Data Fusion for Multi-Sensor Nondestructive Detection of Surface Cracks in Ferromagnetic Materials

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

im Fach Informatik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Dipl.-Inf. René Heideklang

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät
Prof. Dr. Elmar Kulke

Gutachter/innen:	1. Prof. Dr.-Ing. Galina Ivanova
	2. Prof. Dr. rer. nat. Ralf Reulke
	3. Prof. Parisa Shokouhi, Ph.D., P.E.
Tag der mündlichen Prüfung:	16.7.2018

Abstract

Fatigue cracking is a dangerous and cost-intensive phenomenon that requires early detection. But before such cracks grow to a critical size, they originate as micro-defects, and are therefore challenging to detect using standard nondestructive testing approaches. In particular, at high test sensitivity, the abundance of false indications limits the reliability of conventional materials testing. This thesis exploits the diversity of physical principles that different nondestructive surface inspection methods offer, by applying data fusion techniques to increase the reliability of defect detection.

After describing methods for single-sensor defect detection, the first main contribution of this work is to present novel approaches for the fusion of NDT images. These images are formed by nondestructive surface scans, obtained from state-of-the-art inspection procedures in Eddy Current Testing, Thermal Testing and Magnetic Flux Leakage Testing, to detect fatigue cracks and other structural discontinuities.

Two ways of radiometric normalization are proposed to integrate the heterogeneous NDT signals. Results of the implemented image fusion strategy demonstrate that simple algebraic fusion rules are sufficient for high performance, provided that normalization is adequately performed. Fused defect detection successfully outperforms the best individual sensor for shallow surface discontinuities. Accordingly, the rate of false pixels is reduced by a factor of six when detecting a 10 µm deep groove.

Inspired by these positive results, the thesis continues by exploring the utility of state-of-the-art image representations, like the Shearlet domain, for fusion in NDT. Despite extensive treatment of the proposed strategy, the theoretical advantages of such directional transforms over undirectional fusion methods are however not attained in practice with the given data. Nevertheless, the benefit of fusion over single-sensor inspection for the detection of shallow discontinuities is confirmed a second time.

Furthermore, this work proposes novel techniques for fusion at a high level of signal abstraction, that is, after each individual data set has undergone defect detection. A kernel-based approach is introduced to integrate the spatially scattered detection hypotheses. Three mechanisms are proposed to keep the number of false alarms low despite maintaining high sensitivity. Importantly, unlike low-level image fusion, this method explicitly deals with registration errors that are unavoidable in practice. The experimental results show that surface discontinuities as shallow as 30 µm are reliably found by fusion, whereas the best individual sensor requires depths of 40–50 µm for successful detection. The experiment is replicated on a similar second test specimen to corroborate the method's invariantly high performance under different experimental conditions.

In addition to these methodological and experimental contributions, practical guidelines are given at the end of the thesis, and the need for a data sharing initiative is stressed to promote future research on this topic.

Kurzfassung

Ermüdungsrissbildung ist ein gefährliches und kostenintensives Phänomen, welches frühzeitig erkannt werden muss. Doch bevor solche Risse zu einer kritischen Größe heranwachsen, entsehen sie in Form von Mikrofehlern, und sind deshalb mit konventionellen Methoden der Zerstörungsfreien Prüfung schwierig zu erkennen. Insbesondere bei der hohen Testempfindlichkeit, die solch kleine Fehler erfordern, wird die Prüfzuverlässigkeit durch eine große Anzahl von Falschanzeigen vermindert. Diese Arbeit macht sich deshalb die Diversität unterschiedlicher zerstörungsfreier Oberflächenprüfmethoden zu Nutze, indem Techniken der Datenfusion eingesetzt werden, um die Zuverlässigkeit der Fehlererkennung zu erhöhen.

Nachdem zunächst Methoden zur Erkennung mittels Einzelsensoren beschrieben werden, besteht der erste Beitrag dieser Arbeit in neuartigen Ansätzen zur Fusion von Prüfbildern. Diese Bilder werden durch Oberflächenabtastung mittels Wirbelstromprüfung, thermischer Prüfung und magnetischer Streuflussprüfung gewonnen, um Ermüdungsrisse und andere strukturelle Unstetigkeiten zu erkennen. Dazu werden zwei Arten radiometrischer Normalisierung vorgeschlagen, um die heterogenen Prüfsignale zu vereinen. Die Ergebnisse der implementierten Fusionsstrategie zeigen, dass einfache algebraische Fusionsregeln für eine Ergebnisgüte ausreichen, sofern durch Normalisierung adäquat vorverarbeitet wurde. Der Fusionsansatz übertrifft erfolgreich den besten Einzelsensor bei der Erkennung flacher Oberflächenunstetigkeiten. So wird die pixelbasierte Falscherkennungsrate bei einer Nutentiefe von 10 µm um den Faktor sechs reduziert.

Auf Basis dieser Resultate leitet die Arbeit zum Einsatz aktueller Bildrepräsentationen für Fusion in der Zerstörungsfreien Prüfung über, wie z. B. des Shearletbereiches. Trotz intensiver Bearbeitung dieses Ansatzes werden jedoch die theoretischen Vorteile solcher richtungsempfindlichen Transformationen über richtungsunempfindliche Fusionsmethoden in der Praxis mit den vorliegenden Daten nicht erreicht. Nichtsdestotrotz wird der Vorteil der Fusion gegenüber Einzelsensorprüfung zur Ereknnung von flachen Unstetigkeiten auch hier bestätigt.

Weiterhin liefert diese Arbeit neuartige Techniken zur Fusion auch auf höheren Ebenen der Signalabstraktion, also nachdem jeder einzelne Sensordatensatz einer Defekterkennung unterzogen wurde. Ein Ansatz, der auf Kerndichtefunktionen beruht, wird eingeführt, um die örtlich verteilten Detektionshypothesen in Beziehung zu setzen. Drei Mechanismen werden vorgestellt, um die Zahl der Falschanzeigen zu minimieren, während die Detektionsempfindlichkeit für flache Risse möglichst nicht beeinträchtigt wird. Eine wichtige Eigenschaft des vorgestellten Verfahrens ist, dass im Gegensatz zur Fusion auf Signalebene Registrierungsfehler explizit miteinbezogen werden, welche in der Praxis unvermeidbar sind. Die experimentellen Ergebnisse zeigen, dass Oberflächenunstetigkeiten von 30 µm Eindringtiefe zuverlässig durch Fusion gefunden werden, wogegen das beste Einzelverfahren erst Tiefen ab 40–50 µm erfolgreich auffindet. Das Experiment wird auf einem zweiten Prüfkörper repliziert, um die Übertragbarkeit der Ergebnisse unter unterschiedlichen experimentelle Bedingungen zu bestätigen. Zusätzlich zu diesen methodischen und experimentellen Beiträgen, werden am Ende der Arbeit Richtlinien für den Einsatz von Datenfusion in der Praxis gegeben, und die Notwendigkeit einer Initiative zum Teilen von Messdaten wird hervorgehoben, um zukünftige Forschung auf diesem Gebiet zu fördern.

Acknowledgements

I would like to acknowledge a number of people for supporting me while working on this thesis. First, I would like to thank BAM, and in particular Parisa Shokouhi, Werner Daum and Giovanni Bruno, for giving me the opportunity to carry out this project and to learn a lot during this time. The topic of materials testing was entirely new to me, and being a computer scientist, it was an interesting experience to work in a research area that connects people from many diverse disciplines, from engineers to physicists. Special thanks deserve my colleagues who shared their valuable measurements with me: R. Pohl, G. Casperson, R. Casperson and T. Erthner (Eddy Current Testing); R. Stegemann, M. Pelkner, V. Reimund (Magnetic Testing); and M. Ziegler, P. Myrach, D. Mikolai and C. Maierhofer (Thermal Testing). Furthermore, I thank M. Kreutzbruck, T. Heckel and H. Wiggenhauser who advised and supported me. Special thanks to my office mates C. Völker and C. Schöllig for the incredibly nice working atmosphere. More generally, thanks to my fellow PhD students at BAM and to W. Gieschler for many interesting conversations. After writing this thesis, I was glad to have such reliable friends and family Steffi, Simon, Markus and Sebastian, who immediately agreed to proofreading.

My main supervisors P. Shokouhi, G. Ivanova and R. Reulke deserve much credit for always being willing to help whenever needed and for their continued support.

Finally, a big thank you to my wife: For your unconditional support and for your honest interest in discussions about structural noise.

Glossary

AUC Area Under the ROC Curve 46, 49, 64, 115

complementary pieces of information about different aspects of an object 22, *see* redundant

crack flaw that locally separates the surrounding material, having approximately two-dimensional geometry 9, *see* microcrack

CWT Continuous Wavelet Transform 31, 33, 43

detection an indication that satisfies some detection criterion and thus is suspect to represent a flaw *see* indication & flaw

discontinuity a lack of continuity or cohesion; an intentional or unintentional interruption in the physical structure or configuration of a material or component 2

DTCoWT Dual-Tree Complex Wavelet Transform 56

EDM Electrical Discharge Machining 27, 28, 42, 93, 120

ET Eddy Current Testing 2, 10, 12, 13, 27, 42, 94, 133

fatigue degradation of materials under repeated loading 9

flaw an imperfection or discontinuity that may be detectable by nondestructive testing *see* imperfection & discontinuity

FPR False Positive Rate 64, 113, 115

global coordinate system System whose coordinates are identified with physical landmarks on the specimen and can therefore be easily interpreted. This system is used as a reference system for other local coordinate systems. 24, *see* local coordinate system

GMR Giant Magnetoresistance 13, 42, 94

gradiometer sensor that measures the change of a physical quantity, for instance the spatial gradient of magnetic field strength 13, 30

hit 111, *see* detection

imperfection a departure of a quality characteristic from its intended condition 9

indication a significant sensor value with regard to the background signal 1, 20

intensity normalization process of normalizing the signal intensity range 39

KDE Kernel Density Estimation 87

local coordinate system system in which the measurements of an individual inspection are expressed; usually relative to some arbitrary origin and some orientation on the specimen surface, which are defined by the measurement setup. 24, *see*

local ridge detection localization of ridge maxima in inspection images 35

magnitude normalization 39, *see* intensity normalization

MFL Magnetic Flux Leakage Testing 2, 10, 14, 27, 42, 94, 117

MGA Multiscale Geometric Analysis 8, 53

microcrack crack whose depth into the material is in the micrometer range and thus challenges detection sensitivity 9, 13, *see* crack & sensitivity

NDT Nondestructive Testing 1, 39

POD Probability of Detection 8, 114

polarity of a bi-modal peak order of the negative and the positive peak, i.e. hill-valley or valley-hill 11, 13, 31

redundant pieces of information about the same aspect of an object; either agreeing or conflicting 22, *see* complementary

ROC Receiver Operating Characteristic 46, 102, 114

RT Radiographic Testing 1, 117

sensitivity property of a detector to successfully find a high fraction of the actual targets 19, 43, 46, 115, *see* specificity

shape normalization process of converting differential signals to intensity signals 30

SNR Signal-to-Noise Ratio 2, 11, 21, 39, 40, 86

specificity property of a detector to generate only few false alarms 19, 46, 115, *see* sensitivity

ST Shearlet Transform 53, 54

structural noise background signal representing non-defect related variations of material properties 3

SWT Stationary Wavelet Transform 41, 55, 95

TPR True Positive Rate 64, 113, 115

TT Thermographic Testing 2, 5, 10, 27, 42, 45, 94, 117

UT Ultrasonic Testing 1, 27, 117

UWT modified Undecimated Wavelet Transform [1] as an extension of SWT 55

Further acronyms: a.u. = arbitrary units, e.g. = for example, i.e. = that is.

Some definitions from this glossary were adapted from the ASTM 1316 standard [2].

Contents

Chapter 1

Introduction

1.1 Motivation

Nondestructive Testing (NDT) deals with the inspection of materials, parts and structures to assess their condition without compromising their usability or functionality. NDT is important at all stages of the production process – from 100% inspection for quality control during manufacturing, over sample testing after production, to in-service maintenance at regular intervals or even continuous monitoring. Therefore, the multi-disciplinary field greatly contributes to economy and society by reducing costs, maintaining high product quality and ensuring technical safety. Driven by governmental safety regulations and by demands from automotive, aerospace and power generation industries, NDT business is expected to grow during the next years [3].

NDT experts employ different inspection techniques depending on the material and the expected types of defects. For instance, methods such as Ultrasonic Testing (UT) or Radiographic Testing (RT) are well-known from medical examination and are also widely applied in NDT. Traditionally, the single most suitable inspection method for a given task is selected, although multiple methods might qualify. However, single-method inspection is often not reliable enough, for instance with composite materials, complex geometries or miniature flaws. In such settings, the results are often ambiguous and trained experts are required to interpret them. This bears the danger of overlooking critical indications, as they are buried among many false alarms. Missing a critical defect might have catastrophic consequences, costing lives in the worst case. On the other hand, ambiguous indications that are in fact harmless, but can not be identified as such with sufficient confidence, necessitate unnecessary and costly action, like repairs or replacements. For these reasons, a more diversified approach is in demand that does not rely on a single source of information. Through inspection of the same object with different NDT methods, or the same method using different measurement parameters, a more holistic view of the part's condition can be obtained. Especially in safety-critical applications, such as the aerospace and nuclear industries [4, ch. 1.1], there is a great demand for such diversity of information to improve the testing reliability and consequently to promote more substantiated decisions.

In recent years, inspection has become increasingly automatable across various NDT domains [5–8]. This development promotes advanced signal analysis methods to enhance the quality of the results, to ensure repeatability and to extract the relevant information from the extensive[1] data sets. At the same time, the processing power

[1]Especially for volumetric inspection, measurements take up several gigabytes of space, depending on the sampling rates. Data from two-dimensional surface inspection is usually more manageable.

of today's computers is rapidly progressing, which enables developing sophisticated solutions that were not practicable only ten years ago. Only recently, these technical and methodological improvements have been facilitating holistic condition assessment based on multi-method inspection, especially considering the comparably long history of NDT [9] which dates back further than the early 20th century. However, despite multiple NDT methods are already being applied, for instance in civil engineering, often the individual results are only qualitatively compared by a group of experts to reach a conclusion [2]. This practice leaves room for subjectivity and potentially runs the risk of being overwhelmed by too much information. In fact, Vavilov and Burleigh (2015) [11, table 5] characterize the whole research field when they declare that "Data fusion algorithms are not well-explored" in their recent review about thermographic NDT methods. Clearly, for the same reasons that are driving automation forward in single-sensor analysis, there is a need for automated assessment that takes into account all available information. This leads to the incorporation of *Data Fusion* methods into the analysis of NDT inspection data.

To focus this work on a specific type of material flaw, near-surface defects are investigated. More specifically, only ferromagnetic materials are considered here to facilitate magnetization-based test methods, in addition to other more generally applicable techniques. Typical surface flaws, for example in steel, are pores and cracks which are the primary factors that limit the life time of industrial parts such as bearings and turbine blades, but also rails. Under dynamic loads, such microstructural discontinuities may grow to larger cracks that impair the whole part's structural integrity. Therefore, early detection with high reliability allows reducing the required frequency of inspections, which leads to cost savings.

To demonstrate the benefits of multi-sensor defect detection, consider the following inspection result of surface inspection on a steel slab. This test specimen is 10 by 5 by 1 cm large and contains ten artificially introduced discontinuities at its surface. Because the individual discontinuities have different depths, the effect of defect size on each test method can be investigated for this test object. The testing techniques will be briefly compared taking into account a) their sensitivity to shallow defects, b) their tendency to produce false alarms, and c) their defect localization ability.

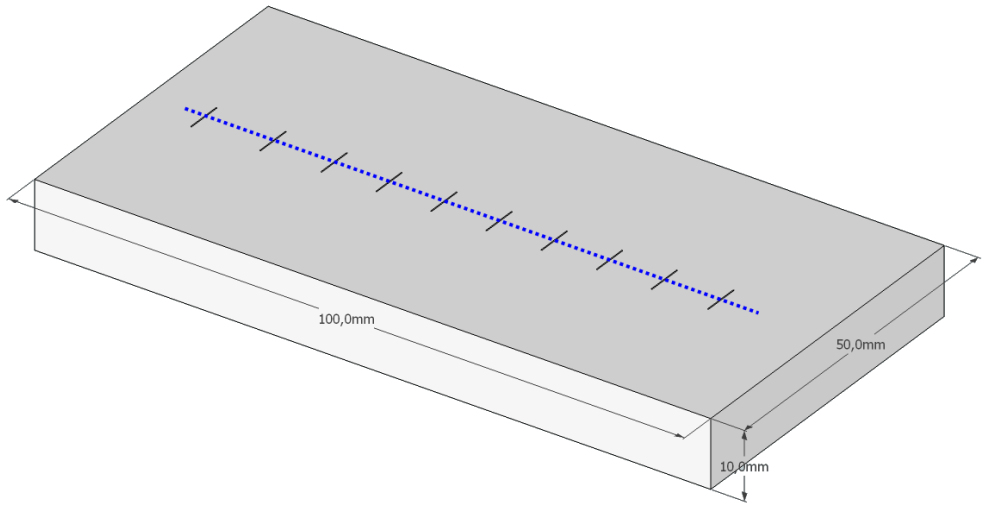
A schema of the part is shown in figure 1.1. The blue dotted line indicates the path on the surface that was inspected using three NDT methods: Eddy Current Testing (ET), Magnetic Flux Leakage Testing (MFL) and Thermographic Testing (TT), which will be explained later on. The inspection results are plotted in figure 1.2. This figure highlights the different characteristics of each test method for this inspection. ET data show sensitivity to most of the tested defect sizes. However, the comparably broad signal peaks degrade the ability to accurately localize any discontinuity, and prevent nearby defects to be resolved individually. Moreover, high signal intensity is not only present near the known groove positions, but also in other regions where material properties change (not seen in the figure), thus producing false alarms. Although MFL inspection yields high Signal-to-Noise Ratio (SNR) for deep grooves and localization is very good, here the shallower discontinuities are not distinguishable from the background signal variations. TT shows particularly high sensitivity and often has superior localization ability compared to ET. On the other hand, the inspection result is overly sensitive

Another type of data complexity is given by the data dimensionality (number of informative features).

[2] [10, sec. 4]: "Although the BetoScan system has a fully automated data acquisition system, data analysis is currently performed manually by direct comparison of the results"

because strong indications away from the known defect positions are present[3]; see the red mark in figure 1.2a.

Figure 1.1: Schema of a test specimen containing ten defects. The blue dotted line indicates the surface inspection path; see figure 1.2.



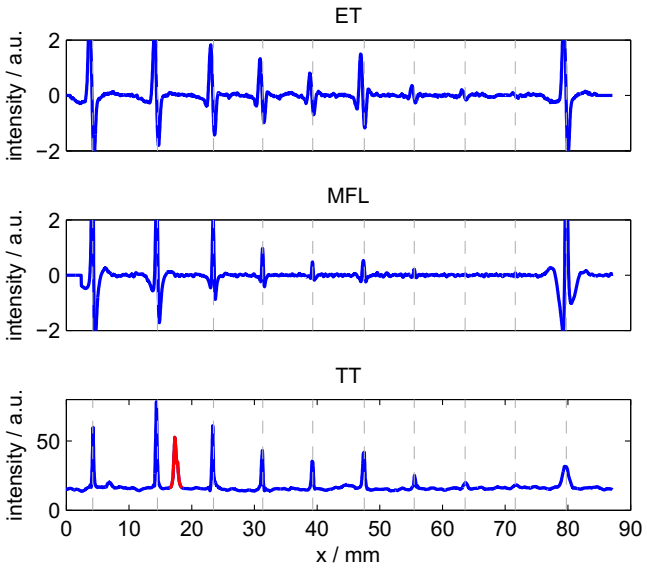
As can be seen, for this test object and the given discontinuities, no individual test method is sufficient regarding all quality criteria. In particular, the false indications degrade the detection performance. Note that such false alarms might have comparable signal intensity even to large true indications, as figure 1.2a shows. Whereas such strong spikes are usually rare in practice and can easily be disproved by consulting one additional NDT technique, the detection of small defects is more challenging.

This is because small defects produce signals that are hardly distinguishable from "normal" background variations, as seen in figure 1.2b). These non-defect related variations reflect the spatial inhomogeneities of the underlying material properties, and are therefore deterministic with regard to multiple measurements. This is unlike the random measurement noise, which is also present but affects the signal only mildly in comparison. In the context of defect detection, the unwanted background signals will therefore be termed *structural noise* in this work. Although structural noise is most pronounced in inhomogeneous materials like composites or concrete, also homogeneous materials like steel produce low signal-to-structural-noise ratios since we are interested in much smaller defects. Because structural noise cannot be identified nor reduced by repeated measurements, additional independent information can only be obtained by considering alternative measurement parameters or inspection techniques.

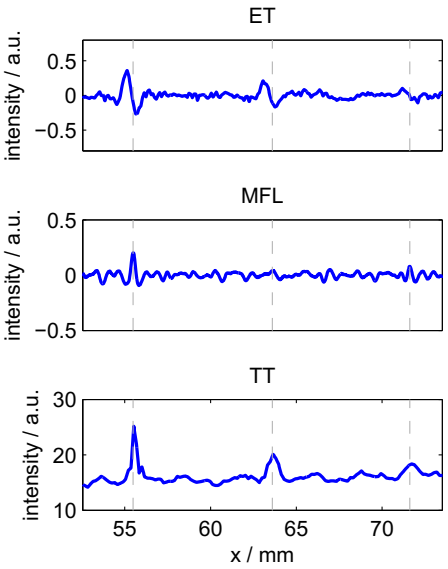
See figure 1.3 as an example. In this figure, a roughly $1\,\mathrm{cm}^2$ large region on the surface of the discussed specimen is shown. Each of the three NDT methods ET, MFL and TT generates a binary image of indications after performing a threshold operation on the respective signal. The threshold was chosen very low, as would be the case to detect small defects. However, this sensitivity to small defects also compromises resistance against structural noise. Consequently, each individual inspection image is filled with numerous false alarms. Without knowing the actual defect positions, there is no way to distinguish true flaws from false indications. Although in this example the indications from structural noise are easily identified based on the image segments'

[3] "[TT] test results can be negatively affected by surface clutter and thermal noise. Therefore, by combining thermal method with other NDT techniques, one may take advantage of both" [12]

Figure 1.2: Example of NDT signals. Known defect positions are indicated by gray dashed lines.



(a) Signals from each NDT method when crossing the ten defects. A false alarm in the TT data set (third row) is highlighted in red.

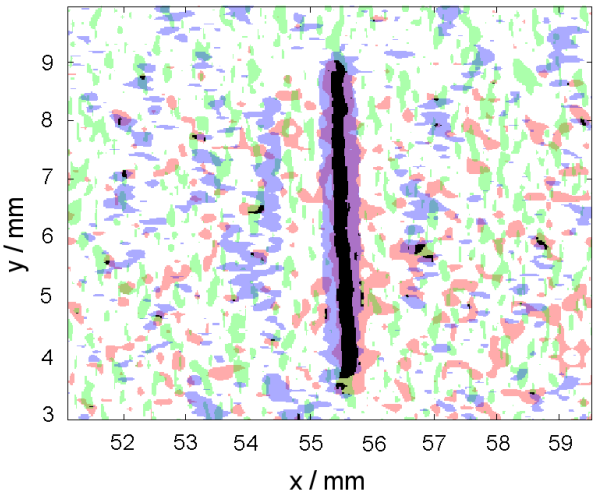


(b) Zoom to three shallowest defects. Defect indications are hardly distinguishable from background signal variations.

shapes, assumptions about shapes are difficult to make in the general case due to the wide natural variation. Nevertheless, at each position on the specimen, the assessment of agreement across different NDT methods clearly identifies the true defect in the center of the figure and retains only a small amount of false alarms.

Challenges for automated nondestructive defect detection To design automatic approaches for both single- and multi-sensory nondestructive defect detection, two main challenges must be overcome. First, there is a vast natural variability of materials and defects, which lead to a diversity of NDT signals. This diversity limits the amount of prior knowledge that can be applied, and thus prohibits making strong assumptions that could otherwise aid the detection procedure. The second challenge is given by the lack of complete understanding about the physical relationships between the test object and the measured signal. Being related to the first reason, this lack of understanding calls for an empirical, data-driven approach. These two issues are now discussed in more detail.

Figure 1.3: Structural noise can be distinguished from true indications by considering the variability of indications across different NDT techniques at each position on the specimen surface. Blue: ET indications. Green: MFL indications. Red: TT indications. Black: regions where all sensors agree.



Considering the first point of vast variability, the following factors influence each test method's ability to indicate defects correctly:

- geometrical properties, e.g. surface roughness
- material properties, e.g. thermal emissivity, electrical conductivity, magnetic permeability
- defect properties, e.g. orientation and size
- measurement conditions, e.g. laser power, sensor-to-surface distance, excitation frequency (see section 2.2)

Among these factors, only the measurement conditions can be controlled by the NDT inspector to achieve the desired performance. Furthermore, due to practical constraints it is sometimes impossible to use the theoretically optimal measurement setup. For instance, high thermal power is always desirable in active TT, but in practice the choice is constrained by requiring a nondestructive solution. Unlike measurement parameters, the other stated factors are not controllable and often unknown, but introduce considerable variability in the inspection results. Due to the low-dimensional inspection output (in the order of 5 features per indication[4] and test method) compared to the higher-dimensional black box system (all unknown factors stated before), the NDT inspector is facing an *ill-posed inverse problem* of estimating the part's state of health from the recorded output signal along with the input to the system, i.e. the known measurement conditions. This problem is ill-posed because usually there are multiple effects that cause similar signal responses. For instance, high signal intensity can be caused by defects or by harmless variations of material or geometrical properties. To approach this problem, it is crucial to obtain as much independent information as possible from the system. This can be achieved by varying the measurement parameters, for instance by multi-frequency eddy current inspection. While this strategy certainly provides additional information about a higher range of depths beneath the material

[4] e.g. signal peak features like position, intensity, shape parameters

surface, it does not change the physical working principle and therefore is susceptible to the same kind of false alarms. To obtain less physically related and therefore more informative measurements, different inspection techniques can be applied to greatly enhance knowledge about the state of the object of interest and thereby to tackle the inverse problem.

Considering the second challenge, one aspect that further complicates solving the inverse problem is that the underlying physical relationships between defect size and the measured signal are not fully understood for some inspection techniques. Whereas physical forward models for ET are readily available [13], models for MFL based on surface-scanning sensors have only recently been developed [14] owing to the novelty of the test method itself. Similarly, laser-induced TT is subject of ongoing research and therefore modeling is still under development [15], let alone being standardized by an international norm. Although such models could in principle be used for defect detection and parameter estimation by fitting them to the measured data [14, 16][11, sec. 6], this approach is infeasible for multi-method NDT data due to the high computational demand of the inversion process, its susceptibility to (structural) noise, and the necessary simplifications that any model implies.

For these reasons, an empirical data-driven approach is taken in this work that makes minimal assumptions about the physical system. At first glance, machine learning techniques seem suitable, since they are able to solve complex tasks based purely on training data. However, those methods are limited by the quantity and the quality of the training data and hence only make sense if a large amount of NDT measurements is available that is representative of all relevant real-world situations. Unfortunately, this assumption conflicts with the high natural variability. It is practically impossible to obtain such a data set because measurements are costly. More importantly, it is technically difficult to produce realistic defects with known characteristics to provide ground truth information for supervised learning. There are two alternatives to the controlled machining of defects. First, natural defects could be created in an uncontrolled way, and after having inspected them, the parts could be dissected for post-hoc ground truth analysis. Although this strategy is feasible in experiments whose scope is limited to a narrow group of materials and defect types, it is costly and time-consuming and cannot be applied to valuable test pieces. The second alternative would be to simulate virtual measurements. While this approach gives the opportunity to create a vast data set, the validity of all results depends on the accuracy of the underlying simulation model, and on all of its simplifications and assumptions, as detailed before. Crucially, such simulations would have to include not only healthy material and discontinuities, but also model the diverse variations in material properties that lead to structural noise. Therefore, even if the lack of availability and other discussed limitations of physical forward models were disregarded, it is questionable if measurements can be realistically simulated in software with acceptable modeling effort. Since none of the two alternatives to the controlled machining of defects appears viable, NDT data are generally scarce. Consequently, one fundamental design principle of the algorithms to be developed in this thesis is to make as few assumptions as possible to avoid overfitting the available set of measurements, and to enable the generalization of the observed detection performances to other defects, test pieces, and materials.

1.2 Aim and scope of this work

The aim of this thesis is to design algorithms to detect near-surface microcracks in ferromagnetic materials, given a set of spatially registered multi-sensor NDT measurements. Despite this restriction concerning the type of material, special emphasis is put on detection algorithms that make minimal assumptions about the measured signals to facilitate wide applicability to other NDT methods and materials. The methods to be developed should yield superior detection performance compared to single-sensor testing by successfully rejecting false alarms such as non-critical indications. This performance improvement should be quantitatively demonstrated using real measurements within a detailed evaluation framework. As a result, the thesis is expected to provide novel techniques and practical guidelines that transfer to other applications of multi-sensor NDT.
The following excerpt from [17, sec. G.2.1] adequately summarizes the relevant guidelines:

> Finding a small flaw is an obvious guideline for any NDE system. While this is necessary, it is not a sufficient condition for effectiveness. Other guidelines include the ability to do this repeatedly under similar but not identical conditions, the ability to distinguish flaws from benign artifacts of similar size, such as microstructure, or surface scratches, and the ability to transition abruptly from passing (nearly) everything smaller than some target size to finding (nearly) everything larger.

1.3 Contributions

Apart from unique experimental contributions to the NDT community, which will be presented in chapter 3, this thesis introduces the following main methodological developments:

1. This thesis presents the first study about the fusion of redundant multi-sensor information (as opposed to complementary; see sec. 2.3) to reduce false alarms using Multiscale Geometric Analysis (MGA) [18]. In this new context, fusion rules that are commonly applied at the signal level are not appropriate anymore. Consequently, more suitable rules were designed in this thesis to successfully reduce the number of false alarms.

2. A new method to fuse spatially scattered locations, here representing flaw indications, is introduced to bypass the need for per-pixel or per-segment fusion at the decision level. Consequently, the method allows to directly account for registration errors, in contrast to per-pixel fusion of decisions. Moreover, avoiding the need for image segmentation obviates inter-sensor segment association, which is typically ambiguous. The proposed method is crucial to enable robust fusion of spatially localized signals, such indications of microcracks.

3. Innovative techniques for evaluation of crack detection when dealing with small sample sizes are developed. Shortcomings of Probability of Detection (POD) analysis, which is traditionally used in NDT, are discussed. Alternative techniques are proposed for quantitative evaluation, which make fewer assumptions than POD while ensuring fair comparison between individual NDT techniques and fusion results. In particular, despite making fewer assumptions, the introduced evaluation framework maintains some of the advantages of POD analysis: It disregards inter-sensor differences in spatial sampling and localization ability, and is tuned to practical applications where it is often sufficient to indicate most of a defect, e.g. without detecting the tip(s) of a crack.

1.4 Thesis outline

This thesis is outlined as follows. The presented detection approaches are systematically organized by their degree of complexity. After giving the necessary background information in chapter 2, a literature overview of data fusion studies in NDT is presented by chapter 3. The methodological part of this thesis starts with a chapter about single-sensor defect detection, which provides basic techniques that will be referenced by the following parts. After that, the first main chapter 5 deals with the fusion of low-level sensory data. To this end, NDT measurements are interpreted as images, and are fused pixel by pixel. This chapter is divided into two sections, the first dealing with fusion techniques that are oblivious to oriented image features, whereas the second section covers more advanced strategies for orientation-aware image fusion using multiscale geometric analysis. After these signal fusion topics, the level of signal abstraction is raised by focusing on the fusion of per-sensor defect detections in chapter 6. The thesis closes with a general discussion about the results presented so far (chapter 7), and gives additional hints from a practical perspective, before summarizing the results and giving an outlook.

Chapter 2

Theoretical Background

2.1 Fatigue cracking

A *crack* is a type of defect that is only vaguely defined as a local separation of the surrounding material, having approximately two-dimensional geometry (very thin opening compared to its length and depth). More specifically, a *microcrack* is operationally defined here as a crack whose penetration depth into the surface is in the micrometer range, for instance as small as 10 µm, and thus challenges detection sensitivity. One major reason for cracking is the phenomenon known as *fatigue*.

Fatigue denotes the degradation of materials under repeated loading, as opposed to monotonic or static load [19]. Such dynamic loads occur for instance in rotating machinery such as bearings, turbines and rotors, but also in rails. The precise definition of fatigue according to ASTM standard [20] is as follows:

> The process of progressive localized permanent structural change occurring in a material subjected to conditions that produce fluctuating stresses and strains at some point or points and that may culminate in cracks or complete fracture after a sufficient number of fluctuations.

Fatigue is relevant because it causes at least half of *all* mechanical failures (including everyday objects)[19]. An example of one (fortunately rare) catastrophic failure is the Eschede train accident in 1998. In this accident, fatigue of one of the wheels triggered a series of events that eventually led to the tragedy in which 101 people died [21]. Moreover, fatigue is responsible for numerous airplane accidents which are listed in [22].

What makes fatigue so dangerous is that objects do not seem to show any sign of warning such as plastic deformation before their sudden fracture. But in fact, most of fatigue life is actually spent on nucleating and growing an initially small fatigue crack well before it reaches its critical size at which the structure is not able to support the applied stress anymore. This characteristic provides a window of opportunity during which the defect is large enough to be detected during inspection, but still small enough to ensure safe operability. In order to detect fatigue, it is first necessary to understand its origins.

During repeated loading, the applied stress is often not evenly distributed across the part, but instead concentrates at certain locations where small cracks are likely to develop as a consequence. Sources of stress concentration are material imperfections such as small cracks, containment particles or voids, and geometrical discontinuities such as sharp edges and corners. Other local influences like corrosive environments or changes of temperature further promote fatigue. But even in the absence of these

factors fatigue may develop when so-called *Persistent Slip Bands* form under cyclic load. These bands, which are only a few micrometers wide, roughen the material surface by creating extrusions and intrusions, and are "likely to be critical precursors to the nucleation of fatigue cracks" [23]. Once a crack has nucleated, it starts penetrating into the material as the external load drives crack growth by periodically opening and closing it. Several models exist to describe crack growth, for instance Paris-Erdogan Law [24]. Because such models describe accelerated rather than constant growth during fatigue life, it is essential to detect cracks as early as possible. This stresses the need for highly sensitive detection methods considering the involved miniature crack sizes.

Importantly, fatigue most often develops at the surface of a component. This is because stress concentrators such as environmental conditions, geometrical discontinuities and slip bands only affect near-surface areas [19]. Likewise, in his historical review [25], J. Schijve concludes that "fatigue crack initiation is a surface phenomenon", because slip bands form more easily at the free surface where there is no material at one side. Therefore, although internal imperfections may also cause fatigue, an important tool for quality assurance and failure prevention is nondestructive surface inspection.

2.2 Nondestructive surface inspection of ferromagnetic parts

Among the NDT methods that qualify for the task of near-surface crack detection, special attention is paid to those that allow automatic data acquisition and provide accurate, objective and reproducible results. In this sense, adequate methods are eddy current testing (ET), magnetic flux leakage (MFL) testing, and thermal testing (TT)[1]. Since each is based on unique physical effects, they provide independent "views" of the tested object. As the fundamental generators of signals to be fused within the scope of this work, these techniques are now briefly introduced and compared at the end of this section.

Eddy current testing (ET) The working principle of this electromagnetic method is depicted in 2.1. An eddy current probe containing an excitation coil is positioned near the specimen's surface. Through this coil, an alternating current ① at an adjustable frequency creates an oscillating magnetic field ②, called the *primary field*. Note that in the figure, only a static field is shown that exists momentarily. The field's oscillations induce voltage in the specimen's near-surface region that creates circular eddy currents ③. These eddy currents, which are an undesirable side effect in many applications outside of NDT but are the key element for this inspection method, create a *secondary magnetic field* themselves ④. By Lenz' law, the secondary field opposes the primary field which is measurable through the complex-valued impedance of the coil ⑤. When an inhomogeneity is located near the probe, the eddy currents are disturbed which also impairs the secondary field. Consequently, the coil impedance is increased, which produces an indication in the measured signal.

To inspect a larger area of the specimen, a mechanical scanner system is installed that moves the sensor over the specimen's surface, while the response is sampled at

[1]Further possibilities are Ultrasonic Testing and Microwave Testing, but including these would exceed this thesis' scope, and would also make the study less practically realizable. Nevertheless, since the methods developed in this thesis make minimal assumptions about the underlying physical processes, they are expected to work with other NDT methods as well.

regular intervals. Often, the scanner follows a meander-like path while collecting line scans. Alternatively, for rotationally symmetric specimens, line scans are more easily obtained by rotating the test object underneath the sensor.

For the detection of small defects, a well-suited type of sensor is the so-called *differential probe*. This probe type consists of a pair of pickup coils that are measured against each other. Rather than an absolute measurement of impedance, this probe assesses only local changes in impedance. Therefore, large-scale variations of electromagnetic properties (which are not indicative of small defects such as cracks) are not reflected in the signal. Moreover, differential sensors provide higher SNR than absolute probes concerning measurement noise by allowing stronger signal amplification without risking saturation effects. But unlike absolute probes, the output of differential ET sensors depends on the defect orientation, which requires two scans using perpendicular probe orientations.

Figure 2.2 shows an exemplary test result from ET of a steel test specimen. The signals were obtained by crossing a machined groove, representing a structural discontinuity, with the sensor. On the left part of the figure, the real part (top) and imaginary part (bottom) of the measured impedance is plotted. Because the differential sensor (red) effectively performs spatial subtraction of the respective impedances, as shown by the absolute sensor (blue), the resulting differential signals resemble the first spatial derivative of the absolute signal for both signal components. Therefore, absolute probes indicate the actual defect position by large signal intensities, whereas differential probes indicate defects by near-zero values in the transition area between the two characteristic peaks. The polarity of the bi-modal peak depends on the orientation of the differential probe. Rotating it by 180° makes the differential probe either produce a forward or backward difference signal. The particular choice is arbitrary but must be noted for subsequent signal processing.

A different form of visualization is presented by the right part of the figure, where a curve is formed by plotting the two-dimensional impedance values for consecutive measurement positions on the specimen. Both sensor responses form tilted lines, with the inductive reactance (vertical axes) exhibiting more variation than the resistance component. In fact, the sensors were calibrated before the measurement so that small defects like cracks would mainly affect the vertically displayed component. This simplifies the defect detection step by being able to work with a one-dimensional signal.

The main measurement parameter, apart from the choice of probe, is the frequency f of the alternating current in the excitation coil. This frequency mainly controls the penetration depth of the eddy currents, according to $\delta(f) \approx 1/\sqrt{\pi f \mu \sigma}$. The symbols μ, σ denote the material's magnetic permeability and electrical conductivity, and $\delta(f)$ is the *standard penetration depth* in mm. This is defined as the depth at which the exponentially decaying eddy current density drops to roughly 37 % of its value at the surface. Although the given formula only holds true under theoretical conditions [26, pp. 31–34], it is commonly used to approximate the relationship between excitation frequency and penetration depth also in practical settings. Because the permeability and the conductivity are fixed material properties, the frequency is the only experimentally adjustable quantity in this relation.

Figure 2.1: Principle of ET.

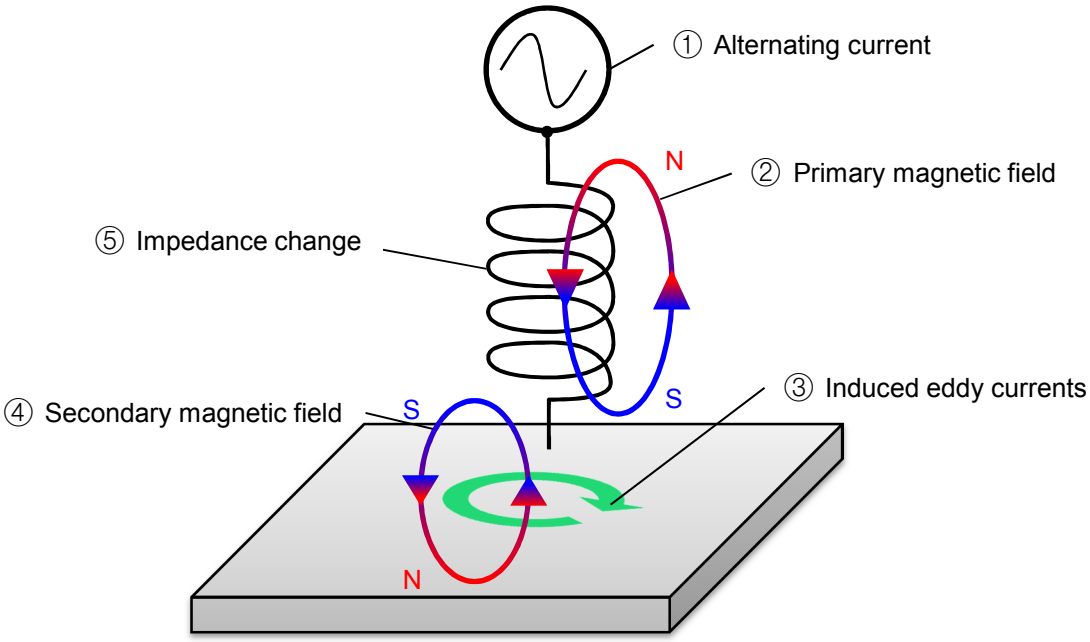
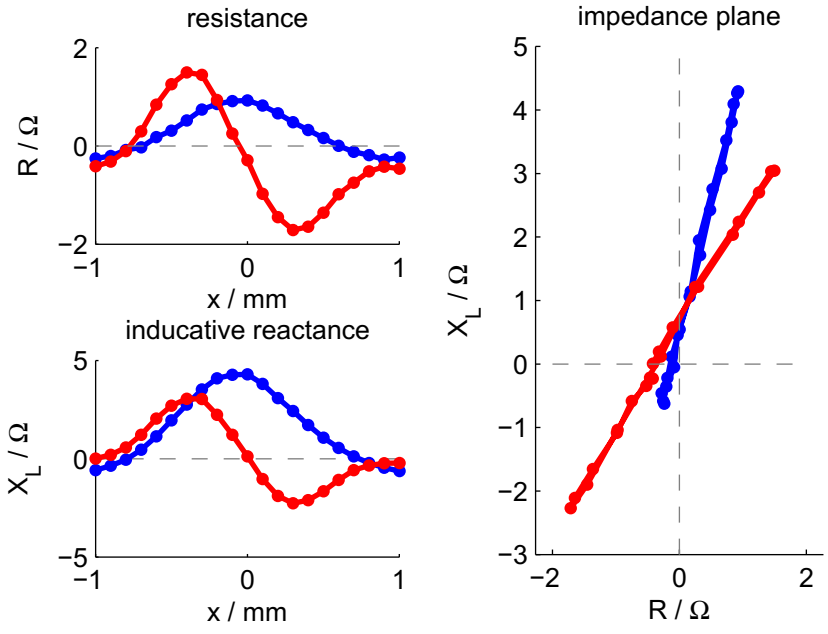


Figure 2.2: Typical signals from ET when moving the prove over a defect. Blue: absolute probe. Red: differential probe. Left: the two components of the complex-valued measured impedance are plotted along a line crossing the specimen surface. The defect is located near the zero position. Signal samples are marked with dots. Right: both components of impedance plotted in the same diagram.



Magnetic flux leakage testing (MFL) Like eddy current testing, magnetic flux leakage testing is based on electromagnetic principles. However, unlike ET, MFL can only be applied to ferromagnetic materials, such as iron, nickel, cobalt and their alloys, for instance steel. In MFL, the specimen is exposed locally or globally to a static magnetic field, which spreads inside the material. See figure 2.3 for an illustration. When structural inhomogeneities are present, they create interfaces between two materials that may have strongly contrasting relative magnetic permeabilities μ_r, like ferromagnetic objects ($\mu_r \gg 1$) and air-filled cracks ($\mu_r \approx 1$). But air cannot support the high magnetic flux density anymore that is present in the surrounding material. Therefore, if such interfaces lie perpendicular to the magnetic field lines, like shown in the figure, physical continuity conditions of the field components force the field to "leak" out of the specimen [27]. The traditional way to detect this magnetic flux leakage is *magnetic particle inspection*. In this approach, a ferromagnetic powder is spread across the magnetized specimen. These particles concentrate near the stray fields and thus indicate inhomogeneities by their distinct color or by their fluorescent properties. However, particle inspection is not quantitative and lacks automation. Therefore, magnetic field sensors, such as Giant Magnetoresistance (GMR) sensors, are an attractive alternative.

Compared to other sensors, this type of magnetic field sensor has considerable advantages for NDT applications [27] due to its miniature size and high sensitivity. Because the sensing elements on the chip have a size of only around $1\,\mu m \times 190\,\mu m$, high spatial resolution can be achieved and close proximity to the specimen's surface facilitates the detection of weak stray fields as produced by microcracks. Similarly to differential eddy current sensors, these GMR sensors may be constructed as *gradiometers* to measure field differences rather than absolute field strength. The three spatial components of the magnetic vector field are measured by separate sensors. For defect detection, the most relevant field component is the normal direction to the specimen's surface, as it is in principle sensitive to arbitrarily rotated defects in the surface plane[2] [27, p. 75]. Moreover, gradiometers in this configuration allow for robustness against changes in the external magnetic field. An example of a differential GMR signal from MFL (normal field component) is shown in figure 2.4. The characteristic pattern resembles the imaginary component of a typical differential eddy current signal (figure 2.2), but can be much narrower (thus necessitating denser spatial sampling) and the GMR peak amplitudes have a higher dynamic range for various crack sizes (not shown in the figure). The peak polarity is determined by the orientation of the gradiometer relative to the direction of the external magnetic field.

For surface inspection, a scanner moves the sensor line-by-line over the specimen's surface, similary to ET.

For MFL using GMR sensors, most measurement parameters are fixed by choosing a GMR chip. These are mostly geometrical issues like the minimal distance between the sensor and the surface, the size of the sensing elements and, in case of a gradiometer, the distance between the two sensors. Depending on surface roughness, the sensor should be placed as close as possible to the surface for maximal sensitivity to weak stray fields, for instance produced by microcracks. The sensing elements should be made large enough to exhibit favorable SNR, but not too large to retain spatial resolution. Moreover, the distance between a pair of gradiometer elements should be made large enough to facilitate significant indications after differentiation, but small enough to minimize the effects of noise from external magnetic influences. The experimental magnetization field

[2]However, the indication strength still depends on the angle between the defect and the magnetic field lines.

should be as strong as possible, up to magnetic saturation, to maximize the stray fields' strengths [28]. Moreover, its direction relates to the orientation of defects to be found. If this orientation is unknown or unconstrained, several inspections at various directions of the external field must be carried out to maximize the reliability of defect detection. Yet, indications can still be obtained for defects that "have an angle of much less than 45 degrees to the direction of the magnetic field [...] Furthermore, most cracks are not really straight, but serrated, so that in practice always certain parts of the cracks can be recognized which is, in most cases. sufficient for the test result." [29, p. 28]

Figure 2.3: Principle of MFL. Note that the course of the field lines is only shown in a schematic, non-realistic way. N and S denote the north and the south magnetic pole of the external magnetization.

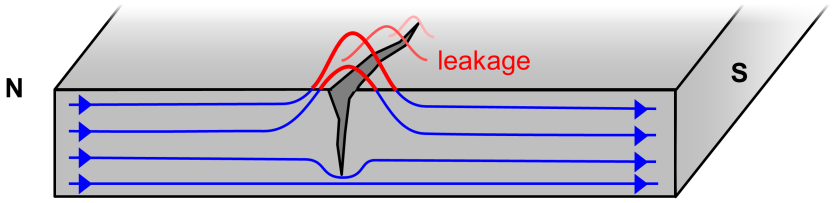
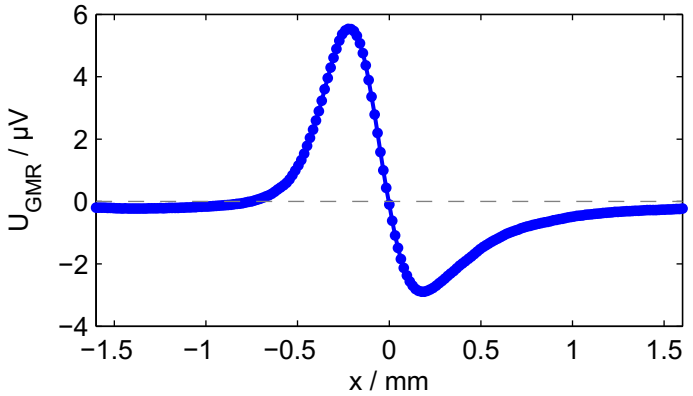


Figure 2.4: Typical signals from MFL (schematic). This is the response of a gradiometer measuring the normal component of the magnetic stray field relative to the specimen surface, while the probe is crossing a defect. The defect is located at the center of the horizontal axis. Signal samples are marked with dots. The measured signal is a voltage, but can be converted to field strength ($\mathrm{A\,m^{-1}}$) after calibration.



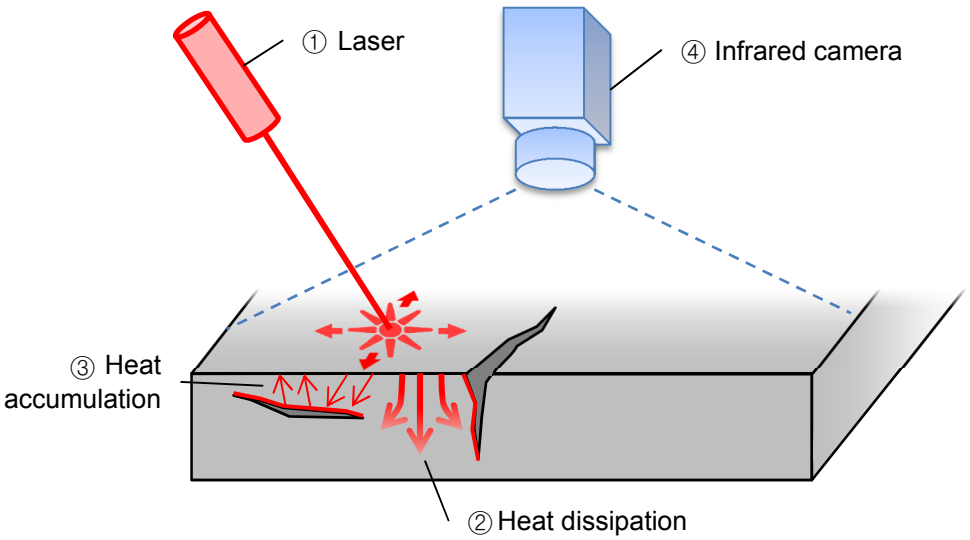
Laser-induced active thermography testing (TT) This nondestructive inspection technique is quite different from the electromagnetic methods mentioned before. The contrast mechanism is based on thermal flow, which facilitates testing of a broader class of materials. The best results are obtained for nonreflective surfaces, which in addition can be blackened to increase the surface's thermal emissivity. The inspection procedure is shown in figure 2.5. Heat is locally induced by a high-power laser ① for crack detection. In defect-free regions, the resulting heat flow is able to dissipate ②, whereas defects cause localized heat accumulation ③. Even non surface breaking defects are detectable, because they reflect the introduced heat flow back to the surface. An

infrared camera ④ monitors the temperature on the specimen's surface and generates a digital image sequence for processing while the laser is systematically moved over the specimen. Note that neither the heat excitation system nor the sensing camera require direct contact with the specimen's surface, which is a practical benefit. Also, no scanner system is necessary because the camera conducts an instantaneous full-field measurement at each frame. However, because different locations on the specimen are visited at different times by the laser, additional processing is required to achieve either temporal alignment or to construct invariant features regarding time [15, 30, 31][12, sec. 6].

Compared to global excitation, for instance by flash lamps, laser-based heating has the advantage of excitation from larger distances while still introducing high energies into the specimen. [32]. More importantly, unlike global excitation, locally excited TT indicates defects that are oriented perpendicularly to the surface, such as cracks. In particular, flash excitation is less suited to identify surface-breaking cracks [31]. Concerning the rotation of cracks around the surface normal vector, TT does not favor any particular defect orientation.

The laser power, speed and the spot's shape and size on the material surface are the most relevant parameters of the excitation. Higher power leads to increased contrast and deeper penetration, or allows to increase the laser's movement speed. However, the power cannot be arbitrarily increased due to technical reasons and to ensure that the material under inspection remains unaffected. Higher speeds facilitate faster inspection, but less energy is directed to each passed position. Similarly, the spot area marks a trade-off between localized energy and inspection speed.

Figure 2.5: Principle of laser-induced active TT.



Comparison In this section, the benefits and drawbacks of the briefly described NDT techniques are summarized; see table 2.1. While all of these methods are sensitive to near-surface cracks, each method is based on a unique physical principle, thus providing independent pieces of information for fusion. Although ET and MFL both make use of electromagnetic mechanisms, they differ in the type of magnetic field that is measured: ET is sensitive to changes in a magnetic field from induced currents, whereas MFL

measures stray fields that exit the surface after the part was subjected to static magnetic saturation.

Each testing method depends on the choice of sensor/actuator and certain measurement parameters. For ET, the main parameter apart from the choice of the probe is the excitation frequency f. In performing MFL with GMR sensors, all relevant parameters are built into the sensor, apart from the orientation and strength of the external magnetic field. A setup for TT involves choosing a laser type and its movement pattern. The type and positioning of the infrared camera determines the achieved spatial resolution.

Whereas the electromagnetic methods excite and sense punctually, in TT the camera obtains a full-field measurement. The infrared camera's pixels can be considered as an array of punctual elements that operate in parallel. Therefore, inspection duration is much shorter for TT than for ET and for MFL with GMR sensors.

Concerning the applicability of the NDT methods to different materials, MFL is certainly the most restricted. While ET is able to handle a broader class of materials, thermal conductivity is the least demanding requirement.

Spatial resolution, that is the ability to discern close-by defects, is comparably better for MFL and TT than for ET due to the physical limitations imposed by the size of the pickup coil. This coil cannot be arbitrarily miniaturized because this would also reduce the probe's sensitivity. One solution is to replace the eddy current pickup coil by a small yet sensitive GMR sensor, similarly to automated MFL [33]. This approach is specifically suited for the inspection of deep flaws, where low excitation frequencies lead to poor SNR in ET, but is also appropriate for the detection of short near-surface cracks. Despite their theoretical advantages, GMR sensors have found only limited applicability in practice, supposedly because accurate localization of near-surface defects is less important than their overall detection. Another reason may be that smaller sensors require finer spatial measurement grids, thus prolonging the inspection duration. In contrast to ET applications, for automated MFL the magnetoresistive sensors are unrivaled. The spatial resolution of thermographic testing can be enhanced by obtaining high-quality cameras and by moving the camera closer to the object, thus reducing the physical area that each pixel covers. Of course, this would narrow the camera's field of view as a negative side effect.

Of the three test methods, only MFL is blind to certain defect orientations. If differential probes are used, then ET must also take defect orientation into account. In contrast, active TT is able to indicate flaws regardless of their orientation.

For near-surface crack detection, the three proposed NDT methods differ in how deep beneath the surface a defect is still detectable. For a given material, in ET this depth mainly depends on the excitation frequency and may range up to 1 mm in iron [26, p. 34]. This enables the technique to detect inhomogeneities even below paint or coating, where visual testing is not applicable. MFL is also sensitive to sub-surface flaws, although sensitivity is limited to the micrometer range due to the weak stray fields. In contrast, TT has the highest potential for the detection of deeply located inhomogeneities among the studied techniques. This is is due to the high thermal energies that are realizable with laser technology, thus generating heat flux in deep regions of the specimen. But despite the high potential for laser-induced active TT, other test methods are still more widely used in practice. The main reasons are that conventional inspection methods are easier to use, cheaper, and do not require safety regulations which apply when working with high-power lasers.

Although the three described inspection techniques differ in several ways, all are

well-suited for fatigue crack detection, because those defects originate directly at the surface, as described in section 2.1. Therefore, they lend themselves to multi-sensor data fusion techniques, which are overviewed in the next section.

Table 2.1: Comparison of three NDT techniques for surface inspection of ferromagnetic materials

	ET	MFL (GMR)	TT (laser)
physical principle	electromagnetism	electromagnetism, giant magnetoresistance	thermal flow
inspection parameters	probe type (absolute, differential, ...), frequency	sensor type, magnetization strength / orientation	laser power, laser speed, laser spot shape / size, camera resolution, camera distance
scan mode	line scan	line scan	full field
scan duration	long	long	short
material requirements	electrically conductive	ferromagnetic	thermoconductive
spatial resolution	coarse (coil size)	fine (sensing element area)	fine (camera's pixel size)
defect orientation determines	differential probe orientation	direction of external magnetization	-
max. defect depth	medium – depends on frequency	shallow	deep
typical test cases	aerospace, tubes (e.g. heat exchangers), welds	pipelines, bearings, tubes	welded joints, engines in aerospace and automotive industry
distance between sensor and specimen	<1 mm	<1 mm	<1 m
standards	[34, 35]	*NYS**; see [35–37]	*NYS**; see [35, 38, 39]
additional remarks		demagnetization may be required before / after testing	safety regulations apply due to high-power laser

**NYS* = Not yet standardized

2.3 Data fusion

Data fusion, also referred to as information fusion[3] , is a multidisciplinary research field that is drawing considerable attention. Because sensors have become ubiquitous in industry but also in our everyday lives, the availability of huge amounts of complex interrelated data challenges our way of information extraction and decision making. Although this challenge is usually problem-specific, common concepts, theories and algorithms have been devised during the last decades to establish an independent field of research. One recent definition of data fusion is given by H.B. Mitchell (2012) [41]:

> [Data fusion denotes] "[...] the theory, techniques and tools which are used for combining sensor data, or data derived from sensory data, into a common representational format". In performing data fusion, our aim is to improve the quality of the information, so that it is, in some sense, better than would be possible if the data sources were used individually

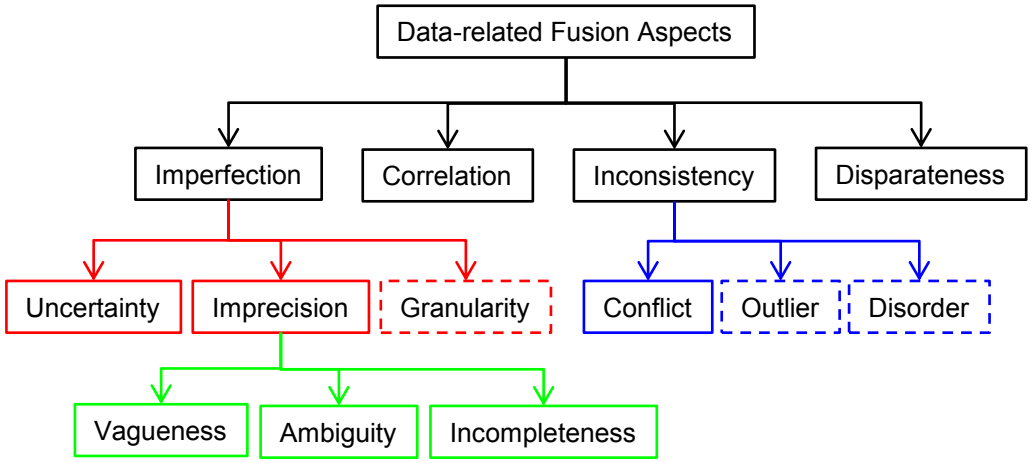
To obtain a 'common representational format', several forms of signal normalization and association are generally necessary to relate the information from the different sources. The abstract notion of *quality improvement* is however application-dependent. Specifically for nondestructive testing, this concept might mean increased sensitivity or specificity in performing defect detection, or more accurate estimates of defect or material characteristics. Furthermore, using data fusion techniques, ambiguities can be resolved and the specimen area that is covered by inspection can be increased. However, to achieve these characteristics of quality, a number of challenges must be overcome.

Challenges of multi-sensor data sets In the review by Khaleghi et al. (2013) [42], a taxonomy of the challenges that multi-sensor data sets bring about is provided, and the authors survey the predominant data fusion algorithms and theories that have been developed to tackle these challenges. In this section, the typical issues are revisited from the perspective of nondestructive testing.

See figure 2.6 for the taxonomy of challenges. Khaleghi et al. identified four major categories. Data *imperfection* is a general notion that comprises the manifold shortcomings of typical sensory data. Among these, *uncertainty* denotes the deviation from the measured or computed value to the true (unknown) value, and has many sources. For instance, the well-known *measurement uncertainty* impairs all sensory output. But there are also other, potentially more severe sources of uncertainty, for instance localization errors due to physical limitations, or unsatisfactory image alignment quality. To face uncertainty, probabilistic (often Bayesian) techniques are usually employed. However, inconsistencies among sensors cannot be handled using the standard approach but require dedicated treatment [43]. Alternatively, the *Dempster-Shafer theory* [44] has been developed to be able to quantify ignorance, that is probability mass that is not assigned to any hypothesis. Concretely, the task of defect detection is a classification problem to assess the probabilities of the two hypotheses: $\mathcal{H}_1$='defect' / $\mathcal{H}_0$='no defect'. However, there might be situations in which a sensor is known to be unreliable, so that its assignments of probability to either of the two classes cannot be fully trusted. The Dempster-Shafer theory therefore allows to reserve a portion of

[3] Although the terms *data* and *information* are in fact not synonyms, it is accepted among researchers in the field to not make a distinction when referring to fusion. Nevertheless, in some applications *information fusion* emphasizes the fact that information was extracted from the data before fusion [40].

Figure 2.6: Challenges of multi-sensor data sets. Adapted from [42, fig. 1]. Colors group entries by same sub-category. Dashed boxed indicate aspects that are less typical for NDT data sets.



the probability mass for the class 'any hypothesis could be true', and provides a rule of combination for fusion.

Another type of imperfection is called *granularity*. By this concept, Khaleghi et al. denote the inability to distinguish between two objects due to lack of sufficient information. For instance, in NDT, a true defect's indication might be indistinguishable from a harmless one, given the results of a set of nondestructive test outcomes. However, the inclusion of a further sensor might yield additional information that was previously lacking. The *rough set theory* [45] is mentioned by Khaleghi et al. as an adequate data fusion technique to tackle knowledge granularity. A third category of multi-sensor data imperfections is denoted by *imprecision*. This covers *vague*, *ambiguous* and *incomplete* information. Vagueness arises when information sources are not able to provide distinct data. In NDT, this might be encountered if test results are informally described by humans, e.g. 'there is something suspicious in the central region of the specimen'. Although the vague expressions 'something suspicious' and 'central region' are imprecise, they might still be very valuable pieces of information, especially if combined with other sources. The fuzzy set theory [46] is well-suited to cope with vague data. When imprecision appears as *ambiguity*, information to be fused may have several alternative interpretations, which is typical for single-sensor indications. A third type of imprecision among imperfect data is *incompleteness*. In multi-method NDT, this can easily occur if the inspected areas among different methods are not completely overlapping, or if a sensor drops out during an automated scan. A common source of partially missing data are bad pixel artifacts caused by imperfect cameras. Moreover, in multi-sensor NDT, often each measurement is carried out at different locations on the specimen, so that each sensor is in fact missing all other sensors' readings at these locations. The latter problem is usually overcome by signal interpolation before fusion. More generally, in such poorly informed situations, the possibility theory [47] is an adequate approach for multi-sensor data fusion.

Correlation is a further characteristic of multi-sensor data sets. This phenomenon occurs when different sources of information yield overlapping pieces of information, i.e. they partially quantify the same underlying cause. It is important to become aware of

correlated sources to correct their impact on the fused result. In the extreme case, if two sources are treated as being independent when they are actually correlated, the same underlying piece of information would enter the fusion rule twice and would therefore have a larger weight compared to the other information sources. In NDT, correlated information sources are for instance inspections that are based on the same physical principle with minor changes in the measurement parameters.

A third category of fusion-related challenges identified by Khaleghi et al. is inconsistency, of which one realization is conflict among information sources. Such disagreement happens in NDT for example when the test methods are sensitive to different types of defects, or when they react to specific geometrical or material variations thus producing false alarms. Interestingly, the conflicts are what creates added value over single-sensor inspection; see also section 2.3. Often, the choice how to cope with disagreement among sensors is at the heart of fusion rule design.

Finally, one fundamental issue with multi-sensor NDT data is their disparateness. Leaving aside differences in data formats, the nature of the quantified information is typically completely unrelated. For instance, inspections might return information about electrical, magnetic and thermal properties. Each of these data sets has individual dimensionality, physical unit and intensity range. This requires dedicated processing and interpretation to transform them into a representational format in which the extracted information is somehow compatible across the data sets.

Fusion at different levels Data fusion can be performed at various levels of signal abstraction, each with specific drawbacks and advantages. Luo and Kay (1990) [48] define the four layers *Signal Level*, *Pixel Level*, *Feature Level* and *Symbol Level*. These stages reflect a prototypical process of information extraction from raw sensory data: First, unrelated individual measurements (e.g. individual samples) are composed to form more structured data (e.g. images). From these data, typically some sort of features are extracted to encode the relevant information. Finally, based on the features, the quantity of interest is estimated, for instance a classification task or parameter estimation procedure is carried out. In this work, the signal and pixel levels are both considered as low levels of signal abstraction, and will be summarized here as the *signal level*. Moreover, the highest level of abstraction, symbol level, will synonymously be denoted here as the *decision level*.

In figure 2.7 the typical processing pipeline is depicted for multi-sensor defect detection at two different levels of signal abstraction: the signal level (top) and the decision (or symbol) level (bottom). Although signals are represented by rectangles, symbolizing images, the illustrated concepts also apply to non-spatial data like dynamic or spectral signals. In both depicted fusion approaches, per-sensor preprocessing is a necessary first step. After that, at signal level, the prepared sensory data are first spatially aligned so that the same pixel in all images reflects the same position on the tested object. Specifically, this step requires signal interpolation to estimate sensory values at common non-measured positions. The result is a set of spatially (and/or temporally, spectrally etc.) aligned signals. After alignment, the signals' intensities must often be made comparable, known as *radiometric normalization*[41, ch. 8], so that fusion rules are meaningful. Fusion is now possible by combining the data among all sensors per pixel to generate a single fused image. This image is supposed to have superior quality over the input data, for example improved SNR. Any further processing, such as flaw detection, is now carried out on the fusion result.

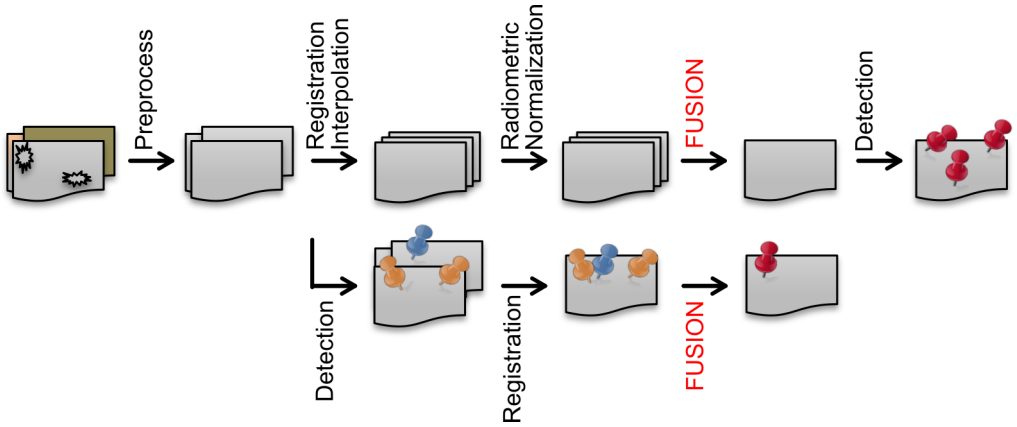
In contrast, at a higher-level fusion approach, the step of data combination is

prolonged towards the end of the processing chain. This means that most of the pipeline is executed for each sensor separately, and only the final results are fused. Likewise, the spatial alignment step does not operate on raw data, but instead associates higher-level results, such as hypothesized defect locations. Therefore, interpolation is usually not necessary at this fusion level so that the application is able to work with the original data. Moreover, radiometric normalization is not required. The result of fusion is a unified outcome based on the results from each individual sensor.

Using diagram 2.7, the various trade-offs between fusion at a lower and a higher level of signal abstraction can be explained. Because low-level fusion approaches are often concerned with improving the overall quality of the data, they are fairly independent from the final processing stage. Therefore, such approaches are most useful as a general preprocessing step, to simplify any (possibly unknown) subsequent operations. The lower the level of fusion, the closer data fusion comes to *data integration* – to combine the heterogeneous source data sets into a fused data set that satisfies a given format, or more generally, some quality criterion. However, this means that higher-level information is not included in the fusion process, potentially leading to suboptimal results. In addition, especially in signal-level fusion, accurate spatial alignment of the input signals is crucial [49–51], because misalignment can hardly be compensated during fusion. Moreover, especially with measurements from different NDT methods, more effort is required to normalize the disparate signal intensities, compared to high-level fusion.

For high-level fusion, the situation is reversed. All relevant analyses have already been carried out for each source data set and only the final results need to be combined. One obvious drawback of this approach is that the fusion rule is unable to access lower-level information to aid the decision. But on the other hand, this gain in independence lets high-level fusion abstract from the specific types of the original source data sets, so that the fusion rule is free of data normalization issues.

Figure 2.7: Fusion at different levels of signal representation. Top row: Fusion at the *signal level*. Bottom row: Fusion at the *decision level*. Grey boxes denote data sets (e.g. inspection images) and pins denote hypothesized defect locations.



Fusion of redundant and complementary information Each inspection result contains a certain amount of (imperfect) information about the health state of the tested object. Different parts of this information can be classified as *redundant* or *complementary*. Complementary information is not shared among different inspection methods and therefore adds to the available information. For instance, whereas radiographic testing like x-ray computed tomography is sensitive to changes in material density, ultrasonic

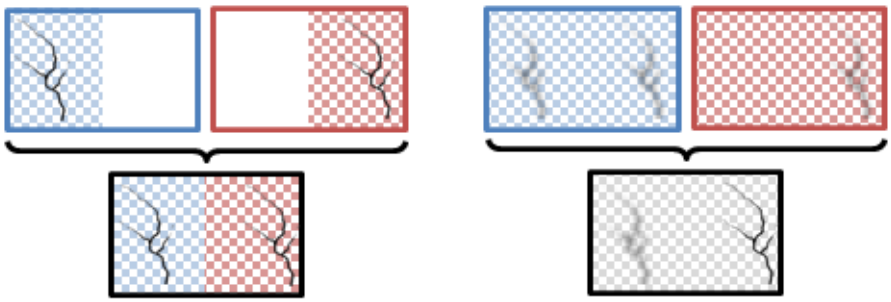
testing produces indications at changes in acoustic properties. Therefore, ultrasonic inspection might indicate narrow or planar discontinuities even when the change in density is not sufficient to facilitate radiographic analysis. The two NDT methods could be combined to form an inspection system that is sensitive to both defect types.

In this work however, it is primarily the redundancy among different NDT methods that is exploited, by *making the fundamental assumption that all inspection methods considered here are sensitive to the same defect types in the same size range.* Whereas inter-sensor agreement directly increases the reliability of an inspection result, disagreement is also valuable in that ambiguities can at least be quantified and reported to the operator. Even more, under the stated assumption, inter-sensor conflict increases the belief in a false alarm. In contrast, even for obvious defects the single-sensor inspections often provide an unsatisfactory assessment of the test object's condition, in the sense that they are lacking information about the variability of the results regarding different test methods.

Figure 2.8 provides a visualization of the described concepts using nondestructive defect detection as an example for two test methods, shown as red and blue. Whereas complementary inspections (left) can be fused to extend the spatial coverage and thus to report both exemplified defects in a single coordinate system, redundant inspections help to reduce the uncertainty about the presence of the two defects. Whereas the left defect is only detected by the blue method and thus the operator is left in doubt[4] of its true presence, the right defect is confirmed by both methods and thus it is detected with high confidence.

In conclusion, the fusion of redundant pieces of information provides the opportunity to distinguish false alarms from true defects and allows more comprehensive assessments based on the variability of the results.

Figure 2.8: Visualization of redundant and complementary information in the context of NDT defect detection. Blue and red denote different NDT inspections. Blurry cracks represent uncertainty.



(a) Fusion of complementary information to extend the spatial coverage.

(b) Fusion of redundant information to reduce the uncertainty

Spatial registration Fusion requires the spatial association between the individual NDT measurements so that it is clear which part of each signal denotes which location on

[4]It might as well be a false alarm.

the specimen surface. Mathematically speaking, the signals of each individual inspection have coordinates that are expressed in some local coordinate system, which is visualized in figure 2.9. Because the individual inspections are sometimes conducted independently and therefore no convention exists as where to start measuring, in what exact direction and to what spatial extent, the geometrical relationship between the different local systems is unknown in general. Especially when the specific choices of origin, orientation, etc. were not logged, the establishment of such a geometric relationship between the systems, which is known as registration, is not trivial. Registration comprises the identification of coordinate transformations with which a fixed point on the specimen surface can be mapped to any coordinate system to arrive at the signal value that was measured at this position. It is often convenient to avoid registering all pairs of local systems, but instead declare one of them as the *global coordinate system*, or *reference system*. As figure 2.10 shows, the geometrical correspondence between any pair of local systems can be defined indirectly via the reference system, thus reducing the required number of pairwise registration operations. Although the choice which of the local systems should be made the reference system is arbitrary, in practice a reference measurement to evaluate the fusion task, if available, is particularly suited.

Figure 2.9: Local coordinate systems overlaid on top of a photograph of a specimen having ten vertical grooves. In this schematic graphic, the relationship between all local systems is known, so that their relative position, orientation and scale can be plotted here. The green system was chosen as the reference system for this figure.

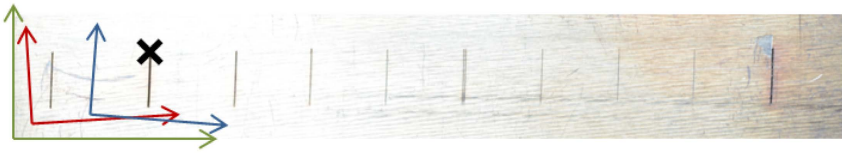
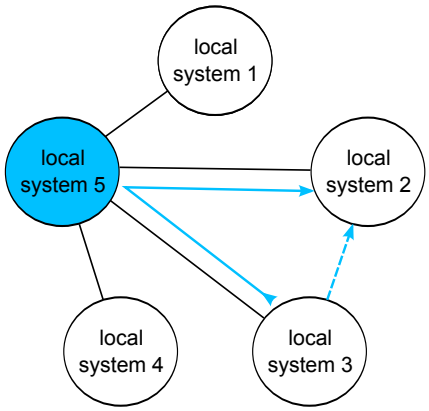


Figure 2.10: Visualization of coordinate transformations after registration. Instead of computing all pairwise mappings, it is sufficient to establish transformations between the reference system (here: system 5) and each of the other systems.



Automatic or semi-automatic registration is a well-explored field of research (see e.g. [41, chapter 5], [51]), because it is required in many applications that deal with spatial information. However, many of the proposed methods rely on signal landmarks that must match between the individual data sets, for example gradient-based image

features [52, 53]. In NDT data, a landmark may be given by any indication that is produced by all test methods, such as geometrical indications or prominent flaws. Geometrical indications are preferable over defects, because they usually produce strong indications in all test methods and have spatially large extent. In contrast, defects are not well-suited, because this would lead to a chicken-and-egg problem: For registration the detection of defects is required, which in turn necessitates registration. Moreover, defects (if any) are usually scarce, and small defects do not generate unique fingerprints in their measured signals which would allow to match them robustly.

Whereas in an industrial setting, registration is readily automatable [54], the registrations in this work were carried out manually. Specifically, initial corresponding landmarks were visually identified based on the vague knowledge about each inspection's geometrical measurement plan. Subsequently, an initial coordinate transformation model (e.g. affine transformation) is fitted to the found landmark correspondences. Based on the transformation, the data are interpolated to generate images whose pixels loosely correspond to the same location on the specimen surface. The registered pair of images often leads to visual clues that help improving the position of the landmarks. This process is iterated until the (visual) registration error is satisfactory or cannot be decreased any further. From there on, automatic registration might help to fine-tune the found transformation parameters by maximizing a measure of similarity between the data sets using numerical optimization techniques. In particular, the data sets investigated here show that gradient-free methods such as the Nelder-Mead simplex algorithm [55, 56] are promising, because gradient-based methods might not find any feasible search direction in the parameter space. Finally, the optimal transformation model is used for spatial association – either during image registration at signal-level fusion, or for alignment of detection results at decision-level fusion.

Because the local coordinate systems usually only differ in position, orientation and scale[5], global transformation models like affine or projective functions are usually suited. In some cases, when for instance a camera is involved, some degree of nonlinear distortion may require correction or should be accounted for in the transformation model itself.

During this thesis, all data sets are assumed to be registered using the just described procedures, that is there exist forward and inverse transformations from each local coordinate system to a designated reference system. Therefore, detailed discussions about registration are omitted in each of the experimental sections of this thesis.

After these concepts of data fusion have been discussed in the context of nondestructive testing, the following section presents an overview of the state of the art in this area.

[5]Scale can be neglected if all local coordinates are specified in physical units, for example in mm. This requires knowing the spatial measurement grid resolution.

Chapter 3

Literature Survey

Although multi-sensor NDT is not yet well-established, a considerable amount of data fusion studies has already been published, as surveyed in the first book about NDT data fusion from 1996 [57], its successor from 2001 [4] and the most recent survey article from 2007 [58]. This section's literature survey adds to these comprehensive works by focusing on studies that were published after the latest review in 2007. Specifically, the period from 2008–2015 is covered. Moreover, only those studies are taken into account that deal with the detection of defects (as opposed to characterization or image reconstruction [59]), exploiting redundant information to increase the reliability of detection (see section 2.3). But to avoid making the scope of the survey too narrow, it is not limited to near-surface crack detection in ferromagnetic materials. Rather, if the published data fusion strategies appear to be relevant for this thesis, then other common defect types like corrosion, concrete honeycombs and impure material are included in the survey as well.

The compiled list of literature is shown in table 3.1. Each study is characterized by the employed NDT methods and the proposed fusion approach, which is further divided into fusion level (see 2.3) and fusion technique. Since it is common to include an experimental section in NDT-related publications, the table also lists the tested material, information about the investigated defects and how the detection results were evaluated. Information about defects comprises the type, location along the surface normal direction (surface / near surface / volume) and number of tested defects. To clarify the meaning of "simulated" defect types in the table, this denotes real measurements of an artificial defect-like structural discontinuity, in contrast to simulations in software.

Certain cells in the table are highlighted. Those cells indicate aspects of studies that directly relate to the work in this thesis. In particular, *NDT methods* are highlighted if inter-modal fusion was carried out, that is, the employed NDT techniques rely on different physical effects. Ferromagnetic *Material*s are also highlighted, because they are amenable to all NDT techniques used in this thesis. Concerning the *Defect* columns, cells are highlighted if cracks are investigated or if the tested imperfections are located near the surface of the material. Lastly, all studies that evaluate their detection method using a known ground truth reference and thus report their results in terms of objective statistics are highlighted in the last column. All other columns are always interesting for this work and therefore do not contain any highlighted cells.

Analyzing the studies' properties column-wise summarizes the current state of the art. Accordingly, only 5/14 studies investigate inter-modal fusion, of which only three deal with crack detection. This small number is explained by the practical challenges of multi-method NDT compared to single-method NDT, and by the additional cost. The

most frequently applied testing techniques among the list are ET (12/14) and UT (5/14). Remarkably, only one study applies automatic MFL testing, and no thermographic data fusion studies were found that meet the requirements to be included in the survey (see above). Despite its high potential [12], TT is rarely used in a multi-sensor context[1]. NDT data sets are fused at all levels, and pixel-level fusion seems to be applied only to physically related signals such as those from electromagnetic testing. This might be due to the additional effort that inter-modal fusion usually requires during signal normalization at the pixel level. Concerning the fusion strategies, nearly every study follows a different approach. A common element seems to be that probabilistic and evidential (Dempster-Shafer) theories are prevalent in high-level fusion. Moving further to *Material*, owing to their industrial relevance, metals are the subject of most of the listed studies, including steel (6/14), aluminum (6/14) and titanium (1/14). In these materials, cracks and corrosion are the most-studied defect types (8/11). It must be emphasized that none of the studies that focus on crack detection actually test real cracks. In fact, notches (induced by Electrical Discharge Machining (EDM)) commonly act as crack mimics to be able to experimentally control the defect size. The few publications that do investigate real cracks are concerned with crack prediction before mechanical loading instead of detection after loading. As the studied locations of defects are mainly determined by the type of defect, 9/14 studies inspect near-surface or at-surface defects. However, the number of imperfections vary strongly among the experiments from a single instance to 18 objects. Yet, even this most extensive study is far from providing results with low statistical uncertainty, let alone from being representative of the wide natural variability of defects. The practical constraints that prohibit investigating a sufficient number of defect cases appears to be typical for NDT studies and therefore represents a major obstacle for research. Finally, the last property that was extracted from each study is the quantitative evaluation of defect detection. A relatively small number of 6/12 studies that aim at defect detection[2] report results based on a rigorous quantitative evaluation scheme. Yet, this would be highly desirable to be able to compare different studies and to objectively assess the actual value of the proposed approaches.

By looking at the highlighted cells row-wise, it becomes clear that the only study which compares to the present thesis in all five aspects is Friedrich et al. (2009)[60]. But the authors do not include TT and concentrate on decision-level fusion only, which distinguishes this thesis from the referred study. Moreover, the shallowest considered defect depth in [60] is 150 µm and its width is 1 cm, which is still much larger than the sizes that will be investigated throughout this thesis.

This overview demonstrates that despite its high potential to improve the reliability of defect detection, multi-sensor data fusion is still a narrow field of research among the NDT community. Although cracking is recognized as a degradation process of high interest for fusion, few studies actually exploit the rich set of information that only inter-modal nondestructive inspection provides. Specifically, the combination of ET, MFL and TT, as proposed in this thesis, is unique.

[1]The recent review [11] (2015) about thermal testing refers to only a single study in the context of data fusion, which dates back to 2001. Likewise, Gros (2001) [4] includes one study that combines ET, TT and UT to inspect carbon fiber reinforced plastic panels.

[2]excluding the two studies that predict defects, but do not actually detect them, as noted in the last column

Table 3.1: List of related literature, sorted by publication date. Unspecified information is represented by *NA*. This table spans multiple pages, and all abbreviations are explained at the bottom.

Reference	NDT methods	Fusion		Material	Defect			Quantitative evaluation
		Level	Technique		Type	Location	Count	
[61, 62] 2015	Radar, UT, Impact Echo	feature	Product, Dempster-Shafer, Clustering (k-means, Fuzzy C-means, DBSCAN)	concrete	simulators of honeycombs	volume	3	ROC space, AUC
[63] 2014	UT	decision	statistical consensus test	titanium	realistic point-like contaminant inclusions	volume	5	ROC space, PFA at TPR=1
[10] 2014	Potential mapping, ET, Microwave moisture testing, Radar	feature, decision	Fuzzy clustering, Dempster-Shafer	concrete	simulators of delamination, honeycombs	volume	8	combined TPR
[64] 2013	ET, UT, Microwave	decision	maximum-likelihood weighted average	aluminium	corrosion	sub-surface	multi-ple corro-ded regions	(no actual crack detection)
[65] 2012	ET excitation, GMR sensing	pixel	algebraic (sum of abs. values of imag. parts)	steel	crack simulated by notch	sub-surface	1	-
[66] 2012	ET at 2 excitation frequencies	pixel	wavelet transform, min/max energy selection	steel	cracks simulated by EDM notches	surface	3	(image quality metrics)
[67] 2012	multi-freq. ET	decision	maximum-likelihood weighted average	aluminium	corrosion, cracks	sub-surface	13 cracks	(no actual crack detection)
[68] 2011	pulsed ET and multi-freq. ET, visual testing	pixel	PCA, IHS, wavelet decomposition, replace details / coeffs entirely (fusion rule is not adaptive)	*NA*	notches?	*NA*	4-8?	(image quality metrics)
[69] 2010	ET	pixel	Spatial Frequency, Wavelet, Bayesian, Dempster-Shafer	steel	*NA*	*NA*	4	(image quality metrics)
[60] 2009	ET+UT, MFL+ET	decision	Bayesian, Dempster-Shafer	steel, alum.	crack simulated by notch	(near-)surface	9 (steel), 9 (alu)	MSE of predicted defect position

[70] 2008	UT (standard & ToFD), multi-freq. ET	decision	Fuzzy inference	alum. alloy	welding: "volume and root defects"	both	>3	-
[71] 2008	large and small ET probe, each for multi-freq. ET	decision	weighted average of a-posteriori probabilities	steel	cracks simulated by notches	surface	*NA*	POD curve
[72] 2008	multi-freq. ET + pulsed ET	decision	Dempster-Shafer, Locally Weighted Regression	aluminium	corrosion	sub-surface	multi-ple cor-roded regions	metrics to compare against X-Ray image: RMSE, PSNR, Corr, DE, MI
[73] 2008	pulsed ET with 3 GMR sensors	decision	boolean rules	steel, alum.	slots simulating cracks, side drilled hole simulating sub-surface defect	(near-) surface	7 (steel), 7 (alu)	-

AUC	=	Area under the ROC Curve		MSE	=	Mean Squared Error
DBSCAN	=	Density-based Clustering		PCA	=	Principal Components Analysis
DE	=	Difference Entropy		PFA	=	Probability of False Alarm
EDM	=	Electrical Discharge Machining		PSNR	=	Peak Signal to Noise Ratio
ET	=	Eddy current Testing		POD	=	Probability of Detection analysis
multi-freq. ET	=	ET with multiple excitation frequencies		RMSE	=	Root Mean Squared Error
GMR	=	Giant Magnetoresistance sensor		ROC	=	Receiver Operating Characteristics
IHS	=	Intensity Hue Saturation transformation		ToFD	=	Time of Flight Diffraction
MFL	=	Magnetic Flux Leakage testing		TPR	=	True Postive Rate
MI	=	Mutual Information		UT	=	Ultrasonic Testing

Chapter 4

Single-sensor Defect Detection

Before the multi-sensor case will be discussed, some general methods for single-method defect detection are introduced. Single-sensor defect detection is relevant at several steps throughout this thesis: as a single-sensor benchmark against fusion, as the prerequisite for decision-level fusion, for scale normalization during signal-level fusion, and for the final detection after fusion at the signal level.

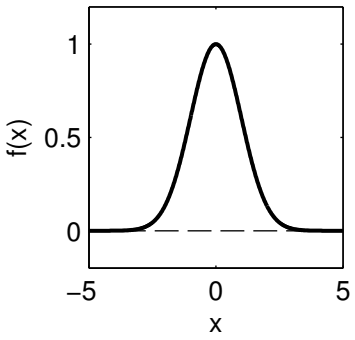
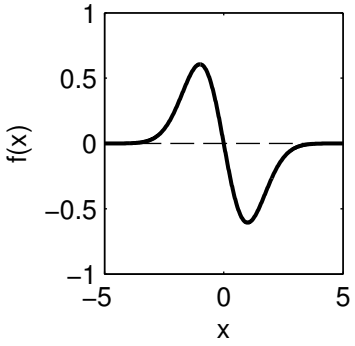
The detection of defect indications depends on the type of signal. As demonstrated in the overview about surface inspection in section 2.2, the NDT signals considered in this work can be categorized into two classes: *intensity signals* and *differential signals*. In line scans of surface inspection, these two categories produce patterns like those shown in table 4.1, where the Gaussian-shaped curve represents an intensity signal and the bi-modal curve models a differential signal. Differential signals are commonly encountered in microcrack detection, because such sharp structural discontinuities are characterized by strong local contrast of measurable material properties, which are effectively captured by differential sensors like magnetic gradiometers or differential eddy current probes. A further source of differential signals is the post-processing of thermal intensity signals. As described in [30], spatial derivatives are formed to clearly indicate microcracks, thus producing differential signals. However, the proposed routine involves additional processing so that the final thermal image can be considered an intensity signal again. Table 4.1 provides a summary. Although differential signals are beneficial for sensing, intensity signals are more favorable for automatic defect detection because the position of an indication is directly identified by regions of high intensity. Therefore, differential signals should be converted to intensity signals before detection. This process of transforming differential to intensity signals will be called *shape normalization* throughout this work, and will be explained next.

4.1 Shape normalization

To invert the differentiation process, the direct approach is to numerically compute the cumulative integral of each line scan: $s_{\text{int}}(t) = \int_{t_0}^{t} s_{\text{diff}}(\tau)\,\mathrm{d}\tau$, where $s_{\text{diff}}(\tau)$ is the differential signal and $s_{\text{int}}(t)$ denotes the estimated intensity signal. Note that any constant term that had been removed by differentiation cannot be recovered by integration, and therefore the original value of the measured physical quantity cannot be reconstructed exactly. Nevertheless, this is not an issue for detection where signal intensity relative to the background noise is more important than absolute intensity[1].

[1]Moreover, the intensity differences among different sensors will be normalized later in the processing chain.

Table 4.1: Intensity and differential signals in nondestructive surface inspection. The figures are schematics and are therefore represented in arbitrary units.

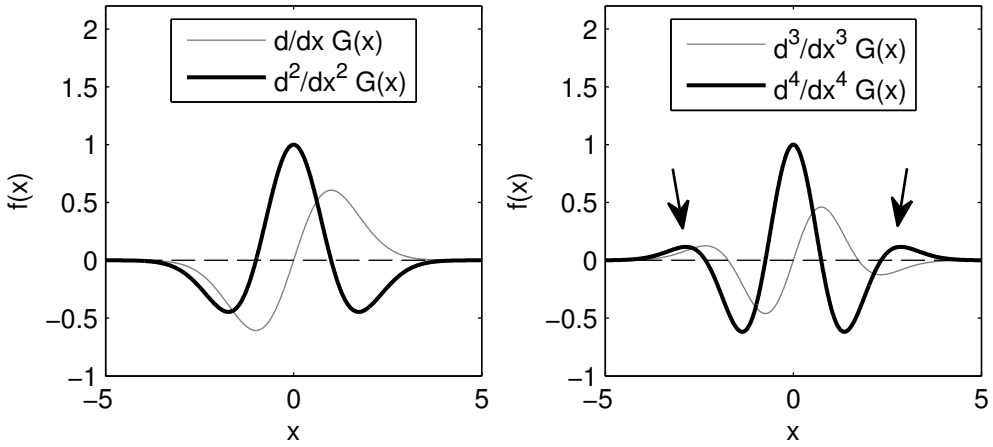
	intensity signals	differential signals
example		
ET	with absolute probe	with differential probe
MFL	with absolute GMR sensor	with gradiometer GMR sensor
TT	raw data	preprocessed data [30]

After numerical integration, the obtained signal must be highpass-filtered to remove large-scale intensity variations that are irrelevant for the detection of small-scale defects.

An alternative to integration for shape normalization is in fact the derivation of the differential signal, thus estimating the second derivative of the original intensity signal. By derivation, the steep flank at the center of a differential peak causes high intensities in the result. However, one important characteristic of NDT signals must be taken into account during derivation: Each NDT technique produces indications at a specific spatial scale, that is the number of measurement samples that the signal peak covers. Depending on the test method's physical resolution, on the spatial sampling density and on the defect size, an indication may be only a few samples large or may stretch across tens of samples. Since the physical resolution and the sampling density are known or controlled, and because the focus of this work is on microcracks, usually a narrow range of suitable scales can be determined. The requirement to operate at a specific scale of interest leads to multi-scale signal analysis, for example by the Continuous Wavelet Transform (CWT)[74]. Since the first derivative of the Gaussian function resembles differential signals quite well, it is thus used here as the mother wavelet function[2] $\psi(x) = -Cx\exp(-x^2)$, with the constant $C = 2\sqrt[4]{\frac{2}{\pi}}$. This choice is also mathematically justified: the computed wavelet coefficients equal the result of spatial derivation after Gaussian smoothing for scale selectivity [75]. The first step is to compute the one-dimensional CWT for each line scan, i.e. image row. The shape-normalized image is then constructed by replacing the measured signal values $I(x, y)$ with their wavelet coefficients $W(x, y) = \frac{1}{\sqrt{b}} \sum_i I(x_i, y)\psi\left(\frac{x_i - x}{b}\right)$ at a pre-determined scale b. The choice of scale in the continuous wavelet analysis is usually not so critical, because the coefficients vary smoothly across scales. The most suitable scale can be determined from prominent indications, for instance by inspecting a prototype defect. Apart from shape normalization, this wavelet-based approach acts also as a band-pass filter and thus conveniently eliminates unwanted background signals such as low-frequency drifts and high-frequency measurement noise. Note that a coefficient's sign indicates the polarity of bi-modal peaks in the data: Positive coefficients reflect the presence of a

[2]This function is implemented in MATLAB as the *gaus1* wavelet.

Figure 4.1: Effect of oscillatory behavior of differential signals (modeled by the first derivative of Gaussian $G(x) = \exp(-0.5x^2)$, gray) on the derivative of the differential signal (black) as computed during defect detection.
Left: No oscillatory behavior in the differential signal. Consequently, the derivative (black) only contains undershoots.
Right: Transient oscillatory behavior in the differential signal leads to overshoots in the derivative signal (arrows), which might produce spurious indications.
Note that the displayed functions are exact derivatives only up to constant scaling factors, which were introduced for demonstration purposes.



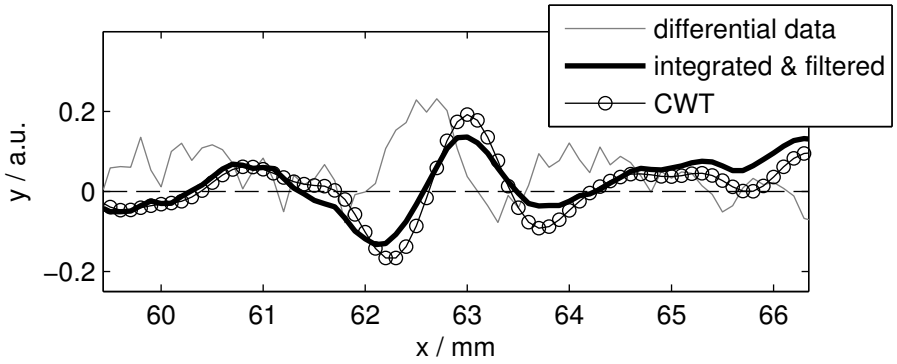
peak whose polarity is equal to the polarity of the mother wavelet. Conversely, negative coefficients occur at signal patterns of the opposite polarity. Because the polarity of the bi-modal signal peaks depends on the measurement setup and is therefore known (see section 2.2), only positive (or only negative) coefficients need to be considered. If the negative coefficients are relevant, then it is suitable to flip all signs so that defects are indicated by positive peaks.

As discussed, the CWT method using the first-order Gaussian wavelet effectively computes numerical derivatives, thus converting the differential data to a second-order derivative signal. Therefore, the computed uni-modal peaks are accompanied by two small undershoots. These artifacts usually do not impair flaw detection, because negative coefficients are irrelevant. However, differential sensor responses often have oscillatory character, that is the signal resembles higher-order derivatives (e.g. 3rd instead of 1st) of a Gaussian peak, which additionally generates positive overshoots in the normalized signal. These overshoots may introduce additional false indications, as is demonstrated in figure 4.1. Nevertheless, their fixed spatial relationship to the true indication is a strong clue that facilitates correct interpretation.

See figure 4.2 for a comparison between the two proposed approaches for shape normalization of line scans, that is numerical integration and (smoothed) differentiation, with a real measurement. The inspected surface is known to have a structural discontinuity at x=63 mm, which is indicated in the measured differential signal (gray) as a bi-modal peak. After shape normalization, both approaches produce relatively high intensity directly at the known position. In fact, the two signals are quite similar in their course. The theoretical advantage of integration over derivation to avoid overshoots is suppressed in the presence of strong structural noise, as demonstrated here.

It is also possible to extend the described techniques to two-dimensional signals, that is images. If line scans are close enough so that their signals are correlated, image

Figure 4.2: Processing of a differential ET signal (gray) to convert it to an intensity signal. Comparison between integration (thick black line) and differentiation (markers). The processed data indicate a true discontinuity around x=63 mm. The signals were scaled for comparability.



processing methods apply. For example, for shape normalization by derivation, the Sobel operator is well-suited, as used in [76]. Although this transformation is indeed sensitive to the steep signal flank at defect locations, it includes only few neighboring pixels and thus is sensitive to noise. To target the specific spatial scale of the indications, the inspection image must be lowpass-filtered by a Gaussian filter before applying the Sobel operator. Equivalently, the Sobel operator can be convolved with a Gaussian low-pass filter[3] to achieve a scale-selective Sobel filter. Note that smoothed Sobel-filtering is conceptually similar to the two-dimensional generalization of the CWT, which was proposed for the one-dimensional case. The signs of the filtered signal must then be handled in the same way as in the one-dimensional case, that is only positive or only negative intensities are relevant after filtering and constitute the shape-normalized image. If a second surface scan with orthogonal defect sensitivity compared to the first scan is available, or if the used sensor is sufficiently sensitive to arbitrarily oriented defects, two orthogonally oriented Sobel filters can be combined to form a gradient vector. Its magnitude provides another possibility to achieve a shape-normalized image. Figure 4.3 illustrates an example of a directionally complementary pair of scale-selective Sobel filters.

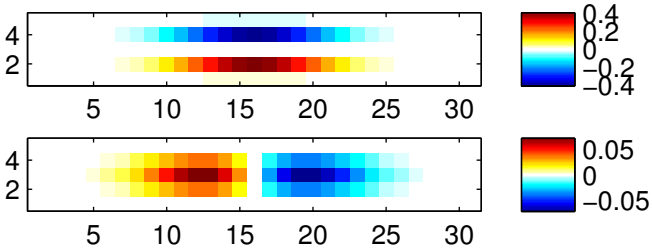
Shape normalization, as was described in this section, is the basis for defect detection, which is covered next.

4.2 Defect detection for intensity signals

After shape normalization, indications are represented by high-intensity signal areas. Typically, they appear as high-intensity ridges in a two-dimensional surface scan, such as shown in figure 4.4. These ridges are oriented along the crack path on the surface and have peak-like cross sections along other directions, e.g. perpendicular to the crack. The width of these cross sections depends on the defect characteristics and on the physical resolution of the sensor. For instance, eddy current measurements show broader ridges than results from MFL-GMR. Single-sensor crack detection amounts to the reconstruction of crack paths along the surface (black lines in figure 4.4) from the given measurements (contours in the figure). This reconstruction is not trivial for two main reasons. First, false indications cannot be reliably distinguished from actual

[3] by associativity of the convolution operation

Figure 4.3: Scale-selective Sobel filter in two directions, for different scales in the horizontal (larger scale) and vertical (smaller scale) direction.
Top: Filter for horizontal defects / differential peaks in the vertical direction.
Bottom: Filter for vertical defects / differential peaks in the horizontal direction.
The filters were created by convolving the standard Sobel filter with scaled Gaussian functions. The axes denote sample indices, that is pixels. Filter intensities are in arbitrary units.



defects using single-method inspection, as detailed in the introduction. But apart from this inherent weakness, already the accurate identification and localization of indications is an ill-posed inverse problem in itself. For instance, if the distance between two defects is smaller than the physical resolution of the sensor, they will appear as a single indication in the measurement and thus reconstruction is ambiguous.

This second problem is addressed by image restoration. In this method, the forward mapping from crack to signal is explicitly modeled to solve the inverse problem. By assuming that the forward process is linear and translation invariant, it can be modeled as a convolution operation. The convolution kernel is either known, or estimated during the inversion (*blind deconvolution*). Deconvolution is beneficial for single-sensor detection and may also serve as a signal normalization step prior to multi-sensor data fusion [77]. However, image restoration often involves many assumptions and is therefore a powerful yet complex tool. Consequently, the measurements appearing in this thesis were not treated with restoration algorithms due to lack of knowledge about the forward model (e.g. convolution kernel), and because the problem of separating closely spaced defects is not a primary objective here. Nevertheless, image restoration of NDT data should be considered before applying one of the following detection techniques.

Given a signal in which indications are assumed to be correctly localized, the simplest operation for single-sensor defect detection is to apply a (possibly adaptive) threshold to the measurements. Thresholding is only based on the assumption that defects are indicated by high-intensity regions and therefore does not take into account the ridge structure.

Whereas thresholding is a localized pointwise operation[4], different detection strategies can be devised by acknowledging that NDT measurements are often surface line scans. These line scans define the preferred direction of analysis, since signal sampling is usually finest along the line scan direction. When such a line scan crosses a ridge-like indication, the one-dimensional signal forms a local maximum. Therefore, the detection of local maxima in line scan signals is another way to identify locations of interest. Note that because many irrelevant local maxima are present in the signal due to background noise, it makes sense to specify a minimal signal intensity, i.e. a threshold, in addition to the peak criterion for detection. Importantly though, this threshold must be small enough to retain indications of micro-defects, because fusion is not able to recover

[4]although adaptive thresholding indeed takes neighboring measurements into account

previously removed indications[5].

Analogous to shape normalization, image processing techniques become applicable if the inter-line sampling distance approaches the intra-line sampling distance. In this case, thresholding results in a binary mask, which can be further processed by morphological filtering, e.g. thinning, to impose ridge-like patterns post-hoc. However, this procedure critically depends on the threshold value. Moreover, thinning assumes that the cross-sections of ridges form symmetric peaks[6]. If this assumption is not met, an indication will be poorly localized in the final output. Because one-dimensional peak detection in line scans does not require such post-processing methods, it seems to be superior in this regard. However, compared to thresholding, noise affects local maximum detection more strongly.

But image processing techniques are not limited to post-processing after per-pixel detection. An alternative procedure is to reverse the operations of thresholding and ridge localization. To this end, detection is based on the assumption that the ridge maxima suffice to localize indications. Throughout this work, this method will be called *local ridge detection*, in analogy to local maximum detection in one-dimensional intensity signals. Thresholding would then be applied after the well-localized ridge maxima had been extracted. In two-dimensional surface scan signals, local ridge detection is more complex than in one-dimensional signals, because of the additional directionality of signal features. Specifically, ridges show a clear local maximum when crossing a crack, but are comparably flat along the defect. Therefore, local ridge detection encompasses two steps for each position in the image: 1) find a suitable direction, and 2) find a local intensity maximum along this direction.

The first step is realized in this work by considering the local Hessian matrix $H_f(x, y)$ for each pixel in the image f, as proposed in [78]. In contrast to edge detection, first derivatives are not useful for ridge detection, because they become unstable at ridge maxima. Note that smoothing estimators of local second derivatives should be used if the data are affected by high-frequency measurement noise. This can be realized by filtering the image with the second derivatives of the two-dimensional Gaussian function $G(x, y)$, where σ_x and σ_y control the degree of smoothing in the respective direction. This filtering operation chooses a specific scale in the scale space representation [79] of the image. See equations 4.1–4.7.

[5] under the assumptions made in this work; see section 2.3

[6] Thinning reduces a broad line in a binary image to a single pixel wide line by successively removing pixels from the borders, so that only the central part remains.

$$G_\sigma(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-0.5\frac{z^2}{\sigma^2}) \tag{4.1}$$

$$G(x,y) = G_{\sigma_x}(x)\, G_{\sigma_y}(y) \tag{4.2}$$

$$\frac{\partial G(x,y)}{\partial x^2} = \frac{(x^2-\sigma_x^2)}{\sigma_x^4}\, G_{\sigma_x}(x)\, G_{\sigma_y}(y) \tag{4.3}$$

$$\frac{\partial G(x,y)}{\partial y^2} = \frac{(y^2-\sigma_y^2)}{\sigma_y^4}\, G_{\sigma_x}(x)\, G_{\sigma_y}(y) \tag{4.4}$$

$$\frac{\partial G(x,y)}{\partial xy} = \frac{xy}{\sigma_x^2\sigma_y^2}\, G_{\sigma_x}(x)\, G_{\sigma_y}(y) \tag{4.5}$$

$$\partial f(x,y) \approx f \overset{conv.}{*} \partial G(x,y) \tag{4.6}$$

$$H_f(x,y) = \begin{bmatrix} \frac{\partial f}{dx^2} & \frac{\partial f}{dxy} \\ \frac{\partial f}{dxy} & \frac{\partial f}{dy^2} \end{bmatrix} \tag{4.7}$$

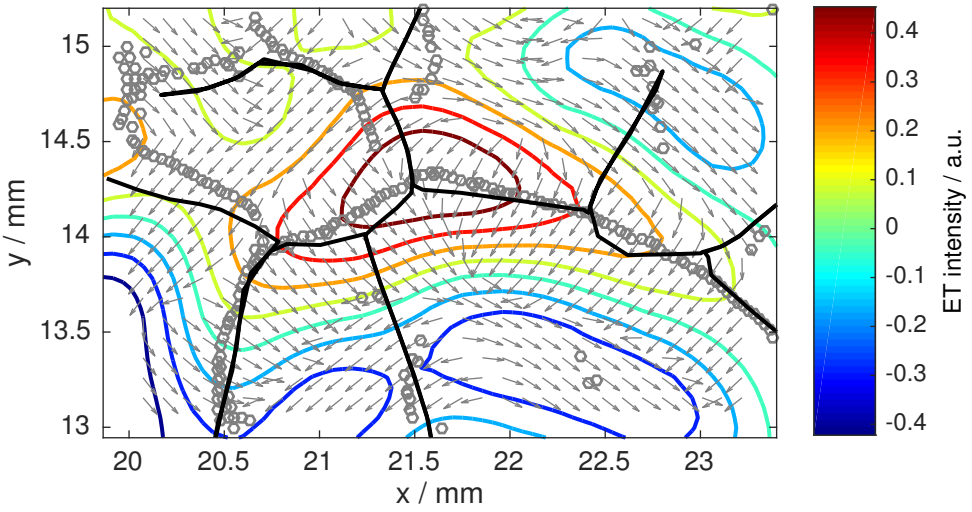
Eigenanalysis of each local hessian $H(x,y)$ yields the desired directions [78]. Specifically, consider a pixel that lies on or near a ridge maximum. The local curvature of the image is described by two main directions. Along the cross-ridge direction, a local maximum is formed which results in strong curvature, as indicated by a negative eigenvalue with high magnitude. In the orthogonal direction, that is along the ridge, the curvature is flatter. The corresponding eigenvalue is near zero and has arbitrary sign. Therefore, the eigenvector corresponding to the minimum eigenvalue defines the direction along which a local maximum is expected that locates the ridge peak.

The second step is straightforward local maximum detection along the direction of the found eigenvector. Because eigenanalysis is carried out for each pixel of the inspection image, line search can be limited to a short range of a single pixel to both sides along the eigenvector. If a local maximum exists, it can be estimated with sub-pixel accuracy by assuming a quadratic peak model. The associated formula of peak position is provided in the appendix (A.5). To reduce the number of ambiguous low-intensity indications during single-sensor crack detection, the image intensities corresponding to the identified positions could additionally be subjected to thresholding.

In the example, the resulting set of ridge locations (without thresholding) are represented by gray circle markers in figure 4.4. Although the physical limitations of ET concerning spatial resolution cannot be overcome completely by ridge detection, that is some cracks were not identified, the found locations are in close proximity to the reference.

Because the concept of a crack itself is not precisely defined in the literature (see section 2.1), making few assumptions during detection is essential to acknowledge the wide range of natural crack shapes and to make the method widely applicable. The benefit of peak / ridge detection over more complex algorithms is that they indeed make minimal assumptions about the shape of the indications, and consequently about the underlying cracks. In addition, peak / ridge detection are sensitive to weak indications, as produced by small defects, and are therefore well-suited for data fusion applications.

Figure 4.4: Ridges from ET (colored contours) indicating natural surface cracks. Deconvolution was not performed. Gray arrows represent eigenvectors of local Hessians. Detected ridge locations are shown as grey circles. Actual crack paths (black lines) were extracted from laser-induced TT for reference.



Chapter 5

Multi-Sensor Defect Detection at the Signal Level

In this chapter, several techniques for *low-level* fusion will be presented, which are organized into two main aspects. First, nondestructive surface scan measurements are treated as images, and fused in a per-pixel manner. This direct approach is also examined in combination with a multi-scale method, thus wrapping a signal transform around the fusion pipeline.[1]

Since, more generally, fusion is possible in any suitable[2] signal transform domain, one particular group of transforms is investigated in the second sub-section. This chosen transform family particularly addresses the fact that cracks produce *elongated* image features, and therefore includes prior knowledge to aid the detection. To distinguish methods that make use of defect shape information from techniques that are oblivious in this regard, the terms *directional* and *undirectional* fusion are adopted in this chapter. Progressing along increased complexity, undirectional fusion will be covered first, before the directional case.

5.1 Undirectional fusion at the signal level

In this section, methods from image fusion are applied to NDT data sets, and fusion rules are applied in a per-pixel manner. But before fusion, radiometric normalization is required to cope with the disparateness of the NDT signals, as discussed in section 2.3. Consequently, three strategies are proposed here to make the inherently unrelated inspection signals compatible for signal integration. These strategies are then applied to real measurements of a test specimen that contains artificial grooves of varying depths and widths. Several fusion techniques are carried out for each of the proposed normalization methods. A detailed quantitative evaluation procedure is given in the form of a ranking strategy based on the reduction of false alarm rates. An overview of all these data processing steps is shown in figure 5.1.

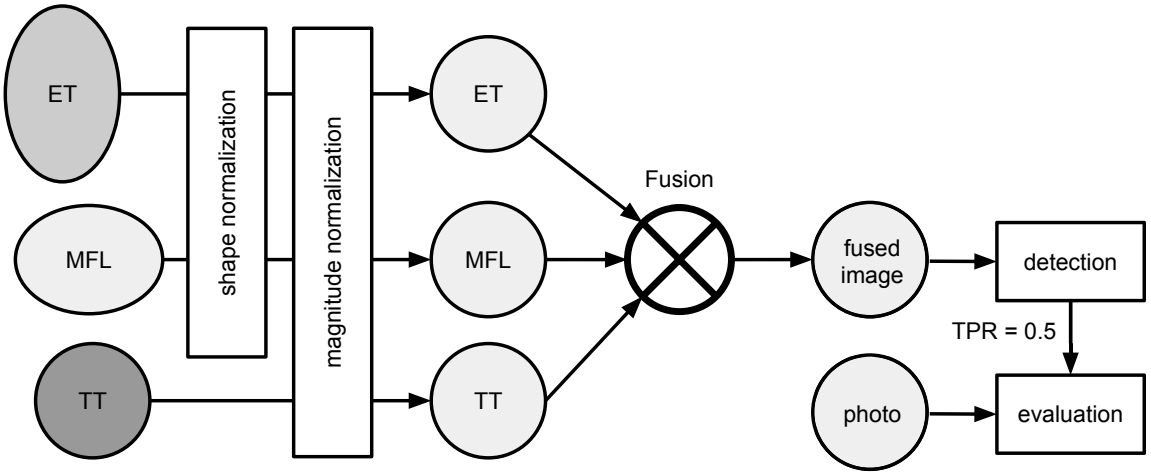
5.1.1 Radiometric normalization

Radiometric normalization includes converting the indications' signal shapes, as well as adjusting their intensities. While methods for shape normalization have already been

[1] Section 5.1 is based on a journal article ([80]) by René Heideklang and Parisa Shokouhi.

[2] in a sense that will be defined later

Figure 5.1: Overview of the fusion process.



covered in section 4.1, intensity normalization is required specifically for fusion at the signal level.

This section assumes that all individual signals have been shape-normalized if necessary, that is, positions of material discontinuities are now directly indicated by higher signal intensity compared to intact regions[3]. The goal of intensity normalization, also denoted magnitude normalization, is then to convert each individual signals' intensity range, which represents the variations of some arbitrarily-scaled physical quantity, to a new number which instead expresses the notion of defect detectability, e.g. SNR or a probability of defect detection. Since all signal sources inherently share this notion, it serves as a common representational format for data fusion.

In this section, two methods are proposed to quantify defect detectability. One method will be a linear function of the original signal intensities, whereas the second method will nonlinearly map the input range to the interval of probabilites $[0 \ldots 1]$. Both strategies assume that the inspections represent strong indications (regardless if defect or not) by high signal intensity, and weaker indications, e.g. structural noise or small defects, by lower signal intensity. Furthermore, both strategies are defined in terms of the structural noise distribution of each signal. This noise distribution can be estimated by sampling each signal at a representative user-defined region of interest within the inspection area where no obvious indications or outliers are present, to avoid skewing the distribution[4]. Spatial low-frequency signal drifts should be eliminated beforehand, if not already filtered out by the shape-normalization procedure described above. In the context of NDT, indications are typically rare (depending on the application), so that a large set of signal samples, denoted $\mathcal{N}$, is available to accurately represent the noise distribution. One important assumption made here concerns the homogeneity of the noise distribution across the inspected area. If this requirement is not given, adaptive techniques [81] are more suited, or alternatively, different regions can be treated separately.

After having obtained a representative sample of noise intensities for each sensor's image, the first proposed method for magnitude normalization is to apply the statistical

[3]That is apart from false alarms.

[4]If this is a concern, then methods of robust statistics are recommended when working with the distribution; see further below.

z-transform to each signal as follows: $I_z(x,y) = \dfrac{I(x,y) - \mathrm{Avg}(\mathcal{N})}{\mathrm{Std}(\mathcal{N})}$. From the noise sample, the noise mean $\mathrm{Avg}(\mathcal{N})$ and standard deviation $\mathrm{Std}(\mathcal{N})$ are obtained. These values are used to standardize the whole image I, including defect indications. Signal magnitudes from different sensors are thus directly comparable as different degrees of significance of an indication with respect to the background noise. Therefore, sensors that provide favorable SNR will produce higher values than lower-quality sensors for the same indication, which is desirable for fusion. It is also possible to consider alternative measures of central tendency and dispersion, for instance robust estimates like *median* and *median absolute deviation* in case of outlier-corrupted data. Moreover, if signal shape normalization was carried out using the CWT, only one half of the distribution is of interest (either positive or negative coefficients; see section 4.1). In this case, one-sided estimates of dispersion can be computed, for instance by $\mathrm{Std}^+(\mathcal{N}) = \mathrm{E}(n^+ - \mathrm{median}(\mathcal{N}))$, where $n^+ = \{\, n \in \mathcal{N} \mid n > \mathrm{median}(\mathcal{N}) \,\}$. This one-sided procedure is especially recommended for asymmetric distributions.

As a second option, in this study each of the individual signals is converted to a probability, in analogy to statistical hypothesis testing. Accordingly, the null hypothesis states that, for each image pixel $I(x,y)$, the measured value is generated by structural noise. The alternative hypothesis states that there is a defect indication. Based on the noise sample[5] $\mathcal{N}$, the one-sided probability of observing an image intensity not larger than $I(x,y)$ will be calculated. To this end, $P_{\mathrm{noise}}(\mathcal{N} \leq I(x,y))$ is estimated for each image pixel $I(x,y)$. For example, this probability approaches 1 for signal intensities that are untypically high with regard to the noise distribution, thus indicating favorable SNR. In comparison to classical hypothesis testing, this probability equals $1-p$, where p is the *p-value*. $1-p$ can be estimated from the empirical cumulative distribution function of the noise samples $\mathcal{N}$, which is a nonlinear and monotonically increasing function of the measured values: $P_{\mathrm{noise}}(\mathcal{N} \leq I(x,y)) \approx \dfrac{|\{\, n \in \mathcal{N} \mid n \leq I(x,y) \,\}|}{|\mathcal{N}|}$, where $|\cdot|$ denotes the number of set elements. Although it is known that, in general, from the null distribution one cannot infer statements about the actual hypothesis of interest (presence of a defect), the computed complementary p-values will be used nonetheless as the input to the data fusion step. This is different from the traditional testing procedure, where probabilities are compared against some significance level, because at the signal level, thresholding (detection) will be performed only after fusion. Since this study does not include any probabilistic method for data fusion (see the following section), the just described transformation can simply be understood as a method for non-linear magnitude normalization which saturates extreme data values.

To sum up, the two proposed methods for magnitude normalization are obtained from a representative set of non-defect related measurements $\mathcal{N}$ by:

- $I_z(x,y) = \dfrac{I(x,y) - \mathrm{Avg}(\mathcal{N})}{\mathrm{Std}(\mathcal{N})}$
- $I_p(x,y) = P_{\mathrm{noise}}(\mathcal{N} \leq I(x,y)) \approx \dfrac{|\{\, n \in \mathcal{N} \mid n \leq I(x,y) \,\}|}{|\mathcal{N}|}$

[5]The symbol $\mathcal{N}$ may denote a set of noise intensities, or a random variable, depending on the context.

5.1.2 Signal fusion

Similarly to [82], straightforward rules are applied to fuse the (normalized) signals. In detail, aggregation of the measurement data of different sensors is done for each pixel independently by taking the *minimum*, *mean*, *maximum*, or the *product*. Concerning the *minimum* rule, the application of a threshold to a fused value for one pixel is equivalent to requiring that all individual sensor values are greater than this threshold value. Therefore, the min-rule is also often used to implement the fuzzy *AND* operator [83, 84]. Similarly, the maximum rule corresponds to a fuzzy *OR* operator, i.e. a single significant sensor value determines the fusion result. The *mean* is just an additive integration in contrast to the multiplicative *product*. In contrast to [69, 82], more involved rules such as probabilistic fusion using Bayes' rule or evidential theory were neglected in this study to avoid the rather subjective decisions about prior probabilities (Bayes) or basic probability assignments (Dempster-Shafer).

As an alternative to per-pixel fusion, one widely-used method for image fusion is based on the two-dimensional discrete wavelet transform [66, 85]. In contrast to treating each pixel separately, the wavelet transform is a multi-scale approach which decomposes images into well-localized details (e.g. single pixels) and larger-scale components at dyadic scale levels. The general workflow for fusion is to compute the wavelet-transform for each individual image, then fuse the obtained decompositions and finally compute the inverse transform to generate a fused representation in the original signal space. The advantage of this approach lies in the adaptability of the fusion rule to different spatial scales and/or positions. Therefore, wavelet-based fusion is also considered in this study.

Because it is known that the standard discrete wavelet transform may produce artifacts[6] in the reconstructed image [85], this study follows the general recommendation to employ a *shift-invariant* wavelet transform, such as the Stationary Wavelet Transform (SWT) [86]. As the mother wavelet, *db2* was chosen for its narrow support, and a maximum decomposition level of 7 was defined. Only shape-normalized and z-transformed signals were subjected to this fusion technique, so as not to lose any edge information by the nonlinear saturation of the probabilistic normalization. Several ways to integrate the approximation coefficients as well as the detail coefficients were implemented. In line with "the majority of image fusion approaches"[87, sec. 2.3], this study follows a *no-grouping scheme*, which means that although any fixed image region is represented by multiple coefficients at each scale and across different scales, these coefficients are nevertheless allowed to be fused independently. Specifically, for the approximation coefficients the following fusion rules are selected: Setting to zero, minimum-rule, mean-rule and median-rule. The detail coefficients were processed by the median-rule as well as the maximum-rule. Additionally, the combination of mean approximation and minimum details coefficients was implemented.

Note that although the individual sensors' signals have no negative values[7], its representations in the wavelet domain are indeed signed. This questions the applicability of the discussed algebraic rules such as sum and product. For example, consider the fusion of coefficients of opposite signs with large magnitudes. The sum-rule would produce a near-zero result, and the product rule would be produce a result whose sign depends mainly on the number of coefficients to be fused, which does not give a reliable

[6] This issue is discussed in more detail in section 5.2.

[7] As explained in section 4.1, shape normalization allows to discard negative values if the peak polarity is known.

fusion rule. Previous studies on image fusion [68],[88, eq. 25] address this problem by considering the following fusion rule:

$$\text{maxAbs\_signed}(C) = c_i \qquad \text{with} \quad i = \arg\max_j \{\, |c_j| \mid c_j \in C \,\} \tag{5.1}$$

In words, the signed coefficient c_i that yields the maximum absolute value among the set C of coefficients to be fused is returned. However, whereas this rule is sensible for complementary image fusion such as multi-focus fusion, the combination of redundant information with regard to a reduction of false alarms requires stricter rules. Consequently, the following fusion methods are also considered in this work:

$$\text{minAbs\_signed}(C) = c_i \qquad \text{with} \quad i = \arg\min_j \{\, |c_j| \mid c_j \in C \,\} \tag{5.2}$$

$$\text{medAbs\_signed}(C) = c_i \qquad \text{with} \quad i = \arg\operatorname{median}_j \{\, |c_j| \mid c_j \in C \,\} \tag{5.3}$$

These three rules were implemented for the fusion of details coefficients, whereas approximation coefficients are fused by the mean rule.

Altogether, this amounts to 12 wavelet-fusion approaches investigated here, in addition to the four basic algebraic rules. The previously discussed pre-processing and fusion approaches were applied to a three-source NDT dataset as described next.

5.1.3 Application to NDT data

Test specimen The test specimen used here has already been briefly introduced in section 1.1, and will be referred to as the steel *slab* from here on. It is a steel block containing ten machined grooves, which simulate near-surface cracks. The specimen has a size of width / height / depth = 10 cm / 5 cm / 1 cm. The grooves were created by electrical discharge machining; they are all about 6 mm long but vary in depth and width, as summarized in table 5.1. The shallowest discontinuity is 10 µm deep and 81 µm wide. Note that groove nr. 6 is actually made of two closely spaced parallel slits, to evaluate the NDT techniques' physical resolutions. However, for the purpose of defect detection, this pair of discontinuities is regarded as a single object.

Table 5.1: Groove dimensions, in mm, for specimen *slab*. The groove nr. 6 is actually made of two grooves.

nr.	1	2	3	4	5	6*	7	8	9	10
width	0.15	0.25	0.19	0.14	0.12	0.095	0.092	0.082	0.081	0.192
depth	1.78	0.85	0.38	0.21	0.11	0.11	0.044	0.03	0.01	2.24

NDT dataset The multi-sensor dataset includes ET, MFL with GMR sensors, and flying laser spot TT [30] data collected on the test specimen. Additionally, a high-resolution photograph (see figure 5.2) of the surface using a digital camera serves as the reference measurement, since the surface-breaking grooves are directly visible.

The ET system consists of a differential probe (Rohmann KDS 2-2) operated at 500 kHz. A low-pass filter with cut-off frequency of 500 Hz was applied. Line scans were obtained at a spatial sampling distance of 0.1 mm for both in-line and between-line directions.

Magnetic flux leakage was induced by subjecting the specimen to magnetic saturation in orthogonal direction to the groove orientation. The resulting stray fields were

Figure 5.2: Photograph of the surface of the test specimen *slab*

measured by a differential GMR sensor which is sensitive to field differences in orthogonal direction to the inspection surface. Line scans were obtained at a spatial sampling distance of 0.016 mm with a between-line distance of 0.1 mm.

It should be noted that the ET and GMR data used for this study represent the best case scenario results. The sensitivity of electromagnetic methods highly depends on the defect orientation. The inspection procedures were optimized here, given the known groove orientation.

Thermography testing involved a laser which runs across the uncoated specimen surface in a raster (93 W power, 1.3 mm spot size, 0.1 m/s speed), while surface temperature is recorded by an infra-red camera. The resulting image sequence has a per-pixel area of about 0.159 mm × 0.159 mm. For pre-processing, the first frame was subtracted from all other frames. To convert the recorded movie to a single image, the temporal dimension was eliminated by computing the mean value per pixel.

Each measurement was carried out automatically to ensure reproducibility and accurate localization. After signal acquisition, the line-scans were composed to create sensor images of the specimen surface. Because data fusion was not a part of the experimental design at the time of the measurements, each inspection was done in its own local coordinate system, meaning at different locations on the surface. To evaluate the sensor data at the same positions, the reference photo was chosen to define a common coordinate system for image registration. Coordinate transformations from the local systems to this global system were computed by manually matching the groove tips with the corresponding locations on the photo. These correspondences determined the parameters of a projective transformation model. Subsequently, each inspection image was interpolated at common locations in the reference system at a spatial grid resolution of 0.02 mm × 0.02 mm.

Shape normalization was conducted for the differential ET and GMR signals by replacing the line scan intensities with coefficients from the one-dimensional CWT at a suitable scale, as described in section 4.1. In this process, negative coefficients were set to zero. Regarding magnitude normalization, the noise distributions in figure 5.3 demonstrate the effectiveness of the proposed strategy. Because the noise distributions are not all symmetric, one-sided estimates of dispersion were computed in the z-transformation as proposed in section 5.1.1. Consequently, all normalized distributions are centered around zero and exhibit similar (one-sided) variance.

The effect of radiometric normalization on the inspection images is shown in figure 5.4. All plots display the groove nr. 7, which is 44 µm deep, as sensed by the three NDT techniques (rows) at various stages of normalization (columns). Shape normalization is best observed in the ET data (top row), where the double-peaked indication (figure 5.4a) is converted to a single peak (figure 5.4b). Structural noise is clearly visible around the groove for all test methods. In particular, the probability normalization (last column) emphasizes the noise relative to the groove by nonlinearly saturating large (relative to the noise distribution) intensities.

Figure 5.3: Noise distribution of the set $\mathcal{N}$ after normalization. Note that the distributions are centered around zero and have similar spread. The differential NDT methods (ET,MFL) are characterized by truncated histograms.

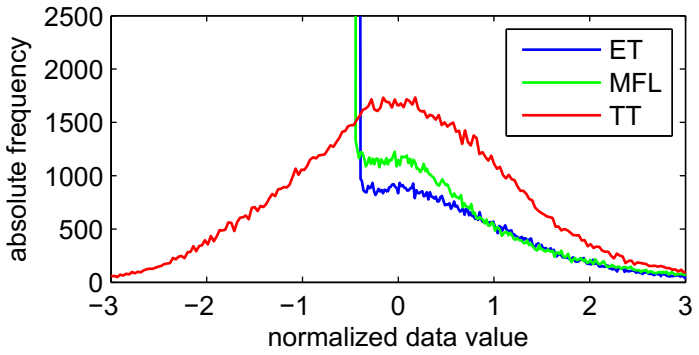
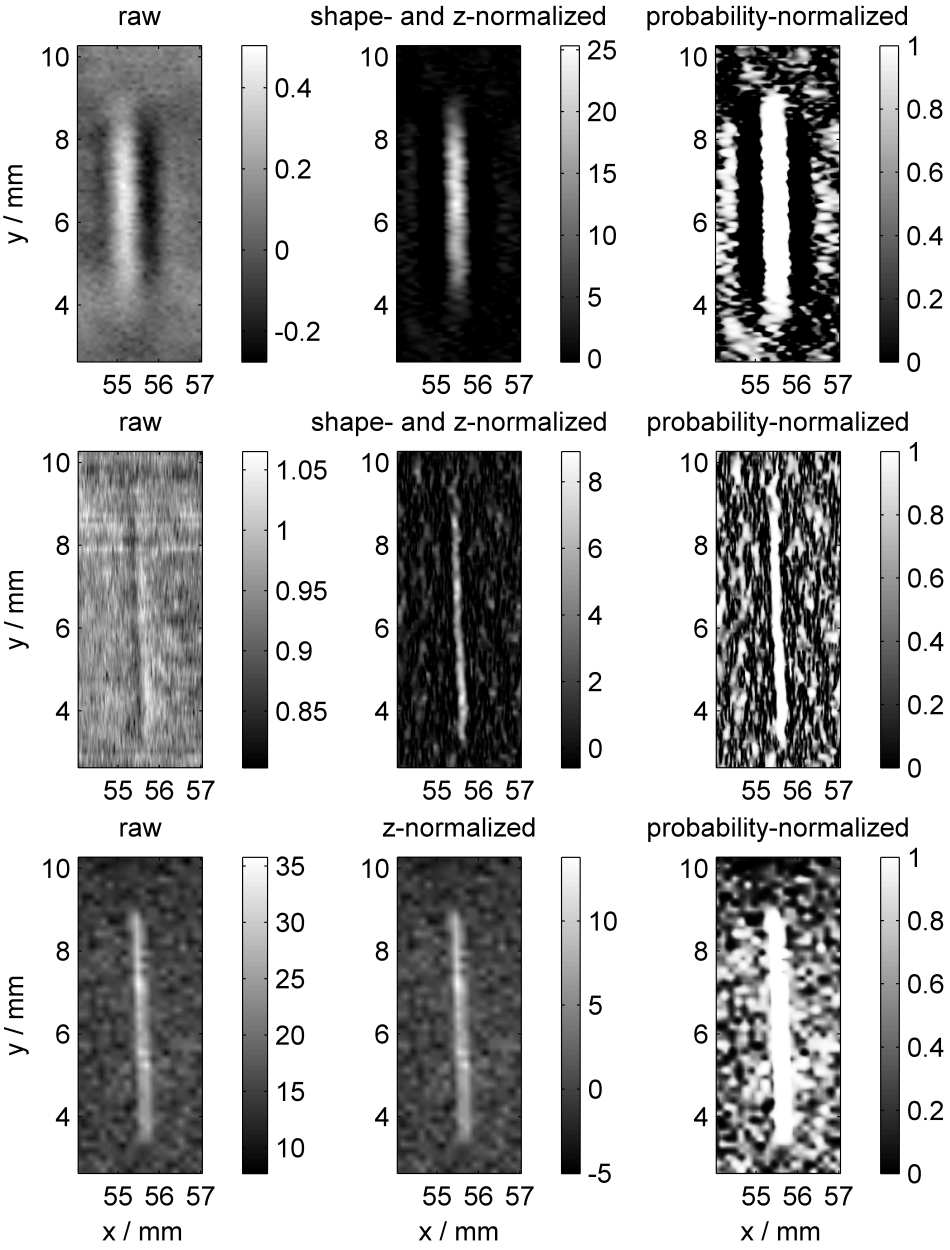


Figure 5.4: Individual signal images of a 44 µm deep groove.
Top: Eddy current testing. Middle: Magnetic flux leakage testing using GMR sensors. Bottom: Thermography testing. Intensities are in arbitrary units, and are only comparable between different sensors after z-normalization or probability-based normalization. To optimize the figure's contrast, color limits are adjusted to the range of signal intensities (arbitrary units) within each shown area.



(a) before shape normalization (does not apply to TT)

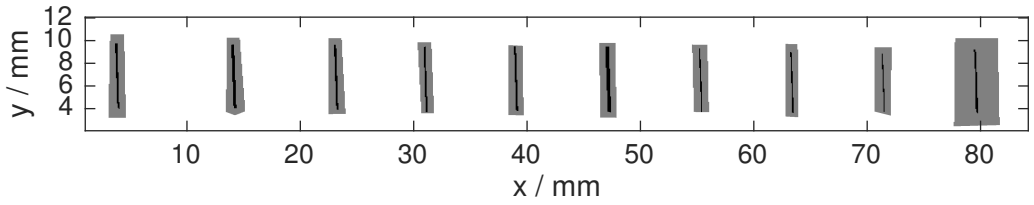
(b) magnitude normalization by z-transformation

(c) magnitude normalization by conversion to probability

5.1.4 Results and discussion

Evaluation strategy The proposed pixel-wise and multi-scale data fusion strategies were applied to the NDT inspections, and their performances were quantitatively evaluated. Based on the reference photo, the "defect" pixels were manually labeled for pixel-wise assessment of detection performance. In contrast to labeling each indication, formed by an arbitrarily-shaped segment of the image, a per-pixel strategy allows automatic judgment of the false positive rate, for which the total number of negatives must be known. However, the per-pixel evaluation puts sensors at a disadvantage if they generate broad signals. For example, the eddy current image contains high signal values not only directly at a groove, but also in its vicinity. For detection (disregarding accurate localization), this is advantageous because the indications are more prominent. Yet, based on the photo in which the grooves are quite thin, the automatic evaluation would report many false positive indications. Therefore, to compensate for this effect and to ensure a fair comparison, image points at a certain distance range around each of the discontinuities were excluded from evaluation. This is visualized in figure 5.5.

Figure 5.5: Areas around the grooves (gray) were ignored during evaluation.



Detection performance is assessed separately for each groove to capture the influence of defect size. Quantitative evaluation is provided by Receiver Operating Characteristic (ROC) curves. Denoting P = positive (detected as "flaw"), N = negative (detected as "not a flaw"), TP = true positive (correct flaw detection), FP = false positive (false alarm), these curves are generated by plotting the true positive rate or sensitivity (TPR = TP/P) versus the false positive rate (FPR = FP/N = 1–specificity) for all sensible detection thresholds. One popular index to compare the detection performances of various methods is the Area Under the ROC Curve (AUC). However, the number of *non-defect* pixels far exceeds the number of *defect* pixels here. In such settings, AUC values are observed to be dominated by ROC regions that correspond to low thresholds, i.e. AUC mainly measures the sensitivity. See for instance the red curve in figure 5.7. The area under this curve is dominated by the sensitivity (TPR) across a broad range of FPR values. Because in contrast, fusion of redundant sensor information is ideal to increase specificity by identifying false positives, a different evaluation strategy is followed here.

At some fixed level of sensitivity, the fusion approaches are expected to produce fewer false indications than traditional single-sensor inspection. This notion provides the basis for the proposed evaluation strategy (see also [89, sec. 2.5]). First, a level of sensitivity is fixed at TPR = 0.5, assuming that finding half of a defect's pixels is sufficient for successful detection[8]. For each groove, all the proposed methods are run and then ranked by their specificity. Each evaluated method comprises a particular normalization step and a fusion rule. The resulting ranking is illustrated in 5.6.

[8]If TPR is fixed to higher values, e.g. 90 %, the corresponding specificity drops to unrealistic detection conditions.

Interpretation Note that for prominent structural discontinuities, depicted in light-gray, all methods perform similarly well and therefore evaluation should concentrate on more challenging grooves. The worst rank per method, which is the sorting criterion for the horizontal-axis here, mostly coincides with the shallowest discontinuity. However, there are also exceptions to this rule, especially among the overall well-performing methods. The bottom subplot visualizes the ranked methods' properties. For instance, fusion rules that compute the minimum or the product require strict agreement among sensors and are termed *conjunctive* here, in contrast to *disjunctive* rules such as the maximum or the mean value. No clear pattern can be observed in the ranking regarding z-normalization vs. probability normalization, or concerning conjunctive vs. disjunctive rules. Also, smoothing the fused image to introduce spatial neighborhood information has no consistent effect. Looking at the lower-end positions, obviously all single-sensor detection methods are outperformed by most of the image fusion approaches. Yet, four of the employed fusion strategies do not seem advantageous for defect detection: The maximum-rule falls behind two of the single sensors (ET, TT), independently from the applied magnitude-normalization strategy. Interestingly, when fusing wavelet coefficients, it is the minimum rule that scores poorly. Nonetheless, all of the other investigated fusion techniques are able to outperform traditional inspection. The best-performing fusion approaches are two of the wavelet methods (median rule for detail coefficients, minimum- or median rule for approximation coefficients), performing exceptionally well by scoring among the top two ranks for every discontinuity. On the third place is detection by z-normalization & product-rule, a method that shows similar detection quality as the first top-ranked methods, but is conceptually much simpler. Therefore, the practical winner is declared z-normalization & product-rule.

The value added by data fusion can further be quantified by computing the reduction of false positives explicitly. The best single sensor (thermography) achieves a false positive rate of 0.0165 at TPR = 0.5 for the shallowest groove. Although this error rate seems small in relative terms, the absolute number of falsely identified defect pixels is large when considering that there are 1.8 million *non-defect* pixels in the data set. In contrast, the FPR for the z-normalized product fused detection method is only 0.0028, almost six times less than the best single sensor.

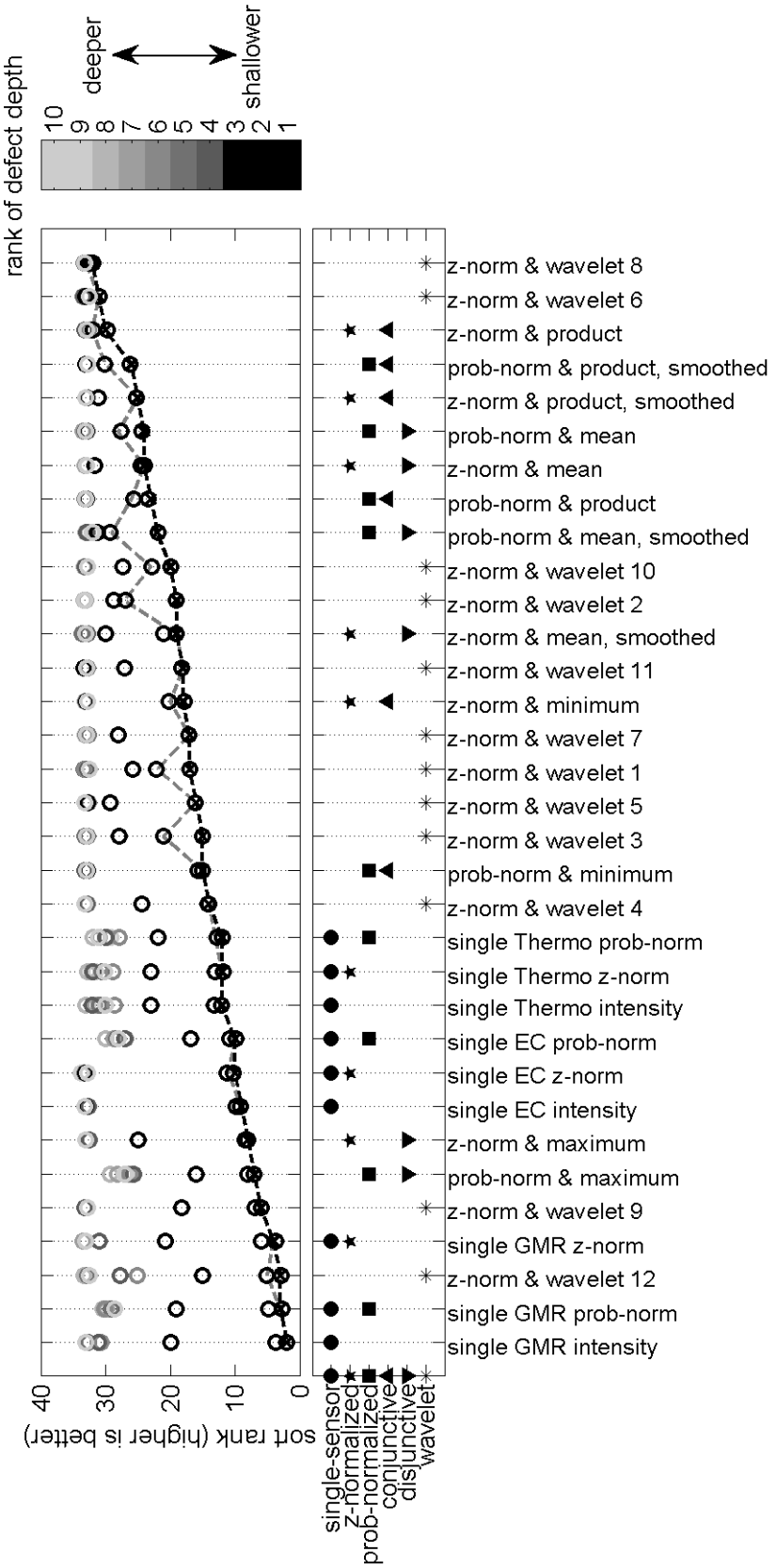
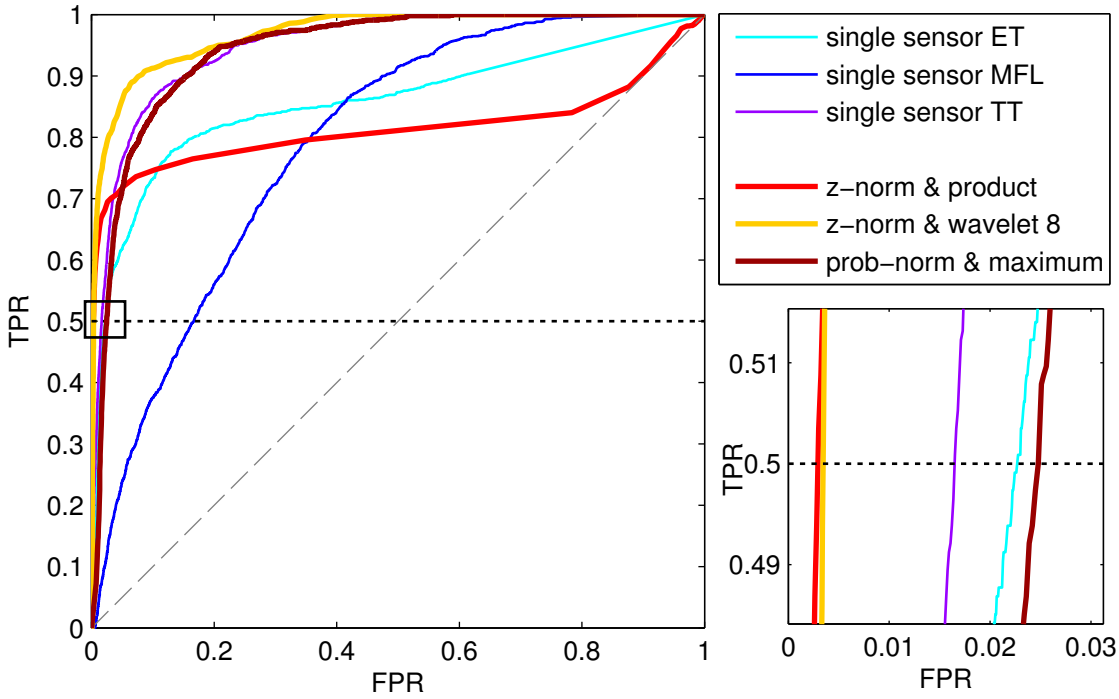


Figure 5.6: Evaluation.
Top: Methods' sort ranks per defect mimic: higher rank = better specificity at fixed TPR = 0.5. The horizontal axis is ordered by worst performance among the grooves, shown by the black dashed line. Each circle represents the sort rank of a detection method for a specific groove, whose depth is coded by gray values (brighter = deeper). The gray dashed line connects circles that represent the shallowest groove. Small jitter is added to the circles to prevent overlapping.
Bottom: A tabular summary of the methods' underlying pre-processing and fusion strategies for improved interpretability. Wavelet methods are intentionally missing the z-normalization symbol, because probability-normalized wavelet fusion was omitted in this study.

To put the evaluation results in perspective, the full ROC curves of some of the evaluated methods at the shallowest groove are presented in figure 5.7. Specifically, the three individual inspection results are shown together with the top two fusion methods (z-norm & product, and z-norm & wavelet with median of details). Additionally, a fusion method that scored at the bottom in the evaluation (prob-norm & max-rule) is also shown. This plot demonstrates the strong influence of choosing a particular evaluation measure on the ranking. For instance, if the detection techniques were ranked by AUC instead of specificity, the individual ET inspection (cyan curve in figure 5.7) would be ranked close to the red curve, which is one of the top scoring methods under the specificity criterion. However, as already explained, the specificity at a fixed sensitivity better expresses the reduction of false alarm rates at a single realistic detection threshold than the AUC, which takes into account the performance for all possible threshold values.

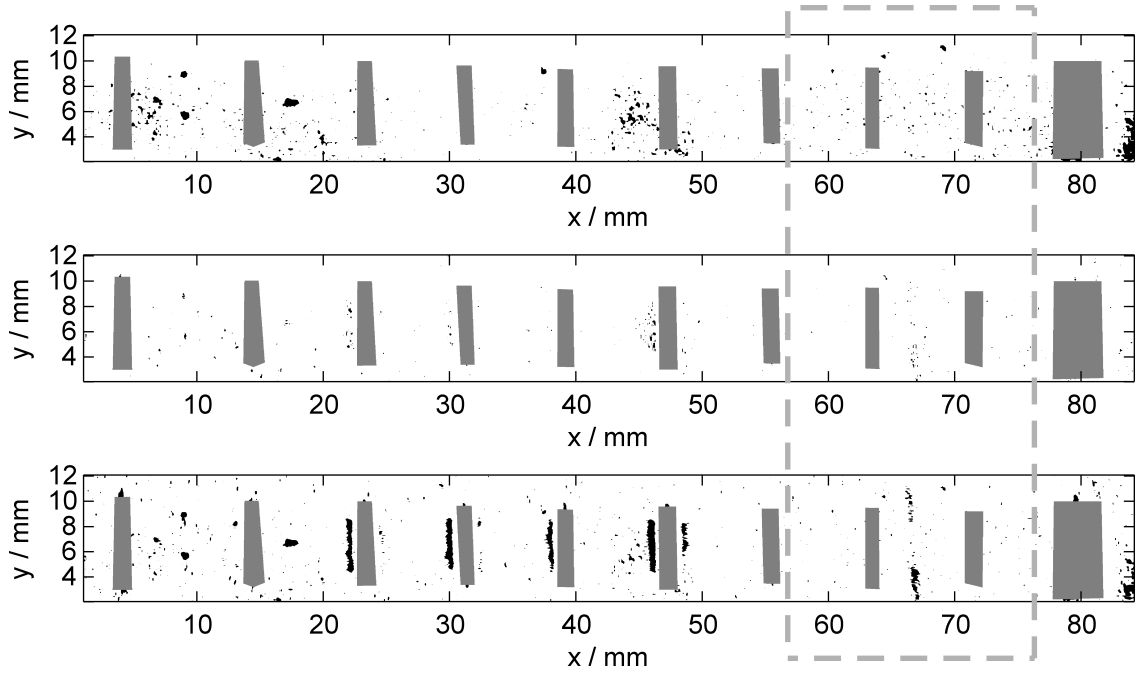
Figure 5.7: ROC curves of select defect detection methods for the shallowest groove (nr. 9, 10 µm deep). In the bottom right a zoom is shown of the region around TPR=0.5 indicated by the black box. The horizontal dashed line indicates the level of sensitivity at which the methods are ranked according to their specificity.



For a more intuitive representation of the performance of three specific detection techniques, see figures 5.8 and 5.9. The selected methods are the best individual sensor (thermography), best fusion method (individual z-normalization & product-rule), and worst fusion method (individual probability-normalization & maximum-rule) according to the above evaluation. The figure displays the positions of false alarms for a threshold that is chosen to achieve a 50 % true positive rate for the shallowest groove. Apparently, the single sensor reports many false indications, which most likely stem from surface particles that cause inhomogeneous heat flow. Additionally to some of these spurious indications, the fusion method shown in the bottom subplot also includes false positives from the other sensors in the final image, which explains its low score. For instance, the

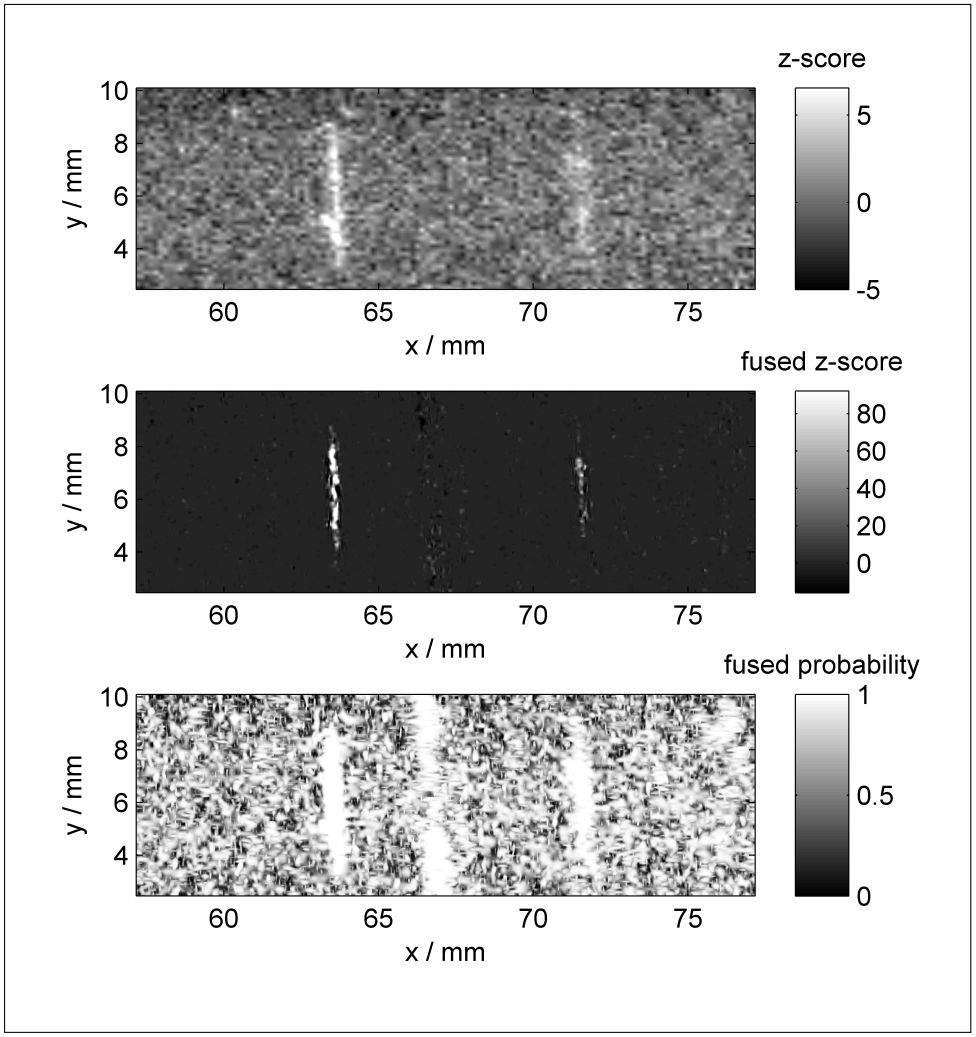
broad indications close to grooves 3–6, counted from left to right, result from the shape normalization method being sensitive to signal undershoots, which often accompany true indications in ET signals; see section 4.1. This low specificity is not surprising: Probability-normalization promotes low-magnitude indications and the maximum-rule carries out a fuzzy OR operation on these defect hypotheses. Whereas for our task of false positive reduction, this fusion rule is not suited, it may perform well for the fusion of complementary data sets provided that the individual sensors are specific enough. Finally, the middle subplot shows the result of the best fusion method as evaluated here. By z-normalization, the individual sensor having the best SNR in a given pixel dominates the other sensors. The product-rule, which acts as the fuzzy AND operator, generates a quite conservative fusion method, while maintaining the same sensitivity level.

Figure 5.8: Spatial plot of false positives. The detection threshold was set to achieve 50% TPR level for groove nr. 9. Detection methods from top to bottom: Best individual sensor (thermography), best fusion method (z-normalization & product-rule), worst fusion method (probability-normalization & max-rule). Black dots indicate false positives; white regions correspond to correct negatives, gray regions are excluded from the evaluation. Pixels marked as *defect* are not shown here for clarity.
The dashed box indicates the region that is displayed in figure 5.9



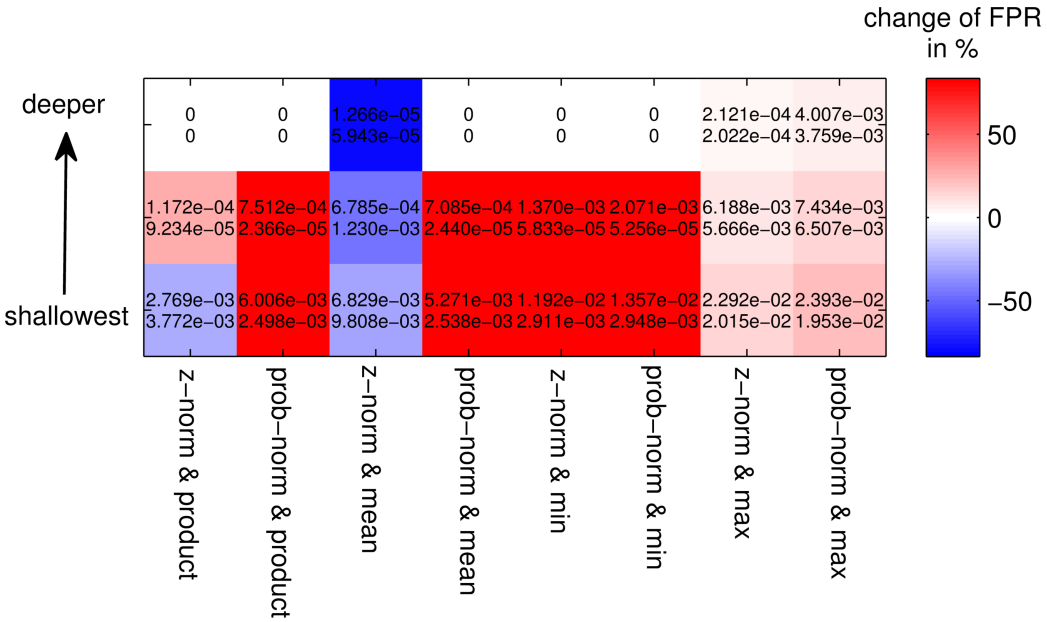
The quantitative evaluation shows that MFL is the worst-performing single sensor among the given inspections for the shallow grooves. It is interesting to assess how including or excluding this source of information from the fusion process affects defect detection. To this end, figure 5.10 compares the false positive rates at TPR=0.5 for the three shallowest grooves and eight fusion strategies when the MFL data are included or excluded from fusion. The experiment suggests that for most fusion rules considered here, including the worst individual source of information degrades the fusion performance when facing small defects. This degradation however is only mild for the maximum-rule, because apparently the MFL data introduces only few high-intensity false alarms. The mean-rule after z-normalization is the one exception where

Figure 5.9: Region around the two shallowest grooves (see figure 5.8), for three detection methods. Top: the best individual sensor (thermography). Middle: the best fusion method (z-normalization & product-rule). Bottom: the worst fusion method (probability-normalization & max-rule). Contrast was stretched to the dynamic range in the vicinity of the right-hand groove.



including the poor data actually improves performance in all three cases. Yet, this relative improvement becomes less notable when considering the absolute FPR values. The top-ranked fusion method z-norm & product still outperforms z-norm & mean in all considered cases, even despite its degraded performance at detecting the second shallowest groove when MFL is excluded. Remarkably, the winning fusion rule is not negatively affected at the shallowest groove when the poor MFL data are used. The technique seems to be fairly robust to ambiguous sources of information. This is rather surprising, because the product rule requires all sources to have high intensity in order to produce a significant indication. An explanation is that the choice of evaluating FPR at TPR=50 % allows missing half of a defect's pixels and therefore compensates for the product rule's stringency. To sum up, while most fusion rules are sensitive to poor-quality sensor data, the z-norm & product rule maintains its position as the top scoring fusion method.

Figure 5.10: Influence of the MFL inspection on the fusion result. Comparison of false positive rates for the three shallowest grooves (rows) and eight fusion methods (columns) when the MFL data are included or excluded during fusion. Red: Including the MFL data impairs detection performance (higher FPR). Blue: Including the MFL data improves detection performance (lower FPR). Whereas blue and red colors indicate the relative change in FPR, the two numbers per cell show the absolute FPR with (top) and without (bottom) the MFL data.



5.2 Directional fusion at the signal level

After having covered the fusion of intensities at zero-dimensional point locations (pixels), now an enhanced scheme is considered in which the fusion rule additionally has access to the two-dimensional *orientation* of indications. Unlike pores, cracks are elongated (see section 2.1) and therefore form oriented signal features. This property may help to distinguish false alarms from true indications. Since the framework of image fusion provides high flexibility through executing the fusion rules in a transform domain, such as the wavelet domain, incorporating orientational information only requires a suitable transform. It would be beneficial to not only apply the fusion rule independently to different locations and scales (as in the wavelet transform), but also to different orientations. In this way, one hopes to reveal conflicts and agreements among the source images that are otherwise not directly accessible in the original domain. This is realized by MGA [18], and in particular by the two-dimensional Discrete Shearlet Transform (ST).

5.2.1 The Shearlet Transform

In line with other multi-resolution analysis techniques, such as wavelets, ridgelets [90], curvelets [91] and contourlets [92], the shearlet transform [93] has been developed to sparsely represent data that include arbitrarily oriented singularities, such as edges in images or narrow cracks in NDT data. Precisely, "compactly supported shearlets can be shown to optimally sparsely approximating cartoon-like functions" [94, p. 3][95]. *Cartoon-like functions* are defined as piecewise smooth functions with discontinuities along twice continuously differentiable curves. Although cracks themselves cannot be assumed to follow differentiable paths, their indications in the inspection images are smoothed by the measurement process. Therefore, the precise technical definition of cartoon-like images can be interpreted in a more general manner, in the sense that "Shearlet transforms can provide almost optimal representation of the anisotropic features of an image" [96]. As stated in [97, p. 6], the particular advantages offered by shearlets comprise:

- A single or a finite set of generating functions
- Optimally sparse approximations of anisotropic features
- Compactly supported analyzing elements
- Fast algorithmic implementations
- A unified treatment of the continuum and digital realms
- Association with classical approximation spaces

Sparse approximation ensures that few shearlet coefficients suffice to encode most of the data, that is the information is localized in few well-interpretable features. Compact support of the analyzing shearlets ensures that salient features in the image do not spread in the transform domain, which "reduces both the introduction of distortions and the loss of contrast information during the fusion process"[98]. As a further advantage, continuous and digital signals are treated in the same way by using a shearing operation instead of a rotation to achieve directional sensitivity (hence the name *shear*let). This key idea allows for an unlimited number of directions to be analyzed while still avoiding interpolation by never querying off-pixel positions in the analyzed signal.

This relatively new technique has already found application in data fusion tasks [99] such as remote sensing [100, 101], multi-focus image fusion [102–104], multi-modal surveillance [105] and medical image fusion [106, 107]. But also among NDT researchers the shearlet transform is known. In [108], surface defects are investigated in the production of continuous casting slabs, hot-rolled steels, and aluminum sheets. Using a shearlet-based feature extraction from visual testing signals, higher classification rates were achieved compared to curvelets or contourlets. Moreover, the shearlet transform was applied to pavement image denoising [109]. The method was shown to reduce noise while retaining cracks. Optical coherence tomography is enhanced in [110] by two- and three-dimensional shearlet transforms. The shearlet technique was used to separate linear structures, such as fibers, from circular indications in the data. This is achieved by posing the separation task as a minimization problem subject to sparsity constraints in the shearlet basis (linear features) and the wavelet basis (spherical features). In particular, the 3D transform was shown to be more effective at reducing unwanted background variations than the 2D variant, given sufficient computational resources (main memory).

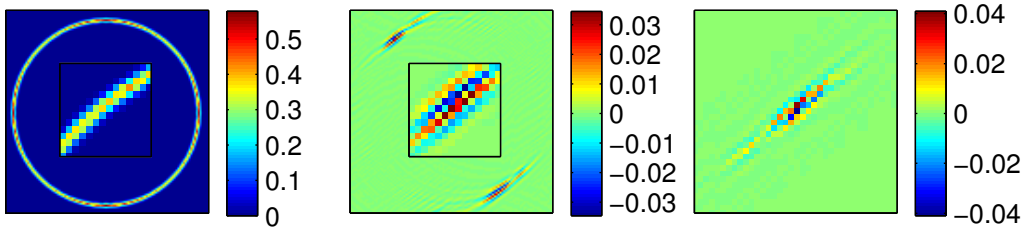
See figure 5.11 for an illustration of the shearlet transform. The left part of the figure shows a test image depicting a circle. This image was chosen because it contains features at all directions, at a fixed scale. In the second sub-figure, the shearlet coefficients are displayed corresponding to a single scale and orientation, for all translations (pixel locations). Note that most coefficients are zero, except those near parts of the circle whose orientation matches that of the chosen shearlet. The analyzing shearlet that was used to compute the coefficients is shown in the third sub-figure (note that the plot is magnified). It can be interpreted as a low-pass filter along the filtered edge direction, and a high-pass filter in the cross-edge direction. Because the shearlet elements are real-valued, the computed coefficients are also real. To illustrate the effect of filtering in the shearlet domain on an input image, consider figure A.1. This test image contains multi-scale features at all orientations. Below, the four subplots show the result of reconstruction after setting all shearlet coefficients to zero, except those at a specific direction and decomposition level (i.e. scale). The results verify that shearlet coefficients capture information at specific scales and directions. Moreover, the transform's redundancy can be observed by noticing that one given region in the test image is represented at multiple decomposition levels.

The ST for bivariate functions $f(x)$, $x \in \mathbb{R}^2$ is defined as follows, adopting the notation of [111]:

$$\begin{aligned} ST(f) &= \langle f, \psi_{a,s,t} \rangle \\ \psi_{a,s,t}(x) &= a^{\frac{-3}{4}} \, \psi \left(A_a^{-1} S_s^{-1} (x - t) \right) \end{aligned}$$

with the anisotropic scaling matrix $A_a = \begin{bmatrix} a & 0 \\ 0 & \sqrt{a} \end{bmatrix}$ and the shear matrix $S_s = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}$. In this formulation, $\langle ., . \rangle$ denotes an inner product. ψ is a two-dimensional shearlet generator, such as a classical shearlet or a nonseparable generator [94, eq. 8]. This generator plays a comparable role as the mother wavelet function in the wavelet transform. The parameters $a \in \mathbb{R}^+, s \in \mathbb{R}$ and $t \in \mathbb{R}^2$ are the scale, shear and translation parameters of the transform. For practical purposes, this transform must be made applicable to digitized functions f (images), which leads to the digital ST analogously to the discrete wavelet transform. Several approaches to ST digitalization have been proposed ([94, sec. 2.3]), and the one in [94, sec. 3] will be used in this work.

Figure 5.11: A circle (left) and its shearlet transform (middle) for one specific scale and orientation. In the center of each figure, a zoom to the top-left region of the circle is presented. Right: The analyzing shearlet in the spatial domain (zoom). Intensities are dimensionless.



5.2.2 Other directional transforms

There are several alternatives to the shearlet transform which will be briefly introduced now. The conceptually most basic method is the two-dimensional shift-invariant wavelet transform [112], sometimes also called undecimated or stationary wavelet transform (SWT). It was developed to address the apparent limitation of conventional digital wavelet transforms of being sensitive to translations in the image domain. That means, an image and a slightly shifted version of itself produce coefficients in the transform domain that are not simply shifted copies of each other. This effect originates from the subsampling and upsampling steps during the decomposition and reconstruction phases. It is detrimental for fusion applications, where shift-variant transforms might cause artifacts in the reconstructed image [85] even at small registration errors. Shift-invariance can be achieved by skipping the subsampling operations, at the cost of increased redundancy. Moreover, this modification leads to a filter design problem that is less constrained, thus facilitating additional design criteria. For example, in [1], Starck, Fadili & Murtagh developed specific filters that are well-suited for image fusion applications. Their analysis filters are very compact to spatially concentrate signal power in neighboring coefficients, and their synthesis filters are regular and all-positive to avoid reconstruction artifacts. SWT decomposition using these filters will be denoted UWT throughout this work. Especially in fusion applications, redundant representations are in fact beneficial if one is willing to accept the higher computational demands, because they "offer design freedom and robustness to corruption or loss of expansion coefficients" [113]. The application of fusion rules is one major source of such mentioned corruption of coefficients, because rules usually are not designed to maintain the kind of regularity or smoothness among spatially neighboring coefficients that was present before fusion[9]. In fact, redundant representations justify the use of such fusion rules. Shrinkage is another source of coefficient corruption, that is setting certain coefficients to zero before reconstruction in order to reduce unwanted signal components. Again, redundant representations help in reconstructing a high-quality image even if some relevant coefficients were accidentally affected by shrinkage.

Shift-invariance was recognized as a desirable property in the subsequently developed transforms, which additionally improve the directional selectivity. Because the wavelet transform is based on separable filters for both spatial dimensions, the data are decomposed into only three directional sub-bands (horizontal, vertical, diagonal). The

[9] "Any wavelet coefficient processing (thresholding, filtering, and quantization [and fusion]) upsets the delicate balance between the forward and inverse transforms, leading to artifacts in the reconstructed signal." [114]

Dual-Tree Complex Wavelet Transform (DTCoWT) [115] achieves higher directional selectivity by implementing a complex-valued filter bank, which is able to separate positive from negative frequencies. Consequently, six orientations are analyzed at $\pm 15°, \pm 45°, \pm 75°$, while being approximately shift-invariant and featuring even smaller redundancy (and hence run time complexity) than the shift-invariant wavelet transform. More recently, other analysis techniques were proposed that explicitly include directional filtering. The most notable in the context of this work, besides the shearlet transform, is the nonsubsampled contourlet transform [116]. Although in contrast to shearlets, contourlets lack theoretical benefits such as unified treatment of continuous and digital signals, they can both be understood as combinations of bandpass and directional filters that extract wedge-shaped sections from the two-dimensional Fourier domain [117][10].

Despite these recent advances, simple techniques like the wavelet transform are still widely applied, and thus the theoretical benefits given by additional directional sensitivity should be investigated in the context of multi-sensor nondestructive defect detection. To this end, the mentioned transforms will be quantitatively compared in the experimental section of this study. But regardless of which multi-scale transform is applied, scale normalization should be considered before fusion, as is explained next.

5.2.3 Scale normalization

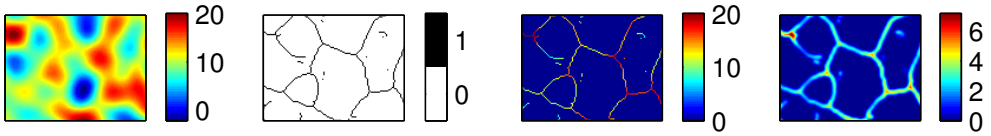
Because NDT methods employ different physical measurement processes, they vary in spatial sensitivity. For instance, the output image of eddy current inspection is much smoother than that of MFL inspection based on GMR sensors. In the language of multi-resolution analysis, signal energy is concentrated at different scales among the NDT methods. In contrast to the previous study in this thesis, which was mainly concerned with per-pixel fusion, scale-related issues move into focus here where we are dealing exclusively with multi-scale image representations. If fusion in the transform domain (e.g. fusion of shearlet coefficients) is carried out for each scale independently, then defect indications from different inspections will not be associated, which might lead to poor fusion performance. Scientific literature is very limited concerning the fusion of images of different resolutions in a redundant manner[11]. One proposed approach [119, 120] is based on numerical optimization to produce a fused image that approximates the fine-resolution image, while including information from the coarse image. This strategy operates directly on the pixel intensities and does not involve any multiscale transforms. In this work, to deal with the different physical image resolutions, a more straightforward approach is proposed that targets a multiscale fusion strategy.

In principle, there are two options to address this issue: 1) Design a fusion strategy that interrelates different scales, or 2) Normalize the data so that the energy distribution across scales is comparable. Although the first approach seems more elegant, it is unclear how to associate information from different scales and to which target scale in the fused image the information should be assigned, so that the inverse transform can be applied. Despite the second strategy's drawback of significantly altering the data, it is adopted in this work because it enables straightforward fusion and reconstruction. The importance of scale normalization will be demonstrated later in the experimental section.

[10] Original quote: "The spatial-frequency tilings of curvelet and shearlet representations are completely different theoretically, yet the implementations of the curvelet transform corresponds to essentially the same tiling as that of the shearlet or contourlet transform."

[11] In contrast, there is a vast amount of work on complementary fusion of images having different resolutions, such as pan sharpening [118]

Figure 5.12: Stages of scale normalization for a section of the ET data.



(a) original data **(b)** ridge positions (binary image) **(c)** ridge intensity **(d)** smoothed ridge intensity

Scale normalization essentially narrows or widens all indications in the image. Such indications are assumed to form ridges in the inspection image. Therefore, shape normalization (see 4.1) might be necessary prior to scale normalization. The proposed normalization process is exemplified in figure 5.12 for the case of eddy current testing. As a first step, regions in the image are identified where ridges should be scaled. This is essentially single-sensor defect detection, which is covered in section 4.2. The result is a set of locations in the inspection image that correspond to peak maxima; see figure 5.12b. Next, a new image is formed that is zero everywhere, except at the ridge positions where the original peak intensities are retained (figure 5.12c). In this image, the signal intensity is mostly concentrated at the finest scale. To produce arbitrarily scaled data, smoothing is carried out with a suitably chosen smoothing kernel size (figure 5.12d). Since all single sensors share the same smoothing kernel size, the indications from each image are projected into the same scale, to be fused according to different rules as described next.

5.2.4 Fusion rules

To fuse the shearlet coefficients from different NDT images at the same scale, orientation and position, similarly to section 5.1 several fusion rules were selected. These entail simple algebraic rules (minimum, median, maximum, sum, product, geometric / harmonic mean). Since shearlet decomposition involves bandpass filtering[12], shearlet coefficients are signed. Therefore, suitable fusion rules have to output sensible results for all combinations of input signs and magnitudes. To this end, the following rules are defined:

$$\text{minAbs\_signed}(C) = c_i \qquad \text{with} \quad i = \arg\min_j \{\, |c_j| \mid c_j \in C \,\} \tag{5.4}$$

$$\text{medAbs\_signed}(C) = c_i \qquad \text{with} \quad i = \arg\operatorname{median}_j \{\, |c_j| \mid c_j \in C \,\} \tag{5.5}$$

$$\text{maxAbs\_signed}(C) = c_i \qquad \text{with} \quad i = \arg\max_j \{\, |c_j| \mid c_j \in C \,\} \tag{5.6}$$

$$\text{minSameSign}(C) = \begin{cases} \text{minAbs\_signed}(C) & \text{signs of all } c_i \in C \text{ equal} \\ 0 & \text{else} \end{cases} \tag{5.7}$$

$$\text{medSameSign}(C) = \begin{cases} \text{medAbs\_signed}(C) & \text{signs of all } c_i \in C \text{ equal} \\ 0 & \text{else} \end{cases} \tag{5.8}$$

$$\text{maxSameSign}(C) = \begin{cases} \text{maxAbs\_signed}(C) & \text{signs of all } c_i \in C \text{ equal} \\ 0 & \text{else} \end{cases} \tag{5.9}$$

[12] Details coefficients have zero mean.

$$\text{prodSameSign}(C) = \begin{cases} s \prod\limits_{c_i \in C} |c_i| & \exists s \, \forall \text{i: } \text{sgn}(c_i \in C) = s \\ 0 & \text{else} \end{cases} \tag{5.10}$$

$$\text{geomeanSameSign}(C) = (\text{prodSameSign}(C))^{\frac{1}{\#C}} \tag{5.11}$$

For complex-valued sets of coefficients, such as those from DTCoWT, the absolute value $|\cdot|$ denotes complex magnitude. Fusion rules based on the product, such as *prodSameSign* and *geomeanSameSign*, are not computed for complex-valued decomposition methods in this study[13]. Note that the rules 5.7–5.11 are newly introduced here, compared to the rules 5.4–5.6 which have been introduced already in section 5.1. **Abs_signed* rules are applicable to fusion rules that select one coefficient from the set C, whereas **SameSign* rules are not restricted in this way. However, they disregard all sets C with contradicting signs, which might result in loss of important information. This danger increases with the number of values to be fused, i.e. the set cardinality $\#C$. Both strategies will be experimentally compared in this study.

5.2.5 Application to NDT data

Need for simulated measurements

To evaluate directionally sensitive detection methods, a specimen having many natural cracks is required. However, defect detection is hard to evaluate with natural microcracks since ground truth information is missing. Although high-resolution methods such as optical microscopy would be able to reliably identify surface-breaking cracks, in practice the effort to cover a sufficiently large surface area is infeasible. Using other inspection methods such as ET, MFL or TT as reference is also suboptimal, because their responses are ambiguous. Moreover, although data fusion can be used to resolve ambiguities, such results of course may not be used as the ground truth to avoid biasing the evaluation in favor of fusion rules. This dilemma is the reason for studying "proto-defects" such as grooves instead of actual cracks, as demonstrated in the previous experiment in this thesis. But whereas grooves are a viable alternative to natural cracks when studying undirectional detection methods, techniques that make use of directional information might have an unfair advantage concerning the grooves' perfectly straight nature. For these reasons, an evaluation approach is required that makes use of real measurements as much as possible, but still allows studying naturally shaped microcracks.

To address this issue, this study empirically "simulates" indications of natural defects by combining measured sensory output from grooves with shapes of microcracks. This is done as illustrated in figure 5.13 a–e. In steps a–b, each sensor's response to discontinuities is determined by extracting measured signals when crossing a groove. Each profile per sensor is generated by fitting a Gaussian peak model to the data to reduce the influence of noise. The idea is to transfer this sensor response to other discontinuities, such as natural cracks, by modeling indications as a result of convolution between the sensor response and the position of the discontinuity. Under this assumption, the groove can simply be replaced with realistic defect shapes, which are for example extracted from a different specimen (steps c–d). Convolution of the groove response with the crack path (step e) results in a simulation of crack inspection in specimen

[13] Although the product of complex numbers is well-defined, this definition does not suit the purpose of amplifying agreeing coefficients and diminishing conflicting values. For example, although complex magnitude of x^N behaves in the same way for both real and complex-valued x, the complex angle of x is altered even if x is multiplied by itself.

1. This noise-free simulation result is then combined with background noise from a groove-free area of specimen 1 to study defect detection under realistic noise conditions.

The proposed empirical simulation strategy has some limitations that should be explicitly stated. First, convolution with a fixed sensor response requires the inspection system to be linear and spatially invariant, analogously to linear time-invariant systems. As a consequence, the same sensor output is assumed along the path of a discontinuity, which is not realistic. Moreover, convolution is only carried out in perpendicular direction to the discontinuity. This results in unnaturally abrupt signal decay at the tips of the discontinuity. Furthermore, although other geometrical defect properties such as angle to the surface normal vector are known to influence inspection results, only depth into the material is accounted for by selecting a specific groove. Another shortcoming is that natural crack paths from other specimens might not be directly transferable to the original test piece, depending on the material and micro structure.

Despite these limitations, the proposed simulation approach has the following benefits:

- Arbitrary real crack shapes can be studied
- Realistic background noise is taken into account
- Ground truth is known
- SNR is controllable
- Localization uncertainty (e.g. registration error) and defect orientation are controllable

The practical implementation of the experiment is detailed next.

Realization

The ring-shaped specimen *SB*, whose properties will be described in more detail in sections 6.2.1 and 6.2.5, is used to extract signal features such as structural noise and groove indications ("Specimen 1" in figure 5.13). Specifically, sensor responses were extracted from ET, MFL/GMR and laser-induced TT inspections of a 13.5 µm deep groove. This discontinuity was chosen because it is the shallowest groove in that specimen which still produces clear indications in all inspection images. For each sensor, the one-dimensional response signal was extracted along a line that perpendicularly crosses the groove at its center, where the indication magnitudes are the strongest. As mentioned before, to each of these three signals, a Gaussian model with zero offset was fitted: $f(x) = p_A \exp(-((x - p_\mu)/p_\sigma)^2)$, where p_* denote the parameters to be fitted. In this process, data points were weighted according to their distance to the groove center, so that the fit is more sensitive to data near the expected peak position than near the tails. Among the found parameters, only the spatial scale p_σ will be used for the combination of sensor response with crack shape. Concretely, estimates of p_σ per sensor are: $p_\sigma^{\mathrm{ET}} = 403.5\,\mu\mathrm{m}$, $p_\sigma^{\mathrm{MFL}} = 66.3\,\mu\mathrm{m}$, $p_\sigma^{\mathrm{TT}} = 461.2\,\mu\mathrm{m}$.

Realistic crack shapes are extracted from TT inspection of a second specimen, called *Vergleichskörper 1* according to DIN EN ISO 9934-2 2003 (see also table A.1). This specimen contains many natural microcracks and corresponds to "Specimen 2" in the figure. Although this test piece contains a wide variety of thousands of real defects, unfortunately it cannot be used itself to evaluate defect detection due to the lack of reliable ground truth, as explained before. Nevertheless, since *Vergleichskörper 1* is

comparable to the Ring specimen in that both are made of ferromagnetic material and both are susceptible to surface-breaking microcracks, its crack paths are transferred to the Ring specimen in this study. This is done by manually marking seven prominent defects based on thermal inspection of *Vergleichskörper 1.* TT was chosen because of its high sensitivity in all orientations[14] and its comparably high spatial resolution[15]. The decision which of the manifold defects to select was guided by the criteria to (1) obtain indications from diverse areas of the surface, to (2) produce high signal to noise ratio for easy shape identification, and to (3) have a minimal length. Consequently, the selected cracks' sorted lengths are: 2.6, 2.8, 3, 3.5, 4.2, 5 and 5.8 mm, which amount to a mean length of 3.8 mm. Each of the chosen defects was approximated by a continuous path consisting of linear segments in real-world spatial units (mm). Figure 5.14 shows inspection results of some of the chosen surface cracks. To make these crack indications mimic the groove indications in the Ring specimen, subsequently these paths were rotated so that their main orientation is parallel to the groove direction before centering them in a groove-free region of the Ring specimen where all present indications represent real structural noise. The result is an image that is zero everywhere, except for the single-pixel-wide crack path along which the image has intensity one.

To convert this image to a more realistic crack indication, it is convolved in the cross-crack direction with each sensor's response pattern, which is represented by p_σ for each sensor. The respective one-dimensional convolution kernel is given by $g(x) = \exp(-x^2/p_\sigma^2)$. After convolution, the simulated inspection images are zero everywhere except for the smoothed crack path, which has intensity one along the path and a Gaussian profile across the path.

This noise-free crack simulation is combined with real noise measurements that were extracted from the groove specimen. Because natural crack paths exceed the inspected noise region, the noise image was extended beyond its boundaries by symmetric boundary value replication. Simulations are produced at various signal-to-noise ratios, based on an additive model: $f_{\text{combined}} = w\, f_{\text{crack}} + (1 - f_{\text{crack}}) f_{\text{noise}}$. This formula applies to each pixel in the simulated image. By weighting f_{noise} with the intensity of the simulated crack, a smooth transition is guaranteed between noise and crack. Note that $f_{\text{crack}} \in [0 \ldots 1]$, and therefore f_{noise} dominates pixels that are far[16] from the crack, whereas it has no effect on the simulated pixel intensity over the crack. This noise model is more realistic than simply adding noise and crack, which is the common additive noise model, because structural noise is not independent from the crack indication. In fact, a crack, being a structural discontinuity itself, rules out the possibility of any other near-surface structural inhomogeneity at the same location. The weight w directly controls amplitude-based SNR (in linear units of intensity) under the following assumptions. f_{crack} is zero at distant locations from the crack, and has unit intensity directly on the crack. f_{noise} denotes real noise measurements after being standardized, thus having zero mean and unit variance. By defining

$$\text{SNR}(f_{\text{signal}}, f_{\text{noise}}) := w \, \max\left(f_{\text{signal}}\right) / \text{Std}\left(f_{\text{noise}}\right) \tag{5.12}$$

we have $\text{SNR}(f_{\text{crack}}, f_{\text{noise}}) = w \, \max\left(f_{\text{crack}}\right) / \text{Std}\left(f_{\text{noise}}\right) = w$. For example, consider figure 5.15. The right column of sub-figures shows actual inspections of the groove that

[14] Because the specimen has a fixed remanent magnetic field according to the norm [121], MFL is less sensitive to cracks that are oriented parallel to the magnetic field lines

[15] Although ET using an absolute probe is also sensitive to cracks of arbitrary rotation, physical spatial resolution is comparably poor.

[16] "Far" is relative to the simulated width of the crack indication, p_σ.

Table 5.2: Decomposition parameters

Method	Levels	Parameters	Reference
SWT	5	Mother wavelet: db2	swt() [122]
UWT	5	Filter: spline_3	[123]
DTCoWT	5	Filter: dtf3	dddtree2() [122]
NSCT	5	Directional filter: dmaxflat7 Low pass filter: 13-tap symmetric maxflat	[124]
ST	5	Directional filter: dmaxflat4 Low pass filter: 9-tap symmetric maxflat	[94]

was used to extract the peak profile per sensor. Also, real noise was captured from these data sets in a off-groove region that is not shown here. The left column of sub-figures contrasts the actual groove images with simulations of natural crack shapes, where the original SNR was reproduced per sensor (figure rows). The simulations represent indications of natural crack shapes using sensor-specific peak profiles across the defects and sensor-specific structural noise in the background. Noise differs between simulations and real measurements, because noise was extracted from a different region than the shown groove neighborhoods.

For the experiments in this section, simulations f_{combined} were computed at the following SNRs: 0.5, 0.75, 1, 1.5, 2, 4. Since these SNRs are lower than those observed in figure 5.15, they represent flaws that are shallower than 13.5 µm.

Subsequently, scale normalization is applied to each simulated image. Because MFL has the best physical resolution among the chosen NDT methods, indications in ET and TT images are thinned to typical widths of MFL indications. Specifically, the size of the smoothing kernel was set to $\sigma_{\text{scale norm.}} = 66.3\,\mu\text{m}$, which has already been determined as the sensor response width p_σ^{MFL} before.

Image decomposition is carried out for each simulated sensor image at each SNR and for each crack, according to the parameters provided by table 5.2. Other choices of filters in the shearlet transform did not have a large effect on the fused images. For each decomposition method SWT, UWT, DTCoWT, NSCT and ST, fusion rules 5.5–5.11 were applied to both the details and approximation coefficients, simultaneously integrating all sensor images, before reconstruction. In addition to the multi-scale transforms, fusion of the original image intensities (without any prior transform, denoted *undirectional* or *per-pixel* fusion) was carried out. Undirectional fusion rules are more straightforward than those designed for transform coefficients, because image intensities can be assumed to be unsigned: After shape normalization (sec. 4.1), indications are represented by high intensities in the image. Furthermore, these intensities vary in the same range by means of radiometric normalization. Each sensor's intensities can therefore be made nonnegative by applying a low threshold (to avoid removing weak indications), and shifting the intensities such that the new minimum value is zero. This operation is notable, because zeroes have a large effect on e.g. the *product* fusion rule.

The next section presents the results, after the evaluation strategy has been clarified.

Figure 5.13: Empirical simulation of natural cracks. a) A groove was inspected by several NDT methods. The dashed line indicates a path on the specimen surface crossing the center of the groove. b) For each inspection method, a peak profile is extracted from the sensor output along the surface path by fitting a Gaussian peak model to the measurements. c) A natural crack is selected from a second specimen. d) Several locations along the natural crack are extracted. e) For each sensor, the extracted peak profile from step b) is convolved with unit impulses whose positions were extracted in step d). The result is a noise-free simulation of natural crack inspection in Specimen 1.

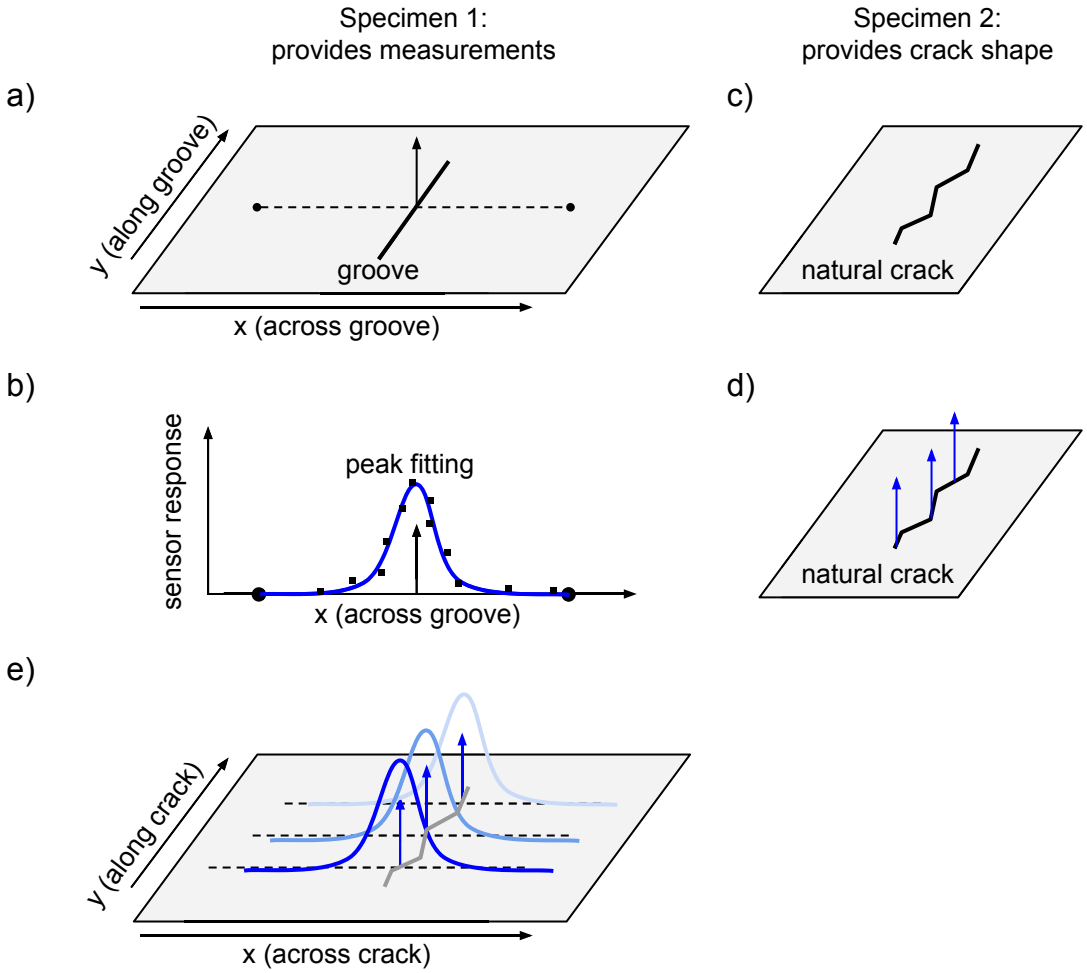


Figure 5.14: Selection of natural crack paths from laser-induced TT of specimen *Vergleichskörper 1.* Black dots indicate vertices of the manually created polygon chains.

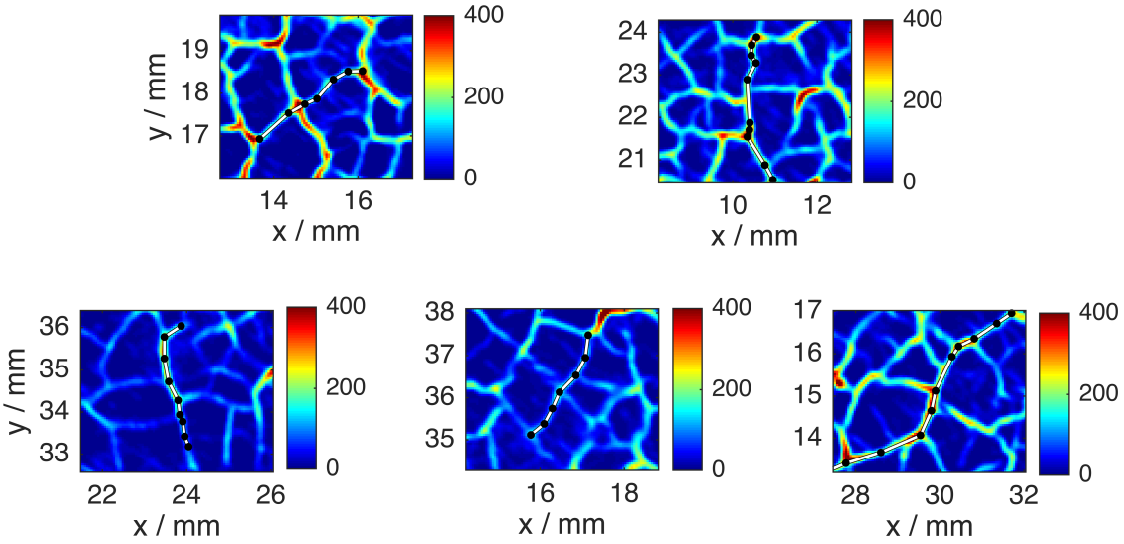
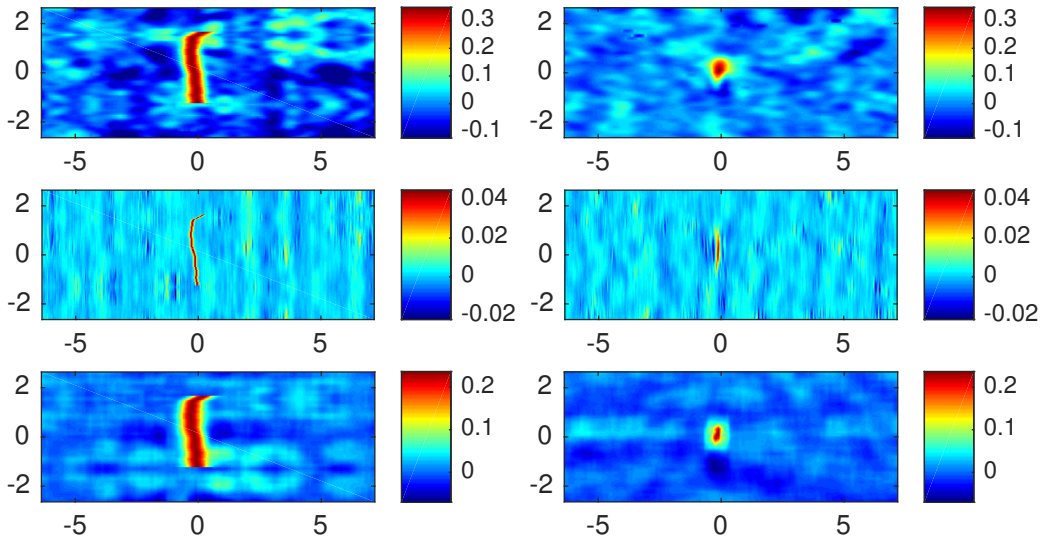


Figure 5.15: Comparison between simulation (left column) and actual inspection (right column). The simulations represent a natural crack shape, whereas actual inspections show a 13.5 µm deep groove. Rows denote different NDT methods: ET (top), MFL/GMR (middle), TT (bottom). For comparability, simulations in this figure reproduce the grooves' respective SNR for each sensor. Specifically, according to equation 5.12: $\mathrm{SNR}_{\mathrm{ET}} = 0.34/0.0563 = 6.035$, $\mathrm{SNR}_{\mathrm{MFL}} = 0.044/0.00315 = 13.99$, $\mathrm{SNR}_{\mathrm{TT}} = 0.23/0.0169 = 13.635$. All intensities are in arbitrary units. Color limits are matched for same sensors.



5.2.6 Results

Evaluation strategy

Classical ROC analysis is carried out. Each pixel is evaluated, and the ground truth is generated from the known simulated crack path. As in undirectional fusion at the signal level (section 5.1.4), a region around the simulated crack pixels is ignored to make the evaluation insensitive to different cross-crack peak widths (physical sensor resolutions), which is important if scale normalization is not carried out. Concretely, a margin of 1.38 mm in all directions away from the crack path, corresponding to $3 \cdot \max_S p_\sigma^S$ among sensors S, was marked to be ignored during evaluation. From the remaining pixels in the simulated image, ROC curves were computed for each combination of SNR, natural crack path, image decomposition technique and fusion rule. Altogether, this study compares up to #(SNRs) · #(Paths) · #(Decomposition and fusion methods) = $4 \cdot 7 \cdot 48 = 1344$ different settings.

To further aggregate the evaluation results for easier comparison of results, two high-level evaluation measures are extracted from each ROC curve. The Area under the ROC Curve (AUC) is a traditional measure of classification performance across the whole range of detection thresholds. But in fact not the whole range of thresholds is of interest in defect detection.

For successful crack detection, in practice it often suffices to set the detection threshold low enough so that a significant part of the defect exceeds the threshold, rather insisting on finding all relevant pixels. Otherwise, if the detection threshold is further lowered, the gain in True Positive Rate (TPR) is not worth sacrificing specificity by introducing many false alarms. Therefore, in the same way as was proposed in section 5.1.4, this study regards cracks as *detected* if at least half of its pixels are correctly found (TPR ≥ 0.5). Consequently, a meaningful performance measure is the False Positive Rate (FPR) that the ROC curve assumes at this fixed TPR value [89, sec. 2.5]. Another alternative to AUC is to integrate the ROC curves not across the whole FPR range $0 \ldots 1$, but only up to a fixed false positive level. Given the much larger number of off-crack pixels (negative class) compared to the crack pixels (positive class), whose ratio is roughly 2000:1 with the present data, only very small FPRs are acceptable. Therefore, the *partial AUC* [89, sec. 2.6] is computed in the FPR subset $0 \ldots 0.01$, and then divided by the chosen upper FPR limit so that the measure is normalized between 0 and 1.

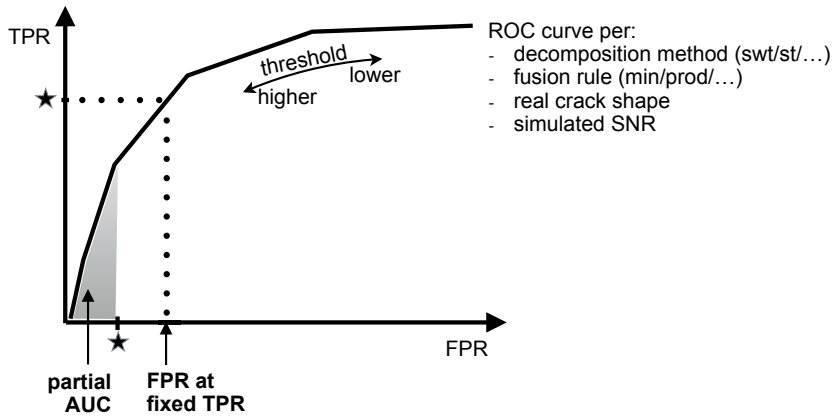
A graphical representation of *FPR at fixed TPR* and *partial AUC* in ROC space is shown in figure 5.16. From the figure, it is clear that both measures are invariant to detection performance in the upper right corner of ROC space, that is the low-threshold region. Consequently, these measures are robust against skewed evaluation results when a defect has high SNR overall but only a small number of its pixels are indistinguishable from noise. The conventional AUC, in contrast, would be strongly affected, because the ROC curve would ascend rapidly but then would flatten out well before the maximum TPR is reached. This flat long part of the curve, however, dominates the curve's integral. Compared to evaluation as carried out for undirectional fusion in section 5.1.4, *partial AUC* is introduced here to provide a complementary evaluation measure that might be less sensitive to small variations, since it represents averaged detection performance. Moreover, *partial AUC* is defined by fixing a certain FPR, in contrast to *FPR at fixed TPR*, and therefore shows a different view of detection performance.

For the given reasons, in the following, quantitative evaluation is based on the two measures *partial AUC up to FPR=0.01*, denoted $pAUC_{0.01}$, and on *FPR at fixed*

TPR=0.5, denoted $FPR_{0.5}$. When it is more appropriate[17] to express detection performance in terms of specificity than in terms of false positive rate, the acronym $Spec_{0.5}$ will be used in the text. Note that the subset numbers do not refer to the same quantity: $FPR_{0.5}$ and $Spec_{0.5}$ are defined by fixing the *true positive rate*, whereas $pAUC_{0.01}$ is defined by fixing the *false positive rate* instead.

The first of the conducted experiments will compare different sign-resolving strategies during fusion.

Figure 5.16: High-level evaluation measures extracted from ROC curves. Star-shaped markers indicate the parameters whose values define the two evaluation measures *partial AUC* and *FPR at fixed TPR*.



Strategies to resolve sign conflicts To limit the number of fusion methods to be evaluated in the next experiments, two categories of fusion rules are compared first. In eq. 5.4–5.9, two strategies for dealing with signed coefficients during fusion are proposed. Rules denoted *SameSign only fuse coefficients whose signs agree, and set the fusion result to zero otherwise. In contrast, *Abs_signed considers all coefficients. It disregards the signs during the fusion rule, and re-assigns the resulting value its original sign. These two strategies are compared for the *minimum*, *median* and *maximum* fusion rules, for all introduced decomposition methods[18] at several simulated SNRs. Comparison is done by means of statistical analysis of the mean performance between the two sign resolution strategies in each case, across different simulated cracks. Specifically, each such pair of strategies was evaluated by means of a one-sided t-test under the null hypothesis that *SameSign does not have higher mean performance score than *Abs_signed. The alternative hypothesis states that *SameSign has a higher mean performance than *Abs_signed. The test assumes normally distributed performance scores across different cracks, but does not require equal variances about the two means. Each test was performed twice, using the two performance measures $pAUC_{0.01}$ and $FPR_{0.5}$. A significance level of 0.01 is adopted in this study. Table 5.3 shows the resulting p-values, which lead to the following observations:

[17] Although specificity is equivalent to FPR (specificity = 1-FPR), unlike FPR it follows the convention that higher evaluation values indicate better performance, similarly to TPR and (partial) AUC.

[18] except for DTCoWT whose complex-valued coefficients are incompatible with *SameSign rules.

- $pAUC_{0.01}$ is more discriminative than $FPR_{0.5}$, because it averages across a broader range and thus varies more
- significant differences are found mostly in a specific range of SNR: if SNR is too small, then all methods perform comparably poorly. If SNR is too high, then all methods are comparably good. There is a niche in which, according to $pAUC_{0.01}$, *SameSign is generally better than *Abs_signed, for all considered decomposition methods.
- considering the *maximum* fusion rule, *SameSign seems to be better at all SNRs. However, as shown later, *max* is an inappropriate rule overall.

Because one-sided tests were carried out, insignificant results in the sense of large p-values might actually represent cases in which the alternate hypothesis is reversed, that is *Abs_signed has a higher mean performance than *SameSign. Regarding this issue, it is particularly interesting to analyze the lowest simulated SNR=0.5 where almost none of the tests rejected the null hypothesis. The *minimum* rules is exemplarily selected for detailed investigation. While results for SWT and UWT show indistinguishable performance between *SameSign and *Abs_signed at SNR=0.5, NSCT and ST both are slightly in favor of *Abs_signed, judging by $FPR_{0.5}$. Also, performance of *Abs_signed is observed to be more stable across different crack shapes than *SameSign. In contrast, results are indistinguishable when considering $pAUC_{0.01}$ as the performance index. In conclusion, although the evidence is scarce, there seems to be an advantage of Abs_signed for the *minimum* rule at very low SNRs, for decomposition methods NSCT and ST.

To simplify the following experiments, according to these findings, analyses will focus on *SameSign fusion rules.

Evaluation of fusion rules for given decomposition methods This section compares the fusion rules *minSameSign*, *medSameSign*, *maxSameSign*, *prodSameSign* and *geomeanSameSign* separately for each decomposition method *SWT*, *UWT*, *DTCoWT*[19], *NSCT* and *ST*. Detailed results are shown in the Appendix in figures A.2–A.6. For each decomposition method, one sub-figure is shown per simulated SNR. Each colored marker represents one simulated crack. Results are colored by same fusion rule (see the legend at the bottom), and the convex hull is plotted for visual guidance. Each axis represents an evaluation measure based on the ROC curve, that is $Spec_{0.5}$ and $pAUC_{0.01}$. Optimal performance is achieved at the top right corner in each sub-figure (coordinates 1/1). Because two evaluation measures are considered, method ranking is ambiguous. In fact, the two measures are complementary, which means that two methods might be inversely ranked with regard to each measure. In such cases, the two concerned methods are assigned the same rank in this study.

A ranking of fusion rules, which summarizes the detailed results, is shown in table 5.4. The two highest simulated SNRs 2 and 4 are not included in the table, because data quality is so good that individual fusion rules often do not differ significantly. Note that the ranking does not convey how far apart individual fusion rules are in terms of performance, and therefore a fourth place might actually not be far away from the top rank in absolute terms. Nevertheless, the reduction of complexity allows to formulate the following conclusions.

Differences between rules are best observable around SNR=1. In this quality range, the performances of different fusion rules usually form distinct clusters, and also a

[19] As mentioned before, product-based rules are not applied to complex coefficients from DTCoWT.

Table 5.3: Comparison of fusion rules for signed coefficients: *SameSign vs. *Abs_signed (see eq. 5.4–5.9). The printed numbers are p-values from t-tests under the null hypothesis that *SameSign does not have higher mean performance score than *Abs_signed among the simulated cracks. In each cell of the table, two p-values correspond to the two performance measures: $pAUC_{0.01}$ / $FPR_{0.5}$. Null probabilities below 0.01 are considered significant, and are printed bold.

	SNR:	0.5	0.75	1	1.5	2	4
SWT	min	0.527 / 0.142	**0.008** / 0.987	**0.003** / 0.998	0.248 / 0.500	0.364 / 0.500	0.500 / 0.500
	med	NaN / **0.008**	**0.001** / 0.922	**0.000** / 1.000	0.579 / 0.969	0.522 / 0.500	0.500 / 0.500
	max	0.118 / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	0.082 / 0.865
UWT	min	0.959 / 0.041	**0.002** / 0.977	0.013 / 0.992	0.344 / 0.500	0.279 / 0.500	0.495 / 0.500
	med	0.820 / 0.103	**0.000** / 0.743	**0.001** / 1.000	0.884 / 0.698	0.515 / 0.500	0.500 / 0.500
	max	**0.008** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	0.109 / 0.862
DTCoWT	min	**0.009** / 0.898	**0.000** / 1.000	**0.000** / 0.999	0.081 / 0.860	0.394 / 0.500	0.500 / 0.500
	med	0.323 / 0.611	**0.000** / 0.980	**0.000** / 1.000	**0.003** / 1.000	0.189 / 0.648	0.500 / 0.500
	max	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	0.098 / 0.178
NSCT	min	0.978 / 0.024	**0.002** / 0.618	**0.001** / 1.000	0.352 / 0.500	0.483 / 0.500	0.500 / 0.500
	med	0.822 / 0.315	**0.005** / 0.377	0.024 / 0.997	0.915 / 0.709	0.574 / 0.500	0.500 / 0.500
	max	0.050 / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	0.145 / 0.796
ST	min	0.735 / 0.016	0.139 / 0.362	0.587 / 0.999	0.287 / 0.500	0.489 / 0.500	0.500 / 0.500
	med	0.822 / 0.340	**0.001** / 0.961	**0.000** / 1.000	0.670 / 0.780	0.457 / 0.564	0.500 / 0.500
	max	0.031 / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	**0.000** / 1.000	0.058 / 0.875

unique ranking is possible because both evaluation measures correlate.

minSameSign is consistently the best rule, across SNRs and decomposition methods. *geomeanSameSign* often shows comparably high performance. This is expected because both rules require agreement in all sensors to produce an output coefficient of high magnitude. However, although the *product* rule was expected to perform similarly to *min* and *geomean*, it is often placed on the last ranks. To investigate this result in detail, figure 5.17 compares the two rules *prodSameSign* and *geomeanSameSign* for the same decomposition method, ST. It is clearly seen that the *product* rule introduces smoother indications than *geomean*, and false alarms have very high intensities. This is explained in the following way. If multiple sensors agree, i.e. their coefficients have the same sign and their magnitudes exceed the structural noise, then the *product* rule strongly amplifies signal power P according to $P^{\#S}$. Other rules, such as the *minimum* rule, are rather passive because they never amplify power. Typically, signal power is higher at coarse spatial scales than at finer scales. Since all considered decomposition methods are not only directionally selective, but also scale-selective, signal power at coarse scales is boosted much more than power at fine scales. Therefore, spurious sensor agreement at coarse scales introduces high-intensity false alarms under the *product* rule. In contrast, although the *geometric mean* is in fact defined in terms of the *product* rule (see equation 5.11), the introduced imbalance between signal powers at different scales is corrected by taking the root. Therefore, while coefficients that express disagreement among sensors are reduced in magnitude by both *product* and *geomean*, only *geomean* retains the original magnitude range of conforming coefficients.

The results further show that at the lowest simulated SNR, the methods that provide high directional selectivity (DTCoWT, NSCT and ST) all show similar performances for *minSameSign* and *maxSameSign*, although these rules contrast each other conceptually. This indicates that the simulated crack does not lead to agreement of signs among the sensor images' coefficients, which produces the same output value of zero.

Table 5.4: Ranking of fusion rules for different image decomposition methods at several SNRs. Rank 1 is best. Ties are comma-separated. To save space, the *geometric mean* fusion rule is denoted *gmean* here.

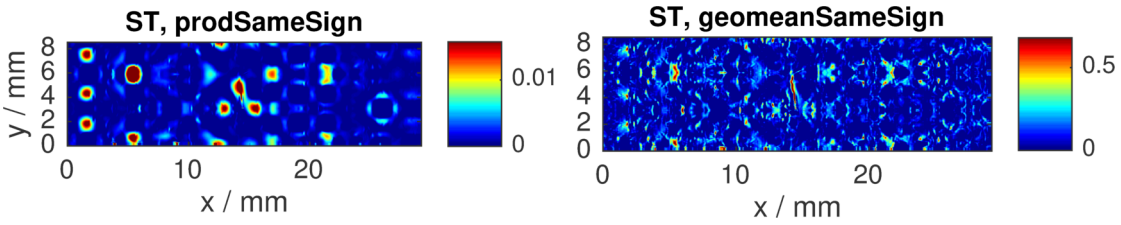
	SNR:	0.5	0.75	1	1.5
	Rank				
SWT	1	min	min, gmean	min, gmean	min, med, gmean
	2	gmean	med, max	med	max, prod
	3	med, max	prod	max	
	4	prod		prod	
UWT	1	min	min	min	min, med, prod, gmean
	2	gmean	gmean	gmean	max
	3	med, max	med, max, prod	med	
	4	prod		prod, max	
DTCoWT	1	min, max	min, med, max	min, med	min, med
	2	min, med		max	max
NSCT	1	min, max	min, gmean	min, gmean	min, med, gmean, max
	2	gmean, med, max	med, max	med, max	prod
	3	prod	prod	prod	
ST	1	min	min, gmean	min, gmean	min, gmean, med
	2	gmean, med, max	med, max	med	max, prod
	3	prod	prod	max	
	4			prod	

Evaluation of fusion rules for per-pixel fusion To assess the benefit of directional decomposition methods, *un*directional per-pixel fusion was evaluated in the same way. The results are shown in figure A.7. A clear hierarchy can be observed among the methods. From best to worst, the ranking is given by *minimum*, *harmonic mean*, *product*, *sum*, *maximum*. Interestingly, this hierarchy is already observed at the lowest simulated SNR, where directional methods are often separated much less clearly in the performance space. Note that in the undirectional case, pixel intensities can be assumed as unsigned and therefore the geometric mean is a monotonic transformation of the product rule, thus resulting in equal performances.

Comparison of decomposition methods After the best-performing fusion rules have been determined for each decomposition method, including undirectional fusion, the different decomposition methods are compared in this analysis. The following hypotheses will be examined:

- Hypothesis 1: All fusion approaches outperform single-sensor inspection

Figure 5.17: Comparison between *product* and *geomean* fusion rules, for ST decomposition method, at simulated SNR of 1, for one of the modeled crack shapes. The simulated crack is located at the center of each figure, in vertical orientation. All other indications are false alarms. Color ranges were set to $[0 \dots K]$, where in each image K denotes the intensity threshold that produces TPR=0.5, i.e. half of the on-crack pixels' intensities are greater than K.



- Hypothesis 2: Directional fusion methods outperform undirectional methods
- Hypothesis 3: Highly directionally sensitive methods (DTCoWT, NSCT, ST) outperform less directionally sensitive methods (SWT, UWT)

To check the first hypothesis, consider figure A.8. At the lowest simulated SNR of 0.5, performance of single-sensors ET and TT is in the same range as the fusion approaches, concerning $Spec_{0.5}$. But the second evaluation measure reveals that all single-sensor images are not able to find any on-crack pixel when requiring FPR ≤ 0.01, because $pAUC_{0.01} = 0$ for all simulated cracks. In contrast, $pAUC_{0.01}$ is non-zero for all fusion approaches at SNR=0.5. On the other end of the range of simulated data quality, there is no difference between single sensor detection and fused detection performance at SNR=4. The only source of variation is simulated crack shape. From simulated individual SNR = 0.75 to 2, all fusion approaches clearly outperform the single sensor images regarding both evaluation measures.

To address the second hypothesis, figure A.9 displays results for all decomposition methods, including undirectional fusion. It can be observed that *pointwise_min* considerably outperforms directional methods at the lowest simulated SNR. Also when increasing the SNR to 0.75, the undirectional approach is among the top methods, although separation between the techniques in performance space is not possible anymore. Interestingly, as SNR increases further, the undirectional fusion method seems to fall slightly behind the directional transforms. Still, practically all methods achieve good results at SNR=1 and above, considering that the false positive rates that have to be accepted to detect at least half of all on-crack pixels is lower than 0.11 %.

Considering the third hypothesis, figure A.9 shows that directional decomposition methods almost always overlap in the evaluation space. Only at the lowest simulated SNR, UWT is clearly the best among the multi-scale methods, even though its directional sensitivity is the most limited (together with SWT). Therefore, it can be concluded that with regard to this experiment, directional sensitivity is not the dominating factor influencing defect detection performance. UWT is unique among the methods due to its specially designed multi-scale filters. Other transforms that rely on less compact analysis filters apparently generate spurious agreement among the NDT sensor images, thus introducing more false indications in the reconstructed images. Therefore, in the next section, modifications are proposed to the basic fusion approach that was followed so far.

5.2.7 Modifications to the fusion approach

To demonstrate the potential shortcomings of directionally sensitive multi-scale transforms compared with undirectional per-pixel fusion, compare figure 5.18 (a) and (e). Whereas the per-pixel minimum rule (e) results in relatively few fused indications, the shearlet-based result (a) contains a multitude of non-zero intensities that also appear much wider. Examination of the fused coefficients confirmed that significant signal power is present at off-groove locations already before reconstruction, thus ruling out problems during image synthesis. Since fusion is carried out for each image pixel independently, coefficients from different image regions are not able to influence each other during fusion. Therefore, it is clear that already during the image analysis step, false indications from the individual sensor images introduce significant wide-spread coefficient intensity. Especially in the presence of low SNR, this effect is detrimental for detection performance because different individual false alarms are falsely related during fusion. This is in contrast to undirectional per-pixel fusion, where a pixel in the source images is guaranteed to influence only the same location in the fused image. The described phenomenon is known as *coefficient spreading problem* [98], and is related to the length of the analysis filters in multi-scale transforms. Note that this study already acknowledges this problem by using an implementation of the shearlet transform that was specifically developed to feature compact decomposition filters [94]. Yet, because the directional filters introduce higher complexity in the filters e.g. compared to the UWT, additional a-priori knowledge is required to correct the spreading problem.

To this end, the following modifications to the fusion approach adopted so far are proposed:

- M1 - Set fused approximation coefficients to zero
- M2 - Apply thresholding to sensor images before decomposition
- M3 - Set details coefficients of single-sensor images to zero at the same locations that were affected by M2

These modifications are not ordered by the time of execution during fusion, but so that their cumulative effects are best seen in figure 5.18. In detail, M1 suppresses large-scale false indications in the fused image. The reason is that approximation coefficients capture coarse-scale background image variations, which are expendable for representing defect indications. Furthermore, even with highly directionally sensitive decomposition methods such as ST, the approximation coefficients do not capture any directed features of the original image, and are therefore irrelevant for the detection of cracks. The effect of M1 is seen in figure 5.18b, which still contains many more false indications than the undirectionally fused image 5.18e. Therefore, M2 further reduces chances for spurious sensor agreement by fusing sensor images that were first made nonnegative, in the same manner as in the per-pixel fusion approach (see the end of section 5.2.5). This is because such preprocessing of the source images incorporates prior knowledge that should be accessible to all detection methods. Note that multi-scale transforms of unsigned images nevertheless produce signed coefficients. However, since negative intensity in the source images is assumed to be irrelevant for defect detection, it is removed to prevent introducing spurious sensor agreement during fusion. Figure 5.18c exemplarily confirms the effectiveness of the modification. Like M2, M3 follows the same logic by removing signal intensity from the same image regions. However, M3 is applied to per-sensor coefficients after image decomposition, whereas M2 operates in the original signal domain. Therefore, M3 directly constrains the fusion result and influences

more pixels in the output image than M2 due to filtering during image reconstruction. Applying severe alterations such as M3 to the coefficient structure is justified by using redundant transforms[20], which still succeed in reconstructing a high-quality image as shown in figure 5.18d. Note that the color scales reflect the loss of overall signal energy as a consequence modifications 1–3.

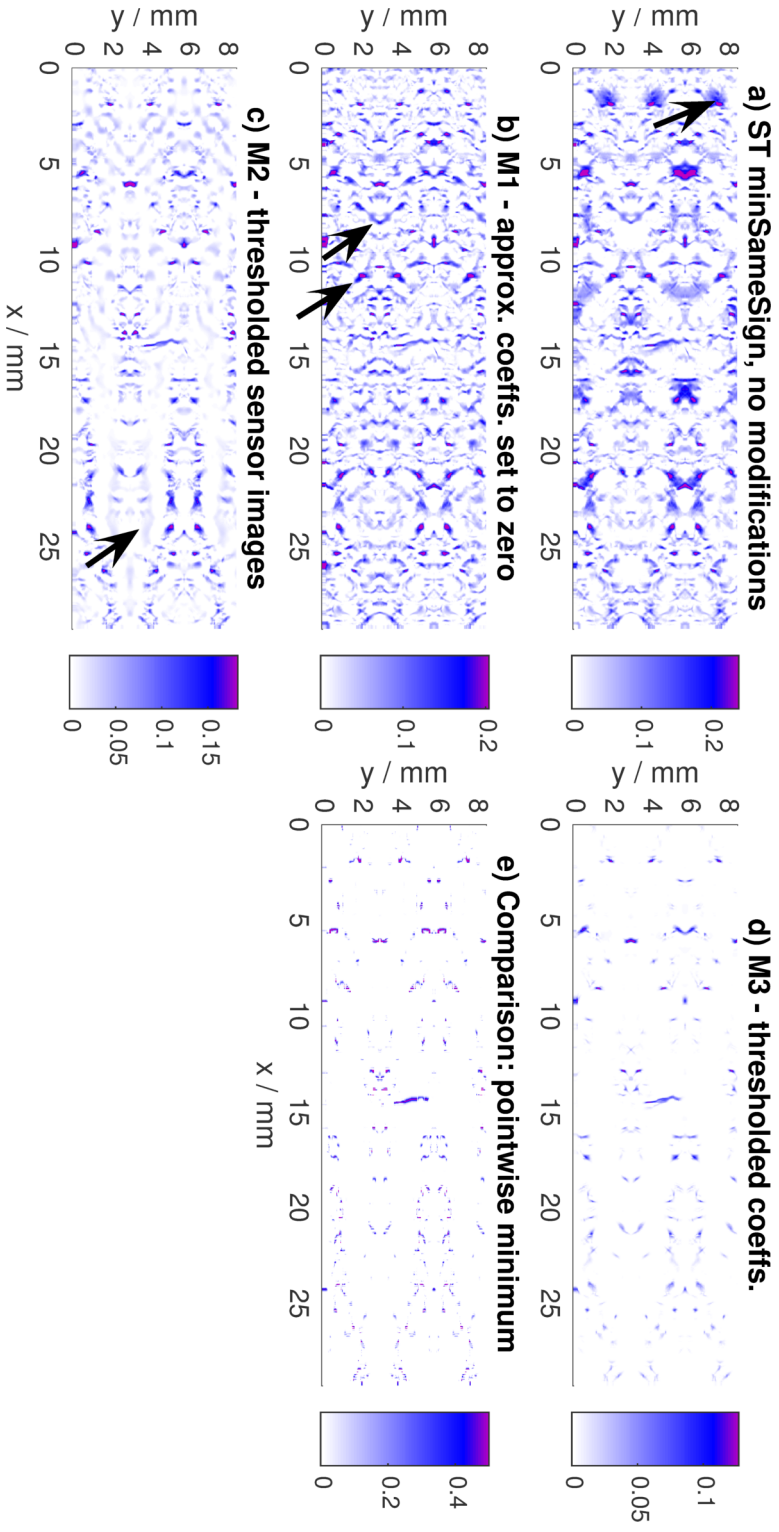
After applying the proposed modifications, both the directional and undirectional fusion approaches produce a comparable set of false indications, quite unlike the non-modified fusion strategy. Still, some differences can be observed. Specifically, the ST-fused image looks smoother than its undirectional counterpart, and its indications have different shapes. Smoothness is explained by the filtered reconstruction process. Moreover, whereas undirectional fusion is unable to analyze or influence the shape of an indication, the Shearlet Transform composes the final indications from directed atoms and therefore favors oriented indications in the fused image (independently from the fusion rule). Most importantly, the crack detectability should be quantified. Do the proposed modifications lead to an improved performance of directionally sensitive fusion strategies, or even to an advantage over the conceptually much simpler per-pixel fusion?

Figure A.10 presents the evaluation results. Note that results are computed for a range of smaller SNRs than before, now starting at 0.25 and ending at 2. To compare against the non-modified algorithm (Fig. A.9), the pointwise fusion rule provides an anchor of reference because modifications M1–M3 do not affect undirected fusion strategies. Accordingly, the proposed modifications improved multi-scale fusion results at low SNRs: Whereas before there was a clear gap in the performance diagram at SNR=0.5 between pointwise fusion and the best directional method, UWT, now the detection ability can be considered equal for UWT, NSCT, ST and pointwise. Improvements of SWT can also be observed, but they do not suffice to challenge undirectional fusion at this SNR. When simulated data quality is further lowered to SNR=0.25[21], performances diversify and it is seen that pointwise fusion is still one of the best methods, together with UWT and ST. Again, directional sensitivity does not seem to be the most influential factor on detection performance, since the two best-performing techniques are at opposite ends in terms of directional resolution. Therefore, although the proposed modifications did not succeed in leveraging the theoretical advantage of directionally sensitive transforms over undirected methods, they are necessary to attain comparable performance.

[20] Whereas SWT, UWT, NSCT and ST are highly redundant, DTCoWT is not. In fact, this method was developed to have approximate shift-invariance despite low redundancy. For this reason, and because M3 is hard to implement because coefficients cannot be uniquely assigned to pixels, M1–M3 were not applied to DTCoWT.

[21] For completeness it should be noted that at this low SNR of 0.25, 49 fusion results produce conventional (non-partial) AUC below 50 %, i.e. worse than chance. Almost all of them were fused by the *maximum* rule, which explains the poor performance. However, also three of the cracks fused by *dtcowt_minSameSign* at SNR=0.25 actually produce worse-than-chance AUC. Higher simulated SNRs are not affected.

Figure 5.18: Effect of proposed modifications on the fusion result. Exemplary results for ST minSameSign, for one crack shape at simulated SNR=0.5. Modifications accumulate, that is M2 includes M1, and M3 includes M1 and M2. Color ranges were adjusted for each sub-figure such that positive intensities in a small region around the simulated defect fill the whole color range. Negative intensities were clipped. Arrows indicate image features that are suppressed by the following modification.



5.2.8 Influence of crack orientation

In the previous section, only vertically oriented defects[22] have been studied, which is the optimal orientation for directionally sensitive NDT methods such as differential ET, and MFL/GMR, which measure contrast of electrical conductivity/field strength along the horizontal direction. But it will also be interesting to study detection performance in the more general case when crack rotations are involved. For this purpose, the same empirical simulation approach is applied, but this time the crack shapes are randomly rotated about their origin. $N = 30$ samples are drawn from a uniform distribution in the interval $[-30° \dots 30°]$. Angles beyond 45° are not of interest because directionally sensitive inspection methods are applied twice using perpendicular orientations and therefore 45° represents the largest relevant angle. Yet, the range of the uniform distribution is made narrower (30° instead of 45°) because the crack paths themselves involve kinks and bends, which add to the simulated amount of rotation. The rotated crack paths represent the ground truth during evaluation. To limit the scope of this evaluation, the NSCT decomposition method is excluded[23], as well as several fusion rules that are known to be prone to false indications (like the *maximum* rule). The modifications M1–M3 are applied as proposed in section 5.2.7.

In a first experiment, the best fusion rule is determined for each decomposition method. Can the results from the purely vertical crack orientation be transferred to the more general case? The results are shown in figure 5.19, for low SNR=0.25 where the differences emerge the most clearly. For all decomposition techniques including per-pixel fusion, the *minimum* rule provides the best detection performance, although $FPR_{0.5}$ is often comparable. This finding agrees with the previous results at vertical defect orientation. Moreover, $pAUC_{0.01}$ indicates that *minAbs_signed* is always as good as or better than its more strict variant, *minSameSign*.

These optimal fusion rules should now be compared for varying amounts of defect rotation. How do different decomposition methods compare for non-vertical defects? Is there a relationship between rotation and detection performance?

Figure 5.20 presents the results at low SNR of 0.25. In both sub-figures, the horizontal axis is absolute deviation from the vertical orientation, to either side. Although performances according to $Spec_{0.5}$ are hardly separated, DT-CoWT and SWT appear to perform worse than UWT, ST and per-pixel fusion, which is consistent with results presented before. $pAUC_{0.01}$ provides more insights. In addition to DT-CoWT and SWT, also per-pixel fusion is clearly separated from the performances of UWT and ST for all simulated crack images. Concerning the influence of rotation angle, it is clear that no relationship exists. Rather, the dominating influence of performance variation is given by the different natural crack shapes that were also randomly sampled.

In conclusion, the experiment shows that per-pixel fusion can be outperformed by directional fusion methods according to evaluation measures that target high image intensities. However, this improvement does not depend on the simulated crack orientation, nor does directionality of the decomposition method seem to play a large role, since UWT has similar directional sensitivity as SWT.

[22] with respect to the chosen joint coordinate system

[23] mainly due to its long execution times, but also because its directional sensitivity is theoretically similar to the Shearlet Transform

Figure 5.19: Detection performance of randomly rotated crack simulations at SNR=0.25. Each sub-figure represents one decomposition method. Each marker denotes one simulated crack and its color represents different fusion rules. Only the best-performing rules are labeled with a text for clarity. Optimal performance is attained at coordinates (1,1) in each diagram, that is, at the top right. Axis scales differ between the sub-figures, because the goal is to compare different fusion rules for each decomposition method independently.

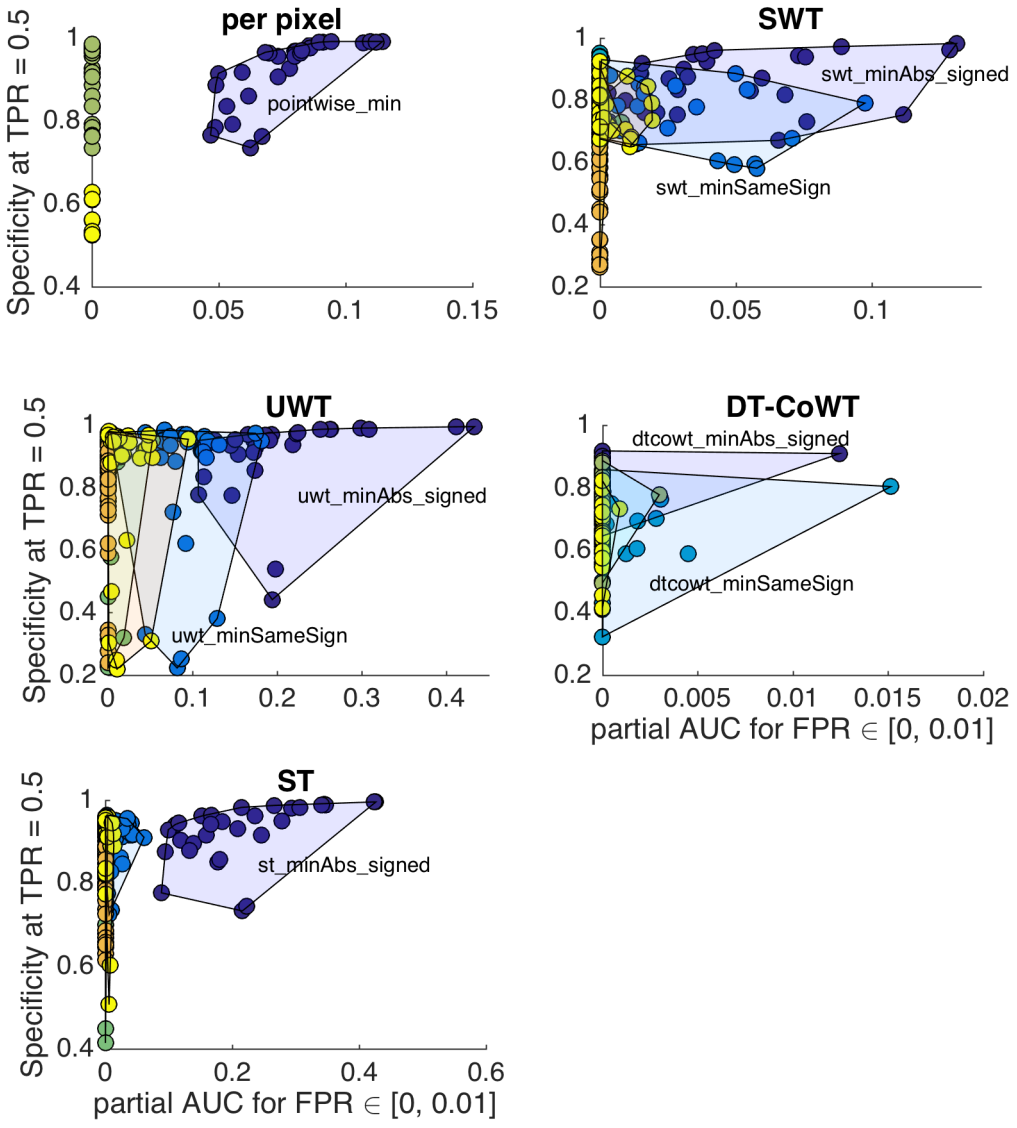
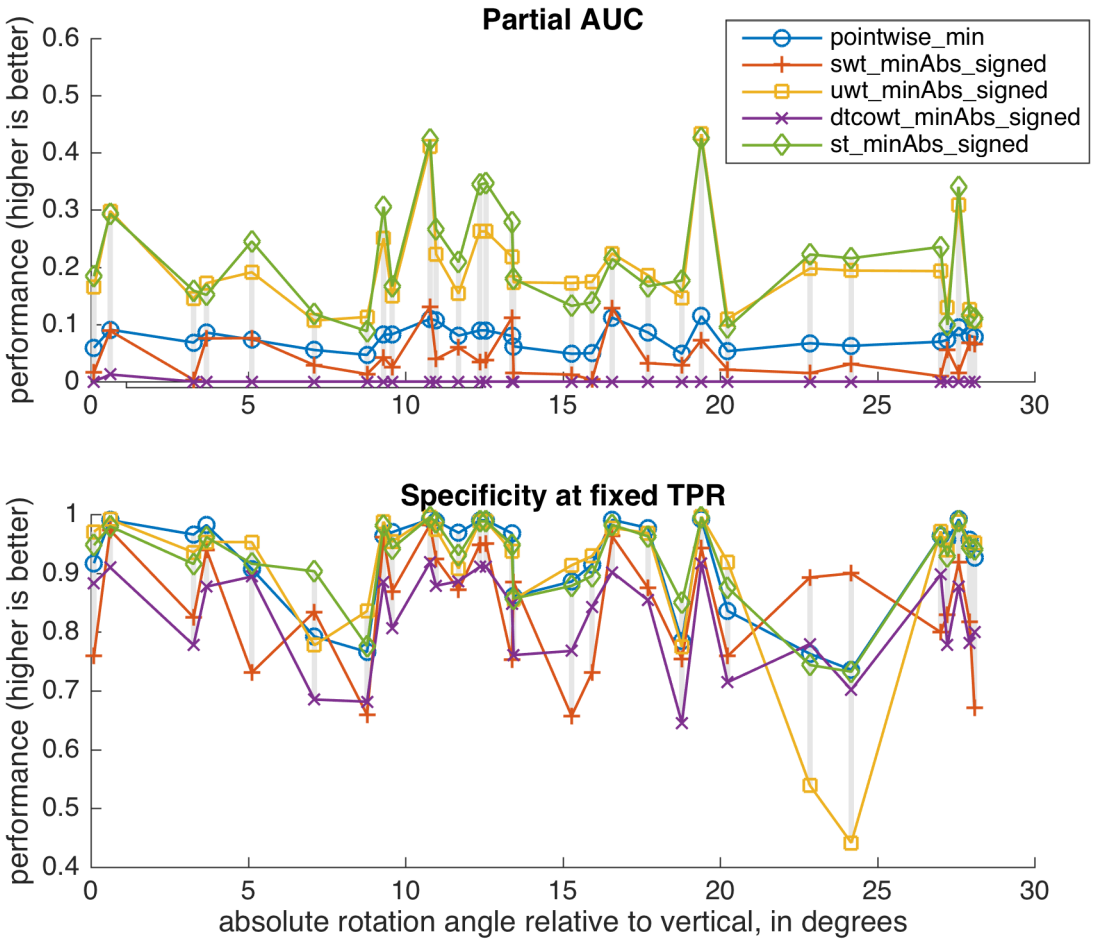


Figure 5.20: Detection performance of randomly rotated crack simulations at SNR=0.25, by absolute rotation angle.



5.2.9 Influence of registration errors

Registration errors are unavoidable in practice. In the best case, these errors can be minimized so that they have negligible influence on the fused result. But otherwise, signal-level fusion approaches are prone to misregistration because it cannot be guaranteed that measurements from the same location on the specimen will be fused. This problem is further aggravated by thin structures, such as cracks, which can be more easily missed than other structures that generate broader indications. For these reasons, the influence of registration error on detection performance should be quantified.

The simulation study is enhanced in the following way for this experiment. A set of crack indications is generated whose horizontal displacement from the original location is randomly varied, to introduce localization uncertainty. Specifically, $N = 30$ inspection results are randomly sampled, where the crack shape was chosen randomly with replacement among the seven natural cracks that have already been used in the previous experiments. In each simulation, the three generated crack indications (one per sensor), using the same crack shape, are shifted in the cross-defect direction by random amounts. These random offsets are drawn from a normal distribution with zero mean and a standard deviation of 0.1 mm. This choice reflects the suboptimal registration quality that is sometimes unavoidable when working with inspection techniques that in principle cannot be carried out in the same coordinate system, e.g. have to use different measurement positions on the specimen. For comparison, the typical cross-defect widths p_σ of single-sensor indications, after scale normalization, were set to 0.066 mm, for a 13.5 µm deep groove (see section 5.2.5). Therefore, the encountered localization uncertainty significantly shifts the indications, possibly preventing spatial overlap. This would impair successful fusion, which was designed assuming that actual defects are indicated by all sensors at the same location. In addition to these N random samples, a $N + 1$ sample is deterministically added that has no localization uncertainty, for comparison.

Randomizing the crack positions for each sensor requires making several changes to the evaluation strategy. Unlike before, on-crack pixels in one sensor might now be off-crack in another sensor.

Where structural noise indications lead to spurious inter-sensor agreement with a simulated crack response, the ground truth label is undefined. This is because "large" simulated crack displacements should be regarded as false alarms, whereas near-zero displacements should still be counted as true indications. Displacements between those two extremes are ambiguous with regard to ground truth. Therefore, in the following evaluation the background noise will be removed from a fixed region around the simulated crack indications. Although this approach produces unrealistic sensor images, it allows investigating the effect of registration errors on detection performance while avoiding to be misled by noise effects.

Another change as a consequence of random crack positions is that the ground truth now has to be adapted to each random sample, whereas a fixed ground truth had been used before where the location of each simulated crack had been known. Since the amount of fused images is too large for manual labeling, a semi-automatic approach is followed instead. The ground truth for each fused image is computed by finding an optimal path through the fused image. "Optimal" means that the sum of image intensities along the path is maximized. Several constraints are implemented to ensure that in fact the whole fused indication is covered by the path, and not just a short segment where fused intensity is the highest. First, the region in which the path is allowed to run is limited by the known crack positions of the individual sensors. It is

assumed that the fused indication does not lie "outside of" the region that is enclosed by these cracks. Formally, the two-dimensional non-convex hull of the set of crack pixels is formed, then extended by 5 pixels by morphological dilation. For computation of the optimal path, image intensities outside of this extended hull are set to $-\infty$. Moreover, since it is known that all simulated cracks are vertically oriented, optimal paths are computed across image rows. Also, because the simulated crack shapes are continuous and have limited curvature, the horizontal stride that the path is allowed in each image row is limited to 0.5 pixels. In practice, the path is allowed to step one image column to the left or right every two rows only, by sub-sampling the image rows (ignoring every second row). To efficiently compute an optimal path[24] under the given regularity constraints, the implementation is based on dynamic programming[25]. Once the ground truth has been adaptively computed in this way, ROC evaluation is carried out as described before, including a margin of ignored pixels around the ground truth crack locations.

To simplify the visualization of results, the three random horizontal crack offsets in each simulated sample are summarized by a single measure of dispersion. Specifically, the actual localization uncertainty in each sample is quantified by computing the *range* of offsets, i.e. the difference between maximum and minimum offset among the three sensors. Samples having a larger range of displacements are expected to show worse detection performance in general, because agreement among sensors is impaired more strongly.

Fusion rules Similarly to the previous experiments, in a first analysis, the effect of simulated registration error on different fusion rules is studied for the Shearlet transform. In all cases, scale normalization is applied before fusion and modifications M1–M3 are implemented as proposed in section 5.2.7. The results are presented in figure 5.21 on page 82, at high and low SNR of 2 and 0.5, respectively. As expected, with increasing localization uncertainty the detection performance becomes less stable and decreases, as indicated by both $pAUC_{0.01}$ and $Spec_{0.5}$. However, some simulated samples contradict this decreasing trend. The unusually high performance at the second largest simulated offset range (0.42 mm) is because the *range* conceals the fact that two of the three simulated indications lie very close in this sample. The same phenomenon occurs at the simulated range of 0.27 mm.

At high SNR, detection performances start to diversify around 0.1 mm of localization uncertainty, as shown by $pAUC_{0.01}$. The second performance measure, $Spec_{0.5}$, shows this effect at slightly larger offset range, from 0.15 to 0.2 mm. According to both measures, fused performance clearly outperforms the best single-sensor up to 0.25–0.3 mm. It is apparent that fusion performance suddenly drops for larger simulated offsets. Concerning the different fusion rules at high SNR, $pAUC_{0.01}$ does not indicate any clear difference. In contrast, specificity gives a more stable picture. However, no plausible explanation for the differences among the fusion rules can be given. For example, the two *median* variants are at opposite ends of the performance scale in the offset region around 0.2–0.25 mm, but the two minimum-variants are not as far apart. It can be concluded that at high SNRs, the choice of fusion rule for ST coefficients does not strongly influence the detection performance in the offset range up to 0.25–0.3 mm.

[24] Although there might be multiple optimal paths, only one of them is arbitrarily chosen in the current implementation.

[25] See for instance [125] for an article that explains dynamic programming to find optimal paths through images, although it implements a different optimality criterion

At higher offsets, both performance measures consistently indicate that *medAbs_signed* and *minAbs_signed* work best and still exceed the best single sensor, which is not the case for any other investigated rule.

Focusing now on low SNR, the individual fusion rules are more clearly distinguished. Consistently with previous results at zero misregistration, the minimum rule is demonstrated to be effective up to 0.25 mm offset range. While the two variants *Abs_signed* and *SameSign* of the *minimum* rule perform similarly up to that point, *minAbs_signed* seems to be advantageous at higher localization uncertainties. In contrast, the product rule shows poor performance, like medAbs_signed, both of which show near-zero $pAUC_{0.01}$ already around 0.1 mm or below, and significantly reduced specificities compared to other methods. Interestingly though, *medSameSign* clearly ranks much higher than *medAbs_signed*, together with the geometric mean. Apparently, the stricter *SameSign* criterion reduces false alarms at low SNR that are not suppressed by the generally mild *median* rule. In conclusion, the *minimum* rule performs comparably well across all SNRs and simulated registration errors and is the recommended fusion method for shearlet coefficients under the influence of localization uncertainty.

The same analysis, carried out for UWT decomposition instead of ST (figure A.11), shows that even at high SNR, some methods are clearly preferable than others when registration error is large. In particular, *minAbs_signed* and *medAbs_signed* look promising. But at reduced SNR, *medAbs_signed* is one of the worst strategies. Nevertheless, *minAbs_signed* is still among the best methods an in particular works much better than its counter-variant *minSameSign*, which is apparently too strict considering UWT's short filter lengths, which aggravate the problem of localization uncertainty. In conclusion, in the context of this simulation, *minAbs_signed* is the preferred fusion rule for UWT decomposition when registration errors are present.

For pointwise fusion, again the *median* seems to dominate especially at high registration errors when SNR is good. At lower registration errors, in contrast, stricter rules like *minimum* or *product* offer much better performance. The turning point regarding which rules are better lies around 0.1 mm offset range. At low SNR, clearly the best fusion rule is the *minimum* for all simulated samples. Therefore, this rule is the recommended first choice.

Based on these findings, the following analyses will focus only on low SNR with fixed fusion rules *minAbs_signed* for ST, *minAbs_signed* for UWT, and *minimum* for pointwise fusion.

Scale normalization and M3 The proposed fusion approach involves other parameters whose optimal settings might be influenced by localization uncertainty. Two of those parameters are whether scale normalization is applied (as proposed in section 5.2.3), and whether modification M3 (section 5.2.7) is used. Both of these decisions might depend on the registration error, because scale normalization narrows indication widths, which possibly impairs detection performance in case of strong misalignment of inspection images. Moreover, M3 aims at preventing false cross-sensor associations by setting coefficients to zero that correspond to locations in the original image where intensity is low. But when large registration errors are present, it is actually desirable to spread indications in space to compensate the localization uncertainty. Therefore, the effect of scale normalization and of M3 on defect detection performance should be evaluated.

Quantitative results are shown in figure 5.22 on page 83, for low SNR=0.5 with the *minAbs_signed* rule applied to shearlet coefficients. Both performance measures are

plotted in separate sub-figures. Each plotted line represents a combination of whether scale normalization is applied (*scaleNorm 1*) or not (*scaleNorm 0*), and whether M3 is used (*M3 1*) or not (*M3 0*). From the figure it is clear that under the stated conditions, both scale normalization and M3 should be applied to maximize the detection performance according to both evaluation measures, across nearly all simulated registration errors. Correspondingly, detection performance is consistently the worst when neither scale normalization nor M3 is activated. The plots further indicate that the influence of scale normalization is stronger than the influence of M3. Further investigations (not shown) indicate that these observations generalize also to higher SNRs, and also for UWT decomposition instead of ST. Moreover, considering undirectional fusion using the per-pixel *minimum* fusion rule, scale normalization improves performance for SNRs of 1 and below. This is explained by the fact that scale normalization creates narrower indications, which reduces the chance for false associations during fusion. Only at high SNR=2, scale normalization is not recommended, because there the effect of weakening the true defect indication is stronger than the effect of removing false alarms. Note that M3 has no effect on per-pixel fusion.

In conclusion, even though the fusion process of scale normalization in combination with modifications M1–M3 was originally proposed without taking into account registration errors, it still performs well when localization uncertainty is introduced.

Directional vs. undirectional fusion In the previous analyses, the optimal fusion settings in the face of localization uncertainty were identified for ST, UWT and undirectional per-pixel fusion. Now these decomposition methods should be quantitatively compared to see if directional fusion is beneficial over undirected fusion in the face of localization uncertainty.

The results are presented in figure 5.23 on page 83. $pAUC_{0.01}$ was omitted as an evaluation measure because it shows very similar results to $Spec_{0.5}$. Moreover, only the low-SNR regime is presented which is the most interesting setting. The shown results were observed to qualitatively match those at higher SNRs. The figure shows that all three depicted fusion approaches perform far better than the best single sensor even up to offset ranges of 0.3 mm. In this interval, the experimental results of directional fusion (ST, UWT) always exceed those of per-pixel fusion. Although it is difficult to make any preference among ST and UWT, the ST results seem to show a more stable influence of localization uncertainty, meaning that ST might be more invariant regarding different simulated crack shapes than UWT. However, the shown results do not allow making substantiated conclusions regarding this hypothesis.

To convey the differences between the fusion methods more directly, examples of fused images are shown in figure 5.24. The images depict the simulated sample that produces a horizontal crack offset range of 0.257 mm at low SNR of 0.5. In this sample, the three sensors' cracks happen to be spatially evenly distributed (not shown). As seen earlier, per-pixel fusion generates very sharp indication boundaries, whereas all transform-based results are smoothed during reconstruction. The effect of localization uncertainty is clearly seen in the image corresponding to *ST_minAbs_signed*, where two individual crack indications are produced. Strangely, although the third simulated crack is centered between the two visible indications, it does not manifest in the fused image. More detailed investigations showed that at fine scales of the shearlet transform, *minAbs_signed* produces negative fused coefficients which suppress the center indication although it exists at coarser scales. In contrast, *ST_minSameSign* generates a more homogeneous fused indication because at fine scales, coefficients whose signs do not

agree do not have the mentioned suppressive effect on the fused indication.

Discussion Overall, the results suggest that in the face of registration errors, per-pixel fusion is still a better defect detector than the evaluated single sensors, even if strict rules such as *minimum* are applied. This was not expected because per-pixel fusion does not operate in scale space where information from nearby pixels could be taken into account. To improve detection results, UWT or ST-based fusion are demonstrated to be feasible alternatives. However, more effort must be invested to achieve these results (modifications M1–M3, scale normalization, computational requirements of the transforms). Another finding of the experiments is that the directionally highly sensitive Shearlet Transform has no clear benefit over the directionally much less sensitive UWT, which confirms the previous analyses.

5.2.10 Discussion of directional fusion

The study shows that detection performance is influenced by many factors (SNR, crack path, rotation, displacement, decomposition method, fusion rule, application of M1–M3, shape normalization), which interact in complex ways. This complexity makes it difficult to give clear recommendations for future applications. Nevertheless, variants of the *minimum* rule consistently perform well in the experiments. Per-pixel fusion is a simple yet effective approach to reduce false indications, which is even possible under the influence of registration errors.

The evaluation further demonstrates that the theoretical benefit of directionally more sensitive representations do not necessarily yield practical guarantees for defect detection. Apparently, the quantification of inter-sensor agreement and conflict does not require a directionally sensitive and/or multi-scale data representation, although these advanced techniques are shown to work well when registration errors are introduced. One point that should be noted is that although cracks are always elongated objects, their indications not necessarily are. In particular, nondestructive inspection methods that lack of fine spatial resolution, like conventional ET probes, produce oriented image features only for cracks that are longer than the extent of sensor smoothing in the cross-crack direction. Deconvolution methods could help in this regard (see section 4.2).

Another possibility for the unmet expectations of Shearlet-based fusion methods could be that they are better suited for *image reconstruction* tasks. For instance, a different way of combining the sensor images in this sense would be to compute the fused image so that it satisfies two optimization goals. Firstly, the image should capture only indications that are confirmed by multiple sensors, to remove false alarms. As the conducted experiments suggest, this goal could be formulated in an undirectional representation. Yet, to include knowledge about the elongated nature of cracks, the desired image should also have a sparse Shearlet representation. Thus, the quantification of inter-sensor agreement would be made possible without the added complications of a multi-scale, multi-directional transform domain, while still solving for oriented fused indications. In the language of numerical optimization, the solution is characterized for instance by

$$\arg\min_{I} \|ST(I)\|_1 + \lambda D\left(I, \text{agreement}\left(\{I_1, \ldots, I_N\}\right)\right)$$

where I is the desired fused image, $ST(I)$ is the Shearlet representation of I, $\|\cdot\|_1$ is the L^1 norm, $D(\cdot)$ is some distance measure to be defined, I_j are the individual sensor images, and agreement$(\cdot)$ denotes a function that quantifies the inter-sensor agreement.

The scalar λ can be chosen to balance the relative importance of the two optimization goals. The L^1 norm constraint is known to favor sparse representations[126], and could be applied to the shearlet coefficients here. Such optimization formulations are highly flexible, and, owing to modern optimization methods, have proven successful in other image processing applications [127, 128]. This direction remains for further research.

One additional interesting approach, which is proposed in [129], could be worth pursuing. Because different image transforms are designed to represent specific image features "well", that is, relatively few coefficients suffice to achieve small approximation error, it makes sense to combine several of these transforms into an overcomplete dictionary. In this way, different parts of an image, e.g. smooth regions, structured regions, oriented features and curved features, could be represented in their respective optimal basis. The study in [129] proposes image fusion while following this approach, where sparse representations of the input images are computed and then fused. The challenge is to represent all input images using the same basis elements, so that fusion can be carried out. This problem is solved by *Simultaneous Orthogonal Matching Pursuit* in the article. In future work, it would be interesting to investigate whether the improvements of fused image quality that were reported in the cited study transfer to improved defect detectability in NDT.

Figure 5.21: Influence of registration error (horizontal axes) on detection performance, by ST fusion rule (curves). Each subplot shows the results for a combination of performance measure ($pAUC_{0.01}$ / specificity at fixed TPR) and SNR (0.5 and 2). In each plot, the vertical axis is scaled to the baseline performance given by the best single sensor (black dashed line).

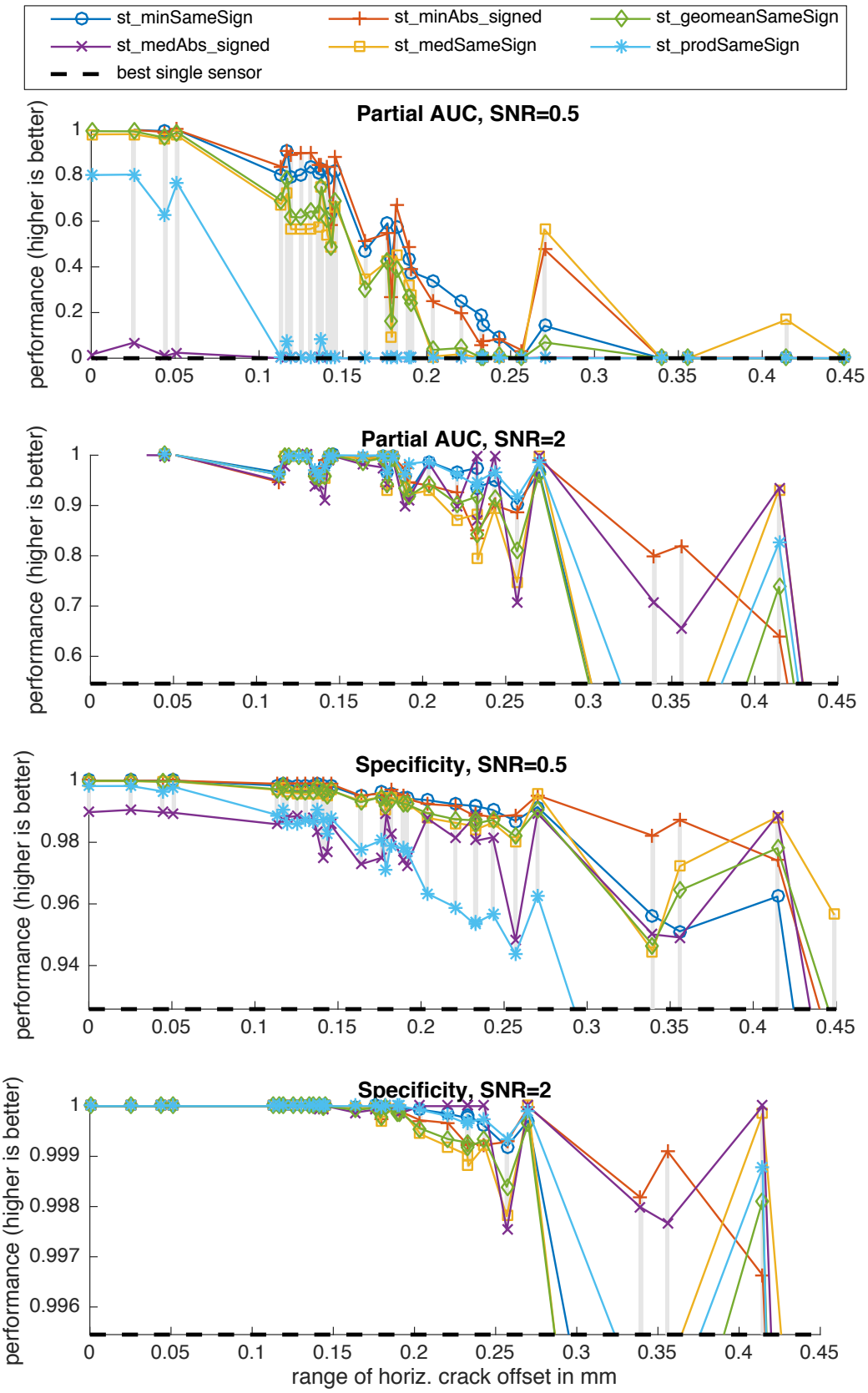


Figure 5.22: Influence of registration error on performance, by algorithm variation. Fusion rule is *st_minAbs_signed*, and simulated SNR=0.5. The two figures show results based on different evaluation measures. Note the different vertical axis scales. In the bottom sub-figure, the inset plot shows a zoom of the top left area.

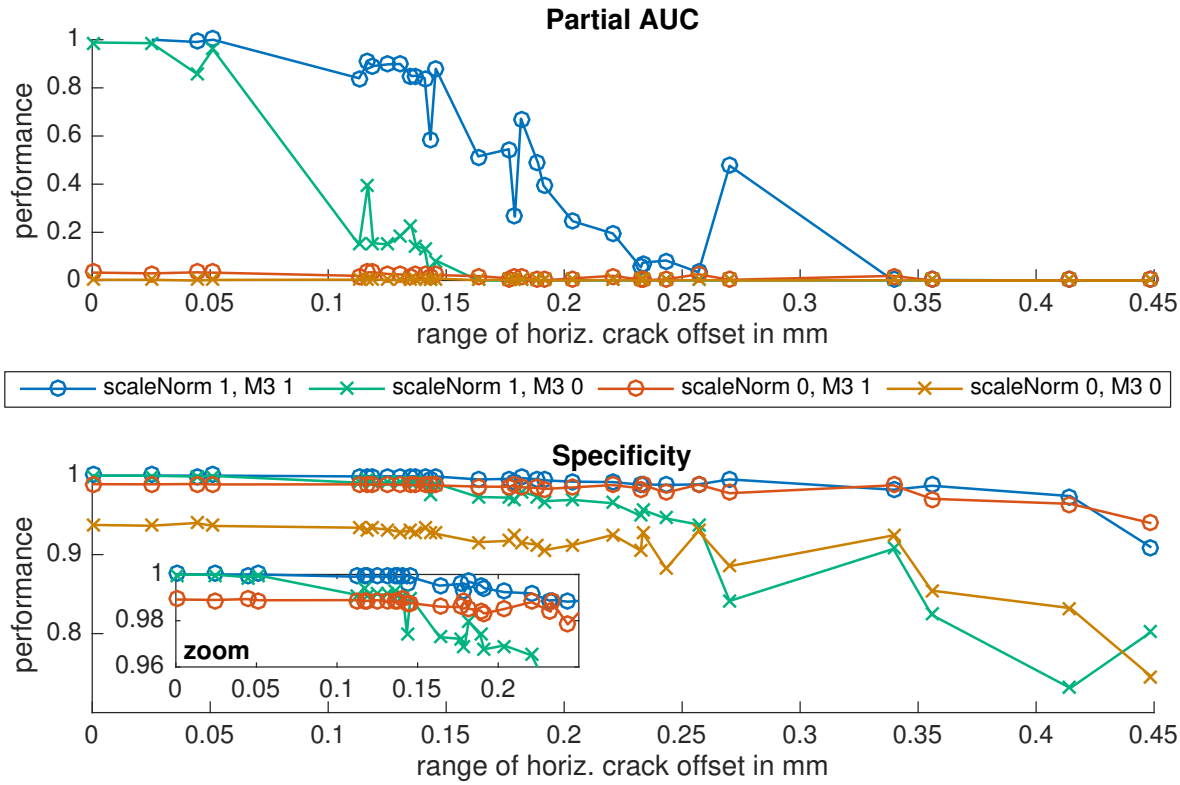


Figure 5.23: Influence of registration error on performance, by algorithm, at SNR=0.5. Scale normalization and M3 were applied. The vertical axis is scaled to the baseline performance given by the best single sensor (black dashed line).

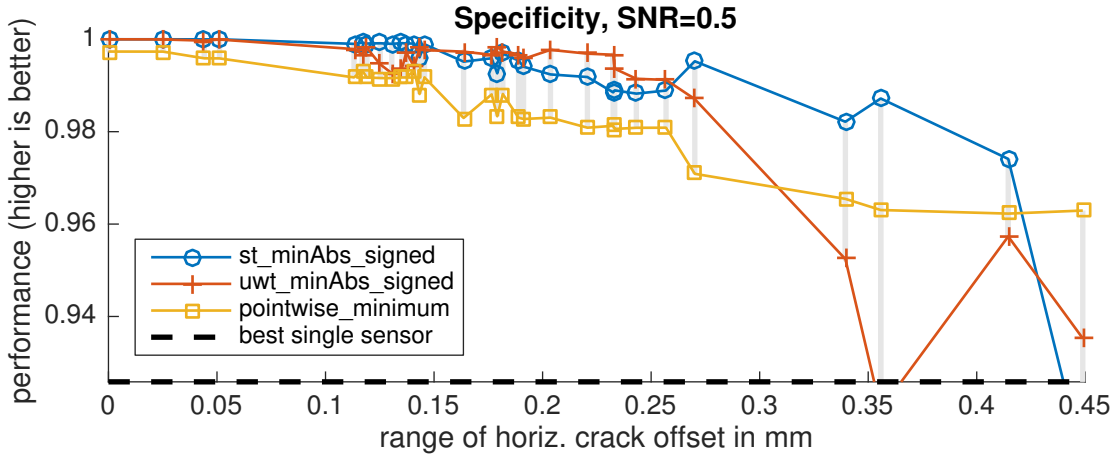
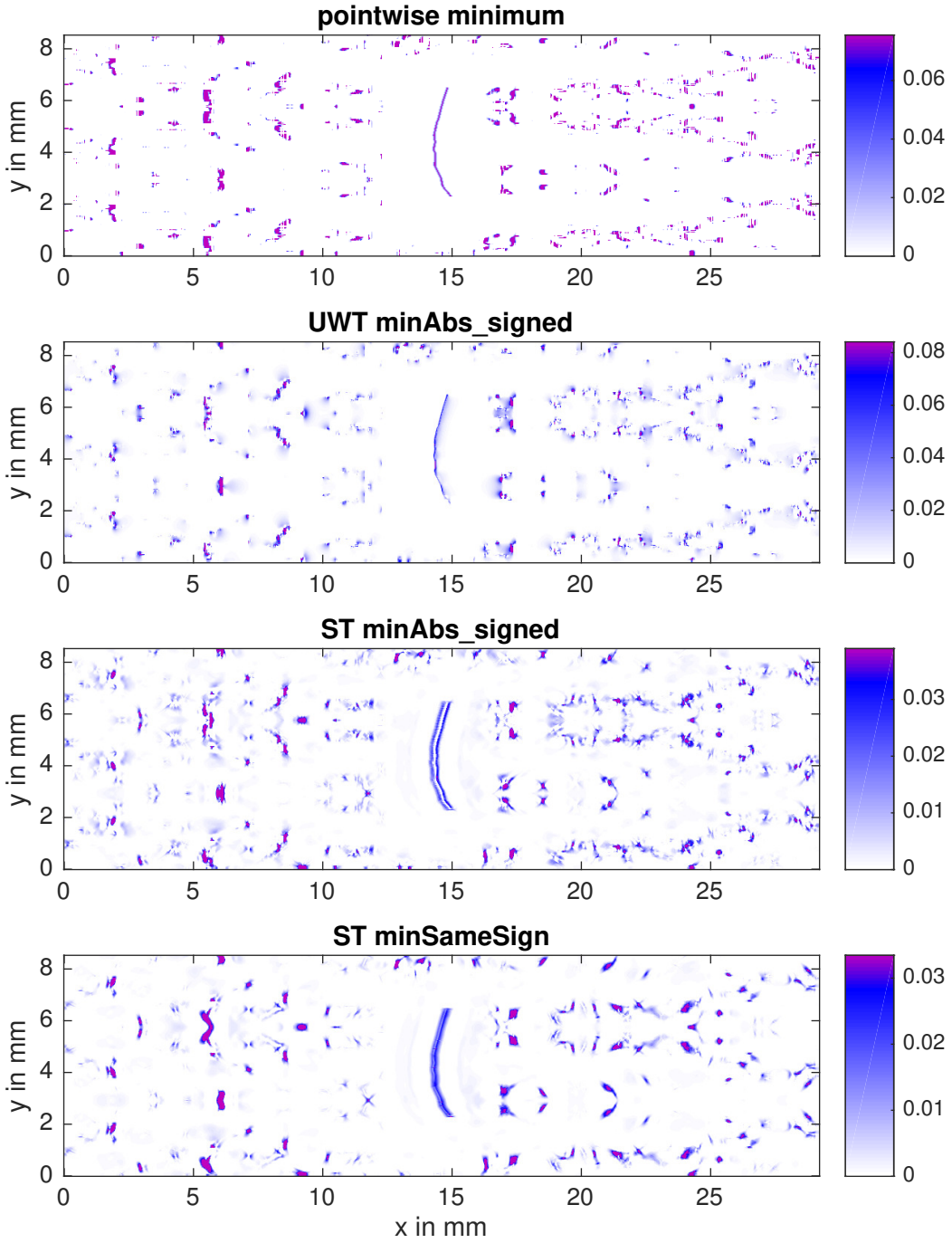


Figure 5.24: Exemplary fused images with simulated registration errors. One fusion rule is shown in each sub-plot. All plots show the same simulated crack sample with horizontal offset range of 0.257 mm. Simulated SNR is low (0.5). Both scale normalization and M3 were applied. Colors are scaled from zero to the maximum crack intensity in each image.



Chapter 6

Multi-sensor Defect Detection at the Decision Level

As discussed in section 2.3, sensor fusion can be performed at various levels of signal abstraction [48], each with specific drawbacks and advantages. In particular, decision-level fusion deals with the higher level aggregation of data after individual detection. That is, each signal is first processed individually, and then fed into the fusion algorithm.

Decision-level fusion has several advantages over signal-level fusion for NDT inspection. First, unlike signal-level fusion, the data to be fused do not have to be interpolated at a common grid of positions, because there is no need for per-pixel[1] fusion. Because the output of fusion at higher levels of signal abstraction is not an entire fused signal, but a set of fused hypotheses about defect locations, the fusion procedure is less constrained and more flexible than at the signal level. Especially when localization uncertainty is involved, e.g. due to registration errors, accurate registration is crucial [49–51] for per-pixel fusion. This is because misalignment can hardly be compensated during per-pixel fusion, regardless of the fusion level. Whereas for larger-scale objects to be detected, such effects are practically negligible given reasonable registration accuracy, strongly localized objects, such as cracks, are severely affected by misregistration when per-pixel fusion is applied. To circumvent problems introduced by per-pixel fusion, an alternative might be to first identify larger *segments* of interest in each signal, e.g. indications, and then to perform per-segment fusion. However, inter-sensor segment association is ambiguous[2] and thus might introduce additional unwanted variability. The method to be proposed in this chapter will solve both the issue of susceptibility to registration errors as well as segmentation ambiguity by fusing sets of spatially scattered locations instead of pixels or segments. As will be demonstrated, this approach allows for explicit handling of localization uncertainty.

A further major benefit of high-level fusion is its modularity. The individual data collection and processing can be carried out independently by the respective experts to tailor the detection process specifically to each inspection method. Consequently, much less effort has to be put into the normalization of the sensor data. One practical benefit of the modularity offered by decision-level fusion is that it allows combining individual results, even if fusion was not envisioned in the original testing plan. This also facilitates independence from the type of data source, making it possible to aggregate

[1] or per-coefficient, when a signal transform is involved

[2] In particular, segments will never match exactly. Their shapes and sizes will vary between sensors, and an indication might be segmented into multiple disjoint areas within one sensor, or even worse, across different sensors.

heterogeneous modalities ranging from manual inspection data to the output of fully automated scanning systems. Consequently, different sources of information can be effortlessly exchanged and the fusion strategy is readily adapted to unknown future changes of input sources.

Although there are recent studies in NDT proposing decision-level fusion, e.g. by weighted averaging [67], hypothesis testing [63], and Bayesian or Dempster-Shafer theory [60, 130], all of these works still rely on image registration and interpolation to perform fusion at common grid points. In fact, thorough literature research did not yield any fusion publication in the field of NDT dealing with scattered decision-level fusion, i.e. using the original measurement locations, despite the aforementioned advantages.

In the following, a new fusion strategy is proposed[3] that combines spatially scattered detection locations to improve the detection performance compared to single-source methods. Central to this method is that registration errors are explicitly accounted for. Although surface inspection will be the primarily addressed problem, the methodology is quite generic and can be easily extended to the three-dimensional case of volume inspection.

6.1 Methodology

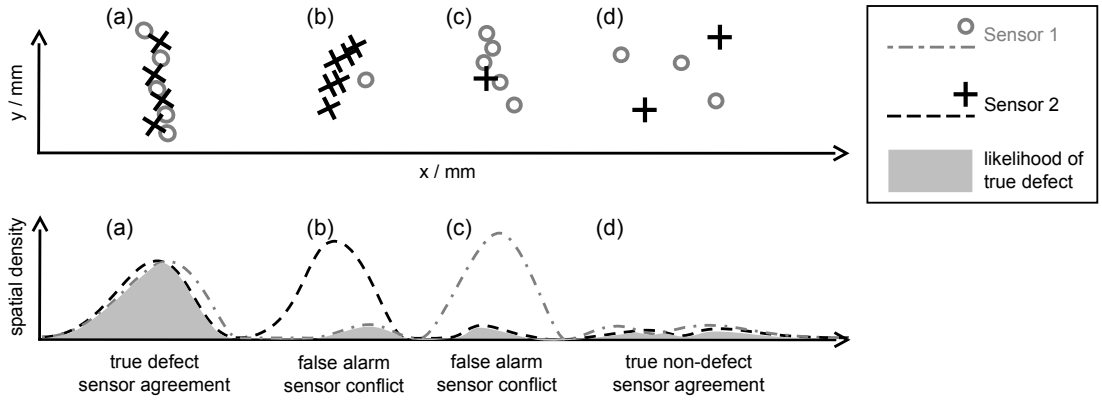
6.1.1 Principle

Before giving a formal definition, the core idea of the proposed approach is now schematically described. To this end, assume that information about potential defect locations $d = (d_x, d_y)$, called *hits* in the following, was obtained from different NDT data sets. For example, consider figure 6.1 for an outcome of individual surface inspection using two NDT methods. In cases (a)—(d), each dot marks a hit generated by some detection rule per sensor. Among the hits, there are also false alarms, for instance indicating changes in material properties unrelated to a defect (structural noise). Such false alarms are illustrated by cases (b) and (c) in figure 6.1. Using single-sensor inspection, these false alarms cannot be distinguished from indications produced by actual defects such as those shown in case (a). A multi-sensor data set, on the other hand, is able to reveal case (a) as a real defect by assessing the agreement among different detection methods. Here, agreement is expressed in terms of joint spatial density of hits, taking into account all sensors. The underlying rationale is that the joint hit density is higher over real defects than in other areas, provided sufficient SNR for at least two sensors. On the contrary, a clear conflict occurs where only one sensor generates hits, and thus the joint density is not significantly increased relative to the individual sensor density. This concept is depicted at the bottom of figure 6.1, where for each sensor the spatial density of hits across the specimen surface is symbolized. Only in case (a) both sensors agree in increased spatial density, whereas in cases (b) and (c) the sensors do not agree. Although there is also agreement in case (d) as well, the joint density is not significant enough, indicating the low likelihood of defect presence. This example demonstrates the potential of the joint spatial density as the basic mechanism for multi-sensor detection.

This study proposes evaluating the local hit density as a measure for multi-sensor data fusion at decision level. Figure 6.2 provides a flow chart of the individual steps. The first step consists of generating hypothesized defect positions from individual NDT

[3]This section is based on a journal article ([131]) that was published by René Heideklang and Parisa Shokouhi.

Figure 6.1: Schematic representation of the principle of the approach. The detection outcomes of two different NDT methods are represented by circles and crosses, for four cases (a)–(d). For each case, the corresponding spatial density per sensor along the x-axis is plotted below. The likelihood of observing a true defect (gray area) depends on both sensors yielding significant hit densities.



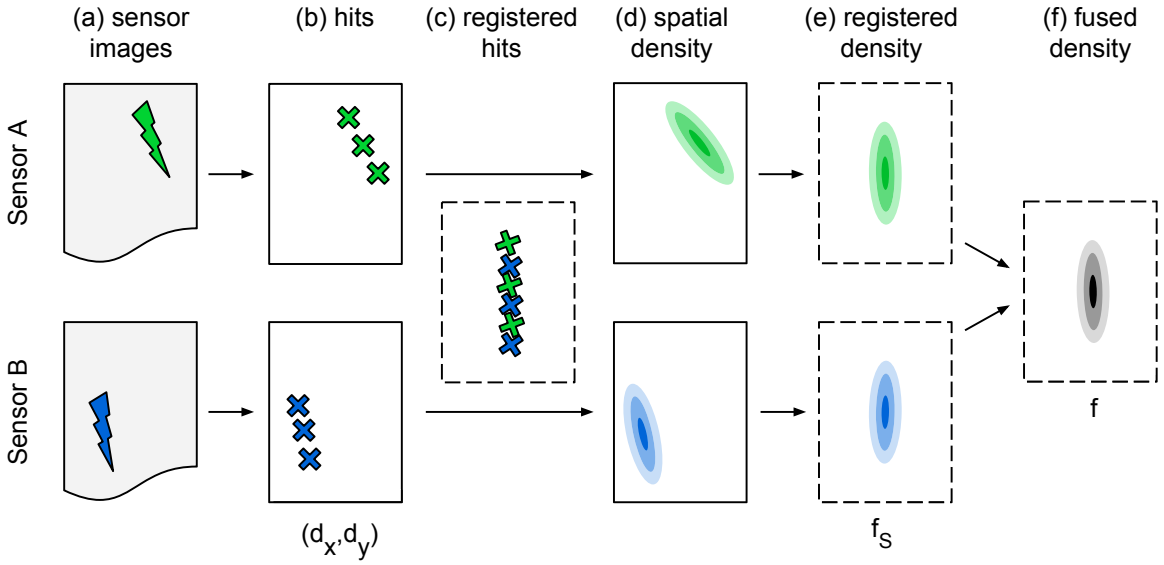
images, for example as discussed in chapter 4. Additionally, each hit is associated with its local signal to noise ratio, which will be used later as a weighting factor (not shown in the flow chart). It is important to make the detection strategies focus on sensitivity rather than specificity, thus ensuring that all (unknown) real defect indications are retained for the final detection by fusion. At the same time, we would still want to discard as many false alarms as possible. The difficulty of facing this trade-off, which is typical for single-sensor detection, is however less critical for multi-sensor detection because the main work is done by the subsequent fusion procedure.

The extracted hit locations from different NDT techniques have to be mapped to a common coordinate system for fusion. Note that, in contrast to fusion at the signal level, this spatial alignment does not entail interpolating the sensor data values. A central property of the hit locations is that although each NDT method usually uses gridded measurement positions, the mapped hit locations after registration are *not* jointly gridded in the common coordinate system, but appear scattered instead. This is visualized in figure 6.2c.

After establishing relationships between the individual coordinate systems, the next question is how to implement the density-based fusion concept already introduced before. One central challenge in the decision-level fusion of non-gridded locations is the uncertainty in localization. Two main factors contribute to this uncertainty. First, each sensor's localization ability is limited by the physical resolution as well as the spatial sampling rate. Second, the coordinate transformations computed during spatial registration might be inaccurate to some degree. To be robust, a fusion approach must adequately cope with the inherent uncertainties about hit positions and must associate nearby hits for the purpose of density quantification. This loose concept of *proximity* should therefore be mathematically formalized.

To that end, various non-parametric techniques have been developed such as Mean shift [132], DBSCAN [133], OPTICS [134], Spectral clustering [135], and Kernel Density Estimation (KDE) [136, 137]. In this study, the framework of kernel density estimation is selected. This choice was motivated by considering that our data space typically has only two or three spatial dimensions, independent from the number of sensors. Therefore, the density can be directly modeled without being affected by the curse of

Figure 6.2: Flow chart of the fusion process. Gray boxes denote original sensor images, each containing an indication (crack symbol). Solid-edge boxes denote local (per sensor) coordinate systems; dashed boxes denote the global (registered) coordinate system. Cross markers denote hits. Two-dimensional spatial densities are indicated by contour plot symbols. Mathematical symbols below the boxes correspond to the notation used throughout this work. (a) Sensor images; (b) Hits; (c) Registered hits; (d) Spatial density; (e) Registered density; (f) Fused density.



dimensionality, which is otherwise known to impair KDE [138, sec. 4.5]. Furthermore, it allows us to evaluate the density at arbitrary positions, not only at the hits.

Returning to the flow chart, KDE is applied to estimate the spatial hit densities from each individual sensor in figure 6.2d–e to associate nearby hits. Finally, figure 6.2f consists of a fusion rule that combines the individual densities and produces a fused image in which higher intensity corresponds to increased estimated likelihood of defect presence.

Next, the proposed method will be formally introduced using ideas from KDE, and different fusion rules are introduced that implement the behavior of the gray shaded area in figure 6.1 to recognize conflicts and agreement between the sensors.

6.1.2 Kernel density estimation (KDE)

Before proposing the developed technique, fundamental concepts of KDE are repeated in this section. These concepts and the associated notation are adopted in the rest of the text.

KDE is a nonparametric statistical method to estimate a probability density function $\hat{f}$ from a set of samples x_i. The result is a continuous function, computed from a weighted sum of kernel functions K_h with an associated bandwidth h, each centered over one of the samples:

$$\hat{f}(y) = \left(\sum_i w_i \right)^{-1} \sum_i w_i K_h(y - x_i)$$

with $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$. Some functions qualifying as a kernel K are the uniform,

triangle, Gaussian or Epanechnikov kernel functions. The bandwidth h controls the size of the neighborhood in which samples influence the density at a specific location. The choice of the bandwidth is critical for the overall performance of the algorithm. If the bandwidth is chosen too wide, KDE results in an overly smoothed density, thus losing important details of the distribution. On the other hand, if it is chosen too narrow, the estimate adapts too much to the specific realization of the sample set, thus missing the global features of the density. This problem has been well-studied, and several solutions have been proposed [139]. This study will describe how to automatically compute a suitable bandwidth for our problem in section 6.1.3.

The general formulation given above for KDE includes the normalization constant $\left(\sum_i w_i\right)^{-1}$, ensuring that the density integrates to one. Since a probabilistic interpretation of the density is not required here, we proceed with a simpler unnormalized version of KDE by dropping the normalization constant. This also simplifies the notation. Furthermore, since the density estimate is a weighted sum with one term per data point, the data set can be partitioned to aggregate the total density function $\hat{f}$ from the sub-densities $\hat{f}_j$, each including only the samples x_i from partition j:

$$\hat{f}(y) = \sum_j \hat{f}_j(y) = \sum_j \sum_{i \in P_j} w_i K_h(y - x_i) \tag{6.1}$$

P_j denotes a subset of all points such that partitions do not overlap and the union of all partitions covers the complete data set. This re-arrangement is taken up in the following section to group hits by each sensor, as is illustrated in figure 6.2d.

KDE can be extended to vector-valued samples. To that end, let $\boldsymbol{x}_i$ denote the ith vector-valued (bold face) sample, and let $\boldsymbol{y}$ denote the vector-valued location where to evaluate the density. Multivariate KDE is computed from multivariate kernel functions and an associated bandwidth matrix H, which describes the scale and the orientation of the kernels. A special kind of multivariate kernel is the *product kernel*, defined by

$$K_H(\boldsymbol{x}) = \prod_j \frac{1}{h_j} K\left(\frac{\boldsymbol{x}_j}{h_j}\right),$$

where one univariate kernel for each dimension j is evaluated. The h_j are the entries of the diagonal bandwidth matrix, that is, product kernels are not arbitrarily oriented in the data space. This property reduces the computational demand, because the data dimensions are considered independently. This study uses product kernels, as is described next.

6.1.3 Scattered decision-level fusion

Overview

In this section, a new fusion method for NDT is developed based on concepts from KDE. Here, the role of the vector-valued data sample $\boldsymbol{x}_i$ is taken by the two-dimensional hit location $\boldsymbol{d}$ as detected by a single sensor S_i during surface inspection. As defined in equation 6.1, the *joint density* that includes the hits from all sensors can be computed from *partial densities* that include only the hits from a single sensor. This property can be extended to a more general framework of density-based fusion. Specifically, two modifications are now introduced: (1) Each partial density is computed in the respective local coordinate system, using sensor-specific KDE parameters; and (2) To merge the partial densities into the joint density, the outer sum from equation 6.1 is generalized to an arbitrary fusion rule F. The approach is outlined by the following steps:

1. Define a grid of discrete locations $\boldsymbol{p}$ where the fused density should be evaluated. These locations are defined in the common coordinate system after registration, so that they refer to the same location on the specimen for all individual inspection methods. Map these locations to each local coordinate system using the coordinate transforms T_S obtained during spatial registration. See figure 6.3 for an illustration of this step.

2. For each individual sensor, compute the spatial density $\hat{f}_S(\boldsymbol{p})$ of single-sensor hits $\boldsymbol{d}$ and evaluate the density at the mapped locations $T_S(\boldsymbol{p})$:

$$\hat{f}_S(\boldsymbol{p}) = N\left(\boldsymbol{h}_S, \Delta^S\right) \sum_{\boldsymbol{d} \in D(S)} w_d K_{\boldsymbol{h}_S}\left(T_S(\boldsymbol{p}) - \boldsymbol{d}\right) \tag{6.2}$$

   The normalization constant $N\left(\boldsymbol{h}_S, \Delta^S\right)$ and the bandwidth $\boldsymbol{h}_S$ are sensor-specific parameters. N depends on the kernel bandwidth and on the sampling distances $\Delta^S = \left(\Delta_x^S, \Delta_y^S\right)$ (pixel dimensions). w_d is a per-hit weighting factor.

3. For each evaluation point $\boldsymbol{p}$, fuse the partial density values $\hat{f}_S(\boldsymbol{p})$ from the different sensors using a fusion rule F:

$$\hat{f}(\boldsymbol{p}) = F\left(\left\{\hat{f}_{S_1}(\boldsymbol{p}), \ldots, \hat{f}_{S_N}(\boldsymbol{p})\right\}\right) \tag{6.3}$$

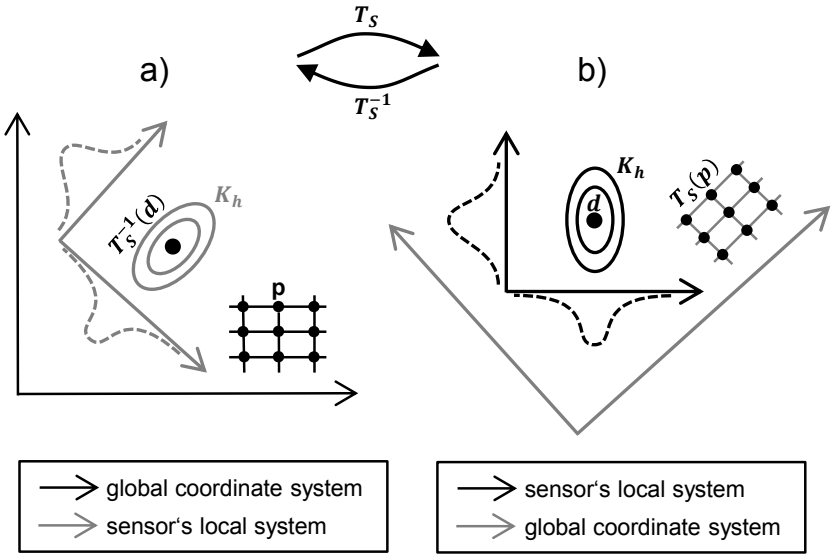
These steps are now explained in detail. The grid defined in Step 1 determines the resolution at which the fused density will be sampled. The grid size depends on the kernel bandwidths, because smoother densities computed from larger kernels facilitate coarser sampling grids to reduce the computational complexity. For optimal resolution however, the grid size should be set according to the sampling distance of the finest-sampled individual sensor. In the second step, the partial densities are computed as explained next.

Estimation of Partial Densities

As noted, the first modification to standard KDE is to carry out density estimation in a per-sensor manner. To this end, the computations in Step 2 are defined in the respective local system for each sensor. Consequently, the kernel function K_S can be defined as an axis-aligned product kernel to reduce the computational cost. The corresponding bandwidth parameters $\boldsymbol{h}_S$ are chosen based on background knowledge about the nature of our data. Intuitively, the density estimator should always be able to "connect" neighboring hit locations. For NDT data, the smallest possible distance between any two hits of the same sensor is given by the known spatial sampling intervals. For a measurement grid per sensor S, the two spatial sampling distances are denoted by Δ_x^S, Δ_y^S in the sensor's coordinate system. To ensure that neighboring line scans crossing the same defect do not form disconnected density peaks, the kernel functions should at least stretch across one pixel in the sensor image. However, to avoid merging two unrelated indications, the kernels should not be made much larger. Therefore, the following minimum bandwidth parameters for product kernels are proposed for each sensor: $\boldsymbol{h}_S = (h_x, h_y) = \left(\Delta_x^S, \Delta_y^S\right)$. It is natural to use product kernels for gridded individual measurements, because the bandwidths directly correspond to the physical pixel dimensions.

The aforementioned kernel size is a minimal setting. In practice, the most significant factor contributing to the localization uncertainty may not be the spatial sampling, but

Figure 6.3: Coordinate transformation during the computation of the fused density. The two black coordinate systems in (a) and (b) are related through the coordinate transform T_S.
(a) Coordinate system in which the fused density will be evaluated at gridded points $\boldsymbol{p}$. The coordinate systems of the individual sensors, where the hits $\boldsymbol{d}$ are defined, are not axis-aligned with the global system. In particular, the kernels K_h would require non-diagonal bandwidth matrices.
(b) Coordinate system of one of the sensors, given by its measurement grid. For single-sensor density estimation, kernels K_h are axis-aligned to the sensor's system, thus facilitating product kernels. The single-sensor density is then evaluated at the transformed points $T_S(\boldsymbol{p})$.



inevitable registration errors. The kernel sizes should be set large enough to smooth out these unwanted variations and to associate poorly registered indications. As a general approach, the following kernel size is proposed:

$$\boldsymbol{h}_S = (h_x, h_y) = \frac{\hat{u}}{\min\left\{\Delta_x^S, \Delta_y^S\right\}} \left(\Delta_x^S, \Delta_y^S\right) \tag{6.4}$$

where $\hat{u}$ denotes an estimate of the localization uncertainty, for instance the mean registration error. This formulation keeps the kernel size ratio $h_x/h_y = \Delta_x^S/\Delta_y^S$ fixed, and scales the two-dimensional kernel size proportionally to $\hat{u}$. Consequently, in the fine-sampled direction corresponding to $\min\left\{\Delta_x^S, \Delta_y^S\right\}$, the kernel will be exactly $\hat{u}$ wide. In the other direction, the kernel is larger to maintain the ratio. Note that with increasing kernel size, the advantage of having spatially accurate sensors may be lost. Also, closely situated defects become harder to separate. Therefore, fusion performance benefits from high-quality registration by facilitating narrow kernels.

Setting the kernel size according to the localization uncertainty implies that sensors with fine spatial sampling produce more hits in the area of influence of a kernel than sensors with coarse sampling. To prevent finely-sampled sensors from having more influence on the fusion process by contributing larger densities, normalization is required. To this end, the normalization factor from equation 6.2 is to be defined as:

$$N\left(\boldsymbol{h}_S, \Delta^S\right) = 1/\max\left\{\frac{h_x}{\Delta_x^S}, \frac{h_y}{\Delta_y^S}\right\} \tag{6.5}$$

This essentially normalizes with regard to the number of pixels per kernel size, which implicitly relates to the potential number of hits in each dimension. Note that if the kernel size $\boldsymbol{h}_S$ is defined according to equation 6.4, then we have $N\left(\boldsymbol{h}_S, \Delta^S\right) = 1/\max\left\{\frac{h_x}{\Delta_x^S}, \frac{h_y}{\Delta_y^S}\right\} = \frac{\Delta_x^S}{h_x} = \frac{\Delta_y^S}{h_y} = \min\left\{\Delta_x^S, \Delta_y^S\right\}/\hat{u}$. This per-sensor normalization factor N replaces the conventional kernel normalization factor for product kernels $N(\boldsymbol{h}_S) = 1/(h_x \cdot h_y)$.

Although the previous considerations are valid for all types of kernel functions, a compactly supported kernel function like the *Epanechnikov* product kernel [138, sec. 4.2.1] is suggested, as defined in equation 6.6. Its compact support has the advantage of limiting the spatial influence area of each hit, which is expressed by the kernel bandwidth parameters $\boldsymbol{h} = (h_x, h_y)$, and thus facilitates faster computation than e.g. the non-vanishing Gaussian kernel:

$$K_{\boldsymbol{h}}(\boldsymbol{v}) = \begin{cases} \left(1 - \left(\frac{v_x}{h_x}\right)^2\right)\left(1 - \left(\frac{v_y}{h_y}\right)^2\right) & \text{if } |v_x| \leq h_x \text{ and } |v_y| \leq h_y \\ 0 & \text{else} \end{cases} \tag{6.6}$$

To further adjust the quantification of density, the kernel functions are scaled according to the weight w_d per hit; see equation 6.2. These weights control the influence of each hit on the final KDE. Each weight should be set proportional to the hit's signal to noise ratio, so that clear indications have more impact on the final density than insignificant ones. Also, the weights offer additional flexibility to regulate the fusion result with regards to specific sensors or different inspection areas.

Fusion of Partial Densities

In equation 6.3 a fusion rule F is introduced that combines the partial densities. The most basic fusion rule is to sum up the individual densities, which in effect computes the total kernel density according to equation 6.1. However, this approach has a major drawback concerning false alarms. High-intensity single sensor contributions have a large effect on sums, even when such indications are not backed up by other sensors. As an extreme example, the *maximum* function is most prone to false alarms. Therefore, more conservative rules are required to capture the agreement among sensors for effective reduction of false alarms. The following fusion rule is conceptually similar to the *sum* rule as used in conventional density estimation, but unlike the sum it guards against single-sensor false alarms:

$$F_{\text{sumIgnoreMax}}\left(\left\{\hat{f}_{S_1}(\boldsymbol{p}), \ldots, \hat{f}_{S_N}(\boldsymbol{p})\right\}\right) = \left(\sum_{S \in \{S_1, \ldots, S_N\}} \hat{f}_S(\boldsymbol{p})\right) - \max_{S \in \{S_1, \ldots, S_N\}} \hat{f}_S(\boldsymbol{p}) \tag{6.7}$$

Note that the *maximum* is evaluated separately for each evaluation point $\boldsymbol{p}$. Equation 6.7 realizes the quantification of agreement among sensors, because a large fused score now requires at least two sensors to produce high individual densities. Thus, $\hat{f}$ is expected to behave similarly to the function indicated by the shaded area in figure 6.1. Note that in the case of only two available sensors for fusion, subtracting the maximum is equal to the *minimum* fusion rule, which is a fuzzy *AND*-operator, and is in fact the operation used to generate the shaded area in the figure. However, as more than two sensors are involved, requiring that all sensors indicate a defect might

be too strict for some applications. Therefore, ignoring the maximum contribution can be viewed as a much milder version of the *AND* fusion rule. Other fusion rules that conceptually differ more from conventional density estimation, but also suit the quantification of agreement, are the *median*, *harmonic mean*, *geometric mean* and the *product*. Theoretical justifications for such basic fusion rules are given by [140], where they are used to combine multiple classifiers. Ensemble learning is similar to multi-sensor NDT, because the cited study assumes that the base classifiers are trained on independent features to predict the same target classes, just like the NDT measurements are carried out independently to indicate the same type of defects. Two differences, however, are that [140] aims at reducing general misclassification, whereas the fundamental assumption of this thesis ignores missed defects and only focuses on reducing false alarms while retaining sensitivity. Moreover, [140] assumes that the sources to be fused are probabilities, whereas this thesis deals with scores that are not necessarily normalized to the unit interval. The study concludes that the *sum* rule is "the most resilient to estimation errors", and provides a plausible theoretical explanation for this observation. Such estimation errors are not accounted for in the basic statistical framework[4], which is why a more sophisticated rule is presented in [43]. The study develops a Bayesian fusion approach that explicitly models inconsistencies among sensors, such as false alarms. However, this thesis focuses on simple algebraic fusion rules to avoid introducing additional parameters by more complex methods.

In total, the proposed fusion approach includes three mechanisms to ensure robustness against false alarms: quantification of density, decision weighting according to significance, and a fusion rule that expresses the agreement among individual sensors.

6.2 Application to experimental data

To demonstrate the fusion technique's performance under realistic conditions, a test specimen containing 15 surface flaws was inspected using three different NDT methods. This section describes the specimen, the individual data collection and processing as well as the application of the fusion algorithm. Finally, the effect of various conditions on the fusion result are quantitatively evaluated, and fused detection is compared against single-sensor detection. To further corroborate the effectiveness of the fusion approach, experiments are replicated using a second specimen at the end of this section.

6.2.1 Specimen

The primary test specimen *Ring SA* is a ring-shaped bearing shell [27, pp. 173–175] made of surface-hardened steel. As illustrated in figure 6.4, it has an outer diameter of 215 mm and is 73 mm long in its axial direction. To study micro-sized elongated faults similarly to cracks, 15 grooves were inserted into the specimen by electrical discharge machining. The reason for choosing grooves over real cracks is that their dimensions can be experimentally controlled. The flaws are regularly spaced across the surface of the specimen and vary in depth from 11 to 354 µm, as detailed in table 6.1. Grooves have constant lengths of 1 mm and their openings vary between 25 µm and 51 µm. The specimen's surface is uncoated and its roughness is very low, thus enabling high-quality near-contact measurement. A secondary specimen *SB* with the same material and

[4] In particular, only the uncertainty about the true class is modeled, but estimation errors render these probabilities themselves uncertain. The authors of [140] deal with second-order uncertainty by introducing an additional error variable.

dimensions, but different groove depths, exists for further analysis; see also table A.1. The first part of this study will focus only on *SA*, whereas *SB* will be investigated thereafter.

Figure 6.4: Schematic view of the ring specimen, not to scale. Top: outer proportions of the ring. Bottom: unrolled outer surface. Short vertical lines indicate the positions of the 15 grooves.

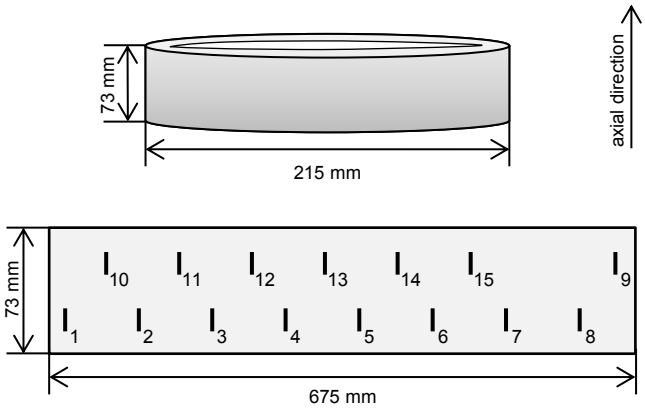


Table 6.1: Groove depths. Labels correspond to those shown in figure 6.4.

Groove	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Depth / µm	354	224	170	105	82	61	57	53	43	40	39	27	29	20	11

6.2.2 Individual measurements and processing

For nondestructive crack detection, inspections were carried out using ET, MFL with GMR sensors and laser-induced TT. The following three inspections were performed sequentially during the course of about one year.

ET was carried out at an excitation frequency of 500 kHz, which is well-suited to inspect surface defects due to the skin effect [13]. An automated scanning device rotates the specimen under the fixed probe. Signal processing is based only on the imaginary part of the measured impedance. The obtained one-dimensional signals are preprocessed by high-pass filtering for trend correction, and by low-pass filtering to improve SNR. An image is formed by stacking the line scans in axial direction of the ring.

The MFL data were collected using the same scanner as for ET, and a GMR sensor array developed at BAM [141]. Using these gradiometers, the normal component of the magnetic stray field was measured while the specimen was locally magnetized. Preprocessing comprised trend correction by high-pass filtering per line scan, and an adaptive 2D wiener filter (See MATLAB's function wiener2 [142] for noise suppression. The image was then Sobel-filtered to highlight the steep flanks that are generated by the gradiometers near the grooves.

Thermography testing was performed by rotating the specimen under a 40 W powered laser while recording with an infrared camera. The movie frames were then composed to form an image of the specimen surface. This image is processed by 2D background

subtraction using median filtering, and noise was suppressed by an adaptive 2D wiener filter.

It must be noted that the presented signal acquisition is tailored to the known groove orientation. In a realistic setting, a second scan should be performed for ET and MFL testing to detect any circumferentially oriented defects as well. It should further be emphasized that for the ring specimen used in this study, the GMR sensors yield far superior results compared to ET and TT, and would suffice by themselves for surface crack detection. Specifically, the MFL data facilitate zero false alarm rate even for the second shallowest of only 20 µm depth. However, such performance is not guaranteed for other materials or in case of suboptimal surface conditions, so that a multi-method approach is still in demand. A practical solution had to be found that allows including the MFL measurements so that a full three-sensor data set can be studied, and that at the same time creates a situation in which single-sensor inspection is not optimal. This dilemma was solved by *intensionally lowering the quality* of the MFL image before preprocessing and detection. This was done by separating the true defect indications from the background signal variations using the shift-invariant wavelet transform SWT [112], and by reconstructing the signal using a much smaller factor for the noise-free component than for the noise component. Specifically, the original MFL image I was treated for each image row separately, knowing that image rows cross through the vertically oriented grooves. Therefore, groove indications manifest themselves as peaks in the image rows' signals. Each row was transformed into wavelet coefficients, and those details coefficients whose absolute values were below an adaptive threshold relative to a manually defined noise region were set to zero before transforming the coefficients back into the image domain. Arranging all rows back into an image yields I_{sig}, the de-noised image that only contains the smooth background and the groove indications. The noise-only image is generated by $I_{noise} = I - I_{sig}$, and the synthesized lower-quality image is given by $I_{noise} + 0.02 I_{sig}$.

Although this process does not simulate a lower-quality MFL measurement in a physically realistic way, it is nevertheless useful to demonstrate the capabilities of the proposed fusion technique in settings where individual inspection is in fact not reliable enough. Therefore, for the rest of the experimental section, only this modified version of the MFL data is considered.

To convey an impression of the signals, an exemplary section of each preprocessed inspection image is shown in figure 6.5. The displayed part of the specimen surface is a 10 mm by 6.5 mm region around groove nr. 13 which is quite shallow, and thus generates relatively weak indications. The figure demonstrates the different signal patterns among sensors, concerning both the groove and the background variations. Also, the different pixel sizes are evident. A related plot is shown in figure 6.6, where one-dimensional line scans crossing the groove reveal more clearly the individual sensor responses. The different spatial sampling positions are demonstrated by the line markers. Table 6.2 offers a quantitative comparison of the individual data sets.

After individual preprocessing, the same detection routine was used for all three images to extract hit locations and confidences, which will later be fused at the decision level. To this end, the signal intensities are converted into confidence values, which are subjected to a threshold to extract only significant indications, as follows. Confidence values are computed by estimating the distribution function of background signal intensities $\mathcal{N}$ from a defect-free area. This estimate serves as the null distribution $P_{\text{noise}}(\mathcal{N})$ in the significance test. For each image pixel's intensity $I(x, y)$, the probability $P_{\text{noise}}(\mathcal{N} \leq I(x, y))$ is computed as proposed in section 5.1.1. But whereas in signal-level

Table 6.2: Quantitative properties of the individual data sets.

	ET	MFL	TT
Δ_x in mm	0.029	0.029	0.469
Δ_y in mm	0.200	0.200	0.126
Width of a typical indication, in mm	$2 \approx 69\Delta_x$	$0.6 \approx 20\Delta_x$	$0.5 \approx \Delta_x$
Avg. nr. of hits per pixel	0.0023	0.0031	0.0068

fusion this probability was designed to be fused with other sensors, here it contributes to a single-sensor detection rule. Pixels are considered significant here, if their confidence exceeds 99 %.[5] Additionally, only those hits that are local maxima with regard to their neighboring pixels along the horizontal axis (thus crossing the grooves) are retained after detection. This constraint further filters many false alarms while making the detection results invariant to different peak widths. Note that more generally, ridge detection (sec. 4.2) might be preferred over local maximum detection when the exact defect orientation is unknown. After detection, each hit is associated with its local signal to noise ratio, which will be used as the weight w_d during density estimation according to equation 6.2. To this end, let $w_d = I_z(x,y) = \dfrac{I(x,y) - \mathrm{Avg}(\mathcal{N})}{\mathrm{Std}(\mathcal{N})}$, as introduced in section 5.1.1. That is, the image intensities $I(x,y)$ are standardized with regard to the null distribution of background signal intensities $\mathcal{N}$ for each sensor.

After registration to a common coordinate system by fitting global transformation models (e.g. affine, projective) to manually defined location correspondences in the data, the final set of hits from all sensors is plotted in figure 6.7. Obviously, the false alarms considerably outnumber the actual groove hits. This is due to the sensitive detection rules, intending that no actual defect is missed during individual processing.

Of course, in a single-sensor inspection task, a much more stringent detection criterion is appropriate to limit the number of false alarms. However, this possibly leads to worse sensitivity to small flaws. In contrast, the presented data fusion approach is supposed to discard most false hits while maintaining high sensitivity to small defects.

The individual sensor results will now be briefly compared. In contrast to ET and TT, the MFL hits cluster spatially. This is because the background variations in this data set are not homogeneous, possibly due to inhomogeneities in the internal magnetic field. MFL data are missing in the strip between the two groove rows. The shallowest groove nr. 15 features low SNR in the ET and TT data due to its shallowness. MFL in contrast is more sensitive. Moreover, grooves nr. 8 and 9 stand out in the TT data, because their confidences are even weaker than the shallower grooves nr. 10–14. Interestingly, in figure 6.7, spatial defect-like patterns are formed, although the specimen is not expected to contain any flaws other than the known grooves. For example, see the vertical lines from TT, or the diagonally oriented lines from ET, as highlighted by the arrows. As previously discussed, using individual inspection, it is not easy to classify these obvious indications as structural indications or flaws. In spite of their regular structure, these patterns are considered *non-defect* indications during the following evaluation, if the multi-sensor data set is not able to give a reliable confirmation. On the other hand, there are a few off-groove locations where different sensors behave consistently. These regions could in fact represent unknown but real defects, and will therefore be excluded from the following evaluation. Note that hits within disregarded

[5] This corresponds to a p-value of 1 % in one-tailed significance testing.

areas are not shown in this figure. Moreover, the confidence associated with each hit is not shown, because all hits exceed the chosen threshold of 99 % as explained before.

6.2.3 Fusion and final detection

To compute the kernel density per sensor, Alexander Ihler's *KDE Toolbox for MATLAB* [143] was used. The fused density, according to equation 6.3, is a continuous function that must be evaluated at discrete locations. In fact, to circumvent the discrete sampling, a multivariate mode-seeking algorithm could be used for detection. However, it is more straightforward to set up a discrete evaluation grid that is designed fine enough to not miss any mode of the density. Modes are then traced similarly to per-sensor detection (see sec. 4.2) by finding local density maxima along parallel lines on the specimen surface. Finding local maxima along one-dimensional lines is straightforward due to the density's smoothness, and also makes the detection results more stable across different kernel sizes.

The final hits after fusion are presented in figure 6.8, where fusion is performed according to equation 6.3 with F equal to the *product* fusion rule. Most of the single-sensor false alarms from figure 6.7 were discarded by the fusion method by recognizing the sensor conflicts. Yet, there are a considerable number of remaining false hits. These spurious hits originate from single-sensor hits that overlap purely by chance. Nevertheless, all grooves but the shallowest, nr. 15, clearly stand out against the false alarms considering the fused density measure, which is represented by the marker colors in figure 6.8. Note that the plotted intensities are all near zero (the color scale is in units of 10^{-14}). This is an effect of multiplying three small individual numbers and may lead to numerical instabilities. For practical implementations, it is suggested to carry out the fusion operation in logarithmic units, based on the identity $\log(a \cdot b) = \log a + \log b$. Because the log function is monotone, it preserves the ranks of the fused intensities, and therefore does not alter the resulting ROC curves.

Since the shallow grooves nr. 13 and 14 are the most interesting, see figure 6.9 for several plots of detection results. In the first subplot, eddy current measurements are presented. While the 29 µm deep groove nr. 13 is distinguished from the structural background noise, the method is not sensitive enough to clearly identify the 20 µm deep groove nr. 14. By per-sensor detection and subsequent computation of the partial density $\hat{f}_{ET}$, which is denoted ET_KDE in the second subplot, already most of the background variations are removed before fusion is carried out. Yet, a multitude of false indications remain in the signal. Although MFL and TT do not produce as many false indications, perfect detection is not possible for these individual sensors either. After fusion (bottom subplot), no false alarms persist in the plotted area.

Because higher values of the fused measure correspond to increased defect likelihood, a threshold can be applied to produce a binary decision. In the following, the detection performance will be quantitatively assessed under various conditions.

Figure 6.5: Preprocessed sensor intensity images, zoomed to a region around the groove nr. 13. Higher intensities correspond to indications. a) ET; b) MFL; c) TT. The vertical black line marks the location of the groove. Each image is shown in the respective sensor's coordinate system, thus explaining the different spatial axis labels.

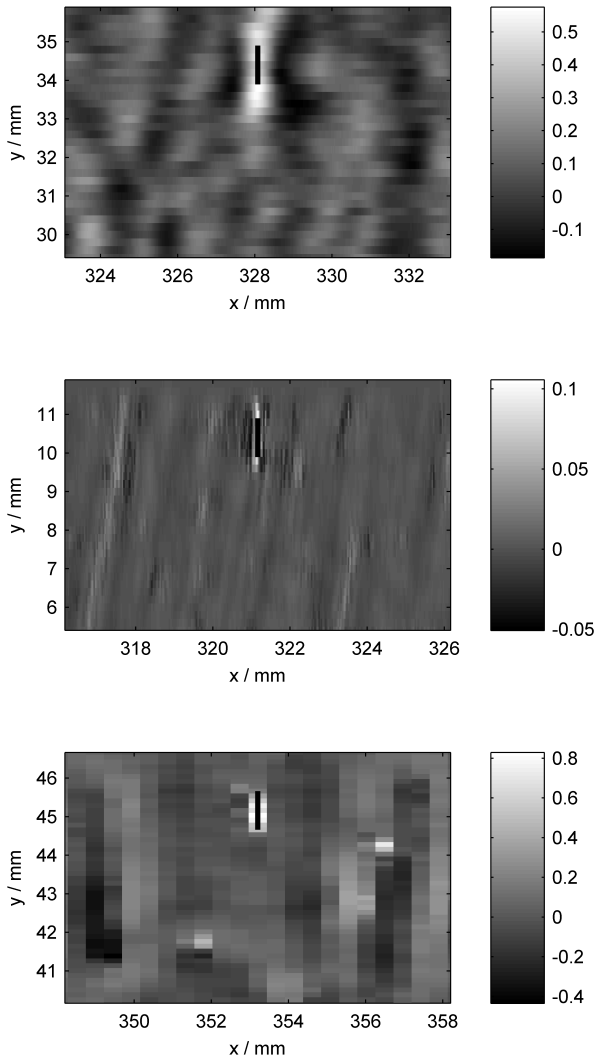


Figure 6.6: Preprocessed line scan per inspection method around groove nr. 13. a) ET; b) MFL; c) TT. The signals are shifted so that each peak value is located at $x = 0$. Note the different intensity scales.

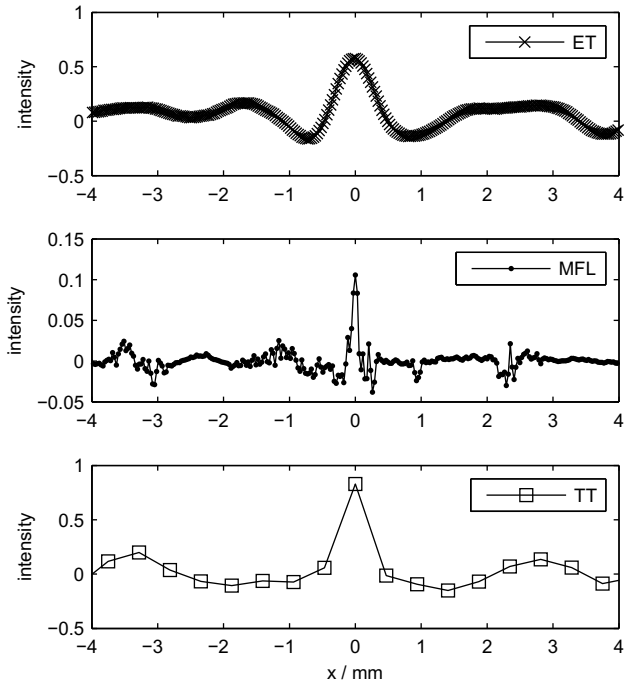


Figure 6.7: Hit locations per sensor in a common coordinate system similar to that in figure 6.4. a) ET; b) MFL; c) TT. Darker colors correspond to higher SNR. Note that the color scale is clipped to a maximum of 10 to prevent non-groove hits from dominating. Axes x and y are not to scale. The tips of the triangular markers indicate the groove positions. The two arrows point to prominent crack-like indications (false positives) in the ET and TT images.

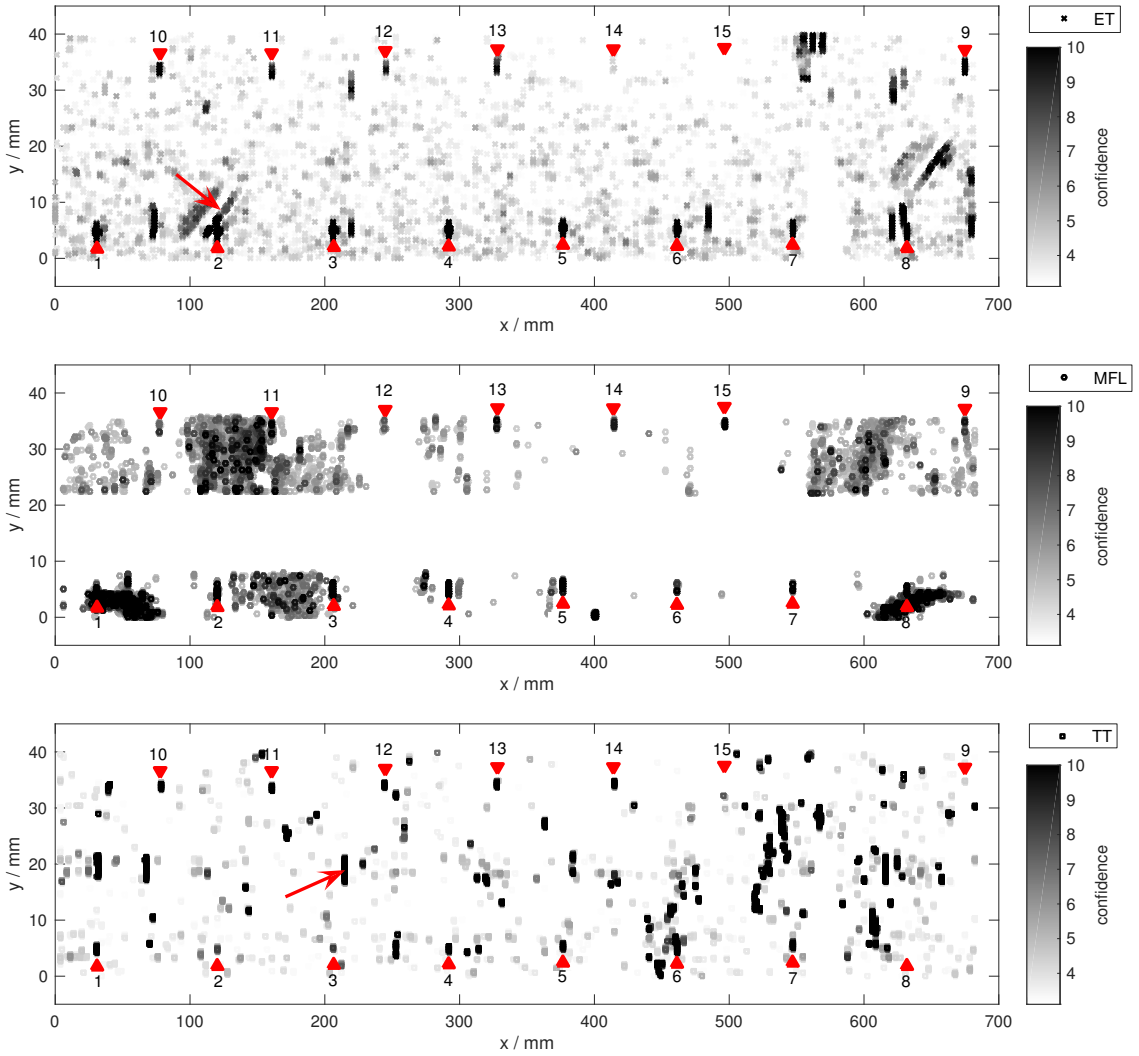


Figure 6.8: Result of decision level fusion using the *product* rule. Darker markers correspond to increased detection confidence. The colors are scaled so that white represents zero fused intensity, and black corresponds to intensities at least as large as at the shallow defect nr. 14. Axes x and y are not to scale.

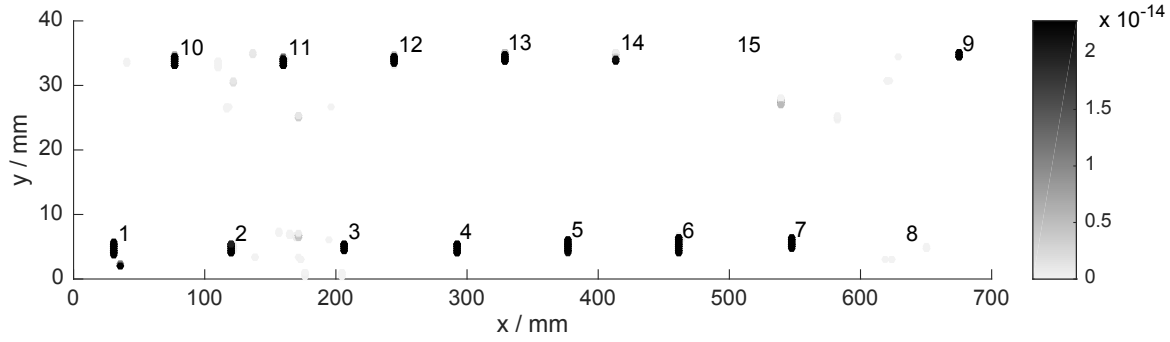
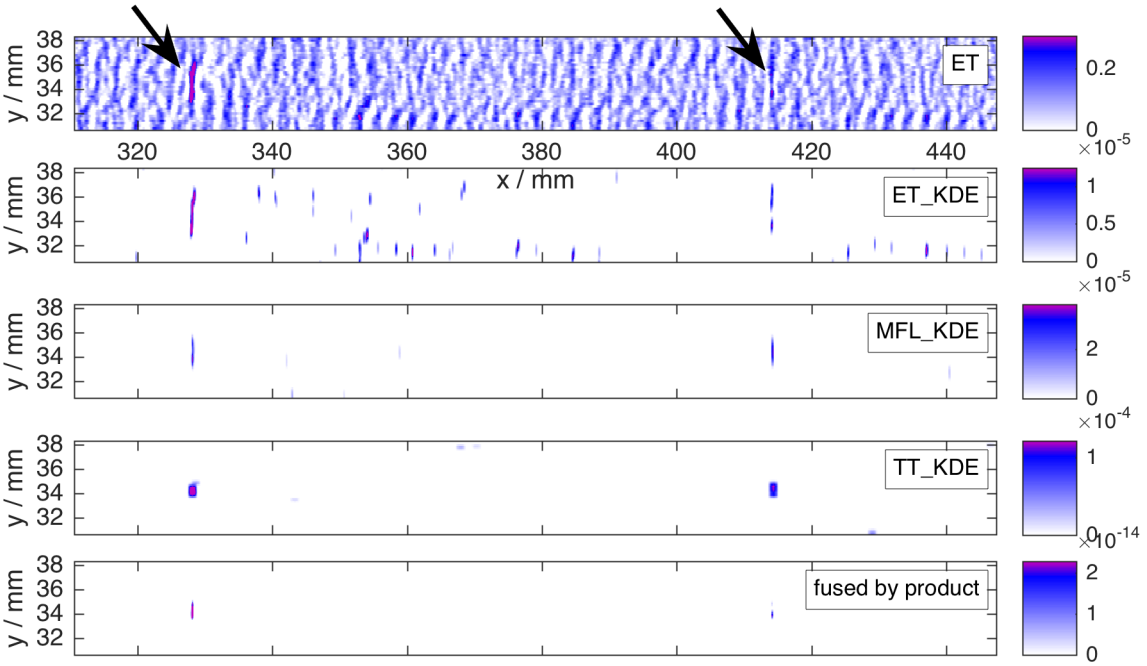


Figure 6.9: Detection results near grooves 13 (left arrow) and 14 (right arrow). In each subplot, the colors are scaled so that white represents zero fused intensity, and magenta corresponds to the maximum intensity of groove nr. 14. Axes x and y are not to scale.



6.2.4 Evaluation

In the following sub-sections, the proposed fusion method is quantitatively evaluated with regard to the presented specimen. This evaluation focuses on detectability, meaning the ability to distinguish between grooves and background in the fusion result. Consequently, the ability to accurately localize a defect after fusion is not a part of this evaluation. For each detection result in the next sections, indications are assigned fuzzy membership values to the two sets *defect* and *non-defect*, based on their distances to the known groove locations. Using this ground truth information, evaluation is carried out automatically by means of precision-recall-curves. Similarly to conventional ROC analysis [144], which is based on recall = true positive rate = *#true hits/#max possible true hits* and false alarm rate = *#false hits/#max possible false hits* for each possible detection threshold, this study replaces false alarm rate with precision = *#true hits/#all hits*. This choice is necessitated by the scattered nature of the hits, which allow an infinite number of possible false alarms, that is, off-groove locations. Precision circumvents this restriction by relating hits to hits, rather than hits to non-hits.

The two evaluation measures precision and recall are fuzzyfied in the evaluation to include the fuzzy membership per hit in the analysis [145, p. 46]. That is, each hit is allowed to be counted partially as a true positive and as a false alarm: Indications near known groove locations are evaluated nearly 100 % as true positives, whereas hits that lie further away have an increasing share as a false alarm. The correspondence between distance to the nearest groove location and fuzzy membership is realized by a Gaussian membership function, whose spread parameter σ =0.2 mm is set equal to the estimated mean registration error of the present data set to account for the localization uncertainty.

Once an evaluation curve in fuzzy ROC space per detection method and per groove is established, the area under each precision-recall-curve quantifies detection performance over the full range of detection thresholds. However, it is preferable to not compute the area under the whole curve, but only for the curve region where recall > 0.5. Denote this measure by AUC-PR-0.5. This focuses evaluation on thresholds that are low enough to ensure that at least half of a groove is detected. Furthermore, a single false alarm hit with higher intensity than the groove suffices to force the curve down to zero precision for small true positive rates, i.e. high thresholds, and therefore dominates the whole AUC measure. This is another reason for ignoring the lower half of the diagram in the computation of AUC-PR-0.5.

Several regions on the specimen surface are marked to be excluded from the evaluation. These are areas near the border of the specimen, indications that result from experimental modification of the specimen surface and off-groove areas where real unplanned defects exist (which would otherwise be counted as false alarms). Not only are all of these disregarded regions removed from evaluation after fusion, but already the hits in these regions are excluded from the density estimation, so that they do not affect the density in the surrounding regions. Furthermore, to evaluate detection performance per flaw depth, after fusion each groove is assessed individually while ignoring all others.

All fusion results are evaluated at the same locations on the specimen surface defined by a dense grid with sampling distances $\Delta_x = 0.0289\,\text{mm}, \Delta_y = 0.1258\,\text{mm}$. This choice of grid resolution is given by the finest spatial sampling among all individual sensors in each spatial dimension. Indications are found by local maximum detection as described in section 6.2.3.

If not stated otherwise, fusion is carried out with a fixed kernel size per sensor

according to equation 6.4, using $\hat{u} = 0.2\,\text{mm}$. For example, the ET sensor would be assigned a kernel size of $(h_x, h_y) = (0.2\,\text{mm}, 1.3793\,\text{mm})$. While this automatic formula ensures that the kernel size exceeds the localization uncertainty in both spatial dimensions and retains the ratio of sampling distances, it might lead to situations in which the kernel is extremely large in the coarsely sampled direction, as is seen here: The kernel size in the y direction is even larger than the 1 mm long grooves themselves, which results from the disproportionate sampling distances in the measurements (see table 6.2). To avoid introducing unrealistically large kernels, this study restricts the kernel size ratio to at most 3. Consequently, for ET and MFL data, the kernel sizes $(h_x, h_y) = (\hat{u}, 3\hat{u})$ are applied. Using this evaluation framework, the performance of the proposed approach is investigated in the following.

Evaluation of fusion rules

As described in section 6.1.3, the three normalized per-sensor densities are fused at each location of interest on the specimen surface using some fusion rule. In this study, the following eight functions are compared: *minimum*, *geometric mean*, *harmonic mean*, *product*, *median*, *sum*, *sumIgnoreMax* according to equation 6.7, and the *maximum*. These are contrasted with single-sensor performance, both before and after individual kernel density estimation. All single-sensor hits are assessed here, in contrast to the fusion methods where hits below the per-sensor thresholds were discarded. For each groove, a separate ROC analysis was carried out to analyze the influence of defect size.

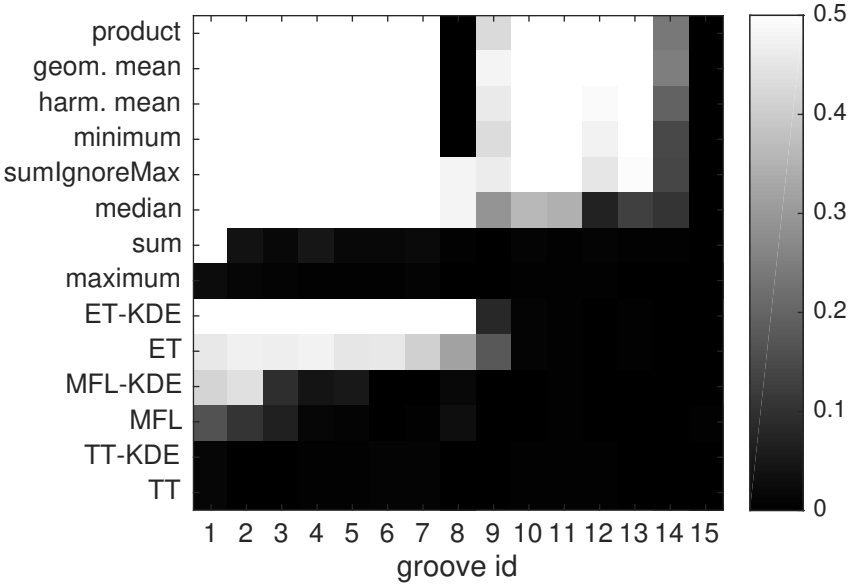
At this point, it is emphasized that the presented results are not representative for the general performance of each individual inspection method. It is possible that better individual results than shown here may be obtained by optimizing e.g. the specimen preparation, the sensors or the processing routines. This is especially true for the MFL data, which were artificially degraded as described in section 6.2.2. Rather, the following experiments demonstrate how the proposed technique copes in the face of imperfect sources of information.

The results are presented in figure 6.10. In agreement with the visual impression from figure 6.7, single-sensor performance (ET, MFL and TT) is unsatisfactory. Although the individual KDEs better pronounce the grooves against more randomly scattered background hits, false alarms still prevent reliable detection even for deep grooves (see TT-KDE and MFL-KDE). After purposely degrading the MFL data (see section 6.2.2), the eddy current technique provides the best single-sensor detection results by reliably indicating groove depths no less than 55 µm (groove nr. 8). In contrast, through multi-sensor fusion, most of the defects can be detected reliably. Two exceptions are the *sum* and the *maximum* rule, which perform poorly over most if not all grooves. These results are explained by the fact that *sum* and *max* do not quantify agreement among sensors, but instead retain all indications from any individual sensor in the fusion result. This is prone to false alarms, which is reflected by low evaluation scores. These results contradict the conclusion in [140], where the *sum* rule was suggested for high-level fusion, as already discussed in section 6.1.3. An explanation for the discrepancy is that this thesis concentrates on the reduction of an abundance of false alarms, for which the *sum* rule is in fact not suited. The strong imbalance of the number of non-defect surface locations compared to flawed surface locations calls for stricter fusion rules than well-balanced situations as are implicitly assumed in [140].

In contrast to these two poorly performing rules, the *minimum*, *geometric* and *harmonic mean* and the *product* rule yield high scores for most defects. Apparently, grooves nr. 8 and 9 are hard to identify across many fusion methods despite the grooves'

midsize depths. This suggests poor single-sensor SNR at these locations, thus leading to inter-sensor conflict, so that this groove is wrongly classified as a false alarm by the strict fusion methods product, *geometric* and *harmonic mean* and *minimum*. On the other hand, the milder fusion rules *median* and *sumIgnoreMax* tolerate unknown single-sensor dropout at the expense of comparably poor detection performance at the shallowest grooves. Specifically, whereas the *median* rule might deteriorate in the face of overall low SNR by permitting too many false alarms, *sumIgnoreMax* offers a good compromise between strictness and tolerance in the evaluation. However, for the detection of very shallow defects like groove nr. 14 in the present specimen, stricter rules appear to offer better performance. The best fusion rule in this evaluation is the *geometric mean*, closely followed by the *harmonic mean* and the *product*. As *product* is conceptually extremely simple yet effective, it is considered the winner. Overall, the shallowest detectable groove depth in this study is given by 29 µm at groove nr. 13. The 20 µm groove nr. 14 could not be found reliably, although fusion offers improved detectability compared to single-sensor detection. The shallowest groove nr. 15 (11 µm) is not distinguishable from background noise due to lack of single-sensor sensitivity.

Figure 6.10: Evaluation of different fusion functions F according to equation 6.3, and of single-sensor detection. For each groove and detection method, the AUC-PR-0.5 is shown in shades of gray. Optimal performance is 0.5. Groove numbers correspond to table 6.1, that is groove nr. 1 is the deepest and nr. 15 is the shallowest.



Additionally, the results are more clearly presented in figure 6.11, where only the *product* fusion is compared against the single-sensor KDEs.

Influence of kernel size

Just like in conventional kernel density estimation, the kernel size is an important factor regarding the detection performance. The sizes assessed in this study are arranged in table 6.3. The *product* fusion rule is selected here due to its strong performance in the previous experiment. Evaluation results are shown in figure 6.12. The results suggest that given a well-performing fusion method and an accurate estimation of the

localization uncertainty $\hat{u}$, a range of kernel sizes around the proposed default setting in equation 6.4 is adequate. Performance only deteriorates for very small kernel sizes like 25 %—50 % of the proposed size. The product rule shows no obvious dependence between kernel size and performance at the shallowest two grooves. However, groove nr. 9 is better identified when larger kernels are used. This could be explained by unusually large registration error at this region, but in this case the reason is that thermography only indicates the top part of groove nr. 9 with large enough SNR to pass the individual detection stage. The results presented here might tempt to favor large kernels. However, large kernels increase the chance of falsely associating spatially nearby false alarms, and thus quantify sensor agreement where there is actually conflict. Therefore, the kernel size proposed in equation 6.4 was found to be effective in this experiment.

Figure 6.11: Evaluation of single-sensor detection (ET-KDE, MFL-KDE, TT-KDE) versus fusion, for the *product* fusion rule and a fixed kernel size. The maximum possible score is 0.5 (left vertical axis). The set of grooves is divided into two sub-figures for clarity. Groove depth is indicated by the dashed blue line corresponding to the right vertical axis. Note the different axis scales for groove depth in the two subplots, required by the comparably small range of the last grooves' depths.

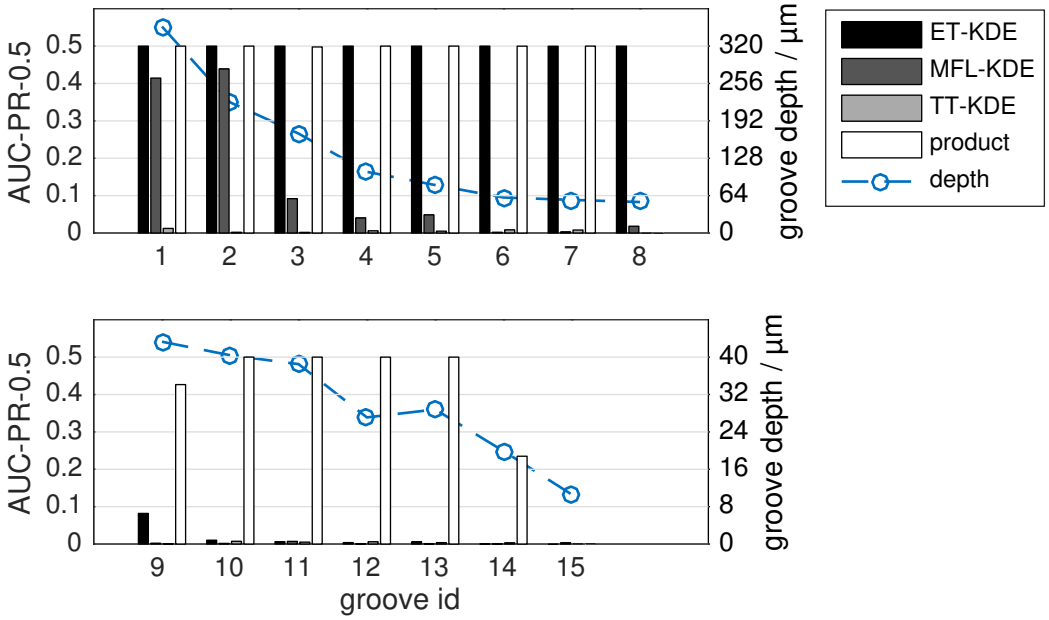


Figure 6.12: Evaluation of different kernel sizes, for the *product* fusion rule. For each groove and fusion method, the AUC-PR-0.5 is shown in shades of gray.

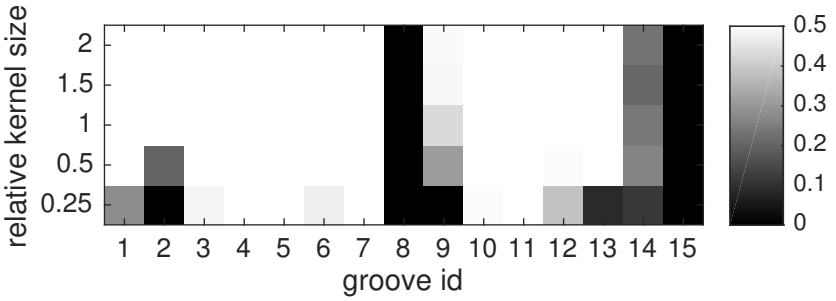


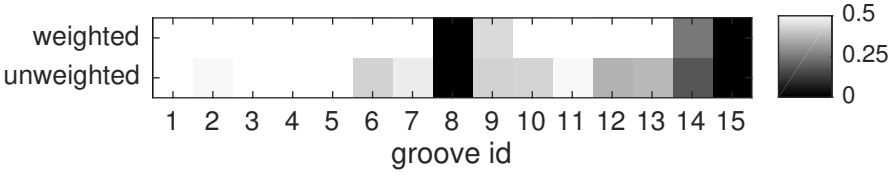
Table 6.3: Kernel sizes used in the experiment. All sizes are in mm. *Relative kernel size* denotes the fraction of $\hat{u} = 0.2\,\text{mm}$ that was used to compute the kernel sizes according to equation 6.4. That is, a range of smaller and larger kernels compared to the default size (relative kernel size = 1, bold faced column) were assessed. To prevent unrealistically large kernels due to the disproportionate spatial sampling distances in our data, kernel sizes were limited to 0.6 mm for ET and MFL, and to 0.746 mm for TT (gray shaded table cells).

		Relative Kernel Size				
		0.25	0.5	1.0	1.5	2
ET	h_x	0.05	0.1	**0.2**	0.3	0.4
	h_y	0.346	0.6	**0.6**	0.6	0.6
MFL	h_x	0.05	0.1	**0.2**	0.3	0.4
	h_y	0.345	0.6	**0.6**	0.6	0.6
TT	h_x	0.186	0.373	**0.746**	0.746	0.746
	h_y	0.05	0.1	**0.2**	0.3	0.4

Influence of weights

In the previous experiments, the individual sensors' hits were weighted by factors w_d in equation 6.2 to take into account the local SNR. This experiment assesses the benefit of these weights over an unweighted approach ($w_d = 1$) in which the densities $\hat{f}_S(\boldsymbol{p})$ are only influenced by the spatial proximity of neighboring hits. This setting represents inspection results for which no measure of confidence is available. In both experimental cases, the *product* fusion rule was chosen and the kernel size was set to the value suggested by equation 6.4. Figure 6.13 illustrates the respective detection performances. According to the results, the unweighted variant never surpasses the proposed weighted density estimation at any groove. Interestingly, although the unweighted method does not take into account the local SNR and therefore is not influenced by defect depth, it is clearly observed that most of the deeper grooves (e.g. nr. 1—5) are more reliably found than the shallower grooves (e.g. nr. 9—15). This is because during the first stage of individual detection before fusion, only parts of the shallower grooves might be retained whereas deeper grooves are completely preserved. Therefore, the weighted approach should be favored over the unweighted method if possible. Otherwise, much effort should be spent on high-quality registration to make the sole feature of spatial proximity of hits across different sensors a reliable indicator of defect presence. Yet, even without weights, product-fusion still outperforms any individual method in the evaluation for grooves shallower than 43 µm (groove nr. 9).

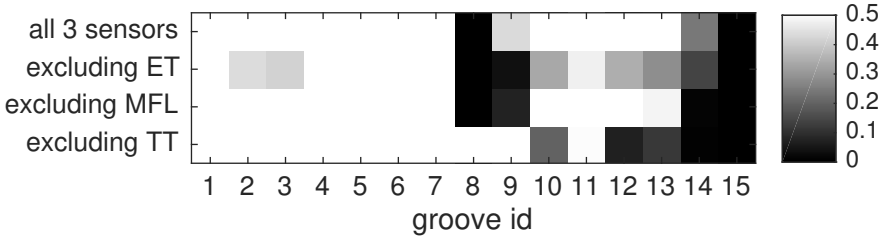
Figure 6.13: Comparison between the weighted approach (as proposed; top row) and the unweighted variant, for the *product* fusion rule.



Influence of individual sensors

The effect of individual sensors on the fused performance is assessed in this experiment. To this end, fusion is carried out three times, while leaving out the hits of one of the sensors in each run. The results are compared to fusing the full data set. Again, the product fusion rule is applied, and hits are weighted. The results are presented in figure 6.14. Each two-sensor subset of inspection methods shows slightly different effects. Apparently, the thermographic data mainly help in detecting the shallow grooves 12–14. However, the same inspection seems to have missed the flaws 8 and 9, because the information from both MFL and ET is crucial for detection here. On the other hand, TT is required to identify most of the shallow grooves in this evaluation. The same observation holds for ET. In contrast, by purposely degrading the MFL data (see section 6.2.2), this inspection method appears less relevant for defect detection. Still, although this low-quality data source has a large set of false alarms, it impairs full three-sensor detection for none of the grooves. On the contrary, MFL improves the detection quality of grooves 9, 13 and 14. Among the deeper grooves, nrs. 1, 4, 5, 6 and 7 are perfectly found using any two-sensor configuration, thus indicating that they are clearly represented in all three measurements. Overall, the evaluation demonstrates that the full set of sensors is required for optimal performance with the given data set. Yet, with the right choice of sensors two-source fusion already has the potential to outperform individual detection.

Figure 6.14: Influence of individual sensors on the fusion result, for the *product* fusion rule and a fixed kernel size.



6.2.5 Replication of results on a second test specimen

In addition to the specimen discussed in the previous section, denoted by *SA*, the presented fusion approach is applied to a second specimen (*SB*) to demonstrate the transferability of results to other samples. The second investigated test specimen is identical to the first bearing shell, thus having the same physical and geometrical properties such as constituent material, shape, size and surface condition. *SB* also contains regularly spaced machined grooves simulating surface cracks. But whereas *SA* has 15 grooves ranging in depth from 10 to 385 µm, *SB* has 16 grooves in a much narrower range between 10 and 50 µm. The detailed specifications of *SB* grooves are given in table 6.4.

Data collection was carried out as described in the main article. However, the analysis differs from that of specimen *SA* in the following aspects:

- The spatial sampling distances differ, as detailed in table 6.5. Specifically, sampling is finer during the inspection of SB by both eddy current and thermal testing.

Table 6.4: The depths of grooves in specimen *SB*. For reference, the IDs of the grooves in *SA* that are of comparable depth are listed in the last row.

Groove nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Depth / µm	54	54	35	33	31	31	28	28	26	24	22	20	19	15	14	12
Groove in *SA*	8		11				13			14					15	

Consequently, the grid where the fused densities are evaluated is refined to $\Delta_x = 0.0288\,\text{mm}, \Delta_y = 0.100\,\text{mm}$, which equals the grid resolution of ET.

- Several regions on the surface of *SA* were identified that had to be excluded from evaluation for reasons given in the previous sections. In contrast, no region was excluded from the evaluation of *SB*. The mean registration error of the inspections of *SB* is slightly higher (around 0.25 mm) than for *SA* (around 0.2 mm).
- Due to the changes in localization uncertainty and in spatial sampling distances, new kernel sizes suggested by equation 6.4 were applied. As for *SA*, kernel bandwidth parameters (h_x, h_y) were restricted to have a ratio of at most 3. For ROC evaluation, the fuzzy membership parameter was set to $\sigma = 0.2\,\text{mm}$ as for *SA*.
- To degrade the quality of the MFL data set for a meaningful assessment of fusion performance, the sensor indications were reduced to 20 % of their original intensities, rather than 2 %. This setting produces roughly comparable signal to noise ratios in *SA* and *SB* at shallow grooves.

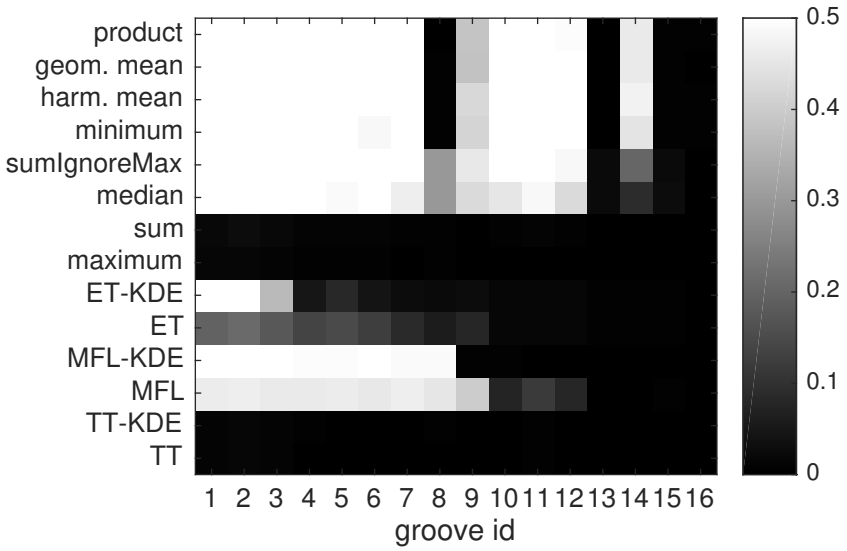
Table 6.5: A comparison of spatial sampling distances during the inspection of specimens *SA* and *SB*. Dissimilar distances for *SB* are in boldface. All measures are in µm.

	SA			*SB*		
ET	28.9	x	200	28.8	x	**100**
MFL	29	x	200	28.9	x	200
TT	469.1	x	125.8	**125.6**	x	125.6

Evaluation of fusion rules The same eight fusion rules applied to the measurements on *SA* are quantitatively compared against single-sensor detections for *SB*. Figure 6.15 presents the results. The results are consistent with those obtained from the first specimen *SA*. Fusion outperforms single-sensor detection in all cases, except for MFL at groove nr. 9. Defect nr. 14 demonstrates the advantage of strict rules (e.g. *product*) over less strict rules (e.g. *median*) to reliably identify shallow defects. Grooves 8 and 9 are hard to find across many detection methods due to poor single-sensor SNR and, in the case of groove nr. 9, due to an unusually large local registration error of 0.75 mm. Note that the mean registration error is about 0.25 mm. The shallowness of grooves nr. 13 and above (shallower than 20 µm) results in an insufficient single-sensor SNR. Yet, groove nr. 14 appears to yield relatively strong indications in the data, which is additionally aided by low local registration error (about 0.2 mm). Whereas in *SA*, the

geometric mean is slightly ahead of *harmonic mean* and *product*, in *SB* the *harmonic mean* takes the first place, followed by *geometric mean* and *product*. Again, the *product* rule can be considered the most basic method that performs best.

Figure 6.15: Evaluation of different fusion functions F according to equation 6.3, and of single-sensor detections. For each groove and detection method, the AUC-PR-0.5 is shown in shades of gray. Optimal performance is 0.5. Groove numbers correspond to those given in table 6.4, that is groove nr. 1 is the deepest and nr. 16 is the shallowest.



6.3 Discussion

The conducted experiments demonstrate that the density-based approach is well-suited to incorporate indications from heterogeneous sensors. The number of false alarms can be strongly reduced relative to single-sensor inspection while retaining most of the defects. In particular, under the chosen evaluation metric, the fusion method performs as well as or better than single-sensor detection for 13 out of 15 grooves. The performance gain is most pronounced for the shallower defects, which usually generate less significant indications.

Note that the principle to quantify agreement among sensors requires that all sensors yield redundant information about the object of interest, e.g. near-surface cracks in this case. That is, in NDT, all sensors must respond to the same flaw type in the same size range. If, in contrast, one of the sensors reports a defect that is not detectable by the other methods, it will be discarded as a "false alarm" by the proposed technique, since it is not designed to fuse complementary information. This is the case for the shallowest investigated defect in the study. Similarly, groove nr. 8 is only found by single-sensor inspection but not by fusion, because TT lacks of a significant indication in that area. To apply a multi-sensor system despite such unexpected effects, mild fusion rules, such as *median* or *sumIgnoreMax*, trade strong reduction of false alarms for the ability to compensate unknown sensor dropout.

Another point concerns the relationship between fusion performance and spatial uncertainty. Specifically, the fusion performance is expected to improve with registration

accuracy. This is because kernels can be made narrower for smaller registration errors, and therefore the likelihood of a non-defect-related indication due to spurious multi-sensor agreement is reduced. In any case, the actual registration error must be quantified to set the kernel size accordingly. Moreover, the fusion technique strongly benefits from realistic estimates of the local signal to noise ratios, which enter the fused density through weights. For the final detection after fusion, a threshold could be chosen to retain only the significant density peaks. Thus, only few parameters (localization uncertainty, fusion rule, density threshold) fully describe the methodology and are usually readily determined. Furthermore, if one is unsure about a fused indication, the original individual hits can always be reconsidered to collect additional evidence for or against the presence of a defect. After all, the density-based approach spatially associates neighboring hits and thus may serve as the basis for multi-sensor detection after feature extraction. For example, the proposed density measure identifies narrow regions of increased defect likelihood. These regions can further be assessed by extracting features from each individual sensor in this region, which could be combined by some classification algorithm to reach a final conclusion.

Concerning the number of sensors, using at least three different sources of information is suggested, as presented here. However, experiments show that improved performance over single-sensor inspection is possible already for two sensors. Note that the more fusion inputs are provided, the higher the likelihood of the purely coincidental agreement between at least two sensors. Therefore, the rule *sumIgnoreMax* (equation 6.7) will have to be extended if even more sensors are included, whereas the *median* rule and *product* rule are expected to perform well regardless of the number of sensors.

This study has certain limitations that should be pointed out. Whereas the eroded grooves facilitate detection assessment for well-defined defect depths, their linear shapes do not resemble natural defects. Also, whereas the orientation of our flaws is well-defined, natural defects often vary in orientation. Therefore, directionally sensitive measurements, such as ET using a differential probe and MFL using gradiometers, must be carried out multiple times in different directions. However, because this issue is only relevant for per-sensor detection prior to fusion, it is not further elaborated here.

6.4 Conclusions and outlook

A density-based method was developed for the fusion of spatially scattered data and it was applied to sensor signals from the nondestructive testing of a bearing shell. This high-level fusion approach has the advantage of being independent from the processes that generate the scattered points, and of directly accounting for registration errors. Three different mechanisms are implemented to ensure robustness against false alarms. Practical suggestions on how to determine the kernel size are given. The technique was quantitatively evaluated using a defect detection experiment. The results demonstrate that single-sensor inspections of the specimen are outperformed by the proposed technique, especially for defects that are too shallow to be reliably indicated otherwise. Moreover, the proposed method is quite generic, as it receives spatial locations from single-source detection routines and returns areas of multi-sensor agreement. Therefore, it may be applied for detection tasks in other domains, such as multi-modal medical image fusion.

Chapter 7

Discussion and Concluding Remarks

After the methodological parts of this thesis, this last chapter widens the focus by discussing some more general questions that might have remained unanswered. The discussion is separated into several paragraphs, each dealing with a specific topic from a practical perspective. Finally, the thesis is summarized and an outlook is presented.

How to make use of directional information in high-level fusion? Although at low-level, this thesis investigates fusion methods that make use of directional spatial information, no directionally sensitive fusion is proposed at the decision level. In fact, much effort went into researching such techniques as part of this work, based on the fact that high-level fusion is well suited to fuse heterogeneous NDT sensors, and on the observation that hits from the individual sensors form coherently directed spatial patterns. However, although ideas from undirectional density-based fusion can be successfully extended to the directional case by elongating the kernel functions, it seems that such techniques provide no benefit over undirectional fusion. But these observations were made based on questionable ground truth data, before the simulation framework for directional indications was developed (see section 5.2.5). Therefore, it would be interesting to see the results of directionally sensitive decision-level fusion applied to these simulated indications. Since the results of low-level directional fusion were discouraging, and due to the necessarily limited scope of this thesis, directionally sensitive fusion at the decision level was not investigated in detail here.

Is there a middle ground between low- and high-level fusion? Since low-level and high-level fusion offer complementary benefits and drawbacks, a cross-level fusion approach seems appropriate. In fact, there are common elements in the methods for both of the proposed fusion levels, thus bridging their apparent conceptual gap. Specifically, low-level information was introduced into the proposed decision-level fusion scheme through weights. Although these weights are in principle optional, the experiments demonstrated (sec. 6.2.4) the substantial boost in detection performance when informative weights are made available. Still, the notion of weights is general enough to clearly separate the two fusion levels, because the developed high-level fusion method is independent from the specific way these weights are determined. For example, weights might directly correspond to local SNR, as proposed, but they could also represent degrees of belief or trust that are attributed to the different sensors, possibly also varying by spatial location. Moreover, weights are only assigned to indications that

have already undergone some form of detection routine, which might have made use of arbitrary high-level information, which is another clear difference to fusion at the signal level. But despite these methodological differences, it is easy to envision a system that integrates detection *results* from fusion at the two levels, thus in turn leading to decision-level fusion. This is realizable in two alternative ways. Either the low-level fusion result is included as an additional "sensor" during fusion at the decision level as proposed in this thesis, or each of the two fused detection results (low and high level) are interpreted as individual sources to be fused once again. It would be interesting to evaluate if there is an additional benefit in this higher-order fusion scheme, but care must be taken to account for the introduced correlation among the fused sources, because the low-level fusion result is clearly not independent from the other sources.

As briefly mentioned in the first part of this thesis, there is a third level of data representation at the transition from signal- to decision-level fusion, denoted as the *feature level*. Although this work did not address fusion at this intermediate level, it is relevant for other fusion applications in NDT [61, 62]. There are several reasons why feature-level fusion was not considered here. First, it is unclear from what object features should be extracted. Taking each signal sample as an object, the measurements already represent low-level features. More generally, features can be extracted from the neighborhoods of each pixel, for instance in a sliding window manner. An alternative is to use image transforms like the wavelet transform, which provide localized features at different spatial scales. Although such techniques have indeed been applied in this thesis, they were attributed to the signal level here, because the inverse transform again yields a signal to finally undergo defect detection. Other definitions of objects for feature extraction are possible apart from pixels, for example by image segmentation. However, image segmentation is a very challenging task in itself, which complicates the primary goal of evaluating data fusion techniques. Moreover, NDT inspection signals typically do not contain any sharp image features like edges, so that the notion of a segment is not clearly defined. One method for feature extraction that was experimented with during the course of this thesis, but was dropped in favor of the nonparametric wavelet analysis, is one-dimensional peak fitting using parametric peak models, like Gaussian or Lorentz functions. The idea was to extract peak parameters as features after fitting. However, the advantages of sub-sample resolution and availability of multiple features did not outweigh the downsides of high computational effort, dependence on initial parameter values, overly strong assumptions of peak shapes, and the difficulty to decide when a fit was successful. Another reason for not specifically addressing the feature level is that feature space analysis is the domain of machine learning methods, which are disregarded here for reasons discussed in the introduction.

Which fusion level should be adopted in practice? This thesis has shown that fusion at both the signal and decision level of data representation improve defect detection over single sensor inspection. Therefore, the question, at which level of abstraction NDT data should be fused, mostly depends on practical considerations. As a general guideline, signal-level fusion is definitely a good choice if the source signals have similar properties, or can be easily converted into the same format. The more disparate the source signals are, the more appropriate a higher-level fusion scheme becomes, because artificially forcing the signals into the same format cannot be justified at some point. Since this thesis is based on the idea that physically complementary sensors are used to search for similar types of defects, pre-processing and detection routines may differ strongly between the individual NDT techniques. In such settings, decision-level

fusion is recommended. In a sense, each sensor votes for the existence of defects, and the final decision is based on all votes. Note that multi-sensor defect detection is not guaranteed to reduce the uncertainty about defect presence. For example, consider the situation in which around half of the sensors vote *for* defect and the other half *against*. In this case, the main benefit of multi-sensor NDT is that this ambiguity is made explicit, and further actions can be taken based on this valuable information. In real applications, the system could decide to measure again (potentially more precisely or with additional sensors), or to notify a human operator. This last example also demonstrates why it is beneficial to retain the raw sensor data, even if they are not required for decision-level fusion. In ambiguous cases, the already available inspection signals may be re-visited to try to find the cause of confusion. Also, even successful fusion situations should be documented in this way, in case doubts occur about the fused decisions in the future.

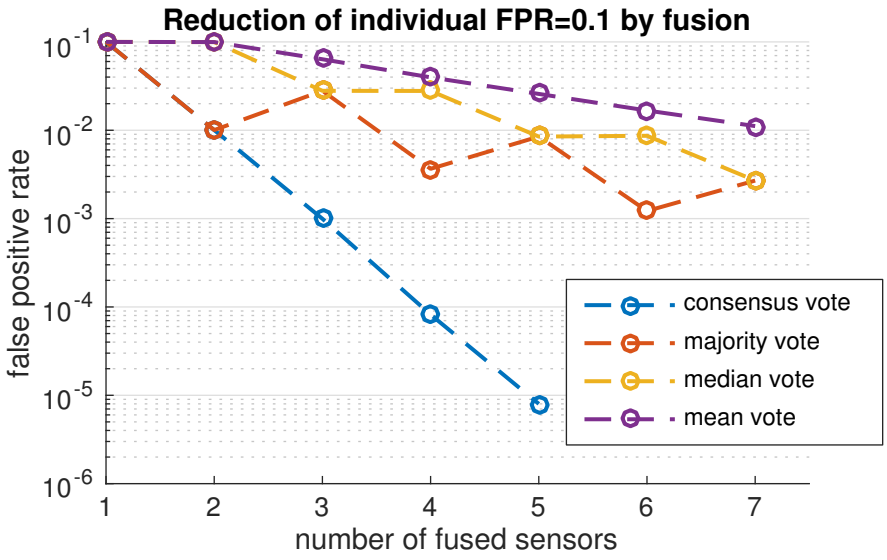
Which and how many sensors should be fused? Although the answer to this question is highly dependent on the task, some guidelines can be given for surface crack detection. The types of applicable sensors are usually determined by practical constraints. Similarly, the number of different NDT techniques to be applied is usually limited by the affordable time and financial resources, unless a dedicated measurement setup for multi-sensor NDT, like in [5–8], is available. However, the experiments in this thesis demonstrate that a higher number of sensors (only 3 were investigated here) consistently improves the results. Moreover, a simulation study shown in 7.1 demonstrates that the false alarm rate is exponentially reduced by fusing more sensors. This study assumes that individual sensors have a fixed FPR of 0.1 (arbitrarily chosen), and each sensor yields a random confidence score about the existence of a defect, to be fused at the decision level. To simulate the desired FPR, a confidence threshold at 0.5 is assumed, and therefore the fixed fraction among the simulated samples that represent false alarms are assigned uniform random scores $\in [0.5, 1]$, whereas samples from the "true negative" class are assigned uniform random scores $\in [0, 0.5]$. For each sample, these random sensor scores are fused according to four different rules: consensus = minimum > 0.5, majority voting, median > 0.5, mean > 0.5. After fusion, the number of remaining false alarms are counted and the false positive rate is computed. This procedure is carried out for different numbers of sensors to be fused, and the results are presented in figure 7.1. Because the vertical axis is scaled logarithmically, and all FPR curves follow linear trends, it is seen that in this simulation, the false positive rate is exponentially reduced by fusing more sensors. One crucial note here is that although the reduction of FPR usually comes at the cost of impairing the true positive rate, the addition of more sensors will *not* decrease the TPR[1], because it is assumed that all sensors are able to indicate the same types of flaws (sec. 2.3).

The shown results are straightforward to interpret. The different slopes of the fusion rules indicate their different degrees of strictness about reporting a fused indication. For example, the consensus rule is the most conservative method, because it reports a fused defect only if all individual sensors reported an indication (score > 0.5). Therefore, this rule allows for the strongest reduction of false alarm rate. Not surprisingly, the fused FPR decays at a rate of 0.1, because the chance that all sensors independently show a false alarm is $(\mathrm{FPR})^{\mathrm{nSensors}}$ with $\mathrm{FPR} < 1$. The shown line would theoretically extend beyond 5 sensors, but due to the limited sample size in this simulation, it was

[1]TPR might only be reduced, if sensors are included that have a high chance of missing a defect. However, such sensors would not be used in NDT anyway.

estimated to be zero, which is not representable on a logarithmic scale. But despite this strong improvement of detection specificity, in case even a single sensor fails to indicate a defect (because its sensitivity is disturbed for some reason), this rule will lead to a missed defect, which might have severe consequences. Therefore, in practice, less conservative rules might be more appropriate, as was discussed in this thesis. Ordering the investigated fusion rules by decreasing ability to reduce FPR, it is seen that majority voting is followed by the median, and the mean fusion rule is last. The oscillating behavior of the majority and median rules are explained by the fact that even and odd numbers of sensors are treated differently by these rules. In summary, the simulation clearly demonstrates that for a wide variety of fusion rules, the addition of more sensors will exponentially decrease the false alarm rate, and therefore improve the reliability of the detection system. It must be questioned, however, at which point the gained benefit is exceeded by the extra effort. Simulation studies, like the one shown here, help to address this question in practical settings, where the FPR of each individual sensor can be defined or roughly estimated, and the decision criterion can be more realistically simulated than the simple cutoff at 0.5 that was implemented here for illustrative purposes.

Figure 7.1: Reduction of FPR by number of fused sensors, assuming individual FPR=0.1. Results are simulated from $N = 500000$ random samples. Note that the line that represents consensus fusion stops early in this figure, because the simulated FPR is zero from that point on, which cannot be represented on a logarithmic scale.



Evaluation of defect detection: ROC space vs. POD In all experimental parts of this thesis, the evaluation of detection results was based on ROC curves. To capture the influence of defect size on detection performance, each known defect was analyzed separately. In NDT, a popular alternative to this strategy is the so-called Probability of Detection (POD) framework [17, 146]. Since POD is not applied in this thesis, a detailed description of the method is omitted here. But instead, based on a comparison between ROC and POD evaluation, which is summarized in table 7.1, this section explains why ROC analysis was chosen over POD in this thesis.

To start with the similarities, both evaluation strategies operate on ordinal data, for example signal intensities from single-sensor or fused inspection, and output evaluation

results based on ground truth knowledge about the presence of defects. Furthermore, both ROC and POD measure the fraction of known defects that were correctly identified (TPR), i.e. whose intensities are above a detection threshold[2]. Also, both methods are able to report confidence intervals about their results. However, there are also fundamental differences. In contrast to ROC analysis, POD is not limited to ordinal input, but can also evaluate results that have already been classified as correctly detected (*hit*) or falsely rejected (*miss*), which makes POD more widely applicable in this regard. In particular, in hit/miss analysis, the detection operation may be arbitrary and need not be known, whereas ROC analysis always requires thresholding. However, if indeed a known threshold was used to generate the hits and misses, then for a fixed defect size, POD is equivalent to a single point in ROC space. Despite POD's wider applicability in the hit/miss case, for the remaining discussion hit/miss analysis is disregarded, because this thesis assumes ordinal input to the evaluation procedure. The other fundamental difference between ROC and POD is that whereas ROC is a general tool for the evaluation of binary classifiers, POD was specifically developed for defect detection. In particular, POD explicitly models defect size, e.g. surface crack depth into the material, by assuming a functional (parametric) relationship between defect size and sensor output intensity. As a result, using statistical regression, POD allows to interpolate the observed evaluation results at defect sizes that were not measured. This leads to the typical output of POD analysis, which summarizes the detection capability of the evaluated system in a single number $a_{90/95}$. This number represents the estimated size at which defects are detectable at a 90 % true positive rate, with 95 % confidence[3]. The $a_{90/95}$ assumes a single fixed detection threshold, which must be defined by the NDT expert based on the noise distribution. In POD, noise is operationally defined as *aberrant signals* that *corrupt* the *target*[17, sec. 4.4]. Essentially, by defining a threshold value, the acceptable FPR is held constant, which provides the basis for the whole evaluation. In contrast, in ROC evaluation, typically one would carry out independent analyses for the different defect sizes, as was done in this thesis. In fact, early POD followed the same approach [17, p. 83], but the high number of required samples led to the development of the regression model described before, by making more assumptions. Because the detection threshold is not fixed in ROC analysis, but rather all possible thresholds are considered, each detection method yields an entire ROC *curve*. These curves are interpreted as the detection system's trade-off between sensitivity and specificity, which cannot be assessed using POD. The resulting nonparametric ROC curves[4] are often summarized by measures such as the (partial) AUC, and these values are reported per defect size.

Overall, although POD evaluation was specifically developed for nondestructive defect detection, its drawbacks led to the decision to favor ROC analysis to evaluate the methods proposed in this thesis. In particular, POD makes strong assumptions, whose violations might invalidate the whole analysis. For example, the functional relationship between defect size and sensor intensity is often assumed to be approximately linear[5].

[2]In fact, although POD is only a synonym for TPR, the term POD is now used to denote the whole evaluation framework that was built around the TPR in NDT, as will be explained.

[3]$a_{90/95}$ is defined as: $p(\text{TPR} > 0.9 \mid \text{size} = a_{90/95}) = 0.975$. Of course, the desired TPR and confidence level are parameters which can be set to other values, depending on the cost of missing a defect. For details, please see [17, sec. G.3.4.4]

[4]Although ROC analysis is usually done nonparametrically, also parametric ROC models exists. In the parametric case, confidence intervals are explicitly computable, whereas confidence bounds are nonparametrically estimated by resampling techniques, e.g. bootstrap.

[5]Other functional forms are possible, too, but the more parameters are required to describe the

Yet, in practice this assumption may not hold, because the sensor output does not only depend on a single factor *defect size*. Especially when inspecting microcracks as in this thesis, the defect responses are weak and are therefore easily impaired by structural noise. Conversely, it was also observed that larger defects unexpectedly did not produce an indication in the expected intensity range, for unclear reasons. To carry out POD analysis nevertheless, robust regression techniques are required, or an impractically large number of inspected defects are necessary. Only if the assumptions required by POD analysis hold, then POD is an adequate and theoretically well-founded framework to evaluate defect detection that allows minimizing the number of required defect inspections. For the current study, it was more practical to apply ROC evaluation, which does not rely on any assumptions about noise, defect sizes or sensory output. Moreover, the goal of this thesis was to increase specificity by suppressing false alarms, so that the focus was put on correctly identifying the *non-defect class*, which is not supported by POD analysis.

In this context, it should be emphasized again that the results reported in this thesis depend on several decisions that were made, for example by focusing on a sub-region in ROC space using partial AUC (sec. 5.2.6). Potential future results are only comparable to those in this thesis, if the same evaluation methods and parameters are used.

Table 7.1: Comparison between the two evaluation frameworks ROC and POD

	ROC	POD
input	ordinal	binary ("hit/miss") or ordinal
parametric	no	yes
confidence bounds	yes (bootstrap)	yes (parametric)
detection model	thresholding a scalar value	thresholding a scalar value
evaluation at threshold	all possible detection thresholds	single detection threshold
evaluation of false alarms	yes: FPR	no
explicit modelling of defect size	no	yes
item under evaluation	signal sample, or whole indication. both non-defects and defects.	signal sample, or whole indication, of a defect of known "size"
comparison between detection methods	(partial) AUC, TPR at fixed FPR, FPR at fixed TPR, classification measures at a fixed threshold	smallest reliably detectable defect size according to $a_{90/95}$
pros	• well-researched (binary classifier evaluation) • robust because no assumptions can be violated • explicit evaluation of "non-defect" class in addition to "defect" class	• theoretically well-founded • specifically designed for NDT → easily interpretable • able to work with binary input • interpolates unobserved defect sizes

relationship, the more data are needed to during regression for high statistical confidence

Practical recommendations From a practical point of view, it is important to consider a few points during the measurements to ease the fusion procedure. This concerns the order of inspections, and spatial sampling schemes. Considering the order in which NDT techniques are applied, it must be ensured that measurements are truly independent, and no side effects are introduced. For example, improper handling of the specimen might introduce small surface scratches, which of course will be only indicated in subsequent inspections, thus contradicting the earlier inspections. Another common source of measurement artifacts is to mark locations on the specimen surface e.g. by a pen, which does not influence electromagnetic measurements. However, these marks will be clearly visible in thermal inspection data, and therefore potentially interfere with defect detection. Consequently, it would be advisable to avoid making any intended changes to the specimen, apart from careful pre-measurement conditioning like surface cleansing. If this is not possible, after each inspection the specimen should be brought into the same state as it was before the measurement. For instance, after MFL, it is necessary to de-magnetize the specimen, so that future electromagnetic inspections are not affected.

Once the order of inspections has been determined, a spatial sampling scheme has to be designed. For the sake of straightforward image registration, all inspections should adhere to the same orientation of the specimen. This is particularly important for rotationally symmetric objects, e.g. discs or rings. If a specimen has multiple surfaces that should be inspected, a common definition for all inspections must be introduced to clearly distinguish them. It is however unavoidable that each NDT technique is applied at different locations on the specimen, owing to their fundamental differences in working principles. In particular, the covered areas (or volumes) will differ, as well as the spatial distance between neighboring sampling locations. It is essential to explicitly store the coordinates of each recorded signal sample together with the measured signal itself. If this is not realizable, the NDT practitioner should at least precisely store the location of the start of scanning, together with the inter-sample and inter-line distances, and the exact direction of scanning, from which each sample's location can be reconstructed. Yet, it should be in the responsibility of each individual NDT technique's expert to provide all necessary information, since all fusion results crucially depend on accurate spatial alignment of the individual signals.

To deal with different spatial sampling densities in this thesis, for signal-level fusion, the signals of coarse resolutions were upsampled to match the sensor that has the finest spatial resolution. This strategy preserves the details of the fine-grained sensor, but leads to a high redundancy of coarse-grained sensor samples. An alternative would be to fuse at the coarse resolution by downsampling the high-resolution sensor, and after defect detection, to accurately localize the fused indications using the fine-grained sensors. Note that differences in spatial resolution do not directly affect fusion at the decision-level, which is a reason to choose high-level fusion in such situations. But, since coarse sampling potentially leads to fewer per-sensor indications than spatially dense sampling, the fusion strategy must balance the individual sensors to avoid putting NDT techniques at a disadvantage that have low spatial resolution. This principle was realized in this thesis by introducing a normalization factor into decision-level fusion.

In addition to differences in spatial sampling, there might also be a difference in data dimensionality, if volume inspection methods like UT or RT are applied in addition to two-dimensional scans. In this thesis, such dimensionality issues did in fact arise, because TT yields a time-varying signal at each spatial location. This issue of data disparateness (sec. 2.3) was resolved here by aggregating the higher-dimensional data.

In other situations, when multiple volumetric signals exist, fusion can in fact be carried out in three dimensions using extensions of the techniques presented in this thesis. In case additional two-dimensional signals are available, such information can still be fused by considering them as constraints to the solution in higher dimensions. For instance, if volumetric fusion is uncertain about an indication (e.g. due to poor SNR), but the two-dimensional signals show a clear indication that corresponds to that area, it could be assumed that all signals indicate the same inhomogeneity and the overall uncertainty would be reduced. Whereas lower-dimensional fusion after data aggregation allows for less complex data processing, easier visualization and interpretation, and might offer better signal quality, a higher-dimensional fusion strategy is less ambiguous, since aggregation merges details. Of course, high-dimensional fusion assumes that all applied inspection techniques are sensitive to the same types of defects. This assumption might not hold, for instance if pure surface-inspection methods are fused with volumetric techniques. In such situations, complementary fusion rules are more appropriate, which were not pursued in this thesis.

Summary and outlook This thesis successfully demonstrated that nondestructive surface inspection based on fundamentally different NDT techniques substantially increases the reliability of fatigue crack detection at the early stages of defect growth. Both at the signal level, as well as the decision level, novel data fusion techniques were developed and applied to real inspection signals. Detailed quantitative evaluation by means of ROC analysis clearly demonstrated the advantages of a multi-sensor system over single-sensor inspection in all tested cases.

For high-level fusion, an independent evaluation on a second test specimen confirmed that by using simple fusion rules over more complex approaches, the developed methods still perform well across different testing conditions. The limitations of complex fusion approaches also became apparent in this thesis. As was seen by example of low-level fusion after directional image transforms like the Shearlet transform, even basic assumptions, such as the elongated shape of cracks, strongly increase the complexity of the signal processing pipeline, but do not necessarily lead to better detection performance compared to undirectional signal representations. Ideas for future research in this direction were given.

This thesis is valuable not only for surface inspection: Because all proposed methods are based on N-dimensional signal representations (Tensor representation, Shearlet coefficients, Kernel Density Estimation), they can be directly extended to volumetric signals as well. In addition, the thesis offers practical hints for setting up a fusion system, and draws attention to potential pitfalls.

So, how has the status of NDT data fusion changed from Gros' book in 2001 [4] to the time of writing this thesis, 15 years later? Gros criticizes a general "skepticism" and "reluctance" about data fusion [4, ch. 1.1], and opposes this attitude by presenting his book itself as a collection of examples for practically relevant fusion applications[6].

[6] "This publication arrives at a time when scepticism and lack of knowledge by decision-making people in the NDT community are reluctant to adopt new ideas. Indeed, it is sometimes believed that the concept of data fusion is a buzzword with no future and substantial applications. The present book not only indicates that data fusion is applied in NDT, but also that it is becoming a major tool in industrial research and development. In 1999, a section exclusively dedicated to data fusion was organised at the famous 'Progress in Quantitative Nondestructive Evaluation' conference. The new journal 'Information Fusion', a scientific magazine aimed at describing theoretical and experimental applications of data fusion, is another evidence of increasing activities in this field." [4, ch. 1.1]

Today, the importance and utility of data fusion in general is unquestionable. However, in the field of NDT, few would actually consider multi-sensor data fusion as a major field of research. In contrast to other domains, where data fusion has become state of the art, NDT in general has not yet made this leap despite constant active research. The biggest challenges to the practical adoption of multi-sensor NDT systems are probably the large associated costs, and the problem of data scarcity. The first challenge is actually only relative, because in the long run, costs might actually be reduced by being able to safely prolong maintenance intervals. Moreover, in safety-critical applications, the costs of missing a dangerous defect by far outweigh the costs of reliable inspection. Data scarcity, however, remains a crucial problem. Although it is definitely possible to collect huge amounts of NDT data, the term *big data* now seems to have taken over the role of the "buzzword" [4, ch. 1.1] that used to be *data fusion*. This comes from the misconception that *big data* promises *big information*. When physically accurate simulations are unavailable, the only way to design a data fusion NDT system and to convincingly demonstrate its reliability is to have access to inspection data of a vast range of materials, defect types and other imperfections. This is practically infeasible for a single organization. Yet, from the methodological point of view, different applications of multi-sensor NDT are readily realizable, as demonstrated in this thesis and all preceding works in this field. The data scarcity problem, however, limits the potential that modern data fusion techniques already offer (machine learning in particular). Moreover, data scarcity legitimately confines the trust that NDT practitioners have in new, complex algorithms. Therefore, to overcome the imbalance between data availability and methodological advances, it is necessary to start a global data sharing initiative in NDT, following the example of the machine learning community. Despite the improving automation of inspection, this requires a change of mentality of the owners of the data, who resist sharing because a lot of their effort and money went into the measurements, and sharing might be seen as additional overhead. Yet, all NDT researchers and practitioners would substantially benefit from openly available and well-documented multi-sensor NDT data sets. Such open data would allow different algorithms to be fairly evaluated on the same data under realistic conditions, and would really give a boost to NDT data fusion.

On this basis, the developments of NDT automation will continue to entice data fusion applications, so that we will see an increasing number of practical implementations of NDT data fusion in the near future.

Appendix A

Appendix

$$a = (x_2^2 - x_3^2)/(x_2 - x_3) \tag{A.1}$$
$$b = (y_2 - y_3)/(x_2 - x_3) \tag{A.2}$$
$$c = (y_1 - y_3 + b(x_3 - x_1))/(x_1^2 - x_3^2 + a(x_3 - x_1)) \tag{A.3}$$
$$d = (y_2 - y_3 - c(x_2^2 - x_3^2))/(x_2 - x_3) \tag{A.4}$$
$$x_{\max} = -0.5\frac{d}{c} \tag{A.5}$$

Equation set 1: Formula for the peak position $x_{\max}$ of a concave quadratic function $y = f(x)$, interpolating three points (x_i, y_i), $i = 1 \ldots 3$

Table A.1: Comparison of specimens used in this work.

name	**Slab**	**Rings *SA*, *SB***	**Vergleichskörper 1**
material	steel	steel (Saarstahl 100Cr6)	steel (90 MnCrV8)
surface dimensions	100 mm x 50 mm	675 mm x 73 mm	outer radius 25 mm, inner radius 5.25 mm
structural discontinuities:			
type	EDM grooves, simulating cracks	EDM grooves, simulating cracks	real stress corrosion cracks
number	10	15 (*SA*), 16 (*SB*)	up tp 5000
lengths	10 mm	1 mm	0.5–10 mm
shallowest depth	10 µm (see table 5.1)	*SA*: 11 µm (see table 6.1) *SB*: 12 µm (see table 6.4)	< 10 µm
comments	produced by BAM	Two industrial parts with identical dimensions and material. Groove depths vary between SA and SB.	specifications according to DIN EN ISO 9934-2_2003; remanent magnetization
appears in sections	5.1.3	*SA*: 6.2.1, *SB*: 5.2.5,6.2.5	5.2.5

Figure A.1: Top: Test image "Zoneplate", which sweeps through all possible spatial frequencies at all directions.
Bottom: Shearlet-filtered image at a single direction, for different decomposition levels. Higher levels represent higher spatial frequencies. Filtering was carried out by setting all shearlet coefficients to zero except those at a fixed direction and detail level, before applying the inverse transform.

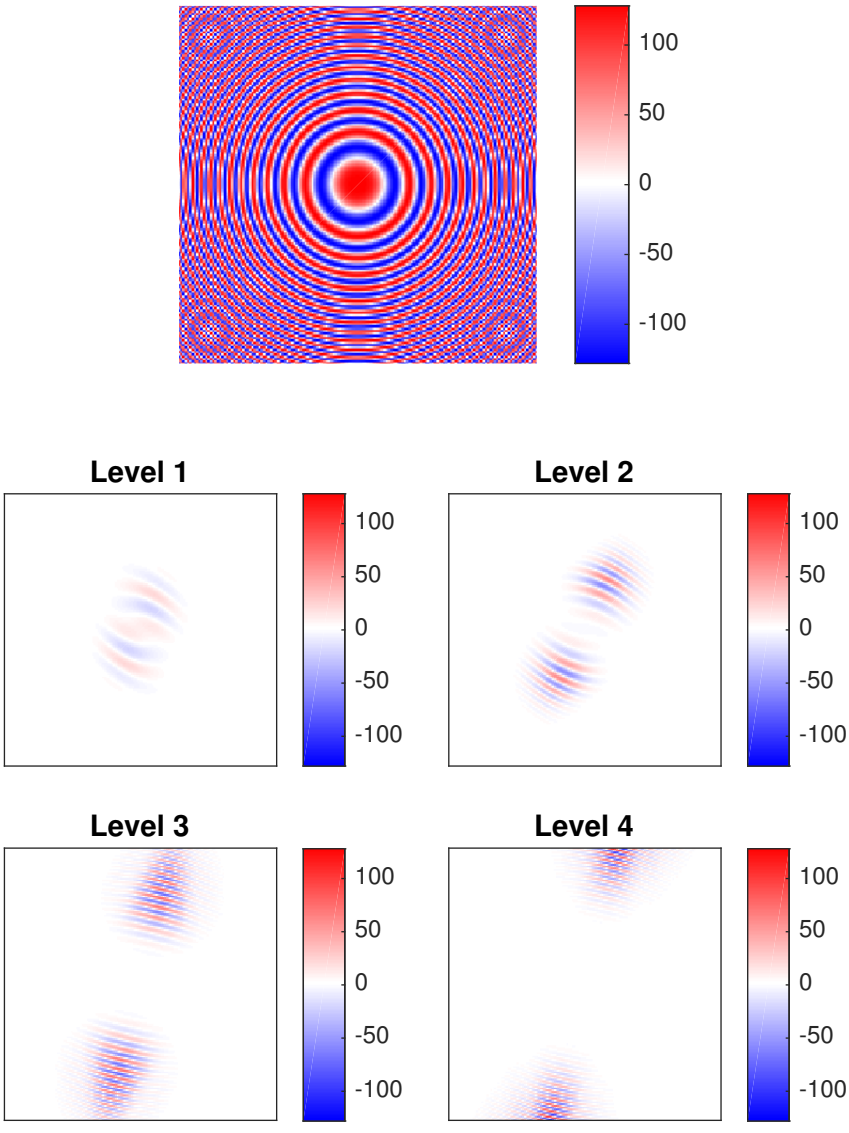


Figure A.2: Evaluation of fusion rules for decomposition method *SWT*.

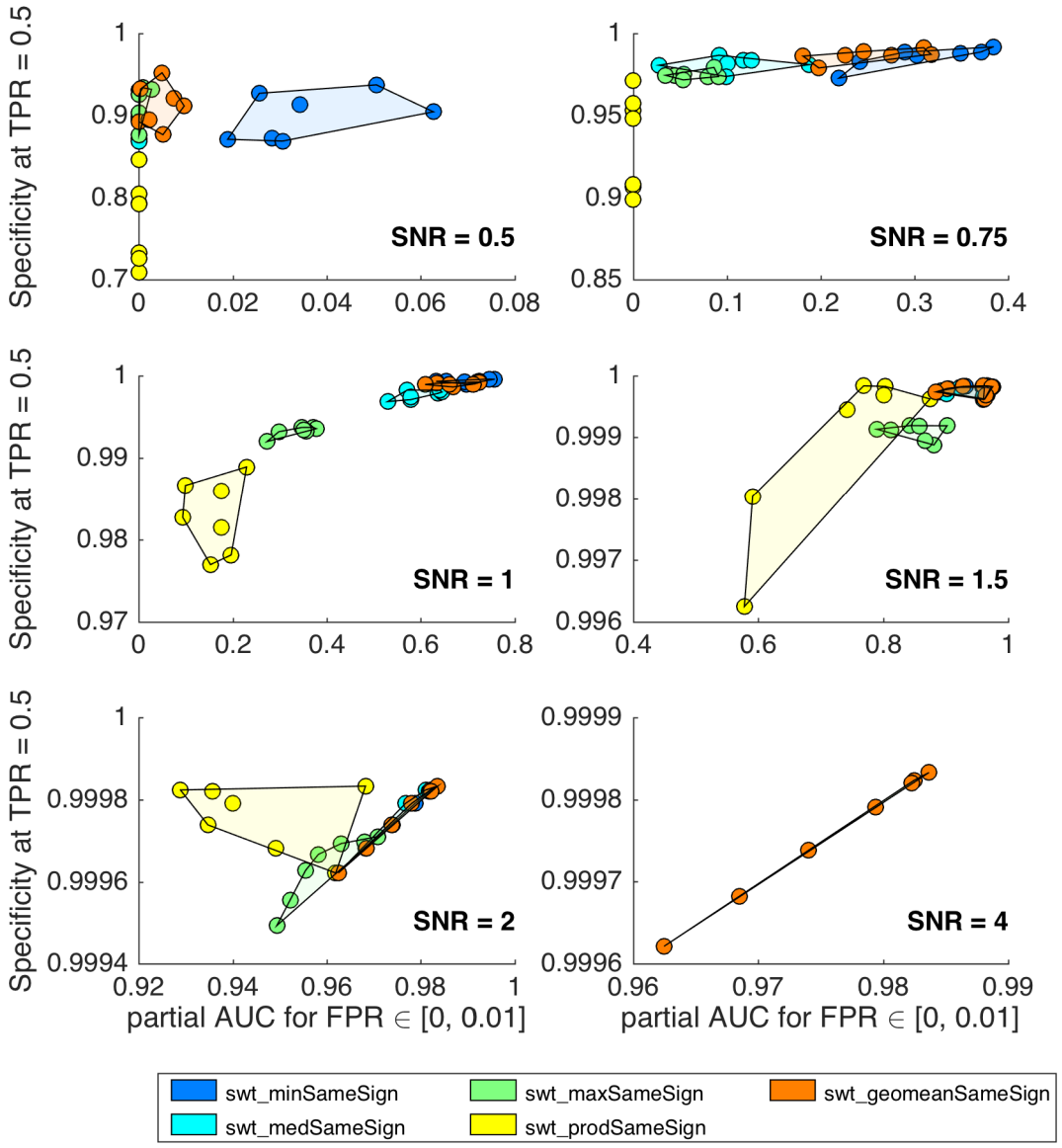


Figure A.3: Evaluation of fusion rules for decomposition method *UWT*.

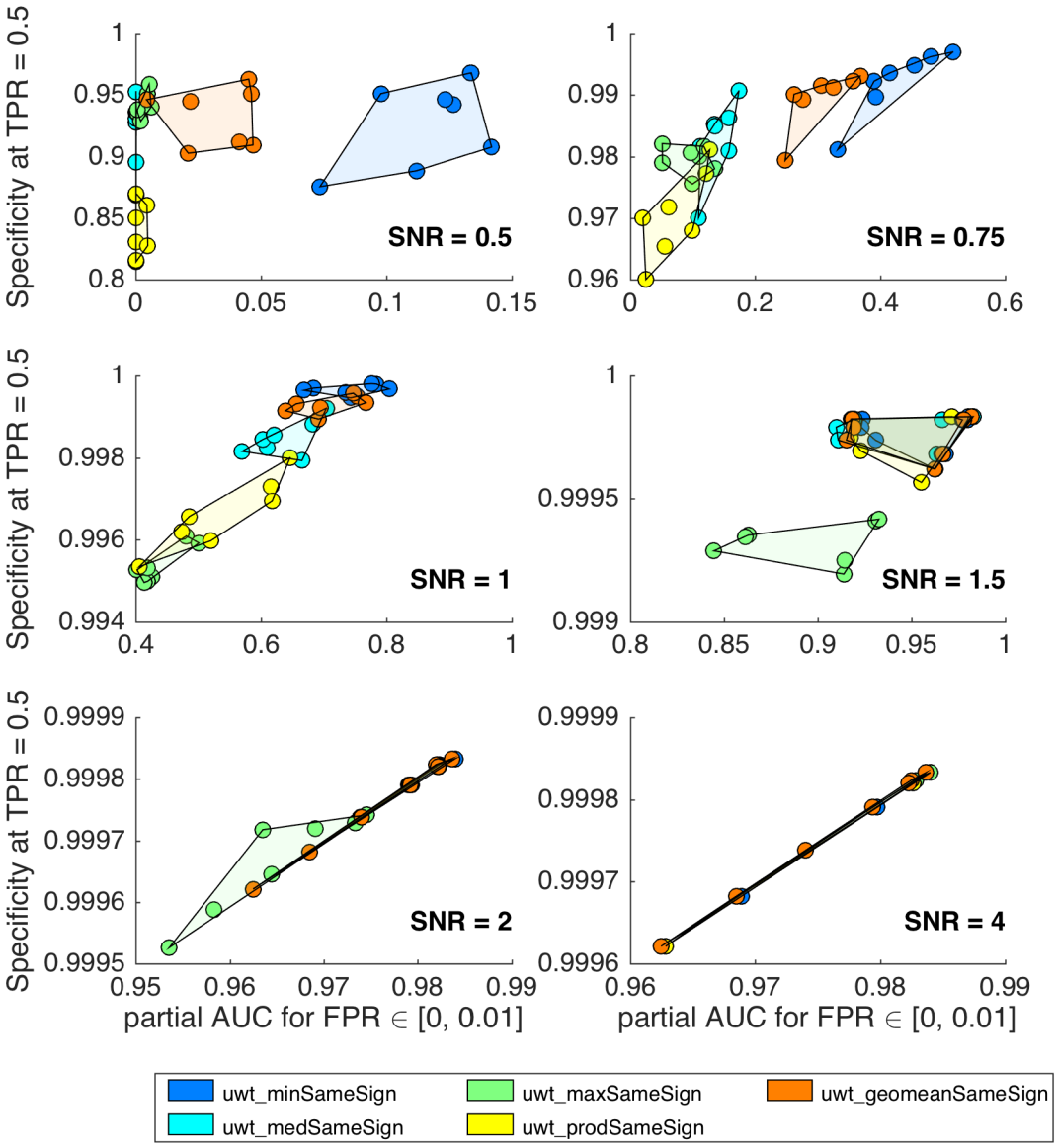


Figure A.4: Evaluation of fusion rules for decomposition method *DTCoWT*. Note that the *product* rule and *geometric mean* are undefined for complex-valued coefficients in this thesis, and are consequently omitted from the plot.

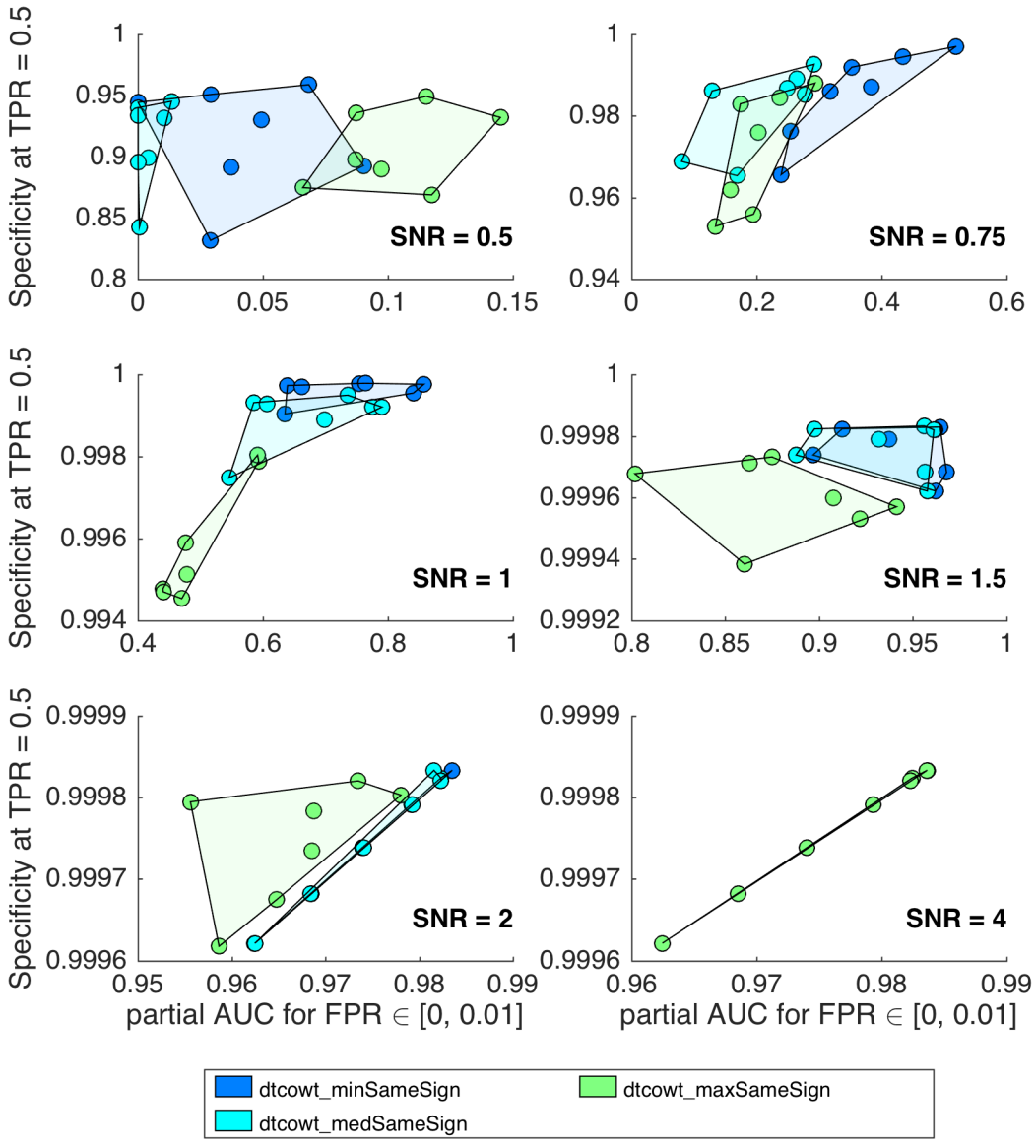


Figure A.5: Evaluation of fusion rules for decomposition method *NSCT*.

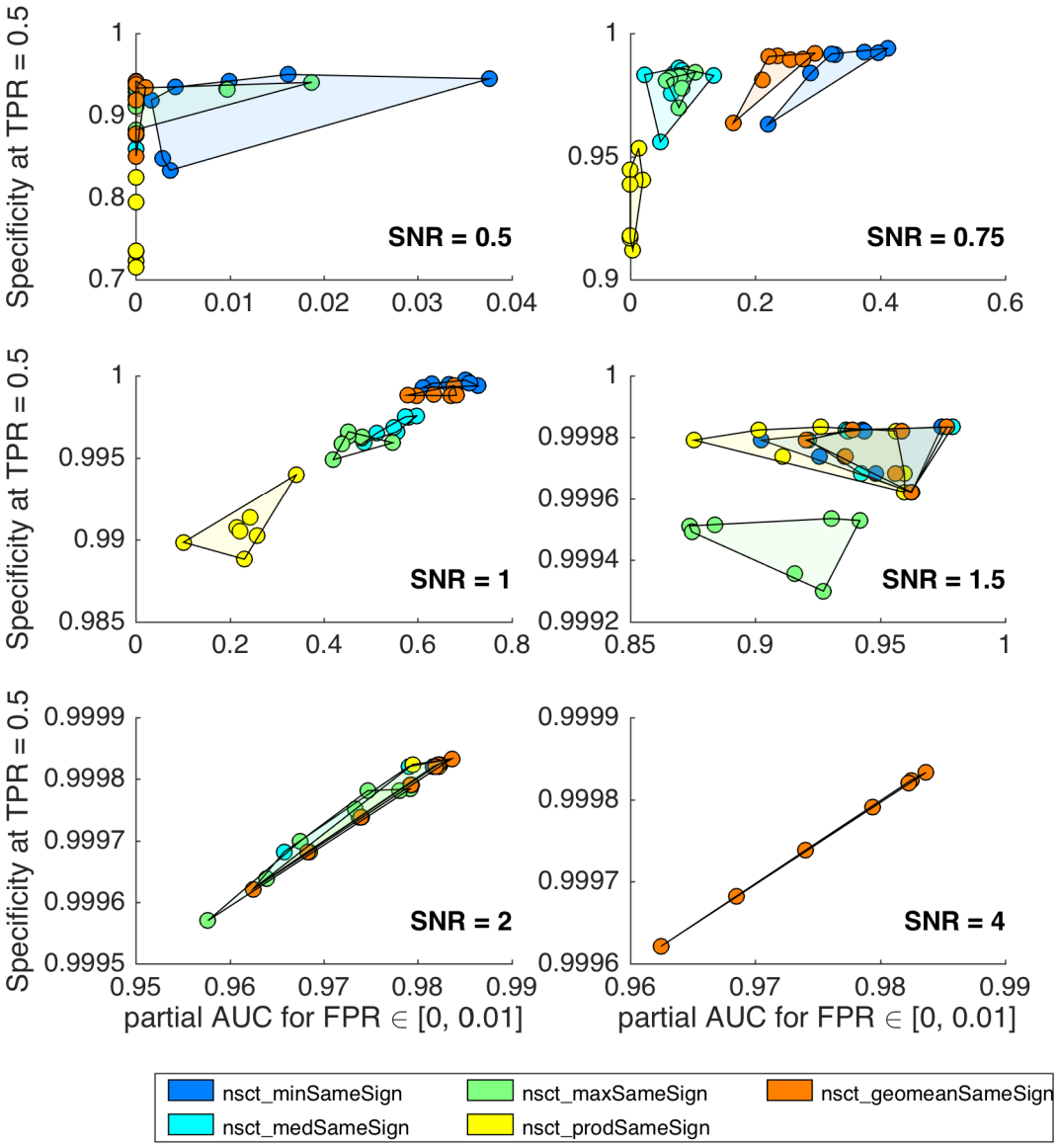


Figure A.6: Evaluation of fusion rules for decomposition method *ST*.

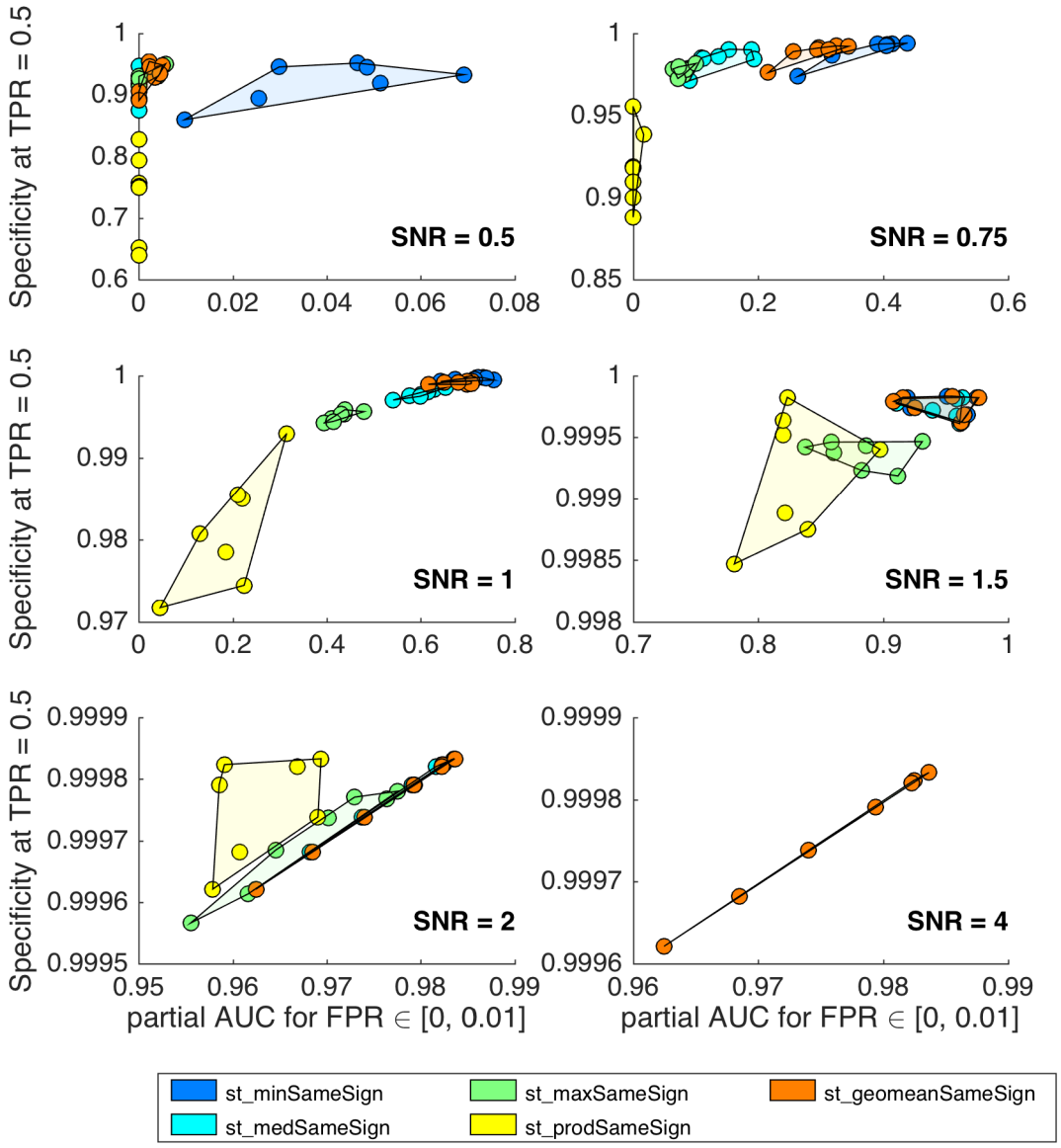


Figure A.7: Evaluation of fusion rules for undirectional (per pixel) fusion. Fusion rules *maximum* and *geometric mean* are not plotted. *Max* is much worse than the other methods and would impair the axis scalings, and *geometric mean* is a monotonic transformation of *product* in the undirectional case, which results in equal ROC curves.

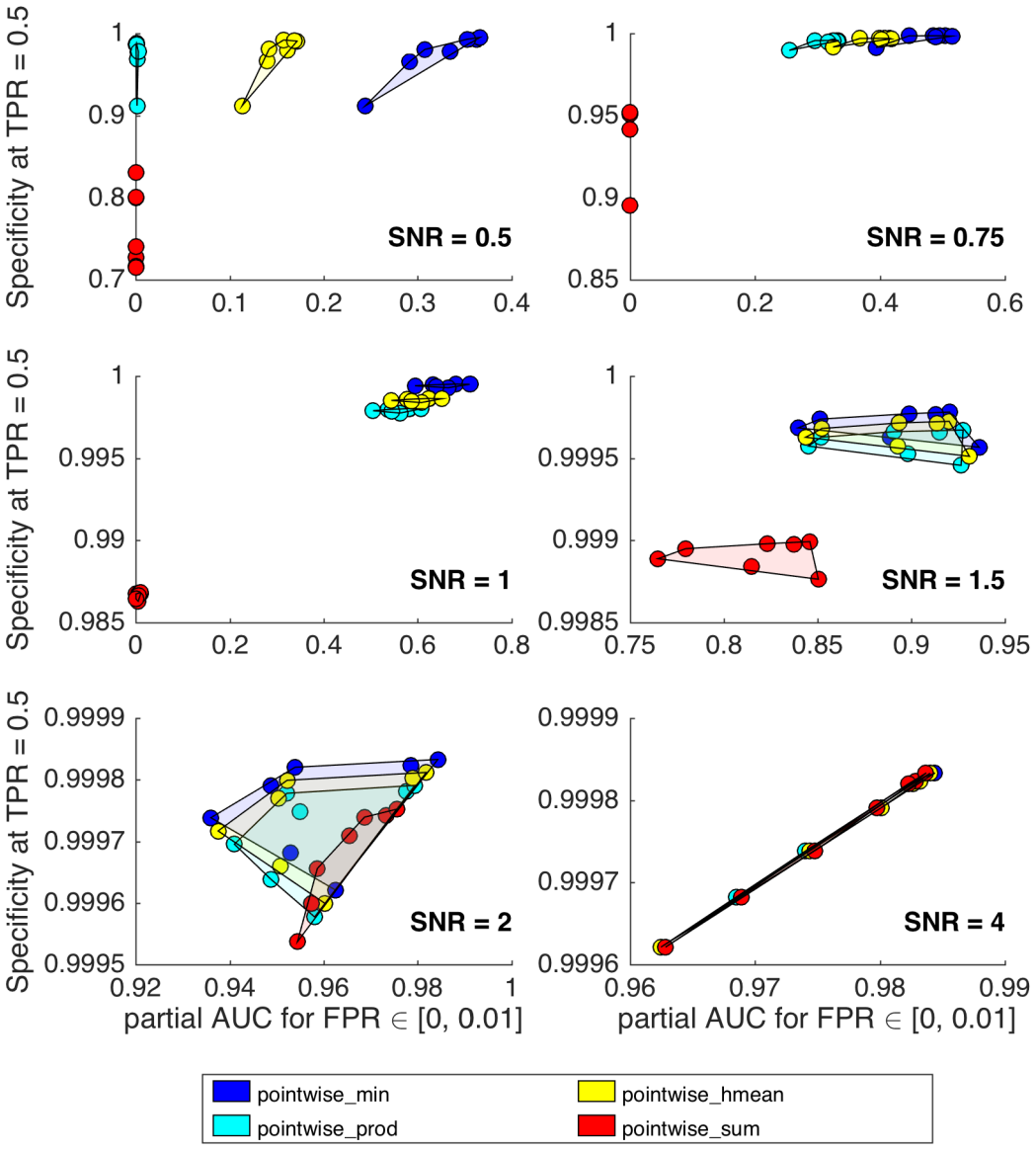


Figure A.8: Evaluation of single sensor performance vs. fusion. At the lowest SNR, only a part of the plot is shown for clarity.

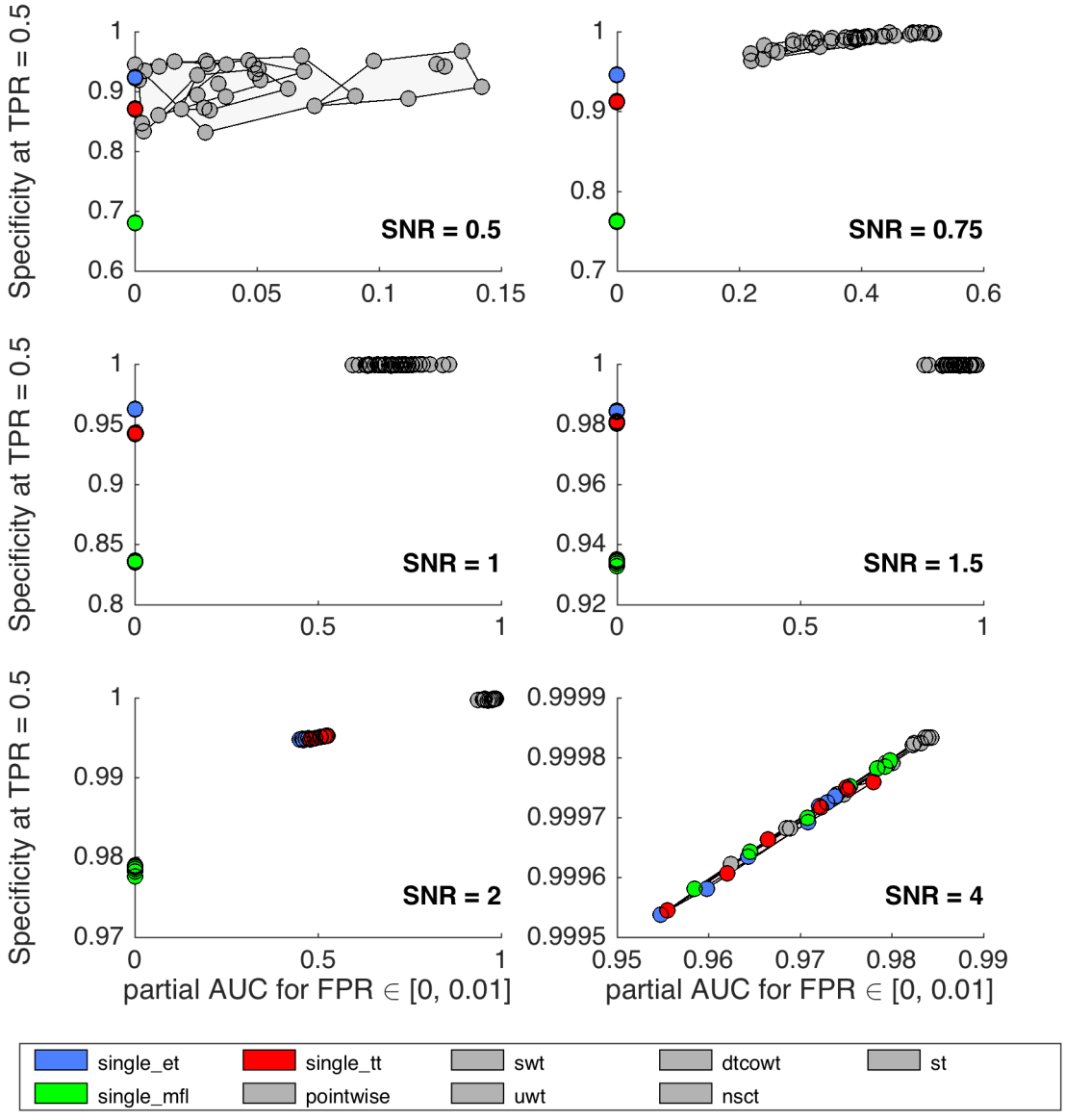


Figure A.9: Evaluation of decomposition methods. Only the best fusion rule is shown per decomposition method. The black arrow indicates *pointwise_min* in each sub-figure. The gray rectangle indicates the region covered by data at the next higher SNR, to visually link the individual plots despite having independent axis scales.

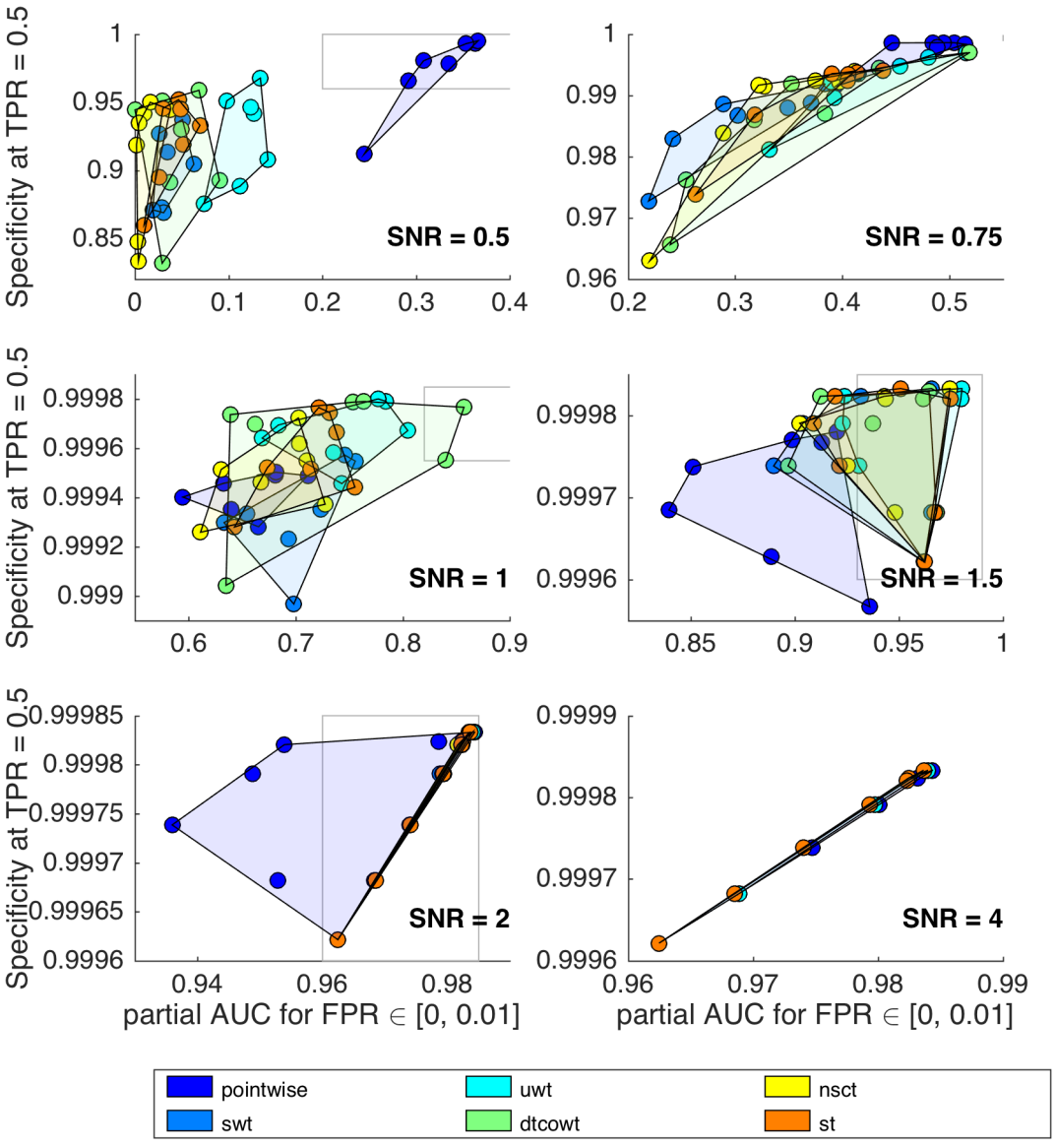


Figure A.10: Evaluation of decomposition methods with the modified fusion approach. Only the best fusion rule is shown per decomposition method. The black arrow indicates *pointwise_min* in each sub-figure. The gray rectangle indicates the region covered by data at the next higher SNR, to visually link the individual plots despite having independent axis scales. Note that in comparison to previous figures A.2–A.8, a different range of SNRs is shown here.

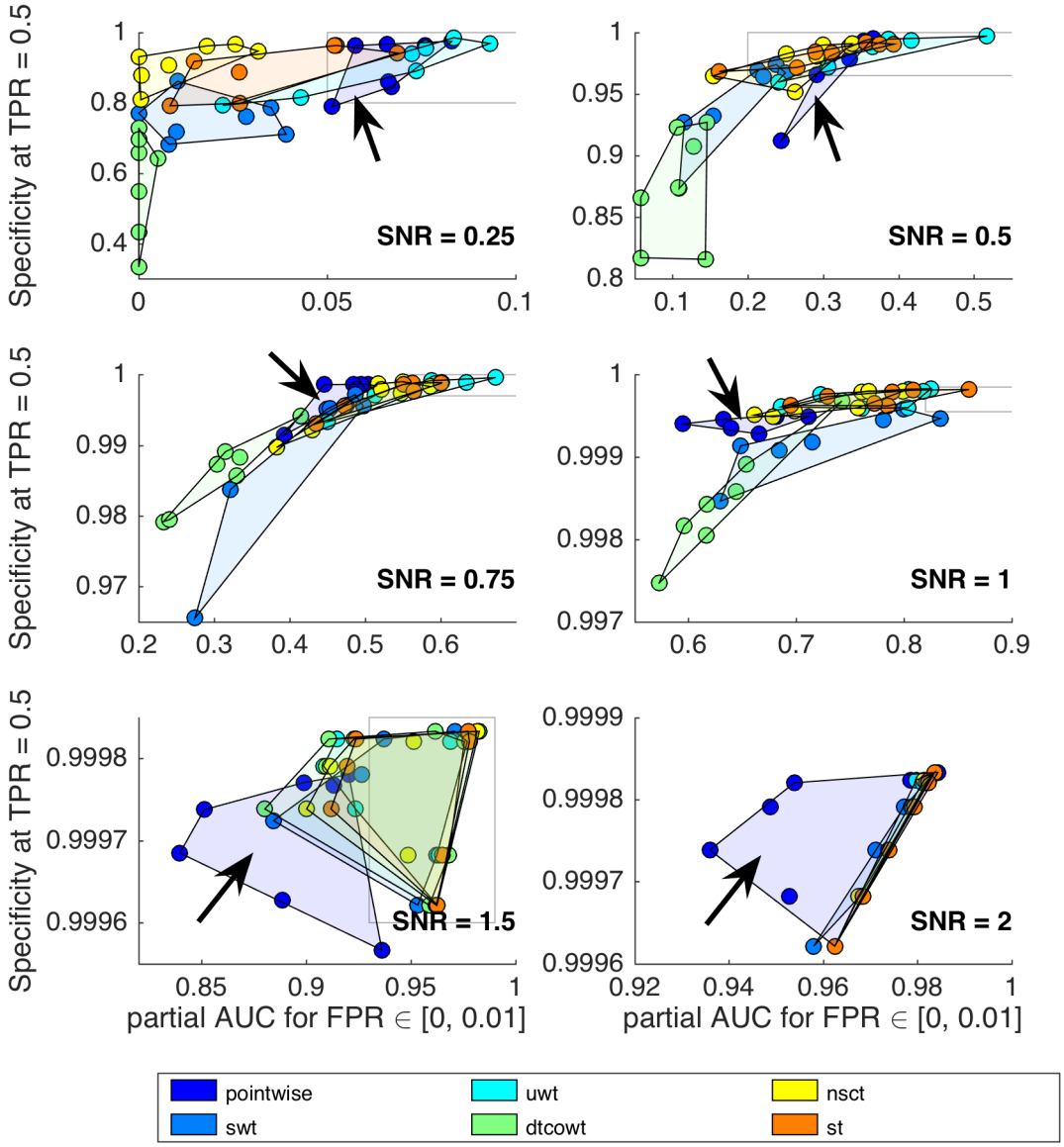


Figure A.11: Influence of registration error (horizontal axes) on detection performance, by **UWT** fusion rule (curves). Each subplot shows the results for a combination of performance measure (partial AUC / specificity at fixed TPR) and SNR (0.5 and 2). In each plot, the vertical axis is scaled to the baseline performance given by the best single sensor (black dashed line).

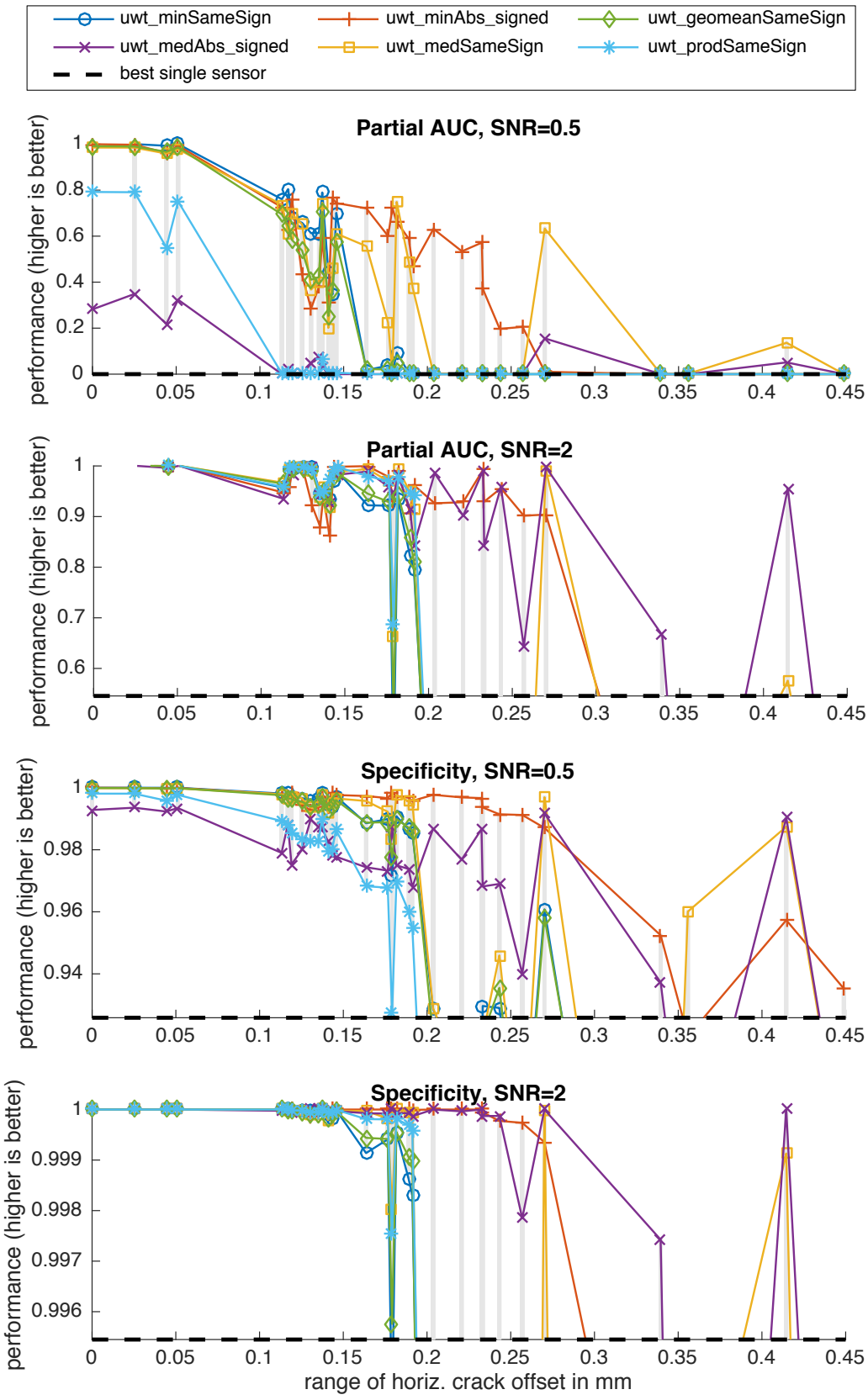
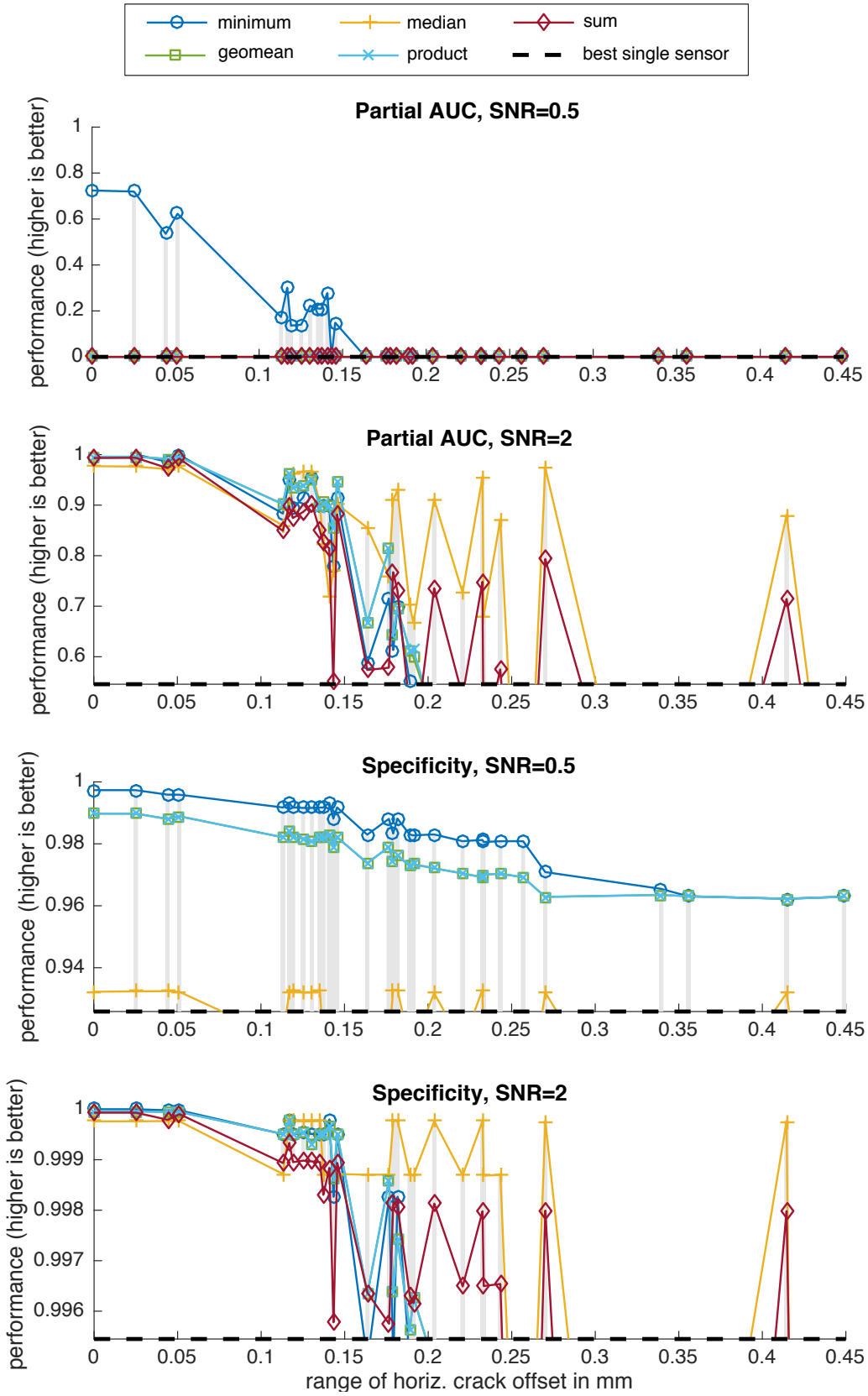


Figure A.12: Influence of registration error (horizontal axes) on detection performance, by **per-pixel** fusion rule (curves). Each subplot shows the results for a combination of performance measure (partial AUC / specificity at fixed TPR) and SNR (0.5 and 2). In each plot, the vertical axis is scaled to the baseline performance given by the best single sensor (black dashed line).



List of Figures

Bibliography

[1] Jean-Luc Starck, Jalal Fadili, and Fionn Murtagh. "The Undecimated Wavelet Decomposition and its Reconstruction". In: *IEEE Transactions on Image Processing* 16.2 (Feb. 2007), pp. 297–309. ISSN: 1057-7149. DOI: 10.1109/TIP.2006.887733.

[2] E07 Committee. *Standard Terminology for Nondestructive Examinations*. en. Standard 1316. ASTM International, 2010.

[3] *North America Electromagnetic NDT Market. Trends & Forecast to 2015–2020*. by Method (Eddy Current Testing, Remote Field Testing, Magnetic Flux Leakage Testing), by Vertical (Oil & Gas, Power Generation, Automotive, Aerospace), by Country (The U.S., Canada, Mexico). SE 3412. Markets and Markets, May 2015. 145 pp.

[4] X. E. [Editor] Gros. *Applications of NDT Data Fusion*. Boston [u.a.]: Kluwer Academic Publ., 2001. ISBN: 0-7923-7412-6.

[5] Thomas Heckel et al. "High speed non-destructive rail testing with advanced ultrasound and eddy-current testing techniques". In: *NDTIP Proceedings, Prague*. NDTIP. 2009.

[6] Kenji Reichling et al. "BETOSCAN–Robot controlled non-destructive diagnosis of reinforced concrete decks". In: *Non-Destructive Testing in Civil Engineering (NDTCE'09)*. NDTCE 2009. 2009.

[7] C Mineo et al. "Robotic non-destructive inspection". In: *51st Annual Conference of the British Institute of Non-Destructive Testing*. 2012, pp. 345–352.

[8] Christos Emmanouilidis, Vasilios Spais, and Kostas Hrissagis. "A mobile robot for automated non-destructive testing of steel plates". In: *Proc. Of the IEEE Mechatronics and Robotics 2004*. IEEE Mechatronics and Robotics 2004. 2004, pp. 871–876.

[9] Volker Deutsch. *NDT: Compact and Understandable: Informative Booklets for Non-destructive Testing*. Vol. 10: The History of NDT-Instrumentation. Castell, 2006. ISBN: 3-934-255-25-6.

[10] Patricia Cotič et al. "Image Fusion for Improved Detection of Near-Surface Defects in NDT-CE Using Unsupervised Clustering Methods". In: *Journal of Nondestructive Evaluation* 33.3 (Sept. 2014), pp. 384–397. ISSN: 0195-9298, 1573-4862. DOI: 10.1007/s10921-014-0232-1.

[11] Vladimir P. Vavilov and Douglas D. Burleigh. "Review of pulsed thermal NDT: Physical principles, theory and data processing". In: *NDT & E International* 73 (July 2015), pp. 28–52. ISSN: 09638695. DOI: 10.1016/j.ndteint.2015.03.003.

[12] D. Balageas et al. "Thermal (IR) and Other NDT Techniques for Improved Material Inspection". In: *Journal of Nondestructive Evaluation* 35.1 (Mar. 2016). ISSN: 0195-9298, 1573-4862. DOI: 10.1007/s10921-015-0331-7.

[13] Javier García-Martín, Jaime Gómez-Gil, and Ernesto Vázquez-Sánchez. "Non-Destructive Techniques Based on Eddy Current Testing". In: *Sensors (Basel, Switzerland)* 11.3 (Feb. 28, 2011), pp. 2525–2565. ISSN: 1424-8220. DOI: 10.3390/s110302525.

[14] V. Reimund et al. “Sensitivity analysis of the non-destructive evaluation of micro-cracks using GMR sensors”. In: *NDT & E International* 64 (June 2014), pp. 21–29. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2014.02.003.

[15] Teng Li, Darryl P. Almond, and D.Andrew S. Rees. “Crack imaging by scanning pulsed laser spot thermography”. In: *NDT & E International* 44.2 (Mar. 2011), pp. 216–225. ISSN: 09638695. DOI: 10.1016/j.ndteint.2010.08.006.

[16] Verena Reimund et al. “Fast defect parameter estimation based on magnetic flux leakage measurements with GMR sensors”. In: *International Journal of Applied Electromagnetics and Mechanics* 37.2 (2011), pp. 199–205. DOI: 10.3233/JAE-2011-1391.

[17] C Annis. “MIL-HDBK-1823A”. In: *Nondestructive Evaluation System Reliability Assessment. Department of Defense Handbook, Wright-Patterson AFB, USA* (2009).

[18] *Workshop: Multiscale Geometric Analysis: Theory, Tools, and Applications*. 2003.

[19] Ralph I Stephens et al. *Metal fatigue in engineering*. John Wiley & Sons, 2000.

[20] E08.02 Committee. *Standard Terminology Relating to Fatigue and Fracture Testing*. en. Standard 1823. ASTM International, 2000.

[21] Dieter Radaj and Michael Vormwald. *Ermüdungsfestigkeit*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. ISBN: 978-3-540-71458-3.

[22] Željko Marušić. “Methods to detect and prevent fatigue in ageing aircraft structures”. In: *Tehnicki vjesnik-Technical Gazette* 22.3 (June 30, 2015), pp. 793–803. ISSN: 13303651, 18486339. DOI: 10.17559/TV-20140702111704.

[23] J. Man, K. Obrtlík, and J. Polák. “Extrusions and intrusions in fatigued metals”. In: *Philosophical Magazine* 89.16 (June 2009), pp. 1295–1336. ISSN: 1478-6435, 1478-6443. DOI: 10.1080/14786430902917616.

[24] PC Paris and Fazil Erdogan. “A critical analysis of crack propagation laws”. In: *Journal of basic engineering* 85.4 (1963), pp. 528–533.

[25] J Schijve. “Fatigue of structures and materials in the 20th century and the state of the art”. In: *International Journal of Fatigue* 25.8 (Aug. 2003), pp. 679–702. ISSN: 01421123. DOI: 10.1016/S0142-1123(03)00051-3.

[26] K. Schiebold and T. Knöll. *NDT: Compact and Understandable: Informative Booklets for Non-destructive Testing*. Vol. 8: Defectoscopy using Eddy Currents. Castell, 2006. ISBN: 3-934-255-27-2.

[27] Matthias Pelkner. “Entwicklung, Untersuchung und Anwendung von GMR-Sensorarrays für die Zerstörungsfreie Prüfung von ferromagnetischen Bauteilen”. Deutsch. Dissertation. Universität des Saarlandes, Sept. 2014.

[28] Matthias Pelkner et al. “Routes for GMR-Sensor Design in Non-Destructive Testing”. In: *Sensors (Basel, Switzerland)* 12.9 (Sept. 5, 2012), pp. 12169–12183. ISSN: 1424-8220. DOI: 10.3390/s120912169.

[29] Vogt Deutsch. *NDT: Compact and Understandable: Informative Booklets for Non-destructive Testing*. Vol. 3: Magnetic Particle Crack Detection. Castell, 2006. ISBN: 3-934-255-25-6.

[30] J. Schlichting et al. “Flying Laser Spot Thermography for the Fast Detection of Surface Breaking Cracks”. In: *Proceedings 18th World Conference on Non-Destructive Testing* (Apr. 2012).

[31] S. E. Burrows et al. “Thermographic detection of surface breaking defects using a scanning laser source”. In: *NDT & E International* 44.7 (Nov. 2011), pp. 589–596. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2011.06.001.

[32] Yun-Kyu An, Ji Min Kim, and Hoon Sohn. "Laser lock-in thermography for fatigue crack detection in an uncoated metallic structure". In: *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2013*. Ed. by Jerome P. Lynch, Chung-Bang Yun, and Kon-Well Wang. Vol. 8692. SPIE-Intl Soc Optical Eng, Apr. 2013, 86921K–86921K–11. DOI: 10.1117/12.2009205.

[33] Teodor Dogaru and Stuart T. Smith. "Giant magnetoresistance-based eddy-current sensor". In: *Magnetics, IEEE Transactions on* 37.5 (2001), pp. 3831–3838.

[34] *Non-destructive testing – Eddy current testing – General principles*. Standard ISO 15549:2008. 2008.

[35] *Non-destructive testing – Qualification and certification of NDT personnel*. Standard ISO 9712. July 2012.

[36] *Non-destructive testing – Magnetic particle testing – Part 1: General principles*. Standard ISO 9934-1. Sept. 2015.

[37] *Non-destructive testing – Magnetic particle testing – Vocabulary*. Standard ISO 12707. Mar. 2016.

[38] *Non-destructive testing – Infrared thermography – Vocabulary*. Standard ISO 10878. Oct. 2013.

[39] *Non-destructive testing – Active thermography*. Standard DIN 54192. Nov. 2010.

[40] Federico Castanedo. "A Review of Data Fusion Techniques". In: *The Scientific World Journal* 2013 (Oct. 27, 2013), e704504. DOI: 10.1155/2013/704504.

[41] H. B. Mitchell. *Data Fusion: Concepts and Ideas*. en. Springer, Feb. 2012. ISBN: 9783642272219.

[42] Bahador Khaleghi et al. "Multisensor data fusion: A review of the state-of-the-art". In: *Information Fusion* 14.1 (Jan. 2013), pp. 28–44. ISSN: 15662535. DOI: 10.1016/j.inffus.2011.08.001.

[43] M. Kumar, D.P. Garg, and R.A. Zachery. "A Method for Judicious Fusion of Inconsistent Multiple Sensor Data". In: *IEEE Sensors Journal* 7.5 (May 2007), pp. 723–733. ISSN: 1530-437X. DOI: 10.1109/JSEN.2007.894905.

[44] Glenn Shafer. *A mathematical theory of evidence*. Vol. 1. Princeton university press Princeton, 1976.

[45] Zdzislaw Pawlak. "Rough sets". In: *International Journal of Computer & Information Sciences* 11.5 (1982), pp. 341–356.

[46] Lotfi A. Zadeh. "The concept of a linguistic variable and its application to approximate reasoning—I". In: *Information sciences* 8.3 (1975), pp. 199–249.

[47] LA Zadeh. "Possibility theory vs. probability theory in decision analysis". In: *Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications, 1977 IEEE Conference on*. IEEE, 1977, pp. 1267–1269. DOI: 10.1109/CDC.1977.271764.

[48] R.C. Luo and M.G. Kay. "A tutorial on multisensor integration and fusion". In: *, 16th Annual Conference of IEEE Industrial Electronics Society, 1990. IECON '90*. Vol. 1. Nov. 1990, pp. 707–722. ISBN: 0-87942-600-4. DOI: 10.1109/IECON.1990.149228.

[49] Barbara Zitová and Jan Flusser. "Image registration methods: a survey". In: *Image and Vision Computing* 21.11 (Oct. 2003), pp. 977–1000. ISSN: 02628856. DOI: 10.1016/S0262-8856(03)00137-9.

[50] Isabelle Bloch and Henri Maître. "Fusion of image information under imprecision". In: *Aggregation and fusion of imperfect information*. Springer, 1998, pp. 189–213.

[51] Lisa Gottesfeld Brown. "A Survey of Image Registration Techniques". In: *ACM Comput. Surv.* 24.4 (Dec. 1992), pp. 325–376. ISSN: 0360-0300. DOI: 10.1145/146370.146374.

[52] David G. Lowe. "Object recognition from local scale-invariant features". In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee, 1999, pp. 1150–1157.

[53] Herbert Bay et al. "Speeded-Up Robust Features (SURF)". In: *Computer Vision and Image Understanding* 110.3 (June 2008), pp. 346–359. ISSN: 10773142. DOI: 10.1016/j.cviu.2007.09.014.

[54] Nicholas Brierley. "The Computational Enhancement of Automated Non-destructive Evaluation". PhD thesis. Imperial College London, Jan. 2014. 258 pp.

[55] John A Nelder and Roger Mead. "A simplex method for function minimization". In: *The computer journal* 7.4 (1965), pp. 308–313.

[56] Jeffrey C. Lagarias et al. "Convergence properties of the Nelder–Mead simplex method in low dimensions". In: *SIAM Journal on optimization* 9.1 (1998), pp. 112–147.

[57] X. E. Gros. *NDT Data Fusion*. Butterworth-Heinemann Ltd, 1996. 205 pp. ISBN: 978-0-340-67648-6.

[58] Zheng Liu et al. "Survey: State of the Art in NDE Data Fusion Techniques". In: *IEEE Transactions on Instrumentation and Measurement* 56.6 (Dec. 2007), pp. 2435–2451. ISSN: 0018-9456. DOI: 10.1109/TIM.2007.908139.

[59] Michael J. Schrapp et al. "Data fusion in neutron and X-ray computed tomography". In: *Journal of Applied Physics* 116.16 (Oct. 28, 2014), p. 163104. ISSN: 0021-8979, 1089-7550. DOI: 10.1063/1.4900515.

[60] M. Friedrich et al. "Miniature Mobile Sensor Platforms for Condition Monitoring of Structures". In: *IEEE Sensors Journal* 9.11 (2009), pp. 1439–1448. ISSN: 1530-437X. DOI: 10.1109/JSEN.2009.2027405.

[61] Christoph Völker and Parisa Shokouhi. "Multi sensor data fusion approach for automatic honeycomb detection in concrete". In: *NDT & E International* 71 (Apr. 2015), pp. 54–60. ISSN: 09638695. DOI: 10.1016/j.ndteint.2015.01.003.

[62] Christoph Völker and Parisa Shokouhi. "Clustering Based Multi Sensor Data Fusion for Honeycomb Detection in Concrete". In: *Journal of Nondestructive Evaluation* 34.4 (Dec. 2015). ISSN: 0195-9298, 1573-4862. DOI: 10.1007/s10921-015-0307-7.

[63] N. Brierley, T. Tippetts, and P. Cawley. "Data fusion for automated non-destructive inspection". en. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 470.2167 (July 8, 2014). ISSN: 1364-5021. DOI: 10.1098/rspa.2014.0167.

[64] S. De et al. "A Comprehensive Multi-Modal NDE Data Fusion Approach for Failure Assessment in Aircraft Lap-Joint Mimics". In: *IEEE Transactions on Instrumentation and Measurement* 62.4 (Apr. 2013), pp. 814–827. ISSN: 0018-9456. DOI: 10.1109/TIM.2013.2240931.

[65] G. Yang et al. "3D EC-GMR sensor system for detection of subsurface defects at steel fastener sites". In: *NDT & E International* 50 (Sept. 2012), pp. 20–28. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2012.04.007.

[66] Sasi Balakrishnan et al. "Development of image fusion methodology using discrete wavelet transform for eddy current images". In: *NDT & E International* 51 (2012), pp. 51–57. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2012.06.006.

[67] Soumya De et al. "A Comprehensive Structural Analysis Process for Failure Assessment in Aircraft Lap-Joint Mimics Using Intramodal Fusion of Eddy Current Data". In: *Research in Nondestructive Evaluation* 23.3 (July 2012), pp. 146–170. ISSN: 0934-9847, 1432-2110. DOI: 10.1080/09349847.2012.660242.

[68] I. Elshafiey, A. Algarni, and M. A. Alkanhal. "Image Fusion Based Enhancement of Nondestructive Evaluation Systems". In: *Image Fusion*. http://www.intechopen.com/books/image-fusion. 2011, pp. 211–236. ISBN: 978-953-307-679-9.

[69] R. Prachetaa and B. P. C. Rao. "Image processing for NDT images". In: *Signal and Image Processing (ICSIP), 2010 International Conference on*. Institute of Electrical and Electronics Engineers (IEEE), 2010, pp. 169–174.

[70] T.G. dos Santos et al. "Data fusion in non destructive testing using fuzzy logic to evaluate friction stir welding". In: *Welding International* 22.12 (Dec. 2008), pp. 826–833. ISSN: 0950-7116, 1754-2138. DOI: 10.1080/09507110802591327.

[71] T. Chady, G. Psuj, and P. Lopato. "Data Fusion of Eddy Current NDT Signals". In: *AIP Conference Proceedings* 975.1 (Feb. 2008), pp. 610–617. ISSN: 0094243X. DOI: doi:10.1063/1.2902718.

[72] Zheng Liu et al. "A Data-Fusion Scheme for Quantitative Image Analysis by Using Locally Weighted Regression and Dempster-Shafer Theory". In: *IEEE Transactions on Instrumentation and Measurement* 57.11 (2008), pp. 2554–2560. ISSN: 0018-9456. DOI: 10.1109/TIM.2008.924933.

[73] R.S. Edwards et al. "Data fusion for defect characterisation using a dual probe system". In: *Sensors and Actuators A: Physical* 144.1 (May 2008), pp. 222–228. ISSN: 09244247. DOI: 10.1016/j.sna.2007.12.020.

[74] X. Ma and A. J. Peyton. "Feature detection and monitoring of eddy current imaging data by means of wavelet based singularity analysis". In: *NDT & E International* 43.8 (Nov. 2010), pp. 687–694. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2010.07.006.

[75] S. Mallat and W.-L. Hwang. "Singularity detection and processing with wavelets". In: *IEEE Transactions on Information Theory* 38.2 (1992), pp. 617–643. ISSN: 0018-9448. DOI: 10.1109/18.119727.

[76] Russell A. Wincheski. *Procedure for Automated Eddy Current Crack Detection in Thin Titanium Plates*. Technical Report NASA/TM-2012-217782, L-20187, NF1676L-15095. This procedure provides the detailed instructions for conducting Eddy Current (EC) inspections of thin (5-30 mils) titanium membranes with thickness and material properties typical of the development of Ultra-Lightweight diaphragm Tanks Technology (ULTT). NASA, Nov. 1, 2012.

[77] D. Bouden and I. Lemahieu. "Use of Blind Deconvolution to Restore Eddy Current Data from Non-Destructive Testing of Defects in Welds". In: *Review of Progress in Quantitative Nondestructive Evaluation*. Ed. by Donald O. Thompson and Dale E. Chimenti. Review of Progress in Quantitative Nondestructive Evaluation 18 A. DOI: 10.1007/978-1-4615-4791-4_95. Springer US, 1999, pp. 743–750. ISBN: 978-1-4613-7170-0 978-1-4615-4791-4.

[78] Cristian Lorenz et al. "Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2D and 3D medical images". In: *CVRMed-MRCAS'97*. Springer, 1997, pp. 233–242.

[79] Tony Lindeberg. "Edge Detection and Ridge Detection with Automatic Scale Selection". In: *International Journal of Computer Vision* 30.2 (Nov. 1, 1998), pp. 117–156. ISSN: 0920-5691, 1573-1405. DOI: 10.1023/A:1008097225773.

[80] René Heideklang and Parisa Shokouhi. "Multi-sensor image fusion at signal level for improved near-surface crack detection". In: *NDT & E International* 71 (Apr. 2015), pp. 16–22. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2014.12.008.

[81] Jong-Sen Lee. "Digital image enhancement and noise filtering by use of local statistics". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1980), pp. 165–168.

[82] X. E. Gros, J. Bousigue, and K. Takahashi. "NDT data fusion at pixel level". In: *NDT & E International* 32.5 (July 1999), pp. 283–292. ISSN: 0963-8695. DOI: 10.1016/S0963-8695(98)00056-5.

[83] Xinsheng Lou and Kenneth A Loparo. "Bearing fault diagnosis based on wavelet transform and fuzzy inference". In: *Mechanical Systems and Signal Processing* 18.5 (Sept. 2004), pp. 1077–1095. ISSN: 0888-3270. DOI: 10.1016/S0888-3270(03)00077-3.

[84] M. Sugeno. "Fuzzy measures and fuzzy integrals: a survey". In: *Fuzzy automata and decision processes*. Ed. by M. M. Gupta, G. N. Saridis, and B. R. Gains. North Holland, Amsterdam, 1977, pp. 89–102.

[85] Krista Amolins, Yun Zhang, and Peter Dare. "Wavelet based image fusion techniques — An introduction, review and comparison". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 62.4 (Sept. 2007), pp. 249–263. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2007.05.009.

[86] Guy P. Nason and Bernard W. Silverman. "The stationary wavelet transform and some statistical applications". In: *LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-* (1995), pp. 281–281.

[87] Zhong Zhang and Rick S. Blum. "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application". In: *Proceedings of the IEEE* 87.8 (1999), pp. 1315–1326.

[88] Gonzalo Pajares and Jesús Manuel de la Cruz. "A wavelet-based image fusion tutorial". In: *Pattern Recognition* 37.9 (Sept. 2004), pp. 1855–1872. ISSN: 00313203. DOI: 10.1016/j.patcog.2004.03.010.

[89] X.H. Zhou, D.K. McClish, and N.A. Obuchowski. *Statistical Methods in Diagnostic Medicine*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN: 978-0-470-31792-1.

[90] Emmanuel J. Candès. "Harmonic Analysis of Neural Networks". In: *Applied and Computational Harmonic Analysis* 6.2 (Mar. 1999), pp. 197–218. ISSN: 1063-5203. DOI: 10.1006/acha.1998.0248.

[91] Emmanuel J. Candès and David L. Donoho. "New tight frames of curvelets and optimal representations of objects with piecewiseC2 singularities". en. In: *Communications on Pure and Applied Mathematics* 57.2 (Feb. 2004), pp. 219–266. ISSN: 0010-3640, 1097-0312. DOI: 10.1002/cpa.10116.

[92] M.N. Do and M. Vetterli. "The contourlet transform: an efficient directional multiresolution image representation". In: *IEEE Transactions on Image Processing* 14.12 (Dec. 2005), pp. 2091–2106. ISSN: 1057-7149. DOI: 10.1109/TIP.2005.859376.

[93] Demetrio Labate et al. "Sparse multidimensional representation using shearlets". In: *Wavelets XI*. Ed. by Manos Papadakis, Andrew F. Laine, and Michael A. Unser. Vol. 5914. SPIE-Intl Soc Optical Eng, Aug. 2005, 59140U–59140U–9. DOI: 10.1117/12.613494.

[94] Gitta Kutyniok, Wang-Q. Lim, and Rafael Reisenhofer. "ShearLab 3D: Faithful Digital Shearlet Transforms based on Compactly Supported Shearlets". In: *arXiv:1402.5670 [math]* (Feb. 2014). arXiv: 1402.5670.

[95] Gitta Kutyniok and Wang-Q Lim. "Compactly supported shearlets are optimally sparse". In: *Journal of Approximation Theory* 163.11 (Nov. 2011), pp. 1564–1589. ISSN: 0021-9045. DOI: 10.1016/j.jat.2011.06.005.

[96] Chang Duan et al. "Remote Sensing Image Fusion Based On IHS and Dual Tree Compactly Supported Shearlet Transform". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 7.5 (2014), pp. 361–374.

[97] Gitta Kutyniok and Demetrio Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Springer Science & Business Media, Mar. 7, 2012. 346 pp. ISBN: 978-0-8176-8316-0.

[98] Andreas Ellmauthaler, Carla L. Pagliari, and Eduardo A. B. da Silva. "Multiscale Image Fusion Using the Undecimated Wavelet Transform With Spectral Factorization and Nonorthogonal Filter Banks". In: *IEEE Transactions on Image Processing* 22.3 (Mar. 2013), pp. 1005–1017. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2012.2226045.

[99] Qi-guang Miao et al. "A novel algorithm of image fusion using shearlets". In: *Optics Communications* 284.6 (Mar. 15, 2011), pp. 1540–1547. ISSN: 0030-4018. DOI: 10.1016/j.optcom.2010.11.048.

[100] Chengzhi Deng, Shengqian Wang, and Xi Chen. "Remote Sensing Images Fusion Algorithm Based on Shearlet Transform". In: *International Conference on Environmental Science and Information Application Technology, 2009. ESIAT 2009*. International Conference on Environmental Science and Information Application Technology, 2009. ESIAT 2009. Vol. 3. July 2009, pp. 451–454. DOI: 10.1109/ESIAT.2009.222.

[101] Yong Chai, You He, and Changwen Qu. "Remote sensing image fusion based on iterative discrete Shearlet transform". In: *Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications)* 47.3 (2011), pp. 174–176.

[102] Li Lü, Jia Zhao, and Hui Sun. "Multi-focus image fusion based on shearlet and local energy". In: *2010 2nd International Conference on Signal Processing Systems (ICSPS)*. 2010 2nd International Conference on Signal Processing Systems (ICSPS). Vol. 1. July 2010, pp. V1–632–V1–635. DOI: 10.1109/ICSPS.2010.5555456.

[103] Jia Zhao, Li Lü, and Hui Sun. "A Novel Multi-Focus Image Fusion Method Using Shearlet Transform". In: *Advanced Materials Research*. Vol. 121. Trans Tech Publ. 2010, pp. 373–378.

[104] Yuan Cao, Shutao Li, and Jianwen Hu. "Multi-focus Image Fusion by Nonsubsampled Shearlet Transform". In: *2011 Sixth International Conference on Image and Graphics (ICIG)*. 2011 Sixth International Conference on Image and Graphics (ICIG). Aug. 2011, pp. 17–21. DOI: 10.1109/ICIG.2011.37.

[105] Weiwei Kong. "Technique for gray-scale visual light and infrared image fusion based on non-subsampled shearlet transform". In: *Infrared Physics & Technology* 63 (Mar. 2014), pp. 110–118. ISSN: 1350-4495. DOI: 10.1016/j.infrared.2013.12.016.

[106] Lei Wang, Bin Li, and Lian-fang Tian. "Multi-modal medical image fusion using the inter-scale and intra-scale dependencies between image shift-invariant shearlet coefficients". In: *Information Fusion*. Special Issue on Information Fusion in Medical Image Computing and Systems 19 (Sept. 2014), pp. 20–28. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2012.03.002.

[107] Lei Wang, Bin Li, and Lianfang Tian. "Multimodal Medical Volumetric Data Fusion Using 3-D Discrete Shearlet Transform and Global-to-Local Rule". In: *IEEE Transactions on Biomedical Engineering* 61.1 (Jan. 2014), pp. 197–206. ISSN: 0018-9294. DOI: 10.1109/TBME.2013.2279301.

[108] Ke Xu, Shunhua Liu, and Yonghao Ai. "Application of Shearlet transform to classification of surface defects for metals". In: *Image and Vision Computing* 35 (Mar. 2015), pp. 23–30. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2015.01.001.

[109] Chengdong Wu et al. "Pavement image denoising based on shearlet treansform". In: *2011 International Conference on Electronics and Optoelectronics (ICEOE)*. 2011 International Conference on Electronics and Optoelectronics (ICEOE). Vol. 3. July 2011, pp. V3–262–V3–265. DOI: 10.1109/ICEOE.2011.6013354.

[110] Bettina Heise et al. "Full-field optical coherence microscopy with a sub-nanosecond supercontinuum light source for material research". In: *Optical Fiber Technology*. Fiber Supercontinuum sources and their applications 18.5 (Sept. 2012), pp. 403–410. ISSN: 1068-5200. DOI: 10.1016/j.yofte.2012.07.011.

[111] Sören Häuser and Gabriele Steidl. "Fast finite shearlet transform: a tutorial". In: *arXiv preprint arXiv:1202.1773* (2012).

[112] Ronald R. Coifman and David L. Donoho. *Translation-invariant de-noising*. Springer, 1995. ISBN: 0-387-94564-4.

[113] Veniamin I. Morgenshtern and Helmut Bölcskei. *A short course on frame theory, E.* Chapter in Mathematical Foundations for Signal Processing, Communications, and Networking: CRC Press, 2012.

[114] Ivan W. Selesnick, Richard G. Baraniuk, and Nick C. Kingsbury. "The dual-tree complex wavelet transform". In: *IEEE signal processing magazine* 22.6 (2005), pp. 123–151.

[115] Nick G. Kingsbury. "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters". In: *IEEE Digital Signal Processing Workshop*. Vol. 86. Citeseer, 1998, pp. 120–131.

[116] A.L. Da Cunha, J. Zhou, and M.N. Do. "The Nonsubsampled Contourlet Transform: Theory, Design, and Applications". In: *IEEE Transactions on Image Processing* 15.10 (Oct. 2006), pp. 3089–3101. ISSN: 1057-7149. DOI: 10.1109/TIP.2006.877507.

[117] V.M. Patel, G.R. Easley, and D.M. Healy. "Shearlet-Based Deconvolution". In: *IEEE Transactions on Image Processing* 18.12 (Dec. 2009), pp. 2673–2685. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2009.2029594.

[118] Israa Amro et al. "A survey of classical methods and new trends in pansharpening of multispectral images". In: *EURASIP Journal on Advances in Signal Processing* 2011.1 (Sept. 30, 2011), p. 79. ISSN: 1687-6180. DOI: 10.1186/1687-6180-2011-79.

[119] M. Costantini, Alfonso Farina, and F. Zirilli. "The fusion of different resolution SAR images". In: *Proceedings of the IEEE* 85.1 (Jan. 1997), pp. 139–146. ISSN: 0018-9219. DOI: 10.1109/5.554214.

[120] Tao Wu, Xiao-Jun Wu, and Xiao-Qing Luo. "A Study on Fusion of Different Resolution Images". In: *Procedia Engineering* 29 (2012), pp. 3980–3985. ISSN: 18777058. DOI: 10.1016/j.proeng.2012.01.605.

[121] ISO. *Non-destructive testing – Magnetic particle testing – Part 2: Detection media*. ISO 9934-2:2002. Geneva, Switzerland: International Organization for Standardization, 2002.

[122] Inc. The MathWorks. *MATLAB R2015b, Natick, Massachusetts, United States.*

[123] Eduardo da Silva. *Undecimated Wavelet Transform for MATLAB.*

[124] Arthur Cunha, Jianping Zhou, and Minh Do. *Nonsubsampled Contourlet Toolbox for MATLAB.*

[125] S. Avidan and A. Shamir. "Seam carving for content-aware image resizing". In: *ACM Transactions on Graphics (TOG)*. Vol. 26. 2007, p. 10.

[126] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[127] Bert Vandeghinste et al. "Iterative CT reconstruction using shearlet-based regularization". In: *IEEE Transactions on Nuclear Science* 60.5 (2013), pp. 3305–3317.

[128] Weihong Guo, Jing Qin, and Wotao Yin. "A new detail-preserving regularization scheme". In: *SIAM Journal on Imaging Sciences* 7.2 (2014), pp. 1309–1334.

[129] Bin Yang and Shutao Li. "Pixel-level image fusion with simultaneous orthogonal matching pursuit". In: *Information Fusion* 13.1 (Jan. 2012), pp. 10–19. ISSN: 15662535. DOI: 10.1016/j.inffus.2010.04.001.

[130] Joseph Moysan et al. "Improvement of the non-destructive evaluation of plasma facing components by data combination of infrared thermal images". In: *NDT & E International* 40.6 (Sept. 2007), pp. 478–485. ISSN: 0963-8695. DOI: 10.1016/j.ndteint.2007.02.003.

[131] René Heideklang and Parisa Shokouhi. "Decision-Level Fusion of Spatially Scattered Multi-Modal Data for Nondestructive Inspection of Surface Defects". In: *Sensors* 16.1 (Jan. 15, 2016), p. 105. DOI: 10.3390/s16010105.

[132] K. Fukunaga and L. Hostetler. "The estimation of the gradient of a density function, with applications in pattern recognition". In: *IEEE Transactions on Information Theory* 21.1 (Jan. 1975), pp. 32–40. ISSN: 0018-9448. DOI: 10.1109/TIT.1975.1055330.

[133] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proc. 2nd int. Conf. on Knowledge Discovery and Data Mining*. KDD '96. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[134] Mihael Ankerst et al. "OPTICS: Ordering Points to Identify the Clustering Structure". In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. SIGMOD '99. New York, NY, USA: ACM, 1999, pp. 49–60. ISBN: 1-58113-084-8. DOI: 10.1145/304182.304187.

[135] Y. Weiss. "Segmentation using eigenvectors: a unifying view". In: *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999*. The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999. Vol. 2. 1999, 975–982 vol.2. DOI: 10.1109/ICCV.1999.790354.

[136] Emanuel Parzen. "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3 (Sept. 1962), pp. 1065–1076. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177704472.

[137] Murray Rosenblatt. "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3 (Sept. 1956), pp. 832–837. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177728190.

[138] Bernard W Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986.

[139] Nils-Bastian Heidenreich, Anja Schindler, and Stefan Sperlich. "Bandwidth selection for kernel density estimation: a review of fully automatic selectors". In: *AStA Advances in Statistical Analysis* 97.4 (Oct. 2013), pp. 403–433. ISSN: 1863-8171, 1863-818X. DOI: 10.1007/s10182-013-0216-y.

[140] Josef Kittler et al. "On combining classifiers". In: *IEEE transactions on pattern analysis and machine intelligence* 20.3 (1998), pp. 226–239.

[141] Matthias Pelkner et al. "Automated inspection of surface breaking cracks using GMR sensor arrays". In: *AIP Conference Proceedings*. 40TH ANNUAL REVIEW OF PROGRESS IN QUANTITATIVE NONDESTRUCTIVE EVALUATION: Incorporating the 10th International Conference on Barkhausen Noise and Micromagnetic Testing. Vol. 1581. AIP Publishing, Feb. 18, 2014, pp. 1393–1399. DOI: 10.1063/1.4864984.

[142] The MathWorks, Inc. MATLAB R2015b. *Wiener2*.

[143] Alexander Ihler. *Kernel Density Estimation Toolbox for MATLAB*.

[144] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

[145] Valerie Cross. "Fuzzy information retrieval". In: *Journal of Intelligent Information Systems* 3.1 (1994), pp. 29–56.

[146] Christina MUELLER et al. "Holistically Evaluating the Reliability of NDE Systems–Paradigm Shift". In: *18th World Conference on Non Destructive Testing* (Apr. 16, 2012).

Selbstständigkeitserklärung

Ich erkläre hiermit, dass

- ich die vorliegende Dissertationsschrift "Data Fusion for Multi-Sensor Nondestructive Detection of Surface Cracks in Ferromagnetic Materials" selbstständig, ohne unerlaubte Hilfe und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe,
- ich mich nicht anderwärts um einen Doktorgrad beworben habe oder einen solchen besitze, und
- mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II der Humboldt-Universität zu Berlin bekannt ist, gemäß Amtl. Mitteilungsblatt Nr. 34/2006.

Berlin, den 29.08.2017 René Heideklang