be selected by the forward selection method, which will instead only indicate a local maximum of optimality.

Backward Features Selection Method

The backward feature selection is the opposite of the forward method. In fact, it is a more accurate feature elimination method. Initially, the entire training data set is considered and the classification performance of the complete set of features is evaluated. The same process is repeated $N-1$ times for the $N-1$ feature sets derived from the complete set with the exclusion of one feature at a time. The feature whose absence from the feature set has the least effect on the classification performance is eliminated and the process continues with the consideration of the $N-2$ possible feature sets formed by excluding the eliminated feature and one of the remaining $N-2$ features at a time. This process of feature elimination continues until none of the remaining features can be eliminated without causing a statistically significant decrease in performance.

The backward feature selection does not suffer from the drawback of the forward selection described earlier. A backward sequential selection has lower computational complexity compared to randomized or exponential algorithms and it has optimality in the subset selection problem [124]. However, it is sub-optimal method as well, for other reasons: the termination criterion may stop the variable elimination process due to a temporary performance degradation because of a local energy function minimum. Also, in the case where highly correlated features exist in the feature set and one of them is eliminated, it may be a feature that would actually prove more useful in a smaller feature set. A final disadvantage of the backwards feature selection method is that it may be much more intensive computation than the forwards one, as it considers larger feature set.

5.4.2 Feature Selection using SVM-based Criteria

In this study, a backward feature selection method, SVM-based feature selection algorithm is applied. The SVM is provided with many statistics that allow one to estimate their generalization performance from bounds on the leave-one-out error L. The leave-one-out error is the number of classification errors produced by the leave-one-out procedure which consists in learning a decision function from $m-1$ examples, testing the remaining one and repeating until all elements served as test example. The leave-one-out error is known to be an unbiased estimator of the generalization performance of a classifier trained on $m-1$ examples. One of the most common L error bounds for SVM is the radius/margin bound (for decision function with non-zero bias b) [40]:

$$L \leq 4R^2\|\mathbf{w}\|^2 \tag{5.33}$$

where R is the radius of the smallest sphere that contains all the mapped data $\phi(\mathbf{x}_i)$. A tighter bound named "span estimate" is also available and is based on the distance S_p between a mapped support vector $\phi(\mathbf{x}_p)$ and the span of all other support vectors [125]. The following equation holds:

$$L \leq \sum_p \alpha_p^* S_p^2 \tag{5.34}$$

where S_p^2 for SVM with quadratic slack variable ξ, is related to the extended matrix of the dot product between support vectors

$$\tilde{K}_{SV} = \begin{pmatrix} K & 1 \\ 1^T & 0 \end{pmatrix} \tag{5.35}$$

The SVM-based feature selection algorithm is as follows:

1. Initialization: $Ranked = [\,]$; $Var = [1, ..., N]$

2. Repeat

1) train an SVM classifier with all the training data and the variables Var.

2) for all variables in Var, evaluate the ranking criterion $R_c(i)$ of variable i end for

3) $best = argmin_i R_c$

4) rank the variable that minimizes $Rc:\ Ranked = [best\ Ranked]$

5) remove the variable that minimizes R_c from the selected variables set: $Var = [1, ..., best-1, ..., N]$

3. Until Var is empty

It requires a ranking criterion to rank variables. The problem of searching for the "best" r variables is solved by means of a greedy algorithm based on backward selection [126].

Hence, the algorithm starts with all features and repeatedly removes a feature until r features are left or all variables have been ranked. Rakotomamonjy [127] presented different criteria for variable selection algorithms. This criteria are derived from the generalization error bounds of SVM theory: weight vector norm $\|\mathrm{w}\|^2$ and upper bounds of the *leave-one-out* error. Drawing inspiration from the neural networks community, a zero-order method and a first-order method were proposed for each criterion. For the zero-order method, the criterion C_t is directly used for variable ranking, and the method consists in identifying the variable that produces the smallest valued of C_t when removed. The ranking criterion then becomes $R_c(i) = C_t^{(i)}$ with $C_t(i)$ being the criterion valued when variable i has been removed. The first-order method uses the derivatives of the criterion C_t with respect to a variable. The ranking criterion is $R_c(i) = |\nabla C_t|$. This method differs from the zero-order method because a variable is ranked according to its influence of the criterion which is measured with the absolute valued of derivative. We have investigated the experimentally performance of the different criteria. The zero-order $\|\mathrm{w}\|^2$ criteria performs consistently well over the dataset we used.

The zero-order $\|\mathrm{w}\|^2$ ranking term is:

$$R_c(i) = \left\|\mathrm{w}^{(i)}\right\|^2 = \sum_{k,j} \alpha_k^{*(i)} \alpha_j^{*(i)} y_k y_j K^{(i)}(\mathrm{x}_k, \mathrm{x}_j) \tag{5.36}$$

where $K^{(i)}$ is again the gram matrix of the training data when the variable i has been removed. In order to reduce time complexity we consider the parameters $\alpha_k^{*(i)}$ equal to α_k^* during the evaluation of $R_c(i)$.

The top 12 features are ranked as shown Table 5.3.

Table 5.3: Top 12 Features Selected using SVM

Ranked No.	1	2	3	4	5	6	7	8	9	10	11	12
Feature	f_{65}	f_{28}	f_{12}	f_{71}	f_{72}	f_{77}	f_{55}	f_{83}	f_{84}	f_{69}	f_{81}	f_5

This algorithm relies on a backward feature selection, which is computationally tractable but not necessarily optimal. Next we improve the performance of the algorithm by combining with the ROC feature selection process.

5.4.3 Feature Selection using the Receiver Operating Characteristic Curves

ROC (receiver operating characteristic) analysis is commonly used to measure the performance of a classifier [128]. An ROC curve depicts the true-positive rate versus the false-positive rate of a classifier for varying classification decision thresholds. In a classification attempt of a two-class data set we can define four rates. If we label one of the classes positive and the other negative the four rates are:

1. True positive rate (TPR), the percentage of positive instances which have been correctly classified as positive.
2. False positive rate (FPR), the percentage of negative instances which have been wrongly classified as positive.
3. True negative rate (TNR), the percentage of negative instances, correctly classified as negative.
4. False negative rate (FNR), the percentage of positive instances, wrongly classified as negative.

It is clear that

$$\begin{aligned} FNR &= 1 - TPR \\ TNR &= 1 - FPR \end{aligned} \tag{5.37}$$

Hence a description of classifier performance may be obtained from the trade-off between either TPR and FPR or TNR and FNR. The defect detection involves problems with only two possible outputs: defect and non-defect. The defect instances are labeled positive while the non-defect ones are labeled negative.

An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC. The AUC is a reasonable performance statistic for classifier systems assuming no knowledge of the true ratio of misclassification costs. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between (0,0) and (1,1), which has

an area of 0.5, no realistic classifier should have an AUC less than 0.5. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

In our case, each feature is analyzed independently using a threshold classifier. For example, Figure 5.11 and Figure 5.12 show the areas under two ROC curves, feature f_{11} and feature f_{87}. Feature f_{11} has greater area and therefore better average performance.

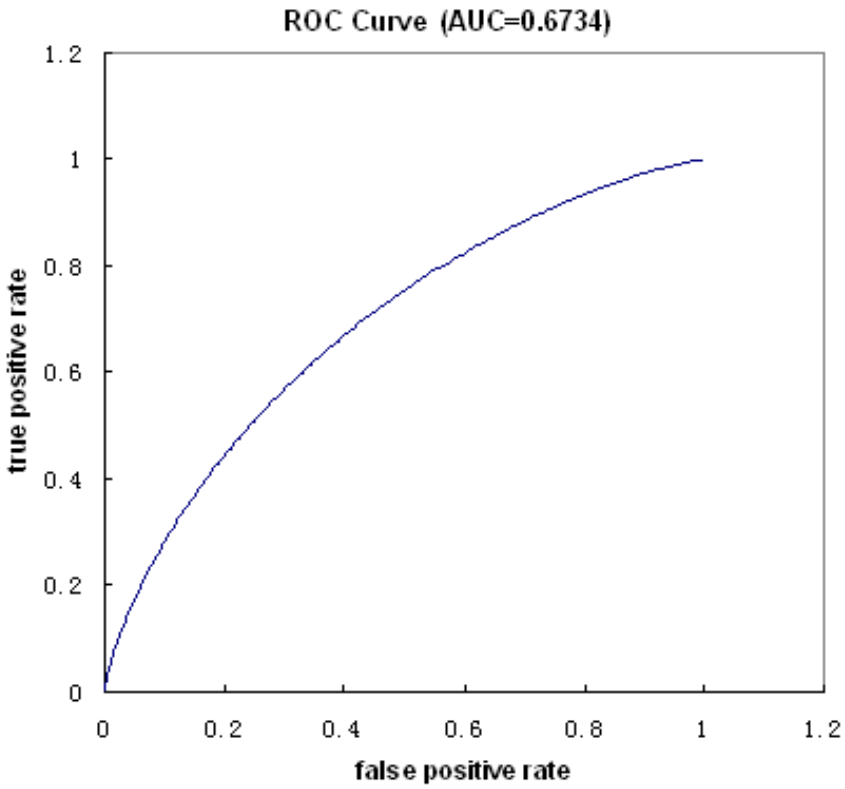


Figure 5.11: ROC of feature f_{12}

Table 1 presents the top 12 valued features obtained by computing the AUC in our data set.

The complete set of selected features is shown in Table 5.5. 16 features are selected by combining the top 12 features selected using SVM and best 12 features selected using ROC.

Table 5.4: ROC analysis

Gabor feature	AUC	Co-occurrence matrix feature	AUC	Morphological feature	AUC
f_5	0.6753	f_{67}	0.6852	f_{81}	0.6977
f_7	0.6751	f_{69}	0.6616	f_{82}	0.6585
f_8	0.67621	f_{71}	0.6792	f_{83}	0.6920
f_{12}	0.6734	f_{79}	0.6852	f_{84}	0.6884

Table 5.5: Selected features

Feature type	Selected feature No.
Gabor features	f_5, f_7, $f_{8,}$, f_{12}, f_{28}, f_{12}, f_{55}
Co-occurrence matrix features	f_{65}, f_{67}, f_{69}, f_{71}, f_{72}, f_{77}, f_{79}
Morphological features	f_{81}, f_{82}, f_{83}, f_{84}

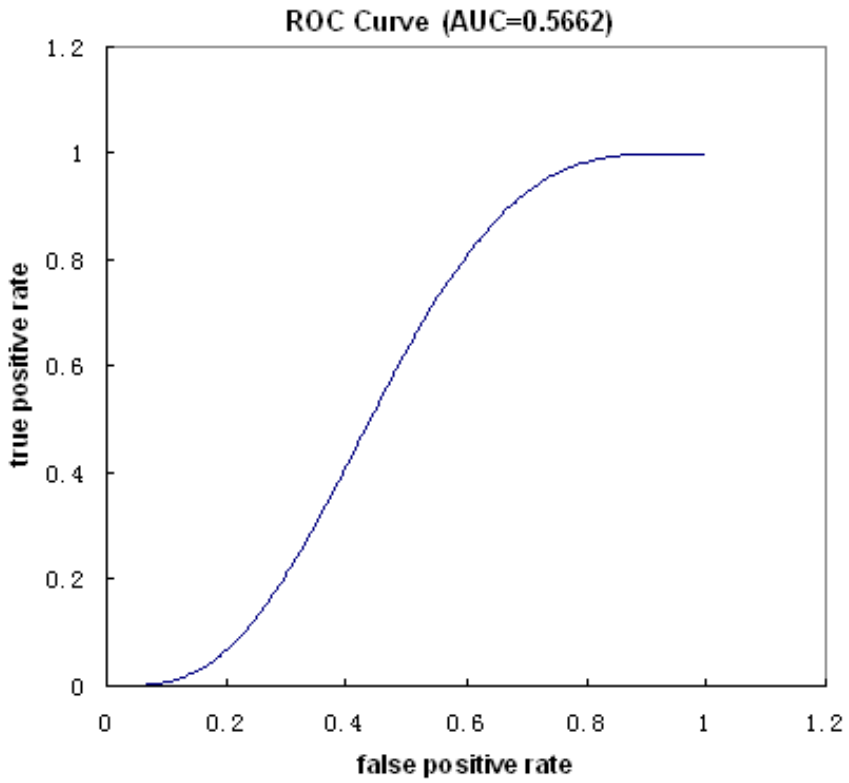


Figure 5.12: ROC of feature f_{87}

5.5 Classification Result using SVM

In this study, 25 radiographic weld images are analyzed. In the segmentation stage, 809 potential discontinuities are obtained, of which 473 are real defects. Types and distribution of welding defects is shown in Figure 5.13. For each potential discontinuity, 87 features are extracted, and only 16 of them are selected. 14 images with 475 potential defects are selected and used to train the SVM classifier designed with a Gaussian RBF kernel. Special attention is paid to ensure that instances of all different defects types are included. The remaining 11 images with 334 potential defects are used to test the classifier. Using this method, 97.99% of the existing flaws are detected with 14.81% of false alarms.

In Table 5.6, the classification results with the selected best sixteen features and no feature selection are compared. The ROC curve obtained by the SVM classification with the sixteen selected features is shown in Figure 5.14. The ROC curve obtained by the SVM classification without feature selection is shown in Figure 5.15. We can see that feature selection can eliminate irrelevant variables to enhance the generalization performance of a given learning algorithm.

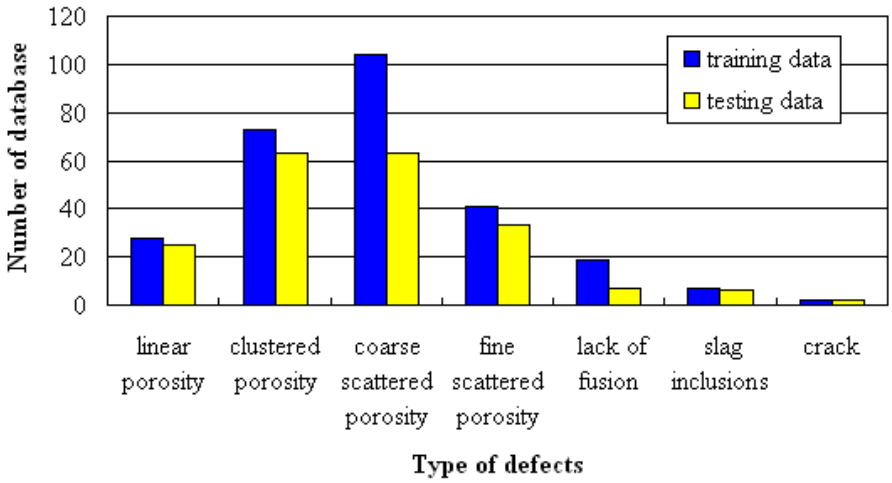


Figure 5.13: Types and distribution of defects

Furthermore, in Table 5.7 the SVM classifier designed with a Gaussian RBF kernel and an SVM classifier designed with polynomial kernel are compared. The best performance is obtained by Gaussian RBF kernel.

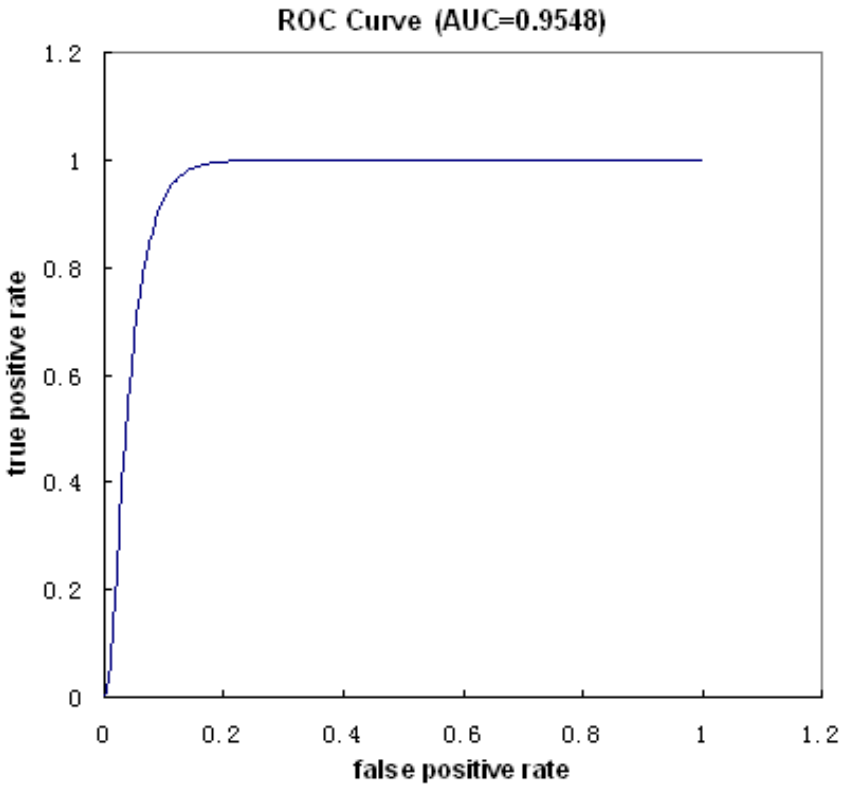


Figure 5.14: ROC curve of SVM classifier designed with Gaussian RBF kernel

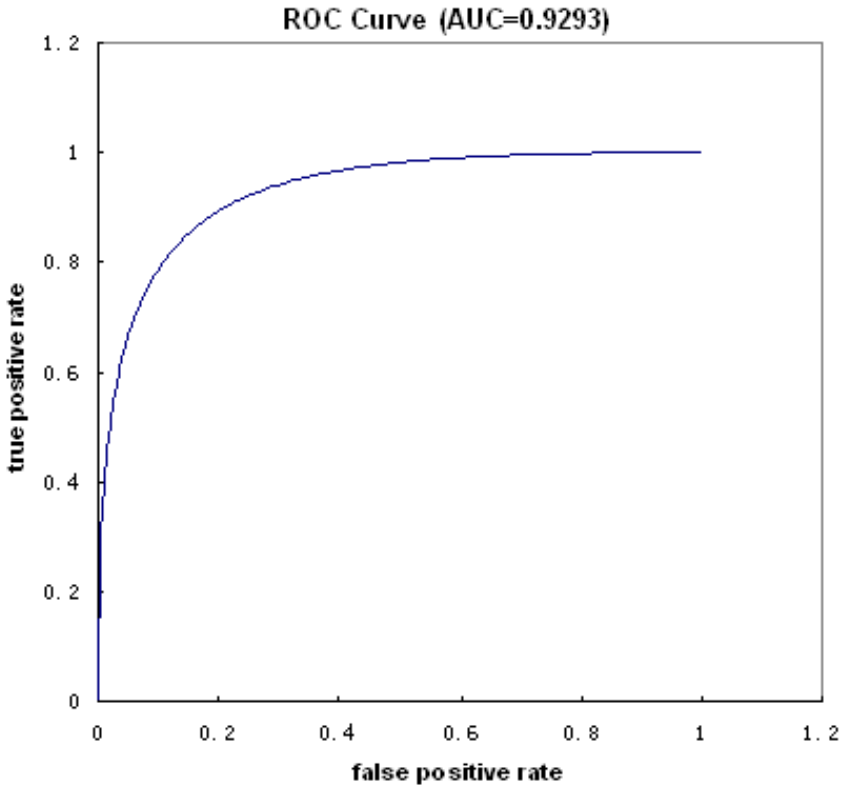


Figure 5.15: ROC curve of SVM without feature selection

Table 5.6: Feature selection influence for classification

Method	error	TP	TN	FP	FN	S_n	$1 - S_p$	AUC
Feature selection is applied before classification	7.19%	195	115	20	4	97.99%	14.81%	0.9548
Feature selection isn't applied before classification	19.96%	150	124	11	49	75.38%	8.15%	0.9293

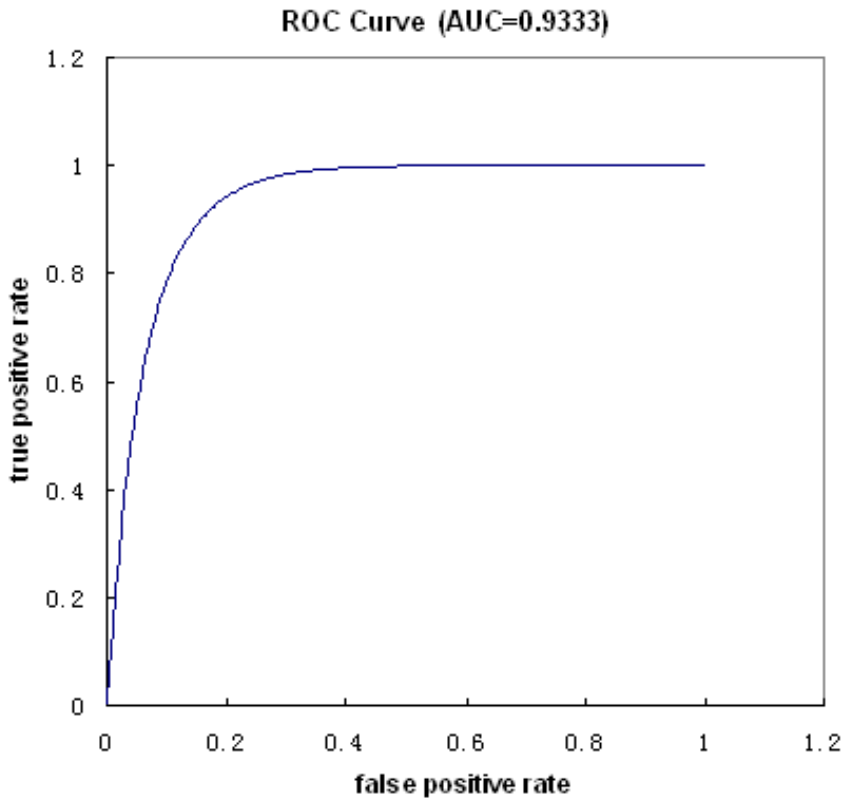


Figure 5.16: ROC curve of SVM classifier designed with Polynomial kernel

$$\begin{aligned} sensitivity \quad S_n &= \frac{TP}{TP+FN} \\ 1-specificity \; 1-S_p &= \frac{FP}{TN+FP} \end{aligned} \tag{5.38}$$

where TP is the number of true positives (correctly detected defects), TN is the number of true negatives (correctly detected non-defects), FP is the number of false positives (non-defects detected as defects), and FN is the number of false negatives (defects detected as non-defects). Ideally, $S_n = 1$ and $1 - S_p = 0$, this means that all defects are found without any false alarms.

5.6 Classification Performance

In this section, an SVM classifier and several types of popular pattern classifiers used for defect detection are implemented on the data set, and the results are compared with each

Table 5.7: SVM classifier with different kernel

Kernel	error	TP	TN	FP	FN	S_n	$1 - S_p$	AUC
Gaussian RBF	7.19%	195	115	20	4	97.99%	14.81%	0.9548
Polynomial	13.17%	170	120	15	29	85.43%	11.11%	0.9333

other. The classifiers compared with SVM in this study are a k-means classifier, a linear discriminant classifier, a k-nearest neighbor classifier and a feed forward neural network. It is noteworthy that the k-means method is the typical unsupervised statistical classifier, the linear discriminant method and k-nearest neighbor classifier are the typical supervised statistical classifiers, and the feed forward network is a typical artificial neural network. In other words, we are comparing the classification performances of the SVM with typical unsupervised/supervised statistical classifiers and artificial neural network in this study.

5.6.1 K-means Classifier

The k-means classifier is a typical unsupervised classifier. In general, the k-means method assumes that the samples belong to k disjoint classes, and the centroid of each of the classes in the feature space can be found in an iterative manner. K-means method will attempt to minimize the following target function [129].

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \tag{5.39}$$

where $\|x_i^{(j)} - c_j\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centres.

The algorithm is composed of the following steps

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

5.6.2 Linear Discriminant Classifier

Linear discriminant method is another widely used method for pattern classification, and it is adopted in this study to classify different patterns. Linear discriminant method can effectively discriminate two multivariate normal populations with equal variance-covariance matrices. The linear discriminant analysis method consists of searching some linear combinations of selected variables, which provide the best separation between the considered classes. These different combinations are called discriminant functions [130]. The fisher linear discriminant is described in the book [1]. In LDA, within-class and between-class scatter are used to formulate criteria for class separability. The class separability in a direction $\mathbf{w} \in \mathbb{R}^n$ is defined as

$$F(\mathbf{w}) = \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_W w}} \tag{5.40}$$

The between-class scatter matrix is defined as

$$\begin{aligned} \mathbf{S_B} &= (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \\ \mu_y &= \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}_i, \;\; y \in 1, 2 \end{aligned} \tag{5.41}$$

The within-class scatter matrix is defined as

$$\begin{aligned} \mathbf{S_W} &= \mathbf{S}_1 + \mathbf{S}_2 \\ \mathbf{S}_y &= \sum_{i \in Y_y} (\mathbf{x}_i - \mu_y)(\mathbf{x}_i - \mu_y)^T, \;\; y \in 1, 2 \end{aligned} \tag{5.42}$$

For the linear discriminant method in two class case, the classification rule can be stated as

- Given a sample feature vector $\mathbf{x}$,
- Choose class I if $\mathbf{w} \cdot \mathbf{x} + b > 0$ and choose class II otherwise

where vector $\mathbf{w} = {\mathbf{S_W}}^{-1}(\mu_1 - \mu_2)$, vector $b = (\mu_1 - \mu_2){\mathbf{S_W}}^{-1}(\mu_1 + \mu_2)/2$. During the training phase, $\mathbf{w}$ and b are learned from the training samples.

5.6.3 K-nearest Neighbors Classification

The k-nearest neighbors (KNN) classification is a method of classification that uses a training set chosen from the data as a point of reference in classifying observations. The idea of the method is to find the k elements of the training set that are closest to the target element to be classified. The target is then classified as belonging to whatever category that is the most frequent among the k-objects.

The KNN method is a widely used technique for solving classification problems. KNN is a non-parametric procedure. This means that it is not necessary to make assumptions about the underlying probability density functions of the d features in the feature vectors. This is a desirable property since in practise the parameters of density functions are often hard to obtain. KNN approximates the density function $p_n(x)$ of the features as:

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n} \tag{5.43}$$

where k_n is the number of observations within a certain volume V_n in the feature space with dimension d and n is the total number of observations. The size of V_n is chosen so that it contains k observations with the feature vector $\mathbf{x}$ as the center point. The approximation gets more accurate when the number of samples increases [1]. When an observation $\mathbf{x}$ is to be classified, a separate approximation is performed for each class ω_i:

$$p_n(\mathbf{x}, \omega_i) = \frac{k_i/n}{V} \tag{5.44}$$

k_i is the number of observations of the class ω_i within the volume V . The probability that an observation $\mathbf{x}$ belongs to a certain class $P_n(\omega_i|\mathbf{x})$ can then be calculated as:

$$P_n(\omega_i|\mathbf{x}) = \frac{p_n(\mathbf{x}, \omega_i)}{\sum_{j=1}^{c} p_n(\mathbf{x}, \omega_j)} = k_i/k \tag{5.45}$$

where c is the number of classes. The probability $P_n(\omega_i|\mathbf{x})$ is evaluated for each class and $\mathbf{x}$ is classified as belonging to the class with the highest probability, i.e. maximizing the

expression ki/k. KNN can be implemented by calculating the distance from the observation $\mathbf{x}$ of unknown class to all the other n observations in the training set and then selecting the k observations with the shortest distance [131]. $\mathbf{x}$ is classified as belonging to the class, which is most common among the k neighbors. In the two-class case the number k should be odd to avoid ties in the classification. The magnitude of k is chosen depending on the number of observations n. k should be increased when n is high and decreased when n is low. The optimal value of k in specific application can be evaluated through testing.

The distance between observations can be calculated with different metrics. A common metric is the Euclidean distance, the L_2 norm. The distance between the observations $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$L_2 = \sqrt{\sum_{m=1}^{d}(x_m - y_m)^2} \tag{5.46}$$

The algorithm can be divided into three steps

1. Calculate the distance from validation sample y to each training sample x using the Euclidean distance.
2. Identify the k training samples with smallest distance to y and check what the majority class is among these.
3. Redo 1-2 for each validation sample

5.6.4 Artificial Neural Network

To assess the performance of the SVMs we also classified our segmented potential using neural networks with a different algorithms and architectures.

Statistical pattern recognition uses statistical properties and criteria to differentiate data patterns. Another type of pattern recognition method which has been widely utilized is the artificial neural network method. Neural networks are designed to have the ability to learn complex nonlinear input-output relationships, use sequential training procedures, and adapt themselves to the input data.

Artificial neural networks result from attempts to modele a system whose performance is analogous to the most basic functions of human brains. As modern computers become ever more powerful, scientists continue to be challenged to use machines effectively for some tasks that are relatively simple to humans. Traditional, sequential, logic-based computers excel in arithmetic, but are less effective than human brains in many fields. Human brains have many other features that would be desirable in artificial systems.

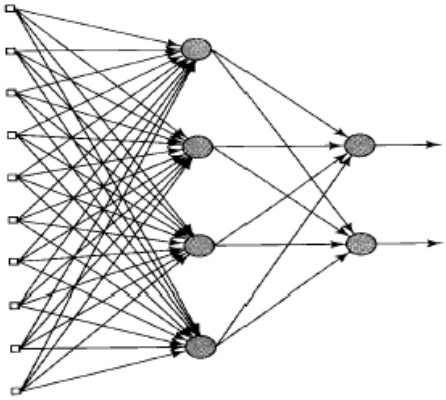


Figure 5.17: The diagram of typical artificial neural network

An artificial neural network is an information-processing system that has certain performance characteristics in common with biological neuron network. Artificial neural networks have been developed as generalizations of mathematical models of human neural biology, based on the assumptions that: information processing occurs at many simple elements called neurons; signals are passed between neurons over connection links; each connection link has an associated weight, which multiplies the signal transmitted; and each neuron applies an activation function to the input to determine the output signal. The typical artificial neural network is shown in figure 5.17.

The artificial neural network used in this study is a multilayer feed forward network with one hidden layer [132], trained using the supervised back propagation approach [132]. The multi-layer feed-forward neural network distinguishes itself from the single layer network by the presence of one or more hidden layers, whose computational units are the hidden layer neurons (nodes). The function of the hidden layer neurons is to intervene between the external input and network output. By adding one or more hidden layer, the network is able to approximate severe non-linearity. Back propagation is one of the simpler members of a family or training algorithms collectively termed gradient descent [133] [134]. The idea is to minimize the network total error by adjusting the weights. Gradient descent, sometimes known as the method of steepest descent, provides a means of doing this. Back

propagation is also the most suitable learning method for a multilayer network. An outline of the back propagation training algorithm is given by Lippmann [135] as follows.

Step1 Initialize weights and offsets. Set all weights and node offsets to small random values.

Step2 Present input and desired output. The desired output is 1. The input could be new on each trial or samples from a training set.

Step3 Calculate actual outputs. Use the sigmoid nonlinearity formulas to calculate outputs.

Step4 Adapt weights.

Step5 Repeat by going to step 2.

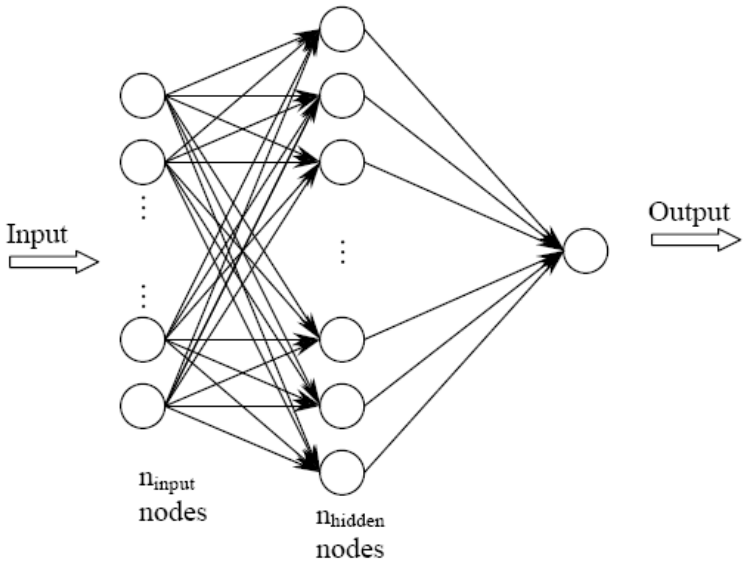


Figure 5.18: The multi-layer feed forward artificial neural network used in this study

The designed neural network is a three-layer feed forward artificial neural network. As shown in Figure 5.18, this ANN has n_{input} nodes for input, n_{hidden} nodes in the hidden layer, and one node for output. Each of the circles in the figure is a neuron, and represents an input-output transfer function. Neural network researchers have chosen from a variety of transfer functions; among the more popular ones are the logistic-sigmoid and the tangent-sigmoid functions. In this study we use the latter, which, mathematically, is:

$$f(n) = tansig(n) = \frac{e^n - e^{-n}}{e^n + e^{-n}} \tag{5.47}$$

where both the input, n, and the output, $f(n)$, are real scalar-values. Layers are connected to each other by a system of weights, which multiplicatively scale the values traversing the links. In the diagram, we observe that there are two sets of weights: one connecting the input to the hidden layer, and the other from the hidden to the output layer. The values from weights converging on a given unit are added to form n [136]. In this study, various selections of n_{input} and n_{hidden} are tested. The best n_{input} is 6 and the best n_{hidden} is 10.

Figure 5.19 shows training curve of the neural network, which is used to classify potential defects.

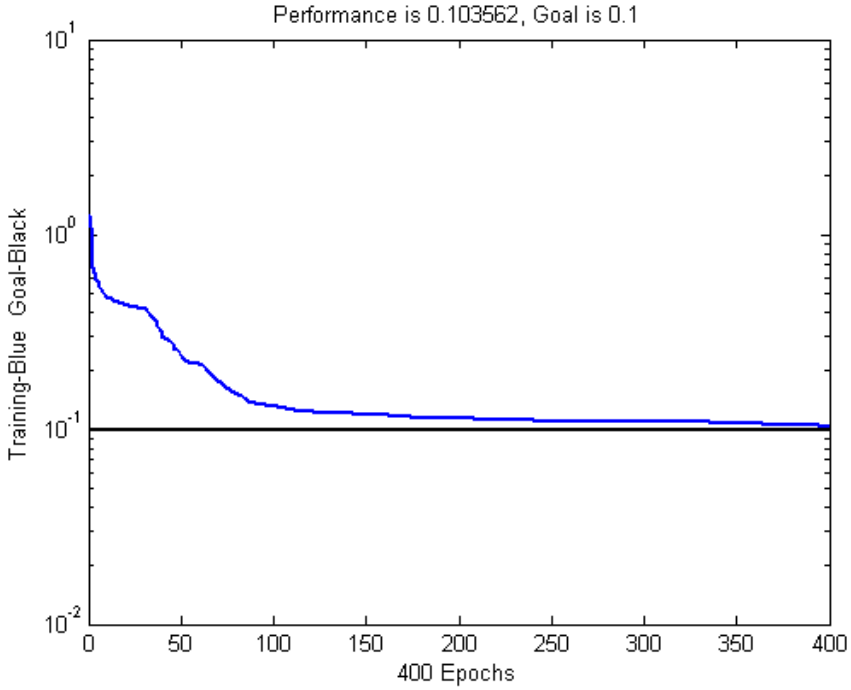


Figure 5.19: Training curve of neural network

5.6.5 Experiments

In these experiments, we use support vector machine, k-means, linear discriminant, k-nearest neighbor classifiers and feed forward neural network. 25 radiographic weld images are analyzed. In the segmentation stage, 809 potential discontinuities are obtained, of which 473 are real defects. For each potential discontinuity, 87 features are extracted, and only 16 of them are selected. 14 images with 475 potential defects are selected and used to train the classifier. Special attention is paid to ensure that instances of all different defects types are included. The remaining 11 images with 334 potential defects are used to test the classifier.

The results are summarized in Table 5.8, where the error, true positives, false positives, false negatives, true negatives, sensibility, 1-specificity, and AUC are tabulated for the mentioned classifiers. The best performance is obtained by the support vector machine, where 97.99% of the existing flaws are detected with 14.81% of false alarms.

Table 5.8: Performance of different classifier

Classifier	Error	TP	TN	FP	FN	S_n	$1 - S_p$	AUC
Support Vector Machine	7.19%	195	115	20	4	97.99%	14.81%	0.9548
Linear Discriminant	19.46%	154	112	20	45	77.39%	14.81%	0.8833
K-nearest Neighbor	12.28%	170	123	12	29	85.43%	8.89%	0.9123
Artificial Neural Network	14.67%	164	121	14	35	82.41%	10.37%	0.9089
K-means	29.34%	113	123	12	86	56.78%	8.89%	0.7042

There are various reasons for preferring SVM to other classification methods. First of all, the most distinguishing property of SVM is that it minimizes the structural risk, given as the probability of misclassifying previously unseen data. Typical pattern classification methods tend to minimize the empirical risk, which is given as the probability of misclassification errors on the training set. More specifically, Vapnik and Chervonenkis (VC) is a well studied theory that places reliable bounds on the generalization of linear classifiers and therefore indicate the control parameters on the complexity of linear functions in kernel spaces [40]. The VC dimension of a function is defined as the maximal number of points that can be shattered by that function. It has been shown that once the VC dimension of the family of decision surfaces is known, so is the upper bound for the probability of misclassification for test data of any possible probability distribution [40]. Second, SVMs pack all the relevant information in the training set into a small number of support vectors. Since the hypersurface is computed with the information from the supports vectors only, the computational efficiency of the classification of a test case is increased by the ratio of the number of data set points over the number of support vectors.

Much research effort in the past ten years has been devoted to the analysis of the performance of artificial neural networks in radiographic image classification. One distinctive advantage of SVM has over traditional neural networks is that support vector machines achieve better generalization performance. While neural networks such as multiple layer

perceptrons can produce low error rate on training data, there is no guarantee that this will translate into good performance on test data. The preferred algorithm of neural network is feed-forward multi-layer perceptron using back-propagation, due to its ability to handle any kind of numerical data, and to its freedom from distributional assumptions. Although neural networks may generally be used to classify data at least as accurately as other statistical classification approaches a number of studies have reported that the users of neural classifiers have problems in setting the choice of various parameters during training. The choice of architecture of the network, the sample size for training, learning algorithms, and number of iterations required for training are some of these problems. The SVM technique is independent of the dimensionality of feature space as the main idea behind this classification technique is to separate the classes with a surface that maximizes the margin between them, using boundary pixels to create the decision surface. The data points that are closest to the hyperplane are support vectors. Another major advantage of support vector classifiers is the use of quadratic programming, which provides global minima only. The absence of local minima is a significant difference from the neural network classifiers.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, an automatic welding defects detection system is proposed. After obtaining digital radiographs, image preprocessing methods are applied to improve the quality of radiographic images. Then the potential defects are extracted using a segmentation method. After extracting features of potential defects, a feature selection and classification system is used to classify the potential defects as the defect or non-defect.

Usually, defects in the original radiographic image are low in number compared with background information, and mixed with noise coming from various processes in the formation of radiographic images. Image preprocessing is employed to lessen the noise effects and to improve the contrast, so that the principal objects in the image can be more apparent than the background. In this thesis, different methods for improving the quality of radiographic images are investigated. Morphological enhancement and adaptive wavelet thresholding are applied to enhance radiographic images. Morphological enhancement can not only improve the local contrast of the radiographic images but also reduce the noise produced in homogeneous areas. The adaptive wavelet thresholding technique can remove the image noise while keeping the sharpness of defects' edges well. Through comparative analysis, morphological enhancement and adaptive wavelet thresholding can greatly enhance radiographic image and are helpful for defect recognition.

Defect segmentation is the most difficult task in the automatic welding defects detection system. It is used to extract the principal objects, which are welding defects in this research, from radiographic images. In this study, after investigating conventional segmen-

tation methods two new segmentation methods are proposed. One method is multiscale edge detection based on wavelet transform (MEWT). According to the wavelet multi-scale character, the coefficients of wavelet transforms are integrated on a series of scales to look for the best scale where the edges are well discriminated from noise to extract edge features. The other method is multi-level thresholding based on fuzzy entropy and genetic algorithm (MTFEGA). The radiographic image is segmented using multi-level thresholding based on maximum fuzzy entropy. The procedure to find the optimal thresholds is implemented by a genetic algorithm. The results show these two algorithms can succeed to segment welding defects present in radiography, contrary to conventional methods. Indeed, the proposed methods give a good basis for the future recognition and classification stage. Then the proposed two segmentation method, MTFEGA and MEWT, are compared with each other. 25 radiographic images are segmented using these two methods. The size of the minimum defect that could be segmented and the contrast between the minimum defect that could be segmented and background for each image are measured. The experimental results show the effectiveness of these two methods for segmenting images. The smallest defect segmented by MTFEGA method is 2.3 both in length and width. The smallest defect segmented by MEWT method is 5.2 both in length and width. The minimum contrast ratio (MCR) of MTFEGA is 3.92% and the MCR of MEWT is 7.06%. MTFEGA approach is a better choice to segment low contrast radiographic images with small size defects.

Feature extraction is necessary to obtain a set of features that can describe the characteristics of welding defects. The author proposes to detect weld defects based on texture features and geometric features. Two groups texture features are extracted: features based on the co-occurrence matrix and features based on 2D Gabor functions, and 7 geometric features of potential defects. For each potential discontinuity, 87 features extracted.

Pattern classification methods are needed to analyze feature data and make a prediction of the defect. A feature selection and classification system based on the support vector machine is proposed to the defect recognition. The top 16 best features are selected based on SVM criteria and ROC and used as inputs to an SVM classifier. In the segmentation stage, 809 potential discontinuities are obtained, of which 473 are real defects. 475 data are selected from the entire set of 809 data and used to train the SVM classifier designed with Gaussian RBF kernel. The remaining 334 data are used to test the classifier. Using this method, 97.99% of the existing flaws are detected with 14.81% false alarms. The author examines the behavior of the proposed classification method and various classi-

fication techniques, k-means, linear discriminant, k-nearest neighbor classifiers and feed forward neural network, which have been used in the past. The proposed system based on the support vector machine has the best result.

6.2 Suggestions for Future Work

In future, there are certainly many open questions related to the work and future improvements may well be made to the methods discussed.

First, a fully automatic segmentation method can be further investigated. The MEFEGA segmentation method still can not extract some fine defects. And the size of defect extracted by MEFEGA is smaller than the actual size of the defect. It is possible to improve the MEFEGA algorithm by combining it with MEWT. What is more, the proposed segmentation methods are still gray level based algorithms. In fact, human decision is based on not only gray levels, but also the position, shapes of defects etc. In order to obtain better result, perhaps these kinds of features can be considered for segmentation.

In this thesis, the SVM algorithm is used for feature selection and classification. More work can be done for tuning and validating the classifier. To become useful for detection, the system needs to be trained and validated on radiographs scanned from different films or digital detectors. This SVM algorithm has higher false positive rate compared with k-nearest neighbor classifier and neural networks. In order to reduce the false alarms, the SVM algorithm might be improved or combined with other classification algorithms. What is more, the SVM algorithm is used for a supervised classification in which a classifier is created based on a priori knowledge of defect features. In contrast with supervised classification, unsupervised classification don't need a priori knowledge. It is worth exploring unsupervised classification methods in defect recognition.

The current system can only detect defects. In fact, whether a weld line is acceptable or not depends not only on the sizes of the defects, but on the positions and types, for example, any size of crack is not acceptable in the application. Hence, it should be established to classify the different type defects and then an expert system can be built to decide if the weld line is a good one. What is more, in order to apply it in the industrial environment, a friendly computer-human interface is necessary.

Bibliography

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, Second Edition*. John Wiley & Sons, 2001.

[2] R. Halmshaw, *Non-destructive Testing*. Edward Arnold, 1988.

[3] G. Edward, "Inspection of welded joints," *ASM Handbook, Welding, Brazing and Soldering*, vol. 6, pp. 1081–1088, 1993.

[4] C. Hayes, "ABC's of nondestructive weld examination," *Weld Journal*, vol. 76, no. 5, pp. 46–51, 1997.

[5] R. J. Ditchburn, S. K. Burke, and C. M. Scala, "Ndt of welds: state of the art," *NDT&E International*, vol. 29, no. 2, pp. 111–117, 1992.

[6] Y. Suga, K. Kojuma, and T. Tominage, "Detection of weld defects by computer-aided x-ray radiography image processing," *International Journal of Offshore and Polar Engineering*, vol. 5, no. 2, June 1995.

[7] T. W. Liao, "Fuzzy reasoning based automatic inspection of radiographic welds: Weld recognition," *Journal of Intelligent Manufacturing*, vol. 15, pp. 69–85, 2004.

[8] D. E. Bray and R. K. Stanley, *Nondestructive Evaluation (A Tool for Design, Manufacturing and Service)*. McGraw-Hill, 1988.

[9] U. Zscherpel, "Film digitisation systems for dir: standards, requirements, archiving and printing," *NDT online Journal*, vol. 5, no. 5, May 2000.

[10] A. W. Davis, P. C. Berry, T. N. Claytor, D. A. Fry, M. H. Jones, and S. M. White, "An analysis of industrial nondestructive testing employing digital radiography as an alternative to film radiography," in *ESA-MT Nondestructive Testing and Evaluation Team Los Alamos National Laboratory*, March 2000.

[11] A. Gayer, A. Saya, and A. Shiloh, "Automatic recognition of welding defects in real-time radiography," *NDT International*, vol. 23, no. 3, pp. 131–136, 1990.

[12] S. Lawson and G. Patker, "Intelligent segmentation of industrial radiographic images using neural networks," in *Machine Vision Applications and Systems Integration III, Proceedings of SPIE*, vol. 2347, 1994, pp. 245–255.

[13] N. Nafaâ, D. Redouane, and B. Amar, "Weld defect extraction and classification in radiographic testing based artificial neural networks," in *Proceedings of the 15th World Conference on Non-Destructive Testing*, 2000.

[14] C. Jacobsen, U. Zscherpel, and C. Nockemann, "Crack detection in digitized radiographs with neuronal methods," in *Proceedings of 7th European Conference on Non-Destructive Testing*, May 1998.

[15] C. Jacobsen and U. Zscherpel, "Automated evaluation of digitized radiographs with neuronal methods," in *International Symposium on Computerized Tomography for Industrial Applications and Image Processing in Radiology*, Berlin, 1999, pp. 141–15.

[16] C. J. P. Perner, U. Zscherpel, "A comparison between neural networks and decision trees based on data from industrial radiographic testing," *Pattern Recognition Letters*, vol. 22, no. 47-54, 2001.

[17] P. Zhao, "Computer aided interpretation of images produced by non-destructive testing procedures," Master Thesis, Nanyang Tological University, Singapore, 2000.

[18] R. Hyatt, G. E. Kechter, and S. Nagashima, "A method for defect segmentation in digital radiographs of pipeline girth welds," *Material Evaluation*, pp. 925–928, 1996.

[19] T. Liao and Y. Li, "An automated radiographic NDT system for weld inspection: Part II. flaw detection," *NDT&E International*, vol. 31, no. 3, pp. 183–192, 1998.

[20] G. Wang and T. W. Liao, "Automatic identification of different types of welding defects in radiographic images," *NDT&E International*, vol. 35, pp. 519–528, 2002.

[21] R. Silva, M. Siqueira, I.Silva, A. Carvalho, and J. Rebello, "Contribution to the development of a radiographic inspection automated system," *NDT Online Journal*, vol. 7, no. 12, December 2002.

[22] M. Sofia and D. Redouane, "Shapes recognition system applied to the non destructive testing," in *Proceedings of the 8th European Conference on Non-Destructive Testing*, Barcelona, June 2002.

[23] D. Mery and M. A. Berti, "Automatic detection of welding defects using texture features," *Insight*, vol. 45, pp. 676–681, 2003.

[24] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, Nov. 1973.

[25] M. Amadasum and R. King, "Textural features corresponding to textural properties," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 19, pp. 1264–1274, 1989.

[26] R. W. Conners and C. A. Harlow, "A theoretical comparison of texture algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 204–222, 1980.

[27] M. M. Galloway, "Texture classification using gray level run length," *Computer Graphics and Image Processing*, vol. 4, pp. 172–179, 1975.

[28] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*. Chapman & Hall,, 1993.

[29] R. Bajcsy and L. Lieberman, "Texture gradient as a depth cue," *Computer Graphics and Image Processing*, vol. 5, pp. 52–67, 1976.

[30] M. Unser, "Local linear transforms for texture measurements," *Signal Processing*, vol. 11, pp. 61–79, 1986.

[31] K. I. Laws, "Rapid texture identification," *Proceedings of Conference on Image Processing for Missile Guidance, SPIE*, vol. 238, pp. 376–380, 1980.

[32] T. Pavlidis, "Structural descriptors and graph grammars," *In Pictorial Information Systems, S. K. Chang and K. S. Fu, editors*, pp. 86–103, 1980.

[33] R. Jain, R. Kasturi, and B. G. Schunk, *Machine Vision*. McGraw-Hall,, 1995.

[34] R. C. Gonzalez and R. Woods, *Digital Image Processing*. Addison-Wesley, 2002.

[35] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer*, vol. 2, no. 7, pp. 1160–1169, 1985.

[36] R. L. Kashyap and R. Chellappa, "Estimation and choice of neighbors in spatial interatction models of images," *IEEE Transactions on Information Theory*, vol. 29, pp. 60–72, 19863.

[37] R. L. Kashyap and A. Khotanzad, "A model-based method for rotation invariant texture classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 472–481, 1986.

[38] R. L. Kashyap and P. M. Lapsa, "Aynthesis and estimation of random fields using long-correclation models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 800–809, 1984.

[39] V. Vapnik, "An overview of statistical learning theory," *IEEE Transaction on Neural Networks*, vol. 10, pp. 989–999, 1999.

[40] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[41] V. Vapnik, *Statistical Learning Theory*. New York: Wiely, 1998.

[42] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.

[43] E. Osuna, R. Freund, and F. Girosi, *Support Vector Machines: Training and Application*. Artifical Intelligence Laboratory, MIT, 1997.

[44] R. M. Gray and L. D. Davisson, *Random Process: A Mathematical Approach for Engineers*. Englewood Cliffs, N. J.: Prentice-Hall, 1986.

[45] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 364–280, 1971.

[46] V. Vapnik and A. Chervonenkis, "The necessary and sufficient conditions for consistency in the empirical risk minimiztion method," *Pattern Recognition and Image Analysis*, vol. 1, pp. 283–305, 1991.

[47] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[48] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 130–136.

[49] P. Soille, *Morphological Image Analysis Principles and Applications*. Berlin, Germany: Springer, 1999.

[50] P. Maragos and R. W. Schafer, "Morphological systems for multidimensional signal processing," *Proc. IEEE*, vol. 78, pp. 690–710, 1990.

[51] E. R. Dougherty, *An Introduction to Morphological Image Processing*. Bellingham, WA: SPIE Press, 1992.

[52] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 4, pp. 532–550, July 1987.

[53] L. Vincent, "Morphological grayscale reconstruction in image analysis: application and efficient algorithms," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[54] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 355–368, 1987.

[55] J. Rosenman, C. A. Roe, R. Cromartie, K. E. Muller, and S. M. Pizer, "Portal film enhancement: technique and clinical utility," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 25, pp. 333–338, 1993.

[56] X. S. adn Brian W. Pogue and S. Jiang, "Automated region detection based on the contrast-to-noise ratio in near-infrared tomography," *Applied Optics*, vol. 43, pp. 1053–1062, 2004.

[57] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM J.Math*, vol. 15, pp. 723–736, 1984.

[58] I. Daubechies, "Orthogonal bases of compactly supported wavelets," *Comm. Pure Appl. Math.*, vol. 41, pp. 909–996, 1988.

[59] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, July 1989.

[60] C. K. Chui, *An Introduction to Wavelets*. Academic Press Inc, 1993.

[61] R. M. Rao and A. S. Bopardikar, *Wavelet Transforms: Introduction to Theory And Applications*. Addison Wesley Longman Inc, 1998.

[62] C. S. Burrus, R. A. Gopinath, and H. Guo, *Intruduction to Wavelet and Wavelet Transforms: A Primer*. Prentice Hall Inc, 1998.

[63] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press Inc, 1998.

[64] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710–732, July 1992.

[65] D. L. Donoho, "Wavelet tresholding and W.V.D.: a 10 minute tour," in *Int. Conf. on Wavelets and Applications*, Toulouse, France, June 1992.

[66] D. L. Donoho and I. M. Johnstone, "Indeal spatial adaptation via wavelet shrinkage," *Biomerika*, vol. 81, pp. 425–455, 1994.

[67] D. L. Donoho, "De-noising by soft-threshholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 425–455, May 1995.

[68] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet tresholding for image denoise and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, Sep 2000.

[69] K. R. Castleman, *Digital Image Processing*. N.J.: Prentice Hall, 1996.

[70] K. S. Fu and J. K. Mui, "A survey on image segmenatation," *Pattern Recognition*, vol. 13, pp. 3–16, 1981.

[71] R. M. Haralick and L. G. Shapiro, "Survey: image segmentation, computer graphics and image processing," *Comput. Vision, Graphics, Image Processing*, vol. 29, pp. 100–132, 1985.

[72] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*. Chapman & Hall, 1993.

[73] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, pp. 1277–1294, 1993.

[74] L. G. Roberts, "Machine perception of three-dimensional solids," in *Optical and Eletro-Optical Information Processing*, 1965, pp. 159–197.

[75] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, November 1986.

[76] V. Berzins, "Accuracy of Laplacian edge detectors," *Computer Vision, Graphics and Image Processing*, vol. 27, pp. 1955–2010, 1984.

[77] A. Jain, *Fundamentals of Digital Image Processing*. Prentice Hall: Prentice Hall Information and System Sciences Series, 1989.

[78] S. L. Horowitz and T. Pavilidis, "Picture segmentation by directed split and merge procedure," in *Proceedings of 2nd International Joint Conference on Pattern Recognition*, 1974, pp. 424–433.

[79] T. Asano and N. Yokoya, "Image segmentation schema for low-level computer vision," *Pattern Recognition*, vol. 14, pp. 267–273, 1981.

[80] T. Pavlidis and Y.-T. Liow, "Integrating region growing and edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 225–233, 1990.

[81] S.-Y. Chen, W.-C. Lin, , and C.-T. Chen, "Split-and-merge image segmentation based on localized feature analysis and statistical tests," *Computer Vision, Graphics, and Image Processing (CVGIP): Graphics, Models Image Processing*, vol. 53, no. 5, pp. 457–475, 1991.

[82] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[83] W. Snyder and H. Qi, *Machine Vision*. Cambridge University Press, 2004.

[84] J. Silverman and D. Cooper, "Bayesian clustering for unsupervised estimation of surface and texture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 57, pp. 373–387, 1993.

[85] M. LaValle and S. Hutchinson, "A bayesian segmentation methodology for parametric image models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 211–217, 1995.

[86] Y. Han, W. Snyder, and G. Bilbro, "Discontinuity-preserving vector smoothing of multivariate mr images using vector mean field annealing," *Journal of Mathematical Imaging and Vision*, vol. 9, no. 3, pp. 199–212, 1998.

[87] D. Geman and S. Geman, "Stochastic relaxation, gibbs distributions and bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[88] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, 1990.

[89] P. S. K. L. G. B. W. Snyder, A. Logenthiran, "Segmentation of magnetic resonance images using mean field annealing," *Image and Vision Computing*, vol. 10, pp. 362–368, 1992.

[90] G. Bilbro and W. Snyder, "Optimization of functions with many minima," *IEEE Transactions on System, Man, and Cybernetics*, vol. 21, pp. 840–849, 1991.

[91] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene and analysis," *IEEE Trans. Computer*, vol. 20, pp. 562–569, 1971.

[92] A. Witkin, "Scale space filtering," in *Proc. Int. Joint Conf. Artificial Intell.*, 1983.

[93] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. Royal Soc. London*, vol. 207, pp. 187–217, 1980.

[94] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Transactions on Transactions on Information Theory*, vol. 38, no. 6, pp. 617–643, Mar 1992.

[95] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[96] L. A. Zadeh, "The concept of a linuistic variable and its application to approximate reasoning, part i," *Information Science*, vol. 8, pp. 199–249, 1975.

[97] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Addision-Wesley, 1976.

[98] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: The University of Michigan Press,, 1989.

[99] L. Davis, *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold, 1991.

[100] N. Ostu, "A threshold selection method from gray level histogram," *IEEE Transactions on Systems Man Cybernet*, vol. SMC-8, pp. 62–66, 1978.

[101] M. Zhao, A. M. N. Fu, and H. Yan, "A technique of three-level thresholding based on probability partition and fuzzy 3-partition," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 3, pp. 469–479, June 2001.

[102] C. Janikow and Z. Michalewicz, "An experimental comparison of binary and floating point representations in genetic algorithm," in *Proceedings of the Fourth International Conference in Genetic Algorithms*, Morgan Kaufmann, 1991, pp. 31–36.

[103] M. Bramlette, "Initialisation, mutation and selection method in genetic algorithms for function optimization," in *Proceedings of the Fourth International Conference in Genetic Algorithms*, Morgan Kaufmann, 1991, pp. 100–107.

[104] D. Whitley and S. Rana, "Search, binary representations, and counting optima," in *Proceeding of a Workshop on Evolutionary Algorithms*, 1998.

[105] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. Berlin: Springer-Verlag, 1996.

[106] N. Ansari and E. Hou, *Computational Intelligence for Optimization.* Norwell, Massachusetts, USA: Kluwer Academic, 1997.

[107] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[108] T. Hastie, R. Tibshirhani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *JASA*, vol. 89, pp. 1255–1270, 1994.

[109] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York: Chapmann and Hall, 1984.

[110] T. Hastie and R. Tibshirani, "Tibshirani discriminant analysis by gaussian mixtures," *J. Royal Statist. Soc. (Series B)*, vol. 58, pp. 155–176, 1996.

[111] J. Morgan and J. Sonquist, "Problems in the analysis of survey data and a proposal," *Journal of the American Statistical Association*, vol. 58, pp. 415–434, 1963.

[112] B. Boster, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, USA, 1992, pp. 144–152.

[113] S. Mallat, "Wavelets for a vision," *Proc. IEEE*, vol. 8, pp. 604–614, April 1996.

[114] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Machines Intell.*, vol. 18, no. 8, pp. 837–842, August 1996.

[115] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM*, vol. 36, no. 4, 1989.

[116] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[117] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optiminzation Methods and Software*, vol. 1, pp. 23–34, 1992.

[118] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[119] N. Cristianini and J. Shawe-Taylor, *Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge University Press, 2000.

[120] B. B. Ripley, *Pattern Recognition and Neural Networks*. Cambridege University Press, 1996.

[121] B. Yu and B. Yuan, "A more efficient branch and bound algorithm for feature selection," *Pattern Recognition*, vol. 26, pp. 883–889, 1993.

[122] P. Devijer and J. Kittler, *Pattern Recognition: A Satistical Approach*. Prentice-Hall, 1982.

[123] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition letters*, vol. 15, pp. 1119–1125, 1994.

[124] C. Couvreur and Y. Bresler, "On the optimality of the backward greedy algorithm for the subset selection problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 3, pp. 797–808, 2000.

[125] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Computation*, vol. 12, no. 9, 2000.

[126] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificaial Intelligence*, vol. 97, pp. 273–324, 1997.

[127] A. Rakotomamonjy, "Variable selection using svm-based critera," *Journal of Machine Learing Research*, vol. 3, pp. 1357–1370, 2003.

[128] T. Fawcett, *ROC Graphs : Notes and Practical Considerations for Researchers*. MS 1143, 1501 Page Mill Road, Palo Alto CA 94304, USA: Technical report, HP Laboratories, 2004.

[129] A. K. Jain, P. Robert, W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Aanalysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[130] M. S. Srivastava and E. M. Carter, *Applied Multivariate Statistics*. North Holland Amsterdam, 1983.

[131] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[132] S. Haykin, *Neural Networks: A Comprehensive Foundation, Second Edition*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[133] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston: PWS Publishing, 1996.

[134] B. D. Ripley, *Pattern Recognition and Neural Networks*. New York: Cambridge University Press, 1996.

[135] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE Acoustics, Speech and Signal Processing*, vol. 4, no. 2, pp. 4–22, 1987.

[136] R. Kasturi and R. C. Jain, *Computer Vision: Principles and Applications*. Los Alamitos, CA: IEEE Computer Society Press, 1991.