Defect Detection in Patterned Silicon Wafers Using Anisotropic Kernels

Maria Zontak

Defect Detection in Patterned Silicon Wafers Using Anisotropic Kernels

Research Thesis

In Partial Fulfillment of the Requirements for
the Degree Master of Science in Electrical Engineering

Maria Zontak

Submitted to the Senate of the Technion—Israel Institute of Technology

Tamuz 5768 Haifa July 2008

The Research Thesis Was Done Under The Supervision of Associate Professor Israel Cohen in the Department of Electrical Engineering.

Acknowledgement

I am grateful to Prof. Israel Cohen for his guidance throughout all stages of this research.

I would also like to thank Gil Sod-Moriah from Applied Materials INC., Rehovot, for his constructive comments and helpful suggestions.

Special thanks to my husband Yuri Pekelny for his love and emotional and technical support.

THE GENEROUS FINANCIAL HELP OF THE TECHNION AND APPLIED MATERIALS INC., REHOVOT, ISRAEL, IS GRATEFULLY ACKNOWLEDGED

Contents

List of Figures

Abstract

This work is focused on the application of anisotropic kernels to defect detection in patterned wafers using Scanning Electron Microscope (SEM) images. Defect detection is a critical component of wafers manufacturing process. Various image processing techniques have been applied to automatic defect detection in wafers, which rely on accurate image registration of source and reference images obtained from neighboring dies. Unfortunately, perfect registration is generally impossible, due to pattern variations between the source and reference images.

In this work, we propose a defect detection procedure for a single image, which avoids image registration and is robust to pattern variations. The proposed method is based on anisotropic kernel reconstruction of the source image using the reference image. The source and reference images are mapped into a feature space, where every feature with origin in the source image is estimated by a weighted sum of neighboring features from the reference image. The set of neighboring features is determined according to the spatial neighborhood in the original image space, and the weights are calculated from exponential distance similarity function. We show that features originating from defect regions are not reconstructible from the reference image, and hence can be identified.

We develop a kernel-based approach to multi-channel defect detection, which relies on the physical similarity relations between multi-channel images

acquired by an SEM tool. We assume that the similarity between pattern-originated regions from the inspected and reference wafers is maintained across the three SEM channels and develop a multi-channel defect detection algorithm. We show that in case of pattern variations, the false detection rate can be significantly reduced by using our kernel-based approach rather than existing methods. We also demonstrate the improved performance of the multi-channel approach in case of non-periodic patterned wafers, compared to using a single-channel approach.

Notations of Channels

External$_1$ Image	indicates the topography of the sample by light and shadows, as if the "light source" is directed to the sample from top-right direction
External$_2$ Image	indicates the topography of the sample by light and shadows, as if the "light source" is directed to the sample from top-left direction
Internal Image	provides information about edges and material of the sample

Abbreviations

DCT	Discrete Cosine Transform
FAR	False Alarm Rate
GLRT	Generalized Likelihood Ratio Test
GPU	graphics processing unit
i.i.d.	independently and identically distributed
MSD	Matched Subspace Detector
NL-means	non local means
PDF	Probability Density Function
ROC	Receiver Operating Characteristics
RX	Reed & Xiaoli detection algorithm
SEM	Scanning Electron Microscope
SHT	Single Hypothesis Testing
SVM	Support Vector Machine

List of Alphabetic Symbols

$\vec{1}_m$	m-dimensional column vector with all the components equal to m^{-1}
$\mathcal{A}$	Set of defect regions
$\mathbf{A}$	Whitening matrix
B	Defect mask
b_ε	Normal density with zero mean and standard deviation ε
$\mathbb{D}$	Diffusion matrix (symmetric positive definite)
$D = I_{\text{ref}} - I_{\text{src}}$	Difference image between the source and aligned reference images
$\mathbf{D}$	Normalization of the weight matrix W
$\mathbf{D}(\mathbf{s})$	Normalizing factor of the similarity weights
$\mathbf{d}^2$	Squared Mahalanobis distance
d	Dimension of a feature space
$E(\cdot)$	Expectancy
$\mathbb{E}$	Estimation operator
f	Real valued bounded function
$G = (g_1, ..., g_d)$	Filter bank
$\tilde{G} = (\tilde{g}_1, ..., \tilde{g}_d)$	Another filter bank
$G(\cdot, \cdot, \sigma), G_\sigma$	Gaussian smoothing kernel (with standard deviation σ)
g	Noisy version of h

H_0	Hypothesis, which indicates absence of anomaly (defect)
H_1	Hypothesis, which indicates presence of anomaly (defect)
h	Noise-free image
$\mathbb{I}$	Identity matrix
I	Image
I_{pat}	Noise-free pattern image
I_{ref}	Reference image
I_{src}	Inspected (source) image
$\hat{I}_{\text{src}}$	Reconstructed source image I_{src}
J_{xy}	Joint similarity measure of two channels, likelihood between the *within-similarity maps* of the patches x and y
J_{xyz}	Joint similarity measure of three channels, likelihood between the *within-similarity maps* of the patches x, y and z
K	Filter of the graph G
$k(\cdot,\cdot)$	Kernel (affinity function)
$\mathbf{k}_j$	A set of N i.i.d. sample vectors having probability density function $p(\cdot\|H_0)$
L_{xyz}	Log-likelihood, positive logarithm of joint similarity measure
$\mathbf{M}$	Expected vector of the vectors $\mathbf{X}$
$\mathcal{M}$	General Riemannian manifold
M	Embedded surface in $\mathbb{R}^3$
m	Number of the reference features (patches), search region size
$N(\mu,\Gamma)$	Normal distribution with mean μ and covariance matrix Γ
$\mathcal{N}_{\mathbf{s}}$	Neighborhood of the pixel $\mathbf{s}$ (usually in a feature space)
N	Number of members in a general data set.
n	Number of pixels in the image

$n(\mathbf{s})$	Noise perturbation at a pixel $\mathbf{s}$
$n_k(\mathbf{s})$	Set of k nearest neighbors of $\mathbf{s}$
$n_{\mathbf{q}}$	Dimension of vector $\mathbf{q}$
$n_{\mathbf{X}}$	Length of vector $\mathbf{X}$
O	Rotation in d dimensions
P_{FA}	Probability of false-alarm (false detection)
$\Pr(\cdot)$	Probability of an occurrence
$p(\cdot)$	Probability density
$\hat{p}_{\varepsilon}(\cdot), \hat{p}(\cdot)$	Parzen's estimator of the probability density $p(\cdot)$
$p(\cdot\|H_0)$	Conditional probability density function (PDF) under H_0
$p(\cdot\|H_1)$	Conditional probability density function (PDF) under H_1
$p_{\mathbf{d}^2}(\zeta)$	PDF of $\mathbf{d}^2$ under H_0
$\mathbf{q}$	General feature vector
$\mathbf{r}$	Translation vector
$r(\mathbf{s})$	Pattern variations perturbation under H_0
	or a defect perturbation under H_1 at a pixel $\mathbf{s}$
$\mathbf{s} = (i, j)$	Index of a pixel in the image (source image)
$\mathbf{s}'$	Index of a pixel in the reference image
$[s_x \times s_y]$	Patch's size
t	Time parameter (in diffusion process)
$\mathbf{u}$	Step function
u	Unknown variable of the heat equation
$u(i, j, t)$	Diffusion solution at a pixel (i, j) after t iterations
V	Set of vertices
Var	Variance
$W(\cdot, \cdot)$	Weighting matrix of the connectivity of graph vertices x and y
$W(\mathbf{s}, \mathbf{s}')$	Similarity weight between the source feature, related
	to the pixel $\mathbf{s}$, and its reference feature, related to the pixel $\mathbf{s}'$
$\mathbf{X}$	Vector constructed from a patch of the difference image D
X	Data set, set of reference features (patches),
	in Chapter 5 relates to $External_1$ channel

x	General member of a data set.
	Patch from $External_1$ image, related to the pixel $\mathbf{s}$
$\{x_i\}_{i=1}^m$	Reference patches of the patch x
x_X	Empirical kernel map of x onto the data set X
Y	Set of reference features (patches) from $External_2$ channel
y	General member of a data set.
	Patch from $External_2$ image, related to the pixel $\mathbf{s}$
Z	Set of reference features (patches) from *Internal* channel
$\mathbf{z}$	Standardized $\mathbf{q}$ feature vector
z (in Chapter 5)	Patch from *Internal* image, related to the pixel $\mathbf{s}$
z_i	i's element of the vector $\mathbf{z}$
$\Gamma(\cdot)$	Gamma Function
γ	Characteristic parameter of the second moment of $\mathbf{d}^2$
$\{\gamma_i\}_{i=1}^m$	Nonuniform weighting coefficients of Parzen's estimator
$\delta(\mathbf{s})$	Independent white noise disturbance at the pixel $\mathbf{s}$
ε	Similarity parameter
	(in de-noising relates to the time parameter t)
η	GLRT or SHT threshold
μ	Distribution of the points on X
μ_0	Mean of the background pattern features (vectors)
μ_1	Mean of the anomaly (defect) features (vectors)
ν	Perturbation function, with small norm
$\xi_G(\mathbf{s})$	Feature, related to the pixel $\mathbf{s}$,
	received by applying the filter bank G to the image
$\hat{\xi}_G(\mathbf{s})$	Estimated feature, related to the pixel $\mathbf{s}$
$\tilde{\xi}_G(\cdot)$	Noise-free value of $\xi_G(\cdot)$
ρ	Distance metric
Σ	Covariance matrix of the vectors $\mathbf{X}$, constructed
	from the overlapping patches of the difference image D
$\hat{\Sigma}$	Sample covariance matrix of the reference data $\{\mathbf{k}_j\}$

σ	Standard deviation, smoothing parameter
τ	Empirical threshold
φ	A confidence level - the probability of correctly deciding on H_0 given H_0 is true
χ_{central}	Characteristic function of the central element of the feature $\xi_G(\mathbf{s})$
Φ	Green's function (impulse response function)
ϕ	Transformation to implicit feature space
$\chi^2_N(0)$	Central chi-square probability distribution function, with N degrees of freedom
Ω	Support of an image

Chapter 1

Introduction

1.1 Motivation and Goals

Defect detection in wafers is a critical component of wafers manufacturing process. Manual defect detection is difficult, time consuming, expensive, and may cause yield ratio loss. Accuracy obtainable by human inspection is often insufficient due to lapses in alertness associated with fatigue, and various image processing techniques have been applied to automatic defect detection in wafers. A common approach for wafer defect detection utilizes a reference image, and applies some detection procedure to the difference between the observed and reference images [17,29,56,61,62]. A semiconductor wafer typically contains many copies of the same electrical component (denoted as "dies"), laid out in a matrix pattern. A reference image for one die is obtained by acquiring an image of the neighboring die, which is verified to be clear of defects. The reference image and the inspected image (further referred to as the "source image") are spatially aligned and subtracted one from another. The resulting difference image is processed for further defect detection. A major drawback of this approach is that the detection performance is very sensitive to image registration inaccuracies between the source and reference images [14,15,30].

Xie and Guan [69] and Guan *et al.* [26] proposed to generate a golden-block database from a periodic patterned image itself, in order to remove

the necessity in *a priori* reference image and registration. A building block, representing the structure of the pattern, is extracted according to the periods of the pattern in both directions and using interpolation to obtain sub-pixel resolution. A new defect-free golden template image, which is built based on the extracted building block, could be modified and refined, when new images with the same pattern arrive. Finally, a pixel-to-pixel comparison is used to find possible defects. A major drawback of this approach is that the template image contains feeble defects, averaged over several building blocks. Hence, the real defects might not be located accurately and only a warning that this is a defective product will be received. This makes the reference-based technique, which relies on another defect-free die from the same wafer, more popular.

Both reference and self-reference based techniques suffer from pattern variations problem. Printed patterns on the source and reference dies may differ slightly, particularly in the neighborhood of their edges. These pattern variations obscure the defects in the difference image and may yield high false detection rate. Onishi *et al.* [44] proposed reference-based method that does not require exact registration and allows slight pattern variations. Grayscale morphological dilation of the reference and inspected images allows dynamic tolerance control, which compensates for slight misregistration. The difference image is calculated according to the minimal distance between the reference and inspected images in the dilation range. However, this technique allows only slight misregistration and pattern variations, because it does not exploit the neighborhood replication of the periodic pattern.

Our goal is to exploit the periodicity of patterned wafers and to develop a defect detection procedure, which avoids image registration and is robust to pattern variations. Moreover, using multi-channel images (images with different perspective of the wafer), we aim to achieve a robustness for pattern variations in non-periodic patterned wafers defect detection. The proposed method is based on anisotropic kernel reconstruction of the source image using

the reference image. The idea of anisotropic kernels was studied by Lafon and Coifman [12, 35], providing a framework for structural multi-scale geometric organization of subsets(data) of $\mathbb{R}^n$ and on graphs. Szlam [59] successfully applied it to image denoising application, which generalized the NL-means algorithm of Morel [5] for edge preserving de-noising. The results, achieved by the proposed method, outperforms the anomaly detection algorithm, reducing false detections caused by pattern variations.

1.2 Overview of Thesis

In this thesis, we introduce kernel-based multi-channel algorithm for defect detection in patterned wafers, which avoids image registration and is robust to pattern variations. The source and reference images are mapped into a feature space, where every feature with origin in the source image is estimated by a weighted sum of neighboring features from the reference image. The set of neighboring features is determined according to the spatial neighborhood in the original image space, and the weights are calculated from exponential distance similarity function. We show that features originating from defect regions are not reconstructible from the reference image, and hence can be identified.

Subsequently, we proceed with the kernel-based approach to multi-channel defect detection, which relies on the simultaneous acquisition of three different images for each sample in the SEM tool. The proposed method assumes that if a pattern-originated region in the source wafer is similar to certain regions in the reference wafer, then this similarity is maintained across the three SEM images. We show that the proposed defect detection under constrained multi-channel reconstruction extends the single-channel defect detection method for non-periodic patterns and outperforms simple union or intersection of single-channel results.

We show that in case of pattern variations, the false detection rate can be significantly reduced by using our kernel-based approach rather than existing methods, based on the difference image. We also demonstrate the improved performance of the multi-channel approach in case of non-periodic patterned wafers, compared to using a single-channel approach.

1.3 Organization

The structure of the thesis is as follows: In Chapter 2 we present a general anomaly detection approach for the case where no *a priory* information about the anomalies is available. We review a single hypothesis testing approach and discuss its drawbacks. In Chapter 3 we present the theory of anisotropic kernels and demonstrate their applications. We show the relation between the kernels methods and the non-parametric statistics estimation. In Chapter 4 we introduce the single channel defect detection algorithm based on anisotropic kernels. We describe in details the detection algorithm and apply it to the periodic patterned wafers with pattern variations. We show that the proposed algorithm is robust to pattern variations. We also evaluate the algorithm by analyzing the receiver operating characteristics (ROC), obtained by simulations on general patterns from Brodatz textures [36]. In Chapter 5 we extend the kernel-based approach to multi-channel detection. The assumption of multi-channel consistent similarity and the relation between the kernel mapping and non-parametric probability density estimation allows us to develop a multi-channel algorithm. We show that the extended approach improves the performances in the case of non-periodic patterned wafers, where the single channel method demonstrates false detections. Finally, in Chapter 6 we summarize and propose subjects for future research.

1.4 Background

In this work, we consider the problem of defect detection in patterned wafers using SEM images. A wafer is irradiated with a focused beam of electrons directed to scan its surface. The analysis is carried out by moving the focused beam of electrons in a sweeping (raster) scan over the surface of the wafer. The energy exchange between the electron beam and the sample generates emission of electrons and electromagnetic radiation which can be detected to produce an image. An SEM tool that is manufactured by Applied Materials and presented in Fig. 1.1 can simultaneously produce three different images for a given sample:

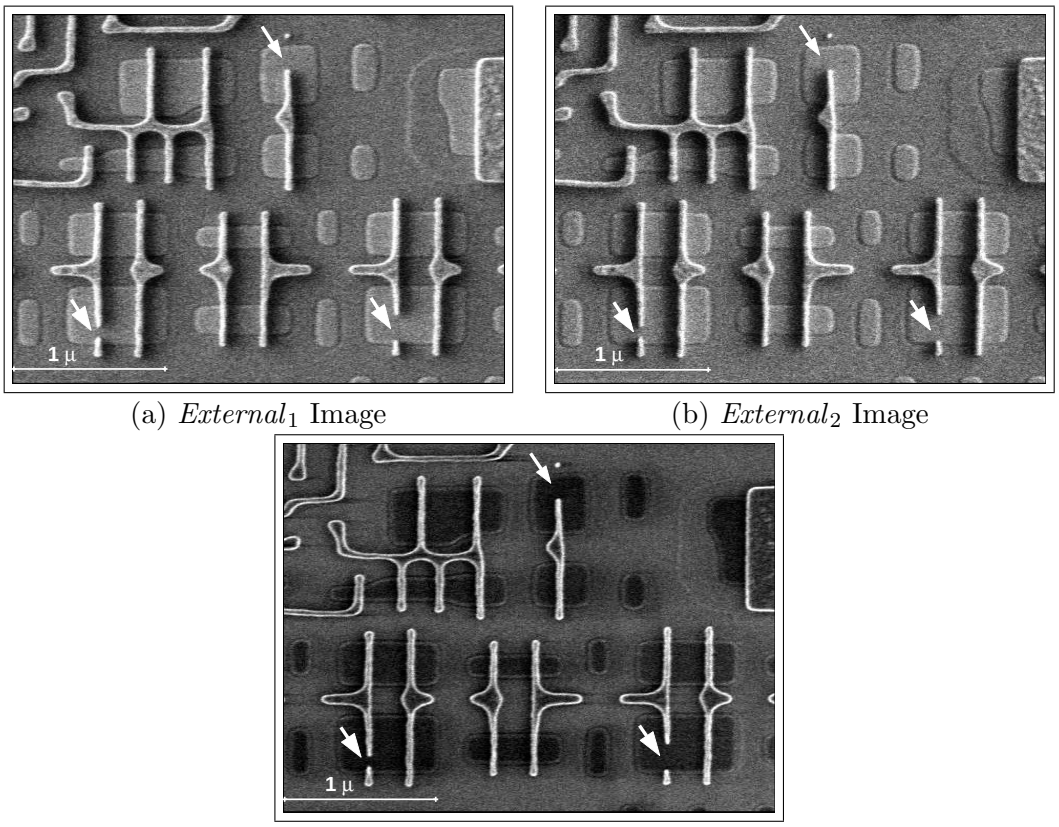
1. An $External_1$ Detector Image, which indicates the topography of the sample by light and shadows, as if the "light source" is directed to the sample from top-left direction.

2. An $External_2$ Detector Image, which also indicates the topography of the sample by light and shadows, but the "light source" is directed to the sample from top-right direction.

3. An *Internal* Detector Image, which provides information about edges and material of the sample.

Figure 1.2 shows examples of $External_1$, $External_2$ and *Internal* images of a patterned wafer. Arrows in the images point to faults in the pattern associated with imperfect connections. $External_1$ and $External_2$ images are acquired by detecting low energy secondary electrons using external detectors at different angles. The brightness of the signal depends on the number of secondary electrons reaching the detectors. If the electron beam is perpendicular to the sample surface, then the activated region is uniform about the axis of the beam and a certain number of electrons "escape" from the sample. As the angle of incidence increases, more secondary electrons will be emitted.

Figure 1.1: Scanning Electron Microscope from Applied Materials Inc., Rehovot, Israel.



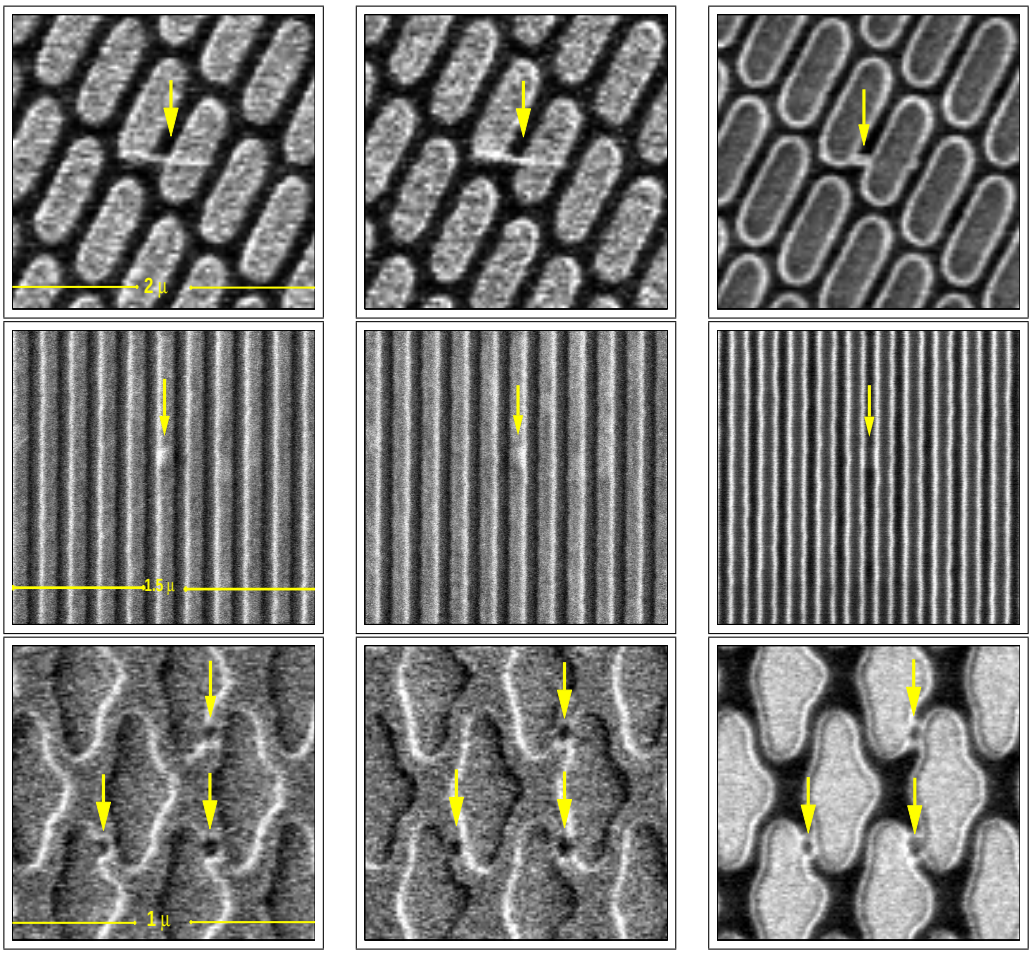
(a) *External*$_1$ Image (b) *External*$_2$ Image

(c) *Internal* Image

Figure 1.2: *External*$_1$ image of a wafer, acquired by an SEM tool from top-right direction; (c) *External*$_2$ image of the same wafer, acquired from top-left direction. (d) *Internal* image of the same wafer, acquired from top direction. Arrows in the images point to defects.

Hence, steep surfaces and edges tend to be brighter than flat surfaces, which provides information about the topography of the sample. The *Internal* image is acquired by detecting high-energy backscattered electrons with an internal detector. Backscattered electrons are used to identify contrast between areas with different chemical compositions and edges.

There are no precise characteristics of the possible defects. Defects may include particles, open lines, shorts between lines, or other problems. Figure 1.3 demonstrates that they may be of various shapes, sizes, may belong to the wafer background or to its pattern. The inspected wafer may contain many defects or no defects at all. The defects may be predominant or scarcely



External$_1$ *External*$_2$ *External*$_3$

Figure 1.3: Examples of defects. The presented defects are of various shapes, sizes, could belong to the wafer background or to its pattern. The number of defects is not known.

noticeable. The described variety makes it difficult or even impossible to perform template matching based on some *a priori* features or training database of detects.

Gleason *et al.* [22] perform the detection without *a priori* characteristics of the defects, modeling self-similarities in the source image with fractal image encoding. The fractal encoding process consists of breaking the image up into different subregions and then determining the set of affine transformations that best map one region to another. The detection is based on the idea that a defect will appear different than the normal image background structure, hence

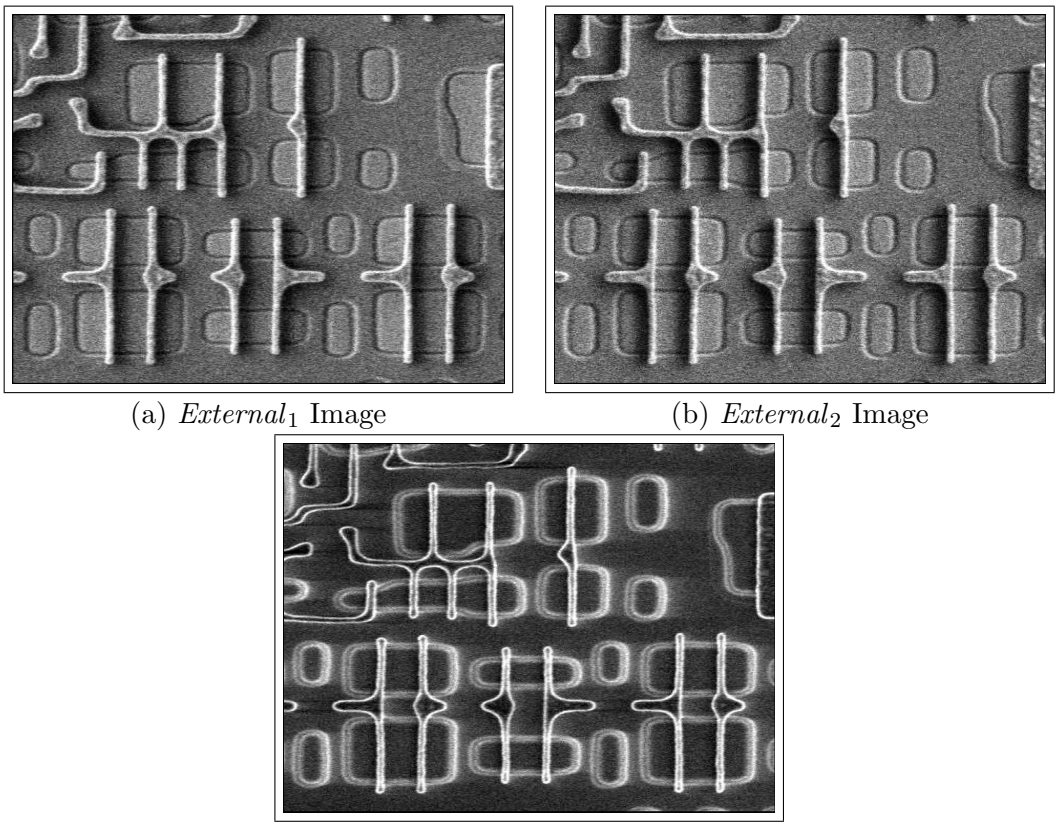
it will be difficult to find an affine transformation that effectively maps another image subregion to the current defective one. This technique has limitations in its ability to locate small defects on complex backgrounds. The authors state that it can be advantageous in the case when no reference image is available. Chang *et al.* [7] use an unsupervised learning by a two-layer competitive Hopfield neural network for defect detection. Their method precludes the necessity of a reference image or a pre-training and enables detection based on the variance of gray level and sharp spatial irregularity. However, the propose method doesn't succeed to detect scarcely noticeable defects.

Pattern to pattern comparison is the most suitable technique for an SEM-based inspection system. This comparison could be performed using reference image that was captured from another wafer's die that was preliminary checked to be clean of defects [17,28–30,61,62] or using self-reference approach based on the golden template construction from the repeating cells in the image [26,69]. Figure 1.4 shows the reference images of the inspected images from Fig. 1.2.

A pure reference system compares every pixel in the inspected image with the corresponding pixel in the reference image, which is assumed to be perfectly registered with the image being analyzed [17, 28–30, 47]. The images are usually registered using various techniques [14, 30], mostly based on maximizing the correlation between blocks. Using the reference image, we want to verify whether a pixel with indices $\mathbf{s} = (i, j)$ from the source image originates from the pattern clutter or not. A pattern-originated pixel from the source or reference image could be viewed as a combination of a noise free pixel from an underlying pattern image and white noise:

$$\begin{aligned} I_{\text{ref}}(\mathbf{s}) &= I_{\text{pat}}(\mathbf{s}) + \delta_1(\mathbf{s}) \quad \forall \mathbf{s} \in \Omega \\ I_{\text{src}}(\mathbf{s}) &= I_{\text{pat}}(\mathbf{s} + \mathbf{r}) + \delta_2(\mathbf{s}) \quad \forall \mathbf{s} \in \Omega \,, \end{aligned} \tag{1.1}$$

where $I_{\text{src}}, I_{\text{pat}}, I_{\text{ref}}$ denote source, noise-free pattern and reference images respectively; Ω denotes a set of indexes in the image domain; $\delta_1(\mathbf{s})$ and $\delta_2(\mathbf{s})$



(a) *External*$_1$ Image (b) *External*$_2$ Image

(c) *Internal* Image

Figure 1.4: Nonregistered reference images of the wafer images presented in Figs. 1.2(b)-(d).

denote independent white noise disturbances; and $\mathbf{r}$ is a translation vector, which is estimated by registration of the reference image to the source image.

The difference image $D = I_{\text{ref}} - I_{\text{src}}$, calculated by subtracting the source image from the aligned reference image, is used in several defect detection applications [2,30,69]. As a preprocessing step, denoising of the source and reference images could be performed. It is important that the denoising procedure will preserve edges and will not blur the defects, for example soft-threshold wavelet denoising [18] could be applied.

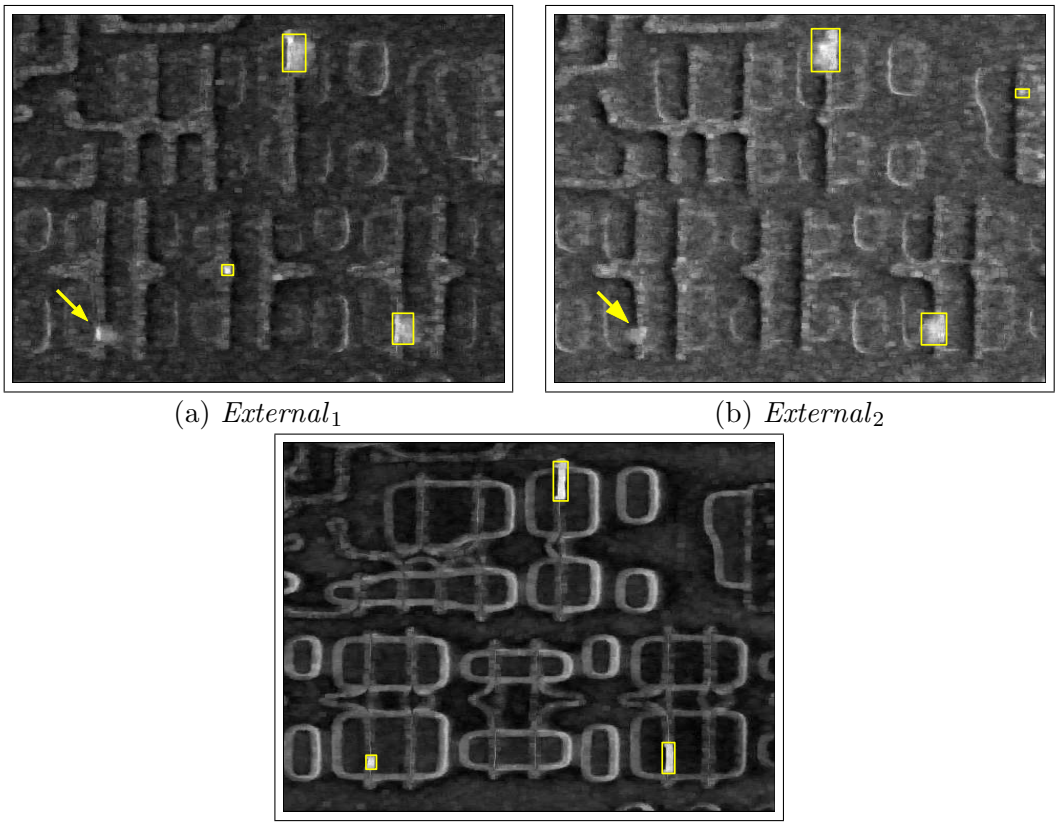
Thresholding of the difference image can reveal the defective regions in the inspected wafer. The defect mask is generated according to the following

decision rule:

$$B(\mathbf{s}) = \begin{cases} 1, & \text{if } |D(\mathbf{s})| > \tau \\ 0, & \text{otherwise} \end{cases} \tag{1.2}$$

Often, the threshold τ is chosen empirically. Rosin [50, 51] surveyed and reported experiments on many different criteria for choosing τ for general change detection applications. However, global threshold of pixel by pixel differencing yield high false alarm rate and is usually outperformed by more advanced statistical algorithms.

Moreover, the pattern in the reference image is generally not identical to the pattern in the source image and even if the alignment is perfect, pattern variations differences are still significant. These differences may be as intense as differences caused by defects and may cause false detections. Onishi *et al.* [44] tried to overcome the problem of slight distortion or rotation misalignment between the source and reference patterns by using grayscale morphological dilation of the reference and inspected images. The difference image is calculated according to the minimal distance between the reference and inspected images in the dilation range. However, this technique can only manipulate slight misregistration and pattern variations. Figure 1.5 shows the differences between the source images from Fig. 1.2 and their reference images after an alignment by the above algorithm. Clearly, defect detection by thresholding the difference images is characterized by high false and missed detection rates due to pattern variations.



(a) *External*$_1$ (b) *External*$_2$

(c) *Internal*

Figure 1.5: Absolute value difference images with overlaid defect detection results (denoted by light rectangulars). Pattern variations between the acquired wafer images (shown in Figs. 1.2(b)-(d)) and their reference images (shown in Fig. 1.4) generate non-negligible differences and high false detection rate under global thresholding.

Chapter 2

Anomaly Detection

2.1 Introduction

The variety and the unpredictability of defects makes it impossible to perform template matching based on some *a priori* learned features and statistics. Hence, general anomaly detection method should be adopted. Algorithms of anomaly detection in images aim to locate elements in a scene which are unlikely to be a part of it. This is a challenging task, mainely due to the large variability of the scenes background clutter and the appearance of the anomalous elements. Anomaly detection techniques are useful in applications such as fault detection [38], radar target detection [6], detection of masses in mammograms [60], e-commerce [38], and other signal processing and image analysis applications.

The decision rule in many anomaly detection algorithms is cast as a statistical hypothesis test [34, 49]. The decision as to whether or not a given pixel arises from an anomaly corresponds to choosing one of two competing hypotheses: the null hypothesis H_0 or the alternative hypothesis H_1, corresponding to no-anomaly and anomaly decisions, respectively. The image pair $(I_{\text{ref}}, I_{\text{src}})$ is viewed as a random vector. Knowledge of the conditional joint probability density functions (PDFs), $p(I_{\text{ref}}, I_{\text{src}}|H_0)$ and $p(I_{\text{ref}}, I_{\text{src}}|H_1)$, allows

to decide upon one of the hypotheses using the classical framework of hypothesis testing [33, 46, 67].

The variety and unpredictability of defects make it impossible to characterize H_1 hypothesis and to construct the respective PDF $p(I_{\text{ref}}, I_{\text{src}}|H_1)$. On contrary, characterizing the null hypothesis, based on the statistics of the background data, is straightforward. In the absence of any defect the difference between the source and aligned reference images can be assumed to be due to noise alone, according to the model presented in eq. (1.1). The simplest statistics could be defined by first two moments of the trained data, in this case the detection is performed by thresholding the distance from the sample to the mean, in terms of number of standard deviations [40]. The distance measure can be Mahalanobis [21] or some other probabilistic distance [66]. An anomaly detection algorithm based on the single hypothesis test (SHT) [21,24] of the difference image $D(\mathbf{s})$ allows to check null hypothesis fulfilment without any statistical knowledge about the defects.

However, the error of SHT increases as the dimensionality of the problem increases. The mapping from the original high dimensional feature space to a one-dimensional feature space destroys valuable classification information, which existed in the original feature space. Goldman and Cohen [25] have shown that if some *a priori* information about the targets is available, the detection performances could be potentially improved using Matched Subspace Detector (MSD). Partial information about the targets defines a subspace in which the targets lie. Hence, the detection procedure compares the projection of the sample onto the interference subspace with the projection onto the combination of the interference and target subspaces. The construction of such target subspace demands learning process, which is not possible when the variety of the targets is high, as in the case of defects detection. Therefore, we refer only to the SHT approach, which doesn't demand *a priori* learning. Next, we discuss SHT formalism and briefly review recently published anomaly

detection algorithms, which incorporate SHT.

2.2 SHT

The Generalized Likelihood Ratio Test (GLRT) for unknown signals is the foundation for many anomaly detection algorithms. Kelly [34] and Reed and Yu [49] developed GLRTs for multidimensional image data assuming that the target signal and the covariance of the background are unknown. Here we describe the single hypothesis test (SHT), formulated based on the GLRT for unknown signals, assuming Gaussian noise.

Let $\mathbf{q}$ denote a multivariate feature vector with $n_{\mathbf{q}}$ variables. Let hypothesis H_0 denote the case, where the feature vector belongs to the background clutter. Let hypothesis H_1 denote the case, where the feature vector does not belong to the background clutter and as such is regarded as an anomaly. Let us assume normal distributions for both hypotheses:

$$\begin{aligned} H_0 \quad &: \quad \mathbf{q} \sim N(\mu_0, \Sigma) \\ H_1 \quad &: \quad \mathbf{q} \sim N(\mu_1, \Sigma) \ , \end{aligned} \tag{2.1}$$

where $N(\mu, \Sigma)$ denotes normal distribution with mean μ and covariance matrix Σ, and μ_1 and Σ are unknown. We denote

$$\hat{\Sigma} = \frac{1}{N} \sum_{j=1}^{N} (\mathbf{k}_j - \mu_0)(\mathbf{k}_j - \mu_0)^{\mathrm{T}} \tag{2.2}$$

to be the sample covariance matrix of the reference data $\{\mathbf{k}_j\}_{j=1}^{N}$. For large number ($N \to \infty$) of i.i.d. sample vectors the Reed Xiaoli (RX) GLRT converges to:

$$RX_{N\to\infty}(\mathbf{q}) = (\mathbf{q} - \mu_0)^{\mathbf{T}} \left(\hat{\Sigma}\right)^{-1} (\mathbf{q} - \mu_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \ , \tag{2.3}$$

which is the Mahalanobis distance [32] of vector $\mathbf{q}$ from μ_0 [39]. It is evident that the GLRT presented in eq. (2.3) does not rely on any information about

the targets but rather on an estimated model of the background. The test is to decide whether vector $\mathbf{q}$ belongs to the background and therefore is a single hypothesis test.

Let us denote a vector $\mathbf{z} = \mathbf{A}^{\mathrm{T}}(\mathbf{q} - \mu_0)$, where $\mathbf{A}$ is a whitening matrix, then:

$$\mathbf{d}^2 = (\mathbf{q} - \mu_0)^{\mathbf{T}} \left(\hat{\Sigma}\right)^{-1} (\mathbf{q} - \mu_0) = \mathbf{z}^{\mathrm{T}}\mathbf{z} = \sum_{i=1}^{n_{\mathbf{q}}} z_i^2 \ . \tag{2.4}$$

Because the expected vector and the covariance matrix of $\mathbf{z}$ are 0 and $\mathbb{I}$ respectively, the z_i's are uncorrelated and $E(z_i) = 0$ and $\mathrm{Var}(z_i) = 1$. Thus, the expected value and variance of $\mathbf{d}^2$ are:

$$\begin{aligned} E(\mathbf{d}^2) &= n_{\mathbf{q}} \\ \mathrm{Var}(\mathbf{d}^2) &= \gamma n_{\mathbf{q}} \ , \end{aligned} \tag{2.5}$$

where $\gamma = E(z_i^4) - 1$. Moreover, if z_i's are normal, the variance becomes $\mathrm{Var}(\mathbf{d}^2) = 2n_{\mathbf{q}}$ and $\mathbf{d}^2$ is distributed $\chi^2_{n_{\mathbf{q}}}(0)$ (central chi-squared distribution with $n_{\mathbf{q}}$ degrees of freedom):

$$p_{\mathbf{d}^2}(\zeta) = \frac{1}{2^{n_{\mathbf{q}}/2}\Gamma(n_{\mathbf{q}}/2)} \zeta^{\frac{n_{\mathbf{q}}-2}{2}} e^{-\frac{\zeta}{2}} \mathbf{u}(\zeta) \ . \tag{2.6}$$

The normality assumption holds under the H_0 hypothesis, hence the detection threshold could be determined according to a specified confidence level, φ, which is the probability of correctly deciding on H_0 given H_0 is true. The threshold, η, and the confidence level, $\varphi = 1 - P_{\mathrm{FA}}$ (where P_{FA} is the maximal allowed probability of false detection), are related by:

$$\varphi \equiv \Pr(H_0|H_0) = p_{\mathbf{d}^2}(\zeta < \eta) \ . \tag{2.7}$$

This relation is a convenient way to determine analytical threshold for detection procedure considering the false detection rate, however it is possible only under gaussian assumption of the data set. Despite its importance, it has been difficult to test whether a given data set is normal or not. Measuring the variance of $\mathbf{d}^2$ provides an estimation of γ via eq. (2.5), which may be used

to test the normality of a high-dimensional distribution [21]. However it must be cautioned that this procedure tests only one marginal aspect of the distribution and doesn't guarantee the overall normality of the distribution even if the samples pass the test.

2.3 Overview of Recent Algorithms

Goldman and Cohen, 2004

Goldman and Cohen have formulated in [24] an anomaly detection approach, which does not rely on an exhaustive statistical model of the targets, but rather on the local statistics of the data and possibly on some a priori information regarding the sizes and shapes of targets. They proposed to perform an iterative procedure of feature extraction, which is based on local statistics and principle components analysis. The background is statistically characterized in a feature space of principle components. A single hypothesis scheme is used for the detection of anomalous pixels in a given region of interest. Subsequently, morphological operators are employed for extracting the sizes and shapes of anomalous clusters in the image domain, and identifying potential targets. The proposed procedure also reduces the SHT error, resulted from mapping high-dimensional feature into one-dimensional distance and classifying by thresholding. Each iteration gradually reduces the false alarm rate while maintaining a high probability of detection. This method was successfully employed on a large data-set of sea-mines sonar images.

Shadhan and Cohen, 2006

Shadhan and Cohen have formulated in [55] an anomaly detection scheme, which relies on a statistical model of textures and is specifically designed for detection of anomalies in textures, utilizing spatial correlation rather than

just energy measures. The authors have introduced a multi-resolution feature space that facilitates anomaly detection with constant false alarm rate. To detect different kinds of targets, the proposed algorithm doesn't rely on the exhaustive statistical model of the targets, but rather on the multi-resolution statistics of the background clutter. Experimental results demonstrated that the proposed algorithm, when applied to images containing background texture, achieves improved detection results and lower false alarm rate than a competitive anomaly detection scheme.

2.4 SHT of Difference Images

Anomalies are often associated with localized groups of pixels, hence it is common for the anomaly decision at a given pixel $\mathbf{s}$ to be based on a small block of pixels in the neighborhood of $\mathbf{s}$ in the image. Accordingly, a data set is constructed from overlapping patches formed around every pixel in the difference image D. Given the expected vector $\mathbf{M}$ and the covariance matrix Σ of the constructed data set, the Mahalanobis distance of any vector $\mathbf{X}$ from $\mathbf{M}$ is obtained by

$$\mathbf{d}^2(\mathbf{X}) = (\mathbf{X} - \mathbf{M})^T \Sigma^{-1} (\mathbf{X} - \mathbf{M}) \ , \tag{2.8}$$

and the SHT is given by

$$\mathbf{d}^2(\mathbf{X}) \underset{H_1}{\overset{H_0}{\lessgtr}} \eta \ , \tag{2.9}$$

where H_1 and H_0 represent hypotheses of anomaly presence and absence, respectively, and η is a distance threshold.

Although there are obvious statistical dependencies within a patch, the observations for each pixel in a patch are typically assumed to be independent and identically distributed (i.i.d.). It is also assumed that the noise in the model presented in eq. (1.1) is Gaussian. Under these assumptions, $\mathbf{d}^2(\mathbf{X})$ is

distributed $\chi^2_{n_{\mathbf{X}}}(0)$ (central chi-squared distribution with $n_{\mathbf{X}}$ degrees of freedom), where $n_{\mathbf{X}}$ is the number of pixels in the patch constructed around pixel $\mathbf{s}$. The decision threshold η for a desired false alarm rate is calculated according to eq. (2.6)

Figure 2.1(d) demonstrates the SHT of the difference image from Fig. 2.1(c), whereas patches of size 7×7 are used. The confidence level $\varphi = 0.99999$ ($P_{\text{FA}} = 10^{-5}$) determines the threshold of SHT, η ,according to the eq. (2.7). The detected anomaly fits the defect from Fig. 2.1(b) exactly and is framed in the difference and SHT images. The histogram of the squared Mahalanobis distance, presented in the Fig. 2.1(e), shows that the empirical distribution fits the theoretical $\chi^2_{n_{\mathbf{X}}}(0)$ distribution with $n_{\mathbf{X}} = 7 * 7 = 49$. Moreover, $\gamma \simeq 2.3$ (according to eq. (2.5)), which is close to the Gaussian value $\gamma = 2$. Hence, the constructed data set from the patches of the difference image is probably Gaussian distributed and thresholding using eq. (2.7) succeeds to detect the defect.

The model presented in the eq. (1.1) handles only the translation differences between the images. However, the source and reference image patterns are not identical and pattern variations may occur. These differences could be as intense as the differences caused by defects and may cause false detections. Strong pattern variations appear in Fig. 2.2(c), which demonstrates absolute difference image between the reference image, Fig. 2.2(a), and the source image, Fig. 2.2(b). Pattern variation differences invalidate the assumption that the constructed feature vector $\mathbf{X}$ of the difference image is Gaussian distributed under the null hypothesis. The distribution of the squared Mahalnobis distance in Fig. 2.2(e) doesn't fit the expected theoretical distribution and the γ value is not close to 2. Hence, the SHT threshold in the eq. (2.9) could not be optimally determined using eq. (2.6). The detection in Fig. 2.2(b), based on the SHT thresholding, reveals false detection and misses.

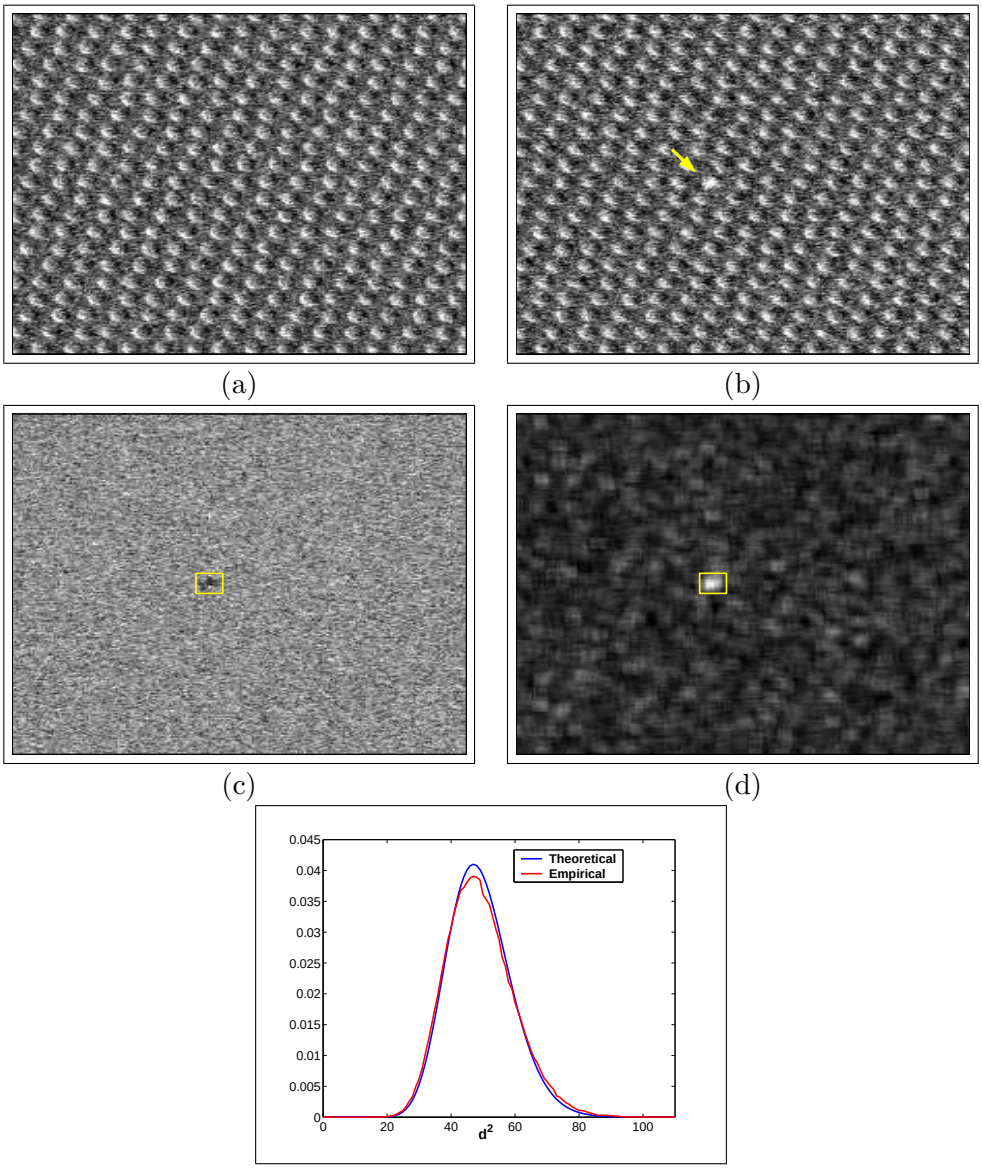


Figure 2.1: (a) Reference image; (b) Source image; (c) Difference image; The arrow in the source and difference images points at the defect; (d) Squared Mahalanobis distance of the difference image with the overlaid exact defect detection based on SHT, eq. 2.9; The threshold is determined to achieve confidence level $\varphi = 0.99999$; (e) Distribution of SHT: the squared Mahalanobis distance (red line) is close to the theoretical curve (blue line) of $\chi^2_{n_{\mathbf{X}}}(0)$- distribution with $n_{\mathbf{X}} = 49$ degrees of freedom and $\gamma(\mathbf{d}^2) \simeq 2.3$. Hence, the data features could arise from Gaussian distribution.

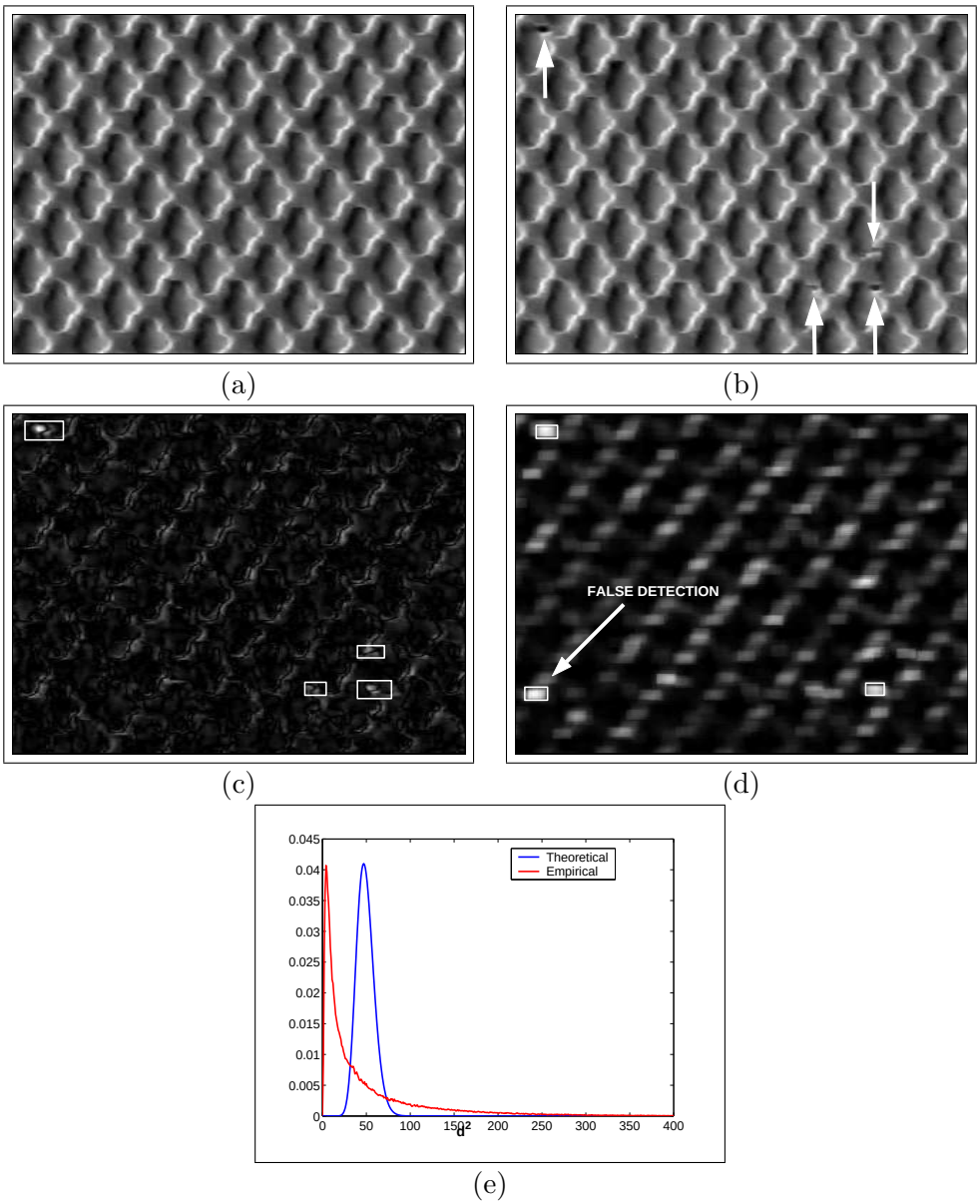


Figure 2.2: (a) Denoised Reference image; (b) Denoised Source image. The arrows in the source image point at the defects; (c) Difference image. Large differences result from pattern variations and obscure the defects (marked with white frames); (d) Detection based on thresholding of SHT of the difference image reveals only two of the four defects and one is falsely detected as defect. Thresholding with lower threshold will detect the missed defects and add more false detections; (e) SHT Distribution: the squared Mahalanobis distance (red line) doesn't fit the theoretical curve (blue line) of $\chi^2(0)$- distribution and $\gamma(\mathbf{d}^2) \simeq 91$. Hence, threshold could not be calculated via eq. (2.7).

2.5 Summary

We have presented the single hypothesis test (SHT), which is based on a statistical model of the background and doesn't use *a priori* information about the targets. We have briefly reviewed recent anomaly detection algorithms which incorporate SHT and applied SHT to defect detection procedure. We have shown that if the differences between the source image and spatially aligned reference image are Gaussian distributed, the detection based on the thresholding of SHT is applicable. However, if pattern variations between the source and the reference images occur, anomaly detection based on the difference image leads to high false detection rate. In Chapter 4 we will introduced a similarity model between the source and the reference image that can significantly conceal the pattern variations disturbance and does nor require precise registration. The anisotropic kernels, which constitute the basis of the novel detection algorithm, are presented in Chapter 3.

Chapter 3

Anisotropic Kernels

3.1 Introduction

Nonlinear information processing algorithms can be designed by means of linear techniques in implicit feature spaces induced by kernel functions. This idea has been successfully applied to the support vector machine (SVM), a learning method with controllable capacity which obtains generalization in high (even infinite) dimensional feature spaces [13,63–65]. Moreover, a great attention has been paid to the kernel methods, such as local linear embedding [52], Laplacian eigenmaps [3], Hessian eigenmaps [19] and local tangent space alignment [70], in the field of dimensionality reduction of the data set [27]. Their nonlinearity as well as their locality-preserving action are generally viewed as a major advantage over classical methods like Principal Component Analysis and classical Multidimensional Scaling. The first aspect is essential as most of the time, in their original form, the data points do not lie on linear manifolds. The second point is the expression of the fact that in many applications, distances of points that are far apart are meaningless, and therefore need not be preserved.

Coifman and Lafon [9] show that all the kernel methods constitute special cases of a general framework based on diffusion processes. They use the eigenfunctions of a Markov matrix defining a random walk on the data to obtain new

descriptions of data sets (subset of $\mathbb{R}^n$, graphs) via a family of mappings that are termed "diffusion maps". These mappings embed the data points into a Euclidean space in which the usual distance describes the relationship between pairs of points in terms of their connectivity. This defines a useful distance between points in the data set that we term "diffusion distance". Different geometric representations of the data set are obtained by iterating the Markov matrix of transition, or equivalently, by running the random walk forward. The diffusion maps are precisely the tools that allow to relate the spectral properties of the diffusion process to the geometry of the data set. Finally, a multiscale family of geometric representations corresponding to descriptions at different scales is obtained.

Various data analysis algorithms have been recently developed based on the diffusion maps and connectivity strengths on graphs. For example, Coifman *et.al* propose to use these tools for compressing and analyzing large and complex data sets, such as those derived from sensor networks or neuronal activity data sets, obtained in the laboratory or through computer modeling [11]. Diffusion geometries can reveal structure in data at different levels of organization. Because many sources of data in neuroscience are high-dimensional, understanding their primary, low-dimensional intrinsic structure can be insightful. Methodologies for using the network of inferences and similarities between the data points (of arbitrary digital data sets) to create robust non-linear estimators for missing or noisy entries are discussed by Coifman *et al.* in [10]. The authors point out that modern sensor systems such as radar, hyperspectral, MRI and others actually do not measure images, but much more elaborate vectors. Hence, the intrinsic geometry of the measurements, which is emerged through a natural process of affinity diffusion, should participate in the information extraction.

Non-linear approach is especially desired in the denoising application, where it is important to preserve the fine structure of the original image. Morel

et. al. have proposed the non local (NL)-means algorithm for de-noising, which takes advantage of the redundancy of the natural images and allows to denoise flat regions, while preserving edges [5]. Szlam and Coifman [59] have shown that NL-means algorithm is a special case of the general anisotropic diffusion approach for denoising, which is presented in the next section. The main observation of the authors is that an isotropic diffusion on a modification of a data set is anisotropic on the original data set. Hence, linear filtering, performed in the features space, applies non-linear de-noising in the original data space.

3.2 Anisotropic Kernels for Denoising

3.2.1 From Gaussian Smoothing to Anisotropic Smoothing

The goal of image denoising methods is to recover the original image from a noisy measurement:

$$g(\mathbf{s}) = h(\mathbf{s}) + n(\mathbf{s}) \ , \tag{3.1}$$

where $g(\mathbf{s})$ is the observed value, $h(\mathbf{s})$ is the "true" value and $n(\mathbf{s})$ is the noise perturbation at a pixel with coordinates $\mathbf{s} = (i, j)$. For simplicity, $n(\mathbf{s})$ is assumed to be i.i.d. gaussian noise.

Isotropic (*i.e.* linear) denoising usually incorporates convolution of the observed image with the Gaussian smoothing kernel with standard deviation σ [37]:

$$I(i, j, \sigma) = G(i, j, \sigma) * I(i, j) \ , \tag{3.2}$$

where $I : \mathbb{R}^2 \to \mathbb{R}, I = g(i,j)$. Smoothing with the Gaussian kernel is equivalent to solving the heat equation:

$$\begin{aligned} \frac{du}{dt} = \Delta u &= \nabla \cdot (\mathbb{I} \nabla u) \\ u(i,j,0) &= g(i,j) \ , \end{aligned} \tag{3.3}$$

where $\mathbb{I}$ is 2×2 identity matrix. The solution is given by the convolution with Green's function (impulse response function):

$$u(i,j,t) = \Phi(i,j,t) * u(i,j) \ . \tag{3.4}$$

The relation between eq. (3.4) and eq. (3.2) is evident, where the related σ and t control the amount of image denoising by averaging. The disadvantage of the Gaussian smoothing is that it blurs or even removes contours and fine structures. Closer inspection of the heat equation allows to develop anisotropic (*i.e.* non linear) filter, which will preserve the edges, while de-noising flat regions. Replacing the identity matrix $\mathbb{I}$ with symmetric positive definite matrix $\mathbb{D}$ in the eq. (3.3), we will achieve anisotropic diffusion, which has major and minor axes of diffusivity given by the eigenvectors of $\mathbb{D}$. If we let $\mathbb{D}$ depend on the function to be denoised, we can adaptively smooth it in a way which better respects the "high frequency" image structures.

For example, the following development could be performed to achieve function-dependent $\mathbb{D}$. The differentiating in eq. (3.3) is preformed in both tangent and normal directions:

$$\Delta u = (\Delta u)_{\text{tangent}} + (\Delta u)_{\text{normal}}. \tag{3.5}$$

Differentiating in normal direction blurs edges and should be removed in order to preserve fine structure:

$$\begin{aligned} \frac{du}{dt} = (\Delta u)_{\text{tangent}} = \Delta u - (\Delta u)_{\text{normal}} &= \\ \Delta u - \frac{1}{\|\nabla u\|^2} \left\langle \nabla^2 u(\nabla u), \nabla u \right\rangle &= \nabla \cdot \left(\frac{\nabla u}{\|\nabla u\|^2} \right) \|\nabla u\| \ . \end{aligned} \tag{3.6}$$

Hence, replacing the identity matrix $\mathbb{I}$ with $\mathbb{D} = \frac{1}{\|u\|}$ in the eq. (3.3) achieves anisotropic filtering. The oriented Laplacian $\Delta u = \nabla \cdot (\mathbb{D}\nabla u)$ slows smoothing against high gradients in the image, which preserves edges.

3.2.2 Denoising using Anisotropic Kernels

Szlam's [59] approach to finding anisotropic filters will be to build a graph, whose vertices are the pixels of the image and whose weights preserve the image structure, and use the isotropic (*i.e.* linear) diffusion on the graph to smooth the image (considered as a function on the graph). Szlam and Coifman believe that the common definition of an image as a function on $\mathbb{R}^2$ is misleading, and that in many cases, it is much more natural to think of the image as being a smooth function on a perhaps more complex set. An isotropic diffusion on a modification of a data set will be anisotropic on the original data set. The following constriction allows to extract a graph G from the processed image f.

Graph Construction

Given an image I, all its pixels $\mathbf{s} = (i, j) \in \mathbb{T}^2$ are mapped into $\mathbb{R}^m$ using a m-vector $G = (g_1, .., g_m)$ (*i.e.* real valued functions on $\mathbb{T}^2$):

$$\mathbf{s} \to \xi_G(\mathbf{s}) = (I * g_1(\mathbf{s}), ..., I * g_m(\mathbf{s})) \ . \tag{3.7}$$

Let

$$\rho(\mathbf{s}, \mathbf{s}') = \rho_G(\mathbf{s}, \mathbf{s}') = \|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\| \ . \tag{3.8}$$

Then, the filter will be:

$$K = \mathbf{D}^{-1} W \ , \tag{3.9}$$

where

$$W(\mathbf{s}, \mathbf{s}') = e^{-\rho(\mathbf{s},\mathbf{s}')^2/2\varepsilon} \ , \tag{3.10}$$

$\mathbf{D}$ is a row- normalizing matrix of W and ε is a kernel width. Hence, a graph on the image is constructed, with pixels as the vertices set V and a matrix of

weights, $W : V \times V \to \mathbb{R}^+$, which preserves the image structure. This matrix is usually symmetric, that is the edges of the graph are undirected. Using the constructed filtering matrix K the anisotropic de-noising is performed.

Choice of Embedding Filters

The construction presented in eq. (3.7) is an extremely flexible, and there are many interesting choices for the filters. One could take a few wavelets at different scales, or edge filters, or patches of texture, or some measure of local statistics. If we take G to be the nine filters

$$g_{1,1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} g_{1,2} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} ... \tag{3.11}$$

where $g_{i,j}$ is the 3×3 matrix with a 1 in the i, j position and zeros elsewhere, then ξ_G is the set of patches of the image embedded in 9 dimensions. The diffusion one gets from this choice of filters is the NL-means filter of Morel *et al.* [5].

Although there are many possible choices of filters, there is the following limitation: if $g_1, ..., g_n$ are a set of orthonormal basis vectors for a subspace in L^2, and $\tilde{g}_1, ..., \tilde{g}_n$ are also a basis for the same subspace, then because the convolution is linear, $\xi_G = O\xi_{\tilde{G}}$, where O is a rotation in d dimensions. Thus the embedding into $s_x \times s_y$ patches is the same embedding (up to a rotation) as into $s_x \times s_y$ DCT coordinates, and so the weight matrices constructed from these embeddings are the same. On the other hand, one may want to attenuate small values in the embedding coordinates, and then the choice of basis matters.

Nearest Neighbors Searches

Szlam and Coifman [59] have found that it is often advantageous to keep some track of the original coordinates from image space domain in the graph we

build from the image. One way to do this, which significantly improves the running time of construction of the graph while also freeing from a nearest neighbors search in high dimension is to do a constrained nearest neighbors search; that is, pick some number of neighbors in the original 2-d Euclidean metric, and then do a full search in the embedding space, but just for the points in the 2-d ball. In many cases, this also improves the quality of the denoising; although you cannot compare as many similar patches, constraining to a small 2-d neighborhood means that patches you decide are similar are much more likely to be similar; *i.e.* you improve the variance of your estimator.

Weights Choice

Equation (3.10) proposes exponential choice of the weights $W(\mathbf{s}, \mathbf{s}')$, however the weight matrix doesn't have to come from an exponentiated distance function and any choice of weights which encodes useful information about the data set could be used. The only restriction upon the weights choice comes from the fact that $W(\mathbf{s}, \mathbf{s}')$ actually represents a "kernel" or "affinity function" $k(\mathbf{s}, \mathbf{s}')$, which verifies the following properties

1. k is symmetric: $k(\mathbf{s}, \mathbf{s}') = k(\mathbf{s}', \mathbf{s})$,

2. k is positivity preserving: for all $\mathbf{s}$ and $\mathbf{s}'$ in the data set, $k(\mathbf{s}, \mathbf{s}') \geq 0$,

3. k is positive semi-definite: for all real-valued bounded functions f defined on any data set X,

$$\int_X \int_X k(x, y) f(x) f(y) d\mu(x) d\mu(y) \geq 0 \ ,$$

   where μ represents the distribution of the points on X.

Equation (3.8) proposes Euclidean metric as the similarity measure, however this is not necessary and for example the "correlation kernel" $k_\varepsilon(\mathbf{s}, \mathbf{s}') = e^{\frac{1-(\xi_G(\mathbf{s}) \cdot \xi_G(\mathbf{s}'))}{\varepsilon}}$ could be used in the case of normalized data set. The chosen

Euclidean distance conserves, in expectation, the order of similarity between the points, in spite of the additive white noise with variance σ^2:

$$E\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_{2,\varepsilon}^2 = \|\tilde{\xi}_G(\mathbf{s}) - \tilde{\xi}_G(\mathbf{s}')\|_{2,\varepsilon}^2 + 2\sigma^2 \ , \tag{3.12}$$

where $\tilde{\xi}_G(\cdot)$ denotes the underling original (without noise) values of $\xi_G(\cdot)$.

Anisotropic Filtering

If we construct K_t with exponential weights,

$$k_t = \exp(-\frac{\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|^2}{t}) \ , \tag{3.13}$$

with (small) time parameter t, the results in [35] show that $(\mathbb{I} - K_t)/t$ tends to Δ_M, where M is the embedded surface in $\mathbb{R}^3$, and K is properly normalized to remove the "density" in the embedding. In other words, an application of (the normalized *a la* Lafon) K_t to the image is a t step in the evolution of the heat equation on the surface, with initial value given by the image (considered as a function on its surface). Hence, for this choice of embedding, the graph method is asymptotically no different than the partial differential equation method introduced before.

3.3 Relation to the Parzen's Probability Density Estimator

This section presents a different aspect of the kernel-methods, which relates them to the non-parametric probability density estimation. The estimation of the PDF of a continuous distribution from a representative sample drawn from the underlying density is a problem of fundamental importance to all aspects of machine learning and pattern recognition. The kernel density estimator, also commonly referred to as the Parzen window estimator [45], can be viewed as the limiting form of a mixture model [42], where the number of

mixture components will equal the number of points in the data sample. The Parzen window form of nonparametric probability density estimation [45] is particularly attractive when no *a priori* information is available to guide the choice of the precise form of density with which to fit the data, for example, the number of components in a mixture model. A probability density estimate $\hat{p}_\varepsilon(x)$ can be obtained from the finite random sample $X = \{x_i\}_{i=1}^m$, drawn from the density $p(x)$ by:

$$p(x) \cong \hat{p}_\varepsilon(x) = \frac{1}{m} \sum_{i=1}^{m} b_\varepsilon(\|x - x_i\|) \ , \tag{3.14}$$

where b_ε is a normal density with zero mean and standard deviation (width) ε. The smoothing parameter ε controls the amount of regularization, and in practice (finite samples) it can be adjusted by cross validation [20, 41, 53].

In the kernels methods, the kernel function k defines the inner product in some implicit feature space with associated transformation ϕ, which need not be computed explicitly

$$k(x, y) = \phi(x) \cdot \phi(y) \ . \tag{3.15}$$

The Gaussian or RBF kernel [64] induces an infinite dimensional feature space in which all image vectors have the same norm and become orthogonal when they are distant in the input space with respect to the scale or regularization parameter ε.

$$k(x, y) \equiv e^{-\frac{1}{2}(\|x-y\|_2^2/\varepsilon)} \ . \tag{3.16}$$

Kernel expansion of the x onto X

$$x_X = (k(x_1, x), k(x_2, x), ..., k(x_m, x)) \tag{3.17}$$

is referred to as *empirical kernel map* [54]. The authors of [53] note that the Parzen estimator can be related to the *empirical kernel map* by

$$\hat{p}_\varepsilon(x) = x_X^T \cdot \vec{1}_m \ , \tag{3.18}$$

where $\vec{1}_m$ denotes m-dimensional column vector with all the components equal to m^{-1}. We will further use this relation for the statistical interpretation of our algorithm.

3.4 Summary

We have presented the theory of kernel-methods for data analysis that constitute special case of a general framework based on diffusion processes. We have reviewed the denoising application that demonstrates the concept of anisotropic kernels, using anisotropic diffusion to preserve edges. The review presents the construction of a graph whose vertices are the pixels of the image and whose weights preserve the image structure, and usage of an isotropic diffusion on the graph to smooth the image. An isotropic diffusion in the feature space of the image will be anisotropic on the original image. We have also discussed the relation of the kernels-methods to the Parzen's non-parametric estimator of PDF. This statistical interpretation of the *empirical kernel map* will facilitate an extension of the single-channel detection method to the multi-channel detection. In Chapters 4 and 5 we propose an adaptation of the presented anisotropic kernels approach for the application of defect detection in patterned wafers.

Chapter 4

Defect Detection Using Anisotropic Kernels

4.1 Introduction

In Chapter 2 we have demonstrated that the defect detection method, which is based on the difference image, suffers from false detections and misses, which are caused by the pattern variations between the reference and the inspected images. Onishi *et al.* [44] tried to overcome the problems of pattern variations and misregistration by using grayscale morphological dilation of the reference and inspected images. The difference image is calculated according to the minimal distance between the reference and inspected images in the dilation range. However, this technique allows only slight misregistration and pattern variations, because it does not exploit the neighborhood replication of the pattern.

The aim of this work is to develop a more flexible similarity model that can significantly conceal the pattern variations disturbance and doesn't require precise registration. The proposed measure takes advantage of the periodicity of patterned wafers images, which results from the replicated circuit pattern. Due to the periodicity and the similarity of the source and reference images patterns, given a patch of the source image we can find similar patches in the

reference image. Hence, under the null hypothesis a pixel $\mathbf{s}$ in the source image could be reconstructed from several pixels in the reference image according to the following model:

$$\hat{I}_{\text{src}}(\mathbf{s}) = \frac{1}{\sum W(\mathbf{s},\mathbf{s}')} \sum_{\mathbf{s}' \in \mathcal{N}_{\mathbf{s}}} W(\mathbf{s},\mathbf{s}') I_{\text{ref}}(\mathbf{s}') \ , \tag{4.1}$$

where $W(\mathbf{s},\mathbf{s}')$ denotes the similarity measure that will be further discussed and the neighborhood $\mathcal{N}_{\mathbf{s}}$ of the pixel $\mathbf{s}$ is given by

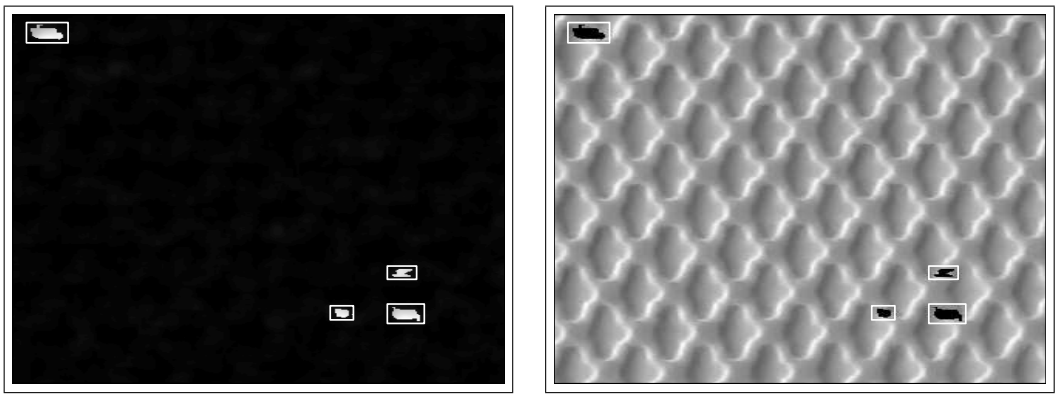
$$\mathcal{N}_{\mathbf{s}} = \{\mathbf{s}' \,|\, \mathbf{s}' \in n_k(\mathbf{s})\} \ , \tag{4.2}$$

$n_k(\mathbf{s})$ is the set of k nearest neighbors of $\mathbf{s}$ (including itself). The neighborhood is determined in the original 2-d Euclidean metric of the image and relates the pixel in the source image only to its spatial neighbors in the reference image.

The model proposed in eq. (4.1) does not assume that a pixel in the source image is related to one specific pixel in the reference image, but originates from a combination of several pixels, according to some similarity measure. This model reduces to the model presented in eq. (1.1), if all the weights $W(\mathbf{s},\mathbf{s}')$ are equal to zeros except the one that relates to $\mathbf{s}' = \mathbf{s} + \mathbf{r}$. The detection is based on the success of source image reconstruction from the reference image. We assume that under the null hypothesis the source image patches can be reconstructed from patches of the reference image due to similarity and periodicity of patterns in the source and reference images. When a patch of the source image contains a defect, there are no similar patches in the reference image, and therefore the patch cannot be reconstructed from patches of the reference image, which indicates the presence of a defect. That is, the detection is obtained by

$$\begin{aligned} H_0 \ &: \ \sum W(\mathbf{s},\mathbf{s}') \neq 0 \\ H_1 \ &: \ \sum_{\forall \mathbf{s}' \in \mathcal{N}_{\mathbf{s}}} W(\mathbf{s},\mathbf{s}') \to 0 \ . \end{aligned} \tag{4.3}$$

Figure 4.1(a) demonstrates the improved difference image (comparing to Fig. 2.2) based on difference between the estimated source image (Fig. 4.1(b))



(a) (b)

Figure 4.1: (a) Difference image between the reconstructed source image and the source image; (d) Detection based on unreconstructed regions of the reconstructed source image

and the original source image (Fig. 2.2(b)). Four unreconstructed (black) regions are exactly the regions of defects, all other parts in the image are reconstructed. The detection procedure we have presented overcomes the pattern variation problem demonstrated in Fig. 2.2(c) and (d) and detects all the four defects without false alarms.

4.2 Detection Procedure

In this section, we discuss the reconstruction procedure of the source image from the reference image and the similarity measure it involves. The reconstruction is performed in a feature space based on the NL-means filters of Morel, which was proposed by Szalm *et al.* [59] for image denoising applications (for details see chapter 3).

4.2.1 Kernel Representation of the Source Image Using a Reference Image

Let us pick a d-vector $G = (g_1, .., g_d)$ of filters and map pixels of source and reference images into $\mathbb{R}^d$ features space ξ_G:

$$\begin{aligned} \mathbf{s} \to \xi_G(\mathbf{s}) &= \{I_{\text{src}} * g_1(\mathbf{s}), ..., I_{\text{src}} * g_d(\mathbf{s})\} \\ \mathbf{s}' \to \xi_G(\mathbf{s}') &= \{I_{\text{ref}} * g_1(\mathbf{s}'), ..., I_{\text{ref}} * g_d(\mathbf{s}')\} \ , \end{aligned} \tag{4.4}$$

where $\mathbf{s}, \mathbf{s}' \in \Omega$ and Ω is a general set of indices in the image space. We omit from features notation labels src and ref, instead the indices $\mathbf{s}$ are associated with features from the source image and $\mathbf{s}'$ are associated with features from the reference image. Given $\xi_G(\mathbf{s}')$ for all $\mathbf{s}' \in \mathcal{N}_\mathbf{s}$, a reconstructed source image is obtained by

$$\hat{I}_{\text{src}}(\mathbf{s}) = \frac{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2 / 2\varepsilon\} I_{\text{ref}}(\mathbf{s}')}{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2 / 2\varepsilon\}} \ . \tag{4.5}$$

where $\mathcal{N}_\mathbf{s}$ is denoted in eq. (4.2). According to [59], we choose

$$W(\mathbf{s}, \mathbf{s}') = \exp^{-\rho(\mathbf{s},\mathbf{s}')^2/2\varepsilon} \ , \tag{4.6}$$

where ρ is a metric in our feature space, ε is a similarity parameter, and $\mathbf{D}(\mathbf{s}) = \sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} W(\mathbf{s}, \mathbf{s}')$ is a normalizing factor. The similarity $W(\mathbf{s}, \mathbf{s}')$ is measured as a decreasing function of the Euclidean distance

$$\rho^2(\mathbf{s}, \mathbf{s}') = \|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2 \ . \tag{4.7}$$

The similarity parameter ε controls the decay of the exponential function and therefore the decay of the weights as a function of the Euclidean distances.

Under the null hypothesis H_0, a source feature is reconstructible from similar reference image features. According to eq. (4.3), if the source feature relates to a defect, there are no similar reference features and the reconstructed pixel is determined to be zero:

$$\begin{aligned} H_0 &: \hat{I}_{\text{src}}(\mathbf{s}) \to I_{\text{src}}(\mathbf{s}) \Rightarrow \mathbf{s} \notin \mathcal{A} \\ H_1 &: \hat{I}_{\text{src}}(\mathbf{s}) \to 0 \Rightarrow \mathbf{s} \in \mathcal{A} \ , \end{aligned} \tag{4.8}$$

where $\mathcal{A}$ denotes a set of defect regions.

4.2.2 Filter Banks

Szalm *et al.* [59] construct G from non local means filters (NL-means) of Morel [5], where $g_{i,j}$ is an $[s_x \times s_y]$ matrix with one in (i, j) position and zeros elsewhere. Thus, ξ_G is the set of overlapping patches of the source and reference images embedded in $d = s_x \times s_y$ dimensions.

There are also other choices of filters G, but different bases may lead to the same similarity measure. As it was explained in Chapter 3 the embedding into $[s_x \times s_y]$ patches is the same embedding (up to a rotation) as into $[s_x \times s_y]$ DCT coordinates, and so the similarity weights constructed from these embeddings are the same. However, emphasizing only specific frequencies in the embedding coordinates leads to different representations. This could be performed by applying a frequency weighting matrix, which is used for example in DCT-based applications like compression [8] and watermarking [68]. Each $[s_x \times s_y]$ DCT coefficient is multiplied by the corresponding element of the frequency weighting matrix, which is usually constructed to reduce the influence of high frequencies on the similarity.

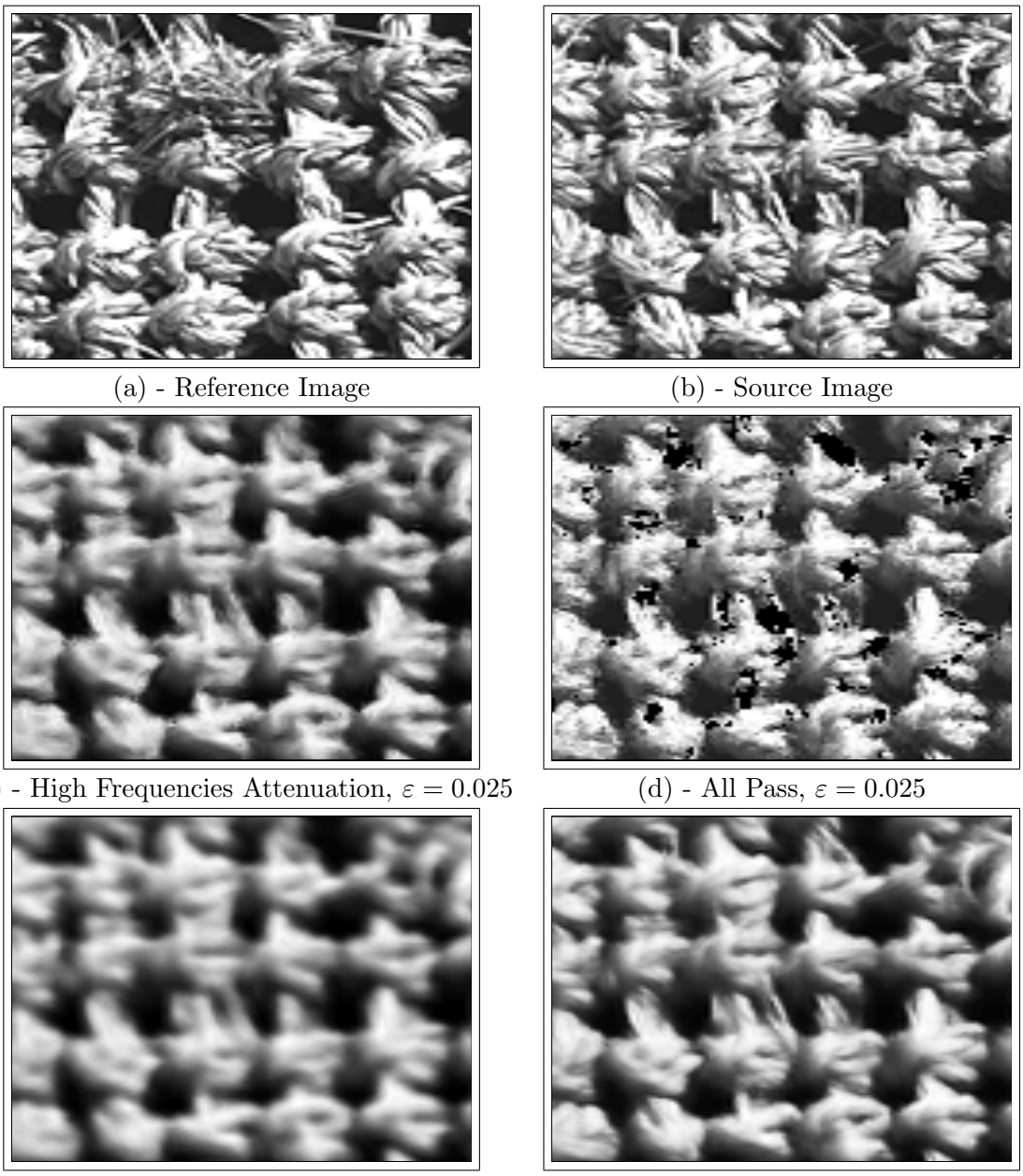
Figures 4.2(a) and (b) present two different parts of Brodatz texture image that are used as source and reference images. The source image is reconstructed from the reference image using $[8 \times 8]$ DCT coordinates feature space and frequency weighting matrix, which is equal either to the human visual frequency matrix given in [48], or all-pass frequency matrix (matrix with all the entries equal to 1). Another varying parameter is the similarity parameter ε. Larger ε results in smoother image and blur of the fine structure. Smaller ε better conserves the image details, but may make the reconstruction of regions with strong pattern variations impossible.

The influence of the frequency weighting matrix could be traced by fixing

ε to be constant and comparing Figs. 4.2(c) with (d) and Figs. 4.2(e) with (f). A feature space based on the DCT transform and all-pass frequency matrix leads to the same similarity relations as the original NL-means feature space, due to the linearity of the DCT. Attenuation of the coefficients, related to certain frequencies, lowers their similarity requirement and enables different reconstructions. In the presented experiment higher frequencies were attenuated, reducing the influence of the details on the similarity measure. Hence, in Fig. 4.2(c) the reconstruction was possible even in the regions that were not reconstructed using NL-means feature space in Fig. 4.2(d) (the black regions). However, lowered similarity requirement of the high frequencies coefficients reduces the reconstruction quality of the pattern details, because smooth regions and fine structures are distinguished less. The loss of details in reconstruction with weighting matrix is especially evident in Fig. 4.2(e), where higher ε is used. On contrary, the NL-means (DCT with all-pass matrix) reconstruction in Fig. 4.2(f) succeeds to preserve more fine structures. To summarize, DCT coordinates feature space with frequency weighting matrix is advantageous in the case of images with high spectral activity. These images are usually characterized by large number of small details with low spatial redundancy. Spectral activity of an image could be examined using a distribution of DCT coefficients that are found by applying DCT to the whole image. In our experiments of defect detection, the NL-means feature space was used due to the periodicity of the details in the inspected patterns.

4.2.3 Analysis of the Representation Error

To analyze the estimation error of eq. (4.5), we construct a set of points $X = \{\xi_G(\mathbf{s}) \bigcup \xi_G(\mathbf{s}') | \forall \mathbf{s}' \in \mathcal{N}_\mathbf{s}\} = \{x_1, x_2, ..., x_N\}$, such that the number of points N (the size of the set) is large, but finite. We assume that the points of the set $X \subseteq \mathcal{M}$ are independent and uniformly distributed on $\mathcal{M}$, a d-dimensional



(a) - Reference Image (b) - Source Image

(c) - High Frequencies Attenuation, $\varepsilon = 0.025$ (d) - All Pass, $\varepsilon = 0.025$

(e) - High Frequencies Attenuation, $\varepsilon = 0.05$ (f) - All Pass, $\varepsilon = 0.05$

Figure 4.2: DCT-based reconstruction: (a) Reference image; (b) source image; (c),(e) images reconstructed using the $[8 \times 8]$ DCT transform with a frequency weighting matrix given in [48] (suppression of the high frequencies), $\varepsilon = 0.025$ and 0.05 respectively; (d),(f) images reconstructed using the $[8 \times 8]$ DCT transform with all pass matrix (all the entries equal 1), $\varepsilon = 0.025$ and 0.05 respectively. A feature space based on the DCT transform and all pass frequency matrix leads to the same similarity relations as the original NL-means feature space, due to the linearity of the DCT. Reduction of the certain frequencies influence resembles adaptive adjustment of the similarity parameter in the frequency space and results in different reconstruction.

compact Riemannian manifold [1] [9, 57]. The last assumption is not trivial under finite set X and will be discussed below.

We assume that there exists a smooth function $f : \mathcal{M} \to \mathbb{R}$, such that $f(\xi_G(\mathbf{s})) = I_{\text{src}}(\mathbf{s})$ and $f(\xi_G(\mathbf{s}')) = I_{\text{ref}}(\mathbf{s}')$. We consider NL-means filter bank, which results in squared patches around the estimated pixels. To estimate a pixel value we calculate the inner product with a characteristic function of the central feature element

$$f(\xi_G(\mathbf{s})) = \langle \xi_G(\mathbf{s}), \chi_{\text{central}} \rangle = I_{\text{src}}(\mathbf{s}) \ , \tag{4.9}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product in the point $\xi_G(\mathbf{s})$ of $\mathcal{M}$, which is a Riemannian manifold and by definition is associated with a smooth inner product. Therefore, the function proposed in eq. (4.9) is smooth as required. The same function is valid for the reference image features construction.

In case of NL-means feature space, eq. 4.5 could be written in a feature space as:

$$\hat{\xi}_G(\mathbf{s}) = \frac{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2 / 2\varepsilon\} \xi_G(\mathbf{s}')}{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2 / 2\varepsilon\}} \ . \tag{4.10}$$

Denoting the estimation operator by $\mathbb{E}$, eq. (4.5) becomes

$$(\mathbb{E}f)(\xi_G(\mathbf{s})) = \frac{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|^2 / 2\varepsilon\} f(\xi_G(\mathbf{s}'))}{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|^2 / 2\varepsilon\}} \ , \tag{4.11}$$

where $(\mathbb{E}f)(\xi_G(\mathbf{s})) = \langle \hat{\xi}_G(\mathbf{s}), \chi_{\text{central}} \rangle = \hat{I}(\mathbf{s})$, which holds due to linearity of the inner product. Under the above assumptions, the estimation error is given by [57]

$$\hat{I}_{\text{src}}(\mathbf{s}) - I_{\text{src}}(\mathbf{s}) = (\mathbb{E}f)(\xi_G(\mathbf{s})) - f(\xi_G(\mathbf{s})) = \frac{\varepsilon}{2} \Delta f(\xi_G(\mathbf{s})) + O(\varepsilon^2) \ , \tag{4.12}$$

where Δ is a Laplace operator, defined for the point $\xi_G(\mathbf{s})$ on the manifold $\mathcal{M}$. Hence, for $\Delta f(\xi_G(\mathbf{s})) \ll 1/\varepsilon$ the estimation error is negligible. Unfortunately,

[1] Riemannian manifold is a real differentiable manifold in which each tangent space is equipped with an inner product in a manner which varies smoothly from point to point.

ε cannot be too small, because eq. (4.12) holds only if the number of points N grows faster than $\varepsilon^{-(\frac{d}{2}+1)}$ [9]. Intuitively the uniform distribution assumption implies that smaller N induces lower density of points around the estimated points. Therefore, if we decrease ε to zero, we will not be able to find any feature similar to the one estimated.

The second aspect that we consider is the fact that the data points X may not lie exactly on $\mathcal{M}$. Suppose that X is a perturbated version of $\mathcal{M}$ and there exists a perturbation function $\nu : \mathcal{M} \to X$, with small norm, such that every point in X can be written as $x + \nu(x)$, for some $x \in \mathcal{M}$. It was shown in [9] that the approximation used for obtaining eq. (4.12) is valid as long as the similarity parameter $\sqrt{\varepsilon}$ remains larger than the size of the perturbation $\|\nu(x)\|$. In our case, any feature can be presented generally as

$$\xi_G(\mathbf{s}) = x(\mathbf{s}) + r(\mathbf{s}) \ , \tag{4.13}$$

where $x(\mathbf{s})$ is an ideal point that belongs to the manifold $\mathcal{M}$ and represents the original pattern, and $r(\mathbf{s})$ denotes pattern variations under H_0 or a defect term under H_1. Hence, the estimation error, presented in eq. (4.12), is valid if

$$\sup_{\forall \mathbf{s} \in \Omega} \|r(\mathbf{s})\| < \sqrt{\varepsilon} \ . \tag{4.14}$$

For a defect-originated point, we assume that $\|r(\mathbf{s})\|$ is larger than $\sqrt{\varepsilon}$, hence the reconstruction in eq. (4.11) does not hold, which indicates presence of a defect.

4.3 Implementation of the Algorithm

Algorithm 1 summarizes the reconstruction and decision procedures for defect detection as described in eqs. (4.4)-(4.8). To verify whether a pixel from the source image belongs to a defect area according to eq. (4.8), we execute the following steps. A patch around every pixel in the source image is transformed

Algorithm 1 Defect Detection Using NL-means Estimation

1: {$\mathbf{s}$ - pixel index, I - source image, $\hat{I}$ - reconstructed source image}
2: **for all** $\mathbf{s} \in I$ **do**
3: $P_{\mathbf{s}} \Leftarrow$ construct a patch of size $[s_x \times s_y]$ around pixel $\mathbf{s}$
4: $i \Leftarrow 1$
5: {$\mathbf{s}'$ - pixel index, I_{ref} - reference image, $\mathcal{N}_{\mathbf{s}}$- search region neighborhood of $\mathbf{s}$}
6: **for all** $\mathbf{s}' \in \mathcal{N}_{\mathbf{s}}$ **do**
7: $P^i_{\mathbf{r}} \Leftarrow$ construct a patch of size $[s_x \times s_y]$ around pixel $\mathbf{s}'$
8: $\mathcal{W}^i \Leftarrow \exp(-\frac{\rho(P_{\mathbf{s}}, P^i_{\mathbf{s}'})^2}{2\varepsilon})$ {ρ - a distance metric}
9: $i \Leftarrow i + 1$
10: **end for**
11: $\mathcal{S}_{\mathcal{W}} \Leftarrow \Sigma_i \mathcal{W}^i$
12: **if** $\mathcal{S}_{\mathcal{W}} \rightarrow 0$ **then**
13: **for all** i **do**
14: $\mathcal{W}^i \Leftarrow 0$
15: **end for**
16: **else**
17: **for all** i **do**
18: $\mathcal{W}^i \Leftarrow \frac{\mathcal{W}^i}{\mathcal{S}_{\mathcal{W}}}$
19: **end for**
20: **end if**
21: $\hat{P}_{\mathbf{s}} \Leftarrow \Sigma_{\forall i} \mathcal{W}^i \cdot P^i_{\mathbf{r}}$ {source image patch estimation using reference neighboring patches}
22: $D(\mathbf{s}) \Leftarrow \|\hat{P}_{\mathbf{s}} - P_{\mathbf{s}}\|_2$ {difference image value at pixel $\mathbf{s}$ calculation}
23: $\hat{I}(\mathbf{s}) \Leftarrow \Sigma_{\forall i} \mathcal{W}^i \cdot I_{\text{ref}}(\mathbf{r}_i)$
24: **if** $\hat{I}(\mathbf{s}) = 0 \& I(\mathbf{s}) \neq 0$ **then**
25: $\mathbf{s} \in \mathcal{A}$ {$\mathcal{A}$ is a set of defect regions}
26: **end if**
27: **end for**

into a vector in the feature space using eq. (4.4), whose dimension is related to the defect size. A patch should be sufficiently large to contain the defect and its nearest surroundings, but not too large, to preserve the dominance of the defect presence. When no *a priori* information about possible defect size is available, the defect detection could be performed several times assuming different sizes in each run. Next, the source feature vector is estimated by feature vectors formed from neighboring patches from the reference image according to eq. (4.10). The neighborhood in the feature space is determined according to the neighborhood in the image space, which is a squared region centered at the tested pixel's spatial location. The neighborhood region must cover at least one period of the pattern to allow estimation without image registration.

The estimation is performed by similarity weights from eq. (4.6) that are calculated from Euclidian distances in the feature space [59]. The parameter ε in eq. (4.6) controls the relation between the distance in feature space and the corresponding weighting factor. It is important to choose a sufficiently large ε (weak similarity requirement) to enable reconstruction of the source image from the reference image even in case of pattern variations. However, ε should be sufficiently small (strong similarity constraint) to prevent reconstruction of defects from the reference image and thereby facilitate the distinction between pattern variations and defects. In our experiments we adjusted this parameter so that the reference image could be reconstructed from itself (a patch was reconstructed only from its neighbors). We chose the minimal ε that provided good reconstruction results.

The calculated weights are normalized only if their sum is larger than zero (in practice, a certain small threshold larger than zero is selected, e.g., 10^{-15}). Zero sum indicates lack of reference patches that are similar to the source patch, which implies that source patch is not reconstructible from the reference patches. Otherwise the weighted average of the neighboring reference patches estimates the source patch via eq. (4.10), and the tested pixel using

eq. (4.5). Hence, black pixels in the reconstructed image can be identified as defects.

In case of pattern variations, the performance is improved by increasing the search region to more than one period, at the expense of increasing the computational complexity. The complexity of the proposed algorithm is $O(n \cdot m \cdot d)$, where n is a data set dimension (number of pixels in the image), m is neighborhood dimension (search region size) and d is a feature space dimension (patch size). Therefore from computational point of view m and d should be as small as possible. In our experiments we used a region that covered 2-3 pattern periods.

4.4 Experimental Results

In this section, we evaluate the proposed algorithm by analyzing the receiver operating characteristics (ROC), and demonstrate its improved performance compared to using an anomaly detection algorithm [23,24]. In all the presented experiments the reconstruction is performed using the NL-means feature space.

The ROC curves are obtained by using images from the Brodatz textures database, where one texture is used for constructing a dictionary of patches (reference patches), and another texture is used for constructing a set of anomalous patches. In our simulations, the reference dictionary contains 5000 patches chosen randomly from one texture image, and additional patches from that texture image (which are different from the patches in the reference dictionary) are used for calculating the false alarm rate. The later set of patches is denoted as FAR-test data set. Each patch in the FAR-test data set is reconstructed from patches in the reference dictionary using eq. (4.10), and if the patch is not reconstructible, it contributes to false alarm rate. Similarly, each anomalous patch is reconstructed from patches in the reference dictionary, and if the patch is not reconstructible, it contributes to detection

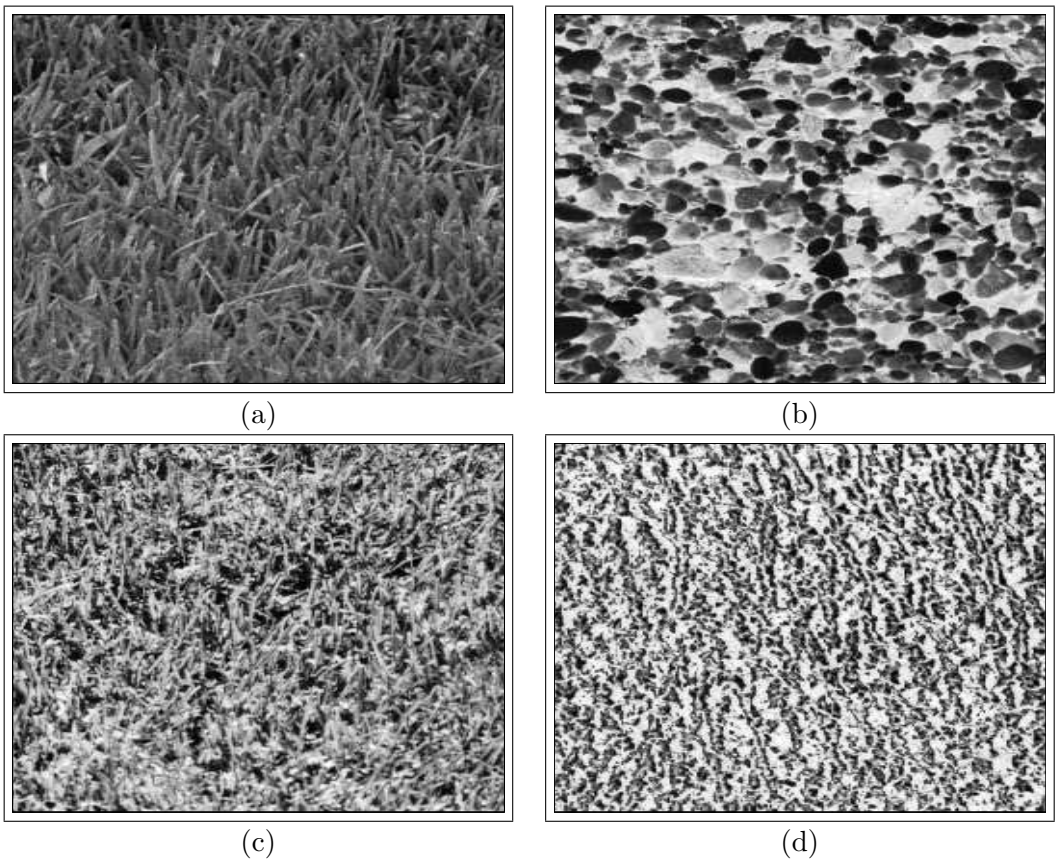
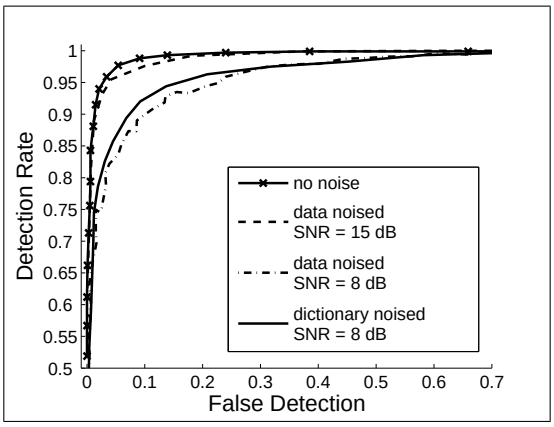


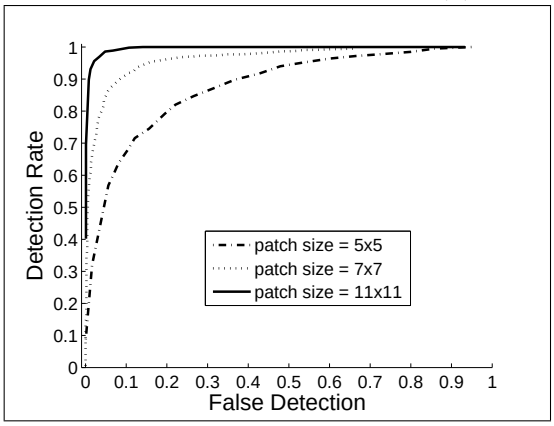
Figure 4.3: Brodatz textures for constructing a reference dictionary and anomalous patches. (a) and (b) Textures for the reference dictionary and FAR-test data sets; (c) and (d) textures for generating anomalous patches.

rate. By varying the value of ε from low values (corresponding to high detection and false-alarm rates) to high values (corresponding to low detection and false-alarm rates), we obtain an ROC curve. If we add noise to the data or change the size of patches, then we obtain different ROC curves.

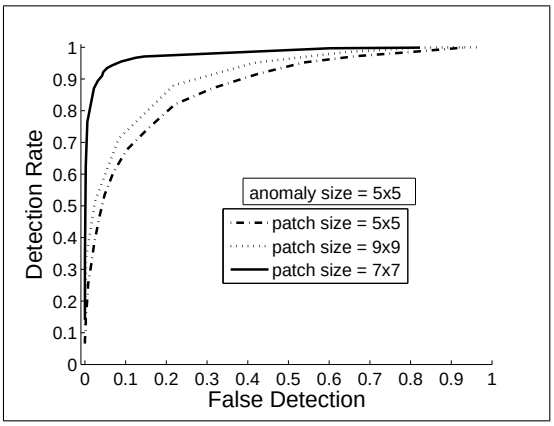
In the first experiment, we choose the texture shown in Fig. 4.3(a) for construction of the reference dictionary, and the texture shown in Fig. 4.3(c) for the anomalous patches. The patch size is 5×5 pixels, and white gaussian noise is added to either patches in the FAR-test data set or the reference dictionary. Figure 4.4(a) shows ROC curves for different signal-to-noise ratios (SNRs), and demonstrates the degradation in performance as the SNR decreases. We observe that noise in the reference dictionary is less significant than noise in the



(a) - Noise Influence



(b) - Anomaly Size Influence



(c) - Patch Size Influence

Figure 4.4: Performances dependence on : (a) additive white gaussian noise; (b) varying anomaly size (patches grow accordingly); (c) varying patch size with constant anomaly.

data, because a given patch may be reconstructed from many patches from the reference dictionary, and thus noise in the reference dictionary is averaged out when reconstructing the source patch, whereas noise in the data is generally not reconstructible from the reference dictionary.

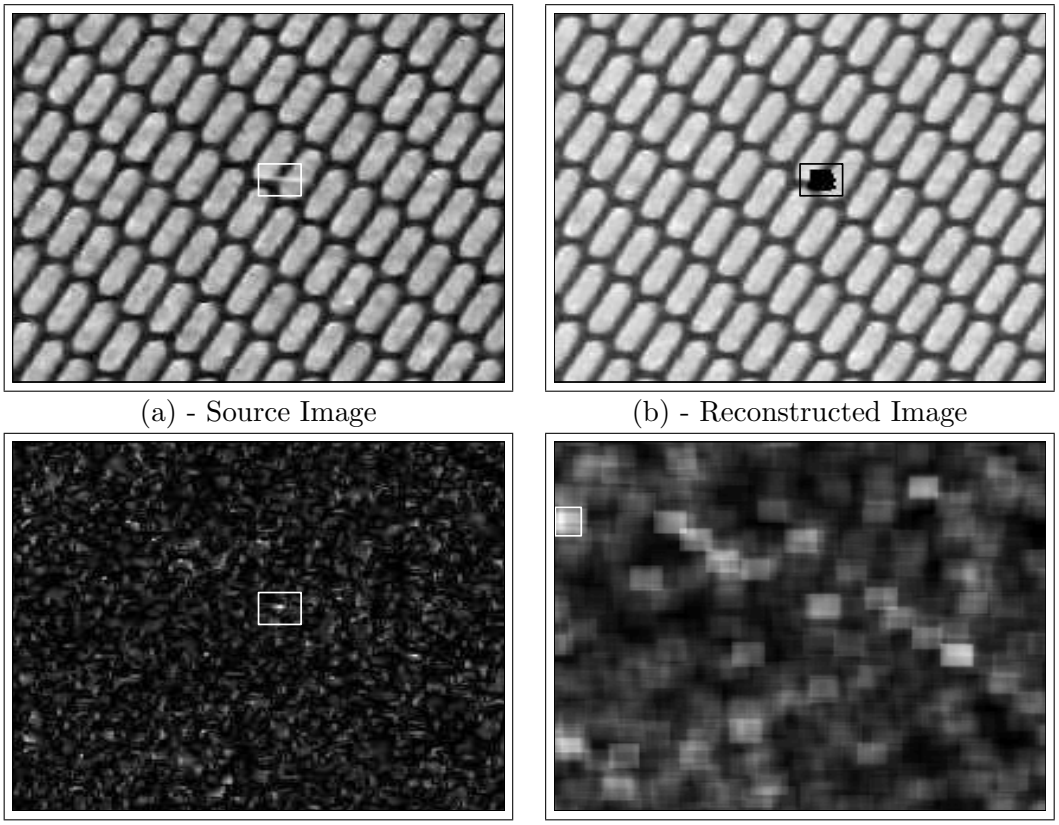
In the second and third experiments, we choose the texture shown in Fig. 4.3(b) for construction of the reference dictionary, and the texture shown in Fig. 4.3(d) for the anomalous patches. The influence of the anomaly size on the detection performances was studied. Figure 4.4(b) shows ROC curves for anomaly sizes of 5×5, 7×7 and 11×11 pixels, where the patch size varies according to the anomaly size. As expected, bigger anomalous patches are more

easily detected than smaller ones. However, detection of bigger anomalous patches involves higher computational complexity.

Figure 4.4(c) shows ROC curves for a constant anomaly size of 5×5 pixels and varying patch sizes around the anomaly of 5×5, 7×7 and 9×9 pixels. The area of the patch that doesn't contain anomaly pattern is filled with pattern from the reference texture. The best performances are achieved when the anomaly fills most of the patch's area (7×7 pixels), but not all of it (5×5) or small part of it (9×9). Hence, a patch should be sufficiently big to contain the anomaly and its nearest surroundings, but not too big, to preserve the dominance of the anomaly presence.

Finally, we apply the proposed algorithm to defect detection in wafers, and compare the results to those obtained by SHT on the difference image. The SHT does not require *a priori* information except rough estimate of defect size. It requires calculation of Mahalanobis distance given in eq. (2.8) in a feature space of the difference image and applying the SHT according to eq. (2.9) to the result.

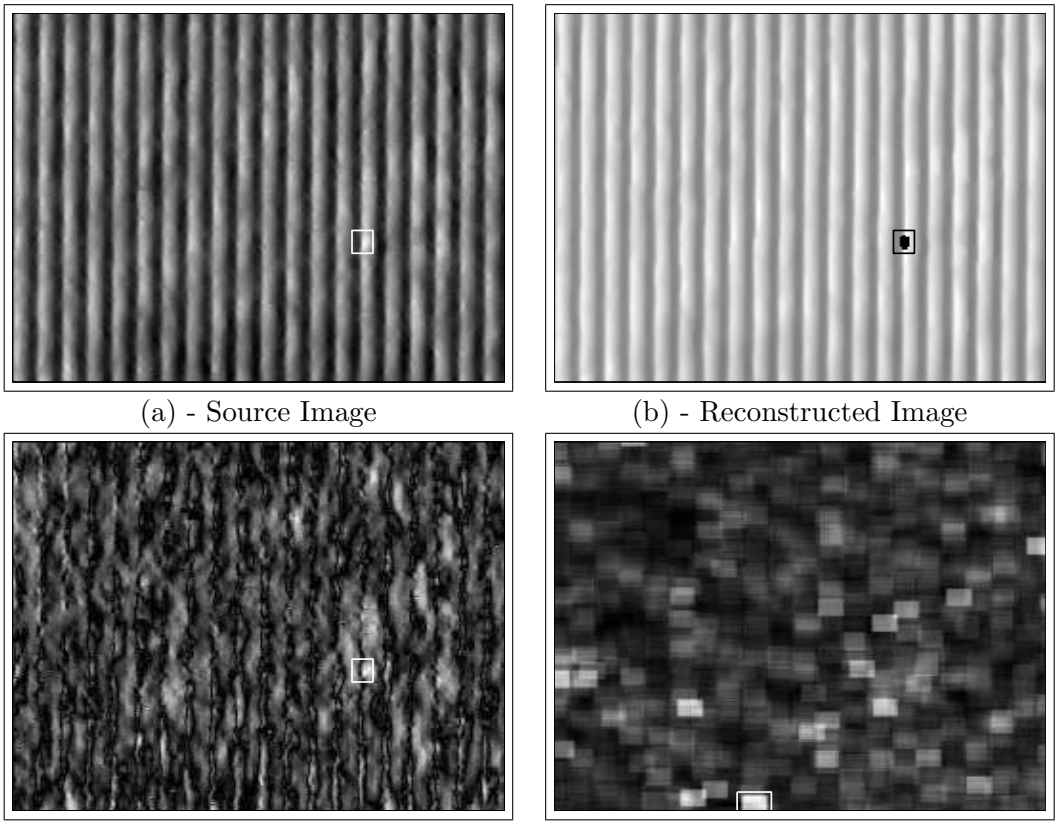
The feature space is constructed from patches formed around every pixel in the difference image, the size of the patches is the same as the size of the patches in the kernel-based detection algorithm. Figures 4.5 and 4.6 demonstrate the poor performance of SHT for wafer defect detection, which is a consequence of pattern variations. By contrast, the proposed approach successfully identifies the detects and is robust to pattern variations. The example presented in previous sections (Fig. 2.2) also demonstrates the advantage of the proposed algorithm in case of multiple defects. In case of a single defect, the SHT threshold in eq. (2.9) could be adjusted for only one detection. However in case of multiple defects, the SHT threshold adjustment becomes more complicated and false detections may appear, especially in the neighborhood of edges with pattern variations.



(a) - Source Image (b) - Reconstructed Image

(c) - Difference Image (d) - SHT on the Difference Image

Figure 4.5: Wafer defect detection. (a) Wafer image containing a defect (designated by white frame); (b) image reconstructed by Algorithm 1 ($\varepsilon = 0.02$, $\mathcal{N}_{\mathbf{s}} = [81 \times 81]$, $s_x \times s_y = [13 \times 13]$); (c) difference image (the white frame is around the defect location); (d) SHT on the difference image yields false detections and misses.



(a) - Source Image (b) - Reconstructed Image

(c) - Difference Image (d) - SHT on the Difference Image

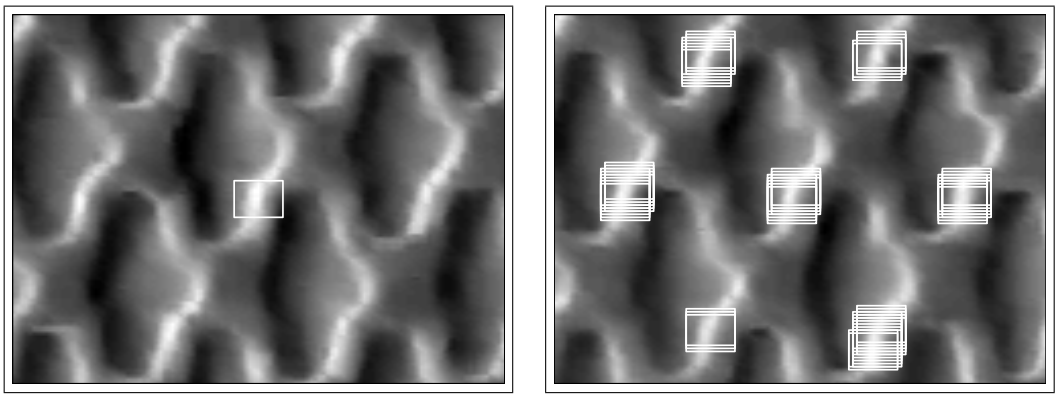
Figure 4.6: Wafer defect detection. (a) Wafer image containing a defect (designated by white frame); (b) image reconstructed by Algorithm 1 ($\varepsilon = 0.015$, $\mathcal{N}_{\mathbf{s}} = [151 \times 151]$, $s_x \times s_y = [13 \times 13]$); (c) difference image (the white frame is around the defect location); (d) SHT on the difference image yields false detections and misses.

4.5 Discussion

In this section we discuss the robustness of the algorithm for pattern variations and misregistration in the case of periodic and non-periodic patterned wafers. Although we refer to the NL-means filters feature space, the conclusions are relevant to other possible feature spaces.

The robustness for the pattern variations is a major advantage of the kernel-based algorithm compared to the difference image approach, as demonstrated in Figs. 4.1,4.5 and 4.6. Figure 4.7 shows the exploitation of pattern periodicity, which allows to overcome the problem of pattern variations. An inspected patch, marked with a white frame in Fig. 4.7(a), does not have to be identical to one reference patch, but could be a combination of several marked patches from Fig. 4.7(b). Moreover, Morel *et al.* [4] considered denoising image sequences using NL-means filters and showed that motion estimation between the sequences is not necessary. Motion estimation between sequences is analogous to image registration between the source and reference images. Hence, the proposed method is robust for misregistration, because similar patches can be found in different regions of the reference image, as it is shown in Fig. 4.7(b). Patches in Fig. 4.7(b) are marked as similar to the inspected patch from Fig. 4.7(a), if their similarity measure according to eq. (4.6) is above a determined threshold.

Due to the nature of the algorithm, a favorable case for NL-means is a periodic case, like periodic patterned wafers images. If the inspected pattern is not periodic, the proposed algorithm will be able to distinguish between the pattern-originated patches and defect-originated patches, only if the search region contains the respective pattern. The misregistration is tolerable only within the search region. Hence, either rough registration should be performed, or the search region should be chosen very large, which is a disadvantageous due to the increased computational complexity. The compensation for pattern



(a) (b)

Figure 4.7: Exploitation of periodicity of the pattern. (a) Region from a source image. A small patch is marked with a white frame in the center; (b) Aligned region from the reference image. Patches in (b) are marked as similar to the inspected patch from (a), if their similarity measure according to eq. (4.6) is above a determined threshold.

variation will be less effective and fusion of detections in different SEM channels may be required to prevent false detections caused by pattern variations. In the next chapter we propose a fusion process based on the joint similarity measure to improve detection performances in the case of non-periodic patterns.

4.6 Summary

We have introduced defect detection algorithm, which avoids registration of the source and reference images and is robust to pattern variations. The detection method is based on the reconstruction of the source image from the reference image, which is carried out in the feature space using anisotropic kernels as the similarity measure between the reconstructed feature of the source image and its reconstructing reference features. We have described the reconstruction procedure and presented analysis of the reconstruction error and validity. We have compared NL-means feature space with feature space based on the DCT combined with frequency weighting matrix. We have also shown simulations based on Brodatz images, which test algorithm ROC-performance dependence

on the size of anomaly and the additional noise. Finally, we have presented experimental results, which demonstrate improved defect detection compared to the detection based on SHT of the difference image.

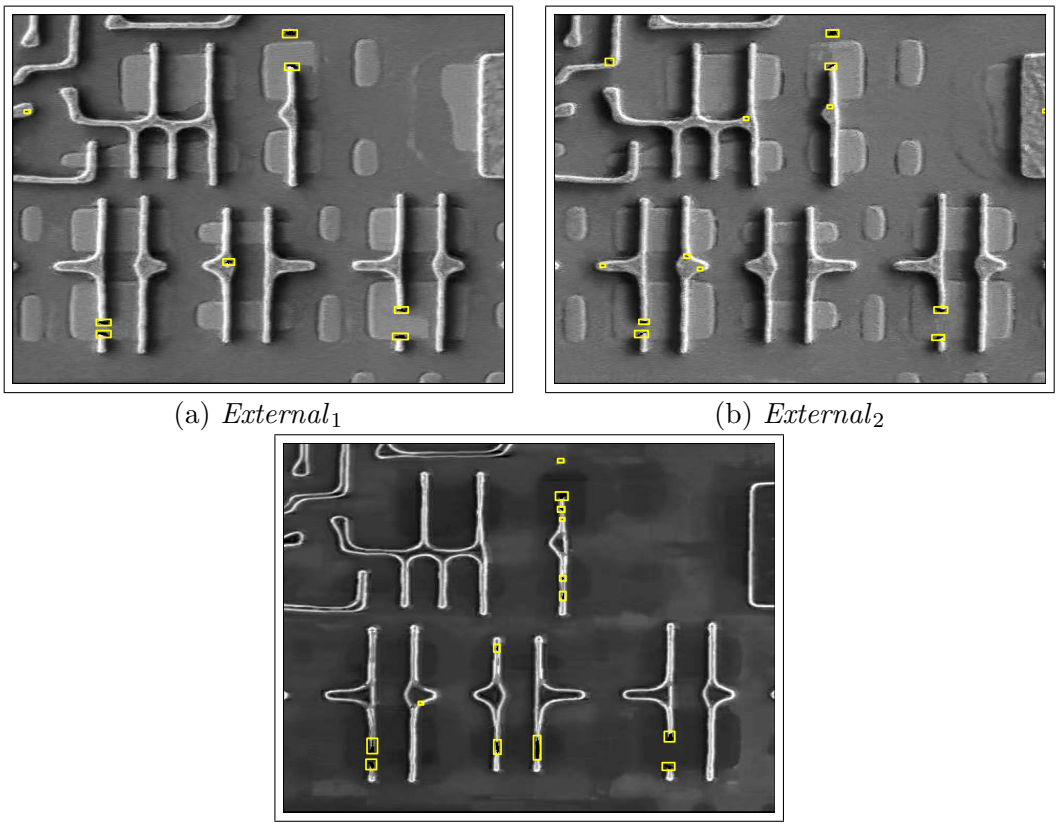
Chapter 5

Multi-Channel Defect Detection

5.1 Introduction

In Chapter 4, we have proposed a novel kernel-based defect detection algorithm, which outperforms the results of existing anomaly detection method. The proposed procedure is based on the reconstruction of the source image from its reference image, and when applied to periodic patterned wafer images, does not require image registration and is robust to pattern variations. Registration of a reference image relatively to the source image is not required, as long as a search region covers at least one pattern period. Furthermore, a source patch is not compared to one reference patch, but to several patches in the search regions. Hence, it doesn't have to be equal to only one patch, but could be equal to a combination of several similar patches, which overcomes the problem of pattern variations. For periodic patterns, the search region is taken often more than one period of the pattern, in order to increase the number of potentially similar reference patches.

If the pattern is non-periodic, the proposed algorithm will be able to reconstruct pattern-originated regions and to distinguish them from defects, only if the reference search region contains the respective pattern. The misregistration is tolerable only within the search region, hence either rough registration should be performed, or the search region should be chosen very large, which is



(a) *External*$_1$ (b) *External*$_2$

(c) *Internal*

Figure 5.1: Reconstructed wafer images with overlaid defect detection results obtained by using single-channel kernel-based algorithm (suspicious regions are denoted by light rectangulars). High false detection rate results from sensitivity to pattern variations.

a disadvantageous due to the increased computational complexity. Moreover, the compensation for pattern variations is insufficient, and false detections occur. For example the source images from Figs. 1.2 are reconstructed using slightly misregistered reference images from Fig. 1.4. The reconstruction results are shown in Fig. 5.1, where non reconstructible regions are marked and identified as defects. In all the three channels, there are some false detections, which are associated with differences between the source and reference images, rather than with defects. Our objective is to eliminate such false detections by exploiting relations between the three channels and performing automatic detection based on these relations.

5.2 Statistical Interpretation

In this section, we discuss the statistical interpretation of the single-channel kernel-based method, which facilitates its extension to multichannel. During the single-channel detection procedure, every pixel in the source image with index $\mathbf{s}$ is represented by a feature x around it. All the pixels in its reference search region $\{\mathbf{s}_i \in \mathcal{N}_\mathbf{s} | i = 1..m\}$ are also represented by features $\{x_i | i = 1..m\}$ around them. The respective search region in the reference image is determined according to a spatial neighborhood of pixel $\mathbf{s}$. Although our discussion refers to the NL-means feature space, which is received using the filters bank given by eq. (3.11), it holds for other possible feature spaces. For every reference feature in the neighborhood, a similarity measure $k(x_i, x)$ is calculated, where $k(x, y) \equiv e^{-\frac{1}{2}(\|x-y\|_2^2/\varepsilon)}$ is the Gaussian kernel [64]. The reconstruction of the source feature x from its reference feature is obtained by

$$x = \sum_{i=1}^{m} \frac{k(x_i, x)}{\sum_j k(x_j, x)} x_i \,, \tag{5.1}$$

Let us denote the kernel-mapping x_X, given by eq. (3.17), as a *within-similarity map* between the elements of X and x. The normalizing factor $\sum_{i=1}^{m} k(x_i, x)$ could be regarded as a L_1 norm of a *within-similarity map* x_X. Normalizing this norm by the number of reference features, we denote it as the total similarity of the source feature to the pattern, which is represented by the reference features:

$$\frac{1}{m} \sum_{i=1}^{m} k(x_i, x) = x_X^T \cdot \vec{1}_m \,, \tag{5.2}$$

where $\vec{1}_m$ denotes an m-dimensional column vector with all components equal to m^{-1}. If the total similarity is close to zero, which indicates that the source feature cannot be well reconstructed from the pattern of the reference image, then presence of a defect is declared.

As it was shown in Chapter 3, the average of the *within-similarity map* relates to the Parzen's nonparametric estimation of a probability density $p(x)$

(see eq. 3.18):

$$\hat{p}_\varepsilon(x) = x_X^T \cdot \vec{1}_m \ . \tag{5.3}$$

Hence, the similarity measure in eq. (5.2) can be considered as the likelihood of the source feature to arise from the pattern statistics, which is represented by reference features from the search region. Features originated form defects will have low similarity to the pattern, and hence can be identified. We aim to perform multichannel detection using joint similarity between the three-channels source features x, y and z, which relate to the same pixel $\mathbf{s}$, and their reference features $X = \{x_i\}_{i=1}^n$, $Y = \{y_i\}_{i=1}^n$ and $Z = \{z_i\}_{i=1}^n$ (from $External_1$, $External_2$ and *Internal* images respectively).

5.3 Consistent Similarity

We assume that if a pattern-originated region in the source wafer is similar to certain regions in the reference wafer, then this similarity is maintained across the three SEM images. NL-means feature space results in overlapping patches around the pixels of the source and reference images. Accordingly, in the image space the similarity between a source patch and its reference patches in the three channels is constrained by a consistency criterion that the locations of reference patches, which are most similar to the source patch, are identical in the three channels.

The consistent similarity concept for three channels is demonstrated in Figs. 5.2 and 5.3. The figures show two columns of $External_1$, $External_2$ and *Internal* images. In the left column of Fig. 5.2, a source patch that is free of defects is delineated by a rectangular. The second column shows the corresponding reference images, and the most similar patches (for each channel, the patches in the reference image that are most similar to the source patch are identified by rectangulars). In the left column of Figs. 5.3, a source patch that contains a defect is delineated by a rectangular. The second column

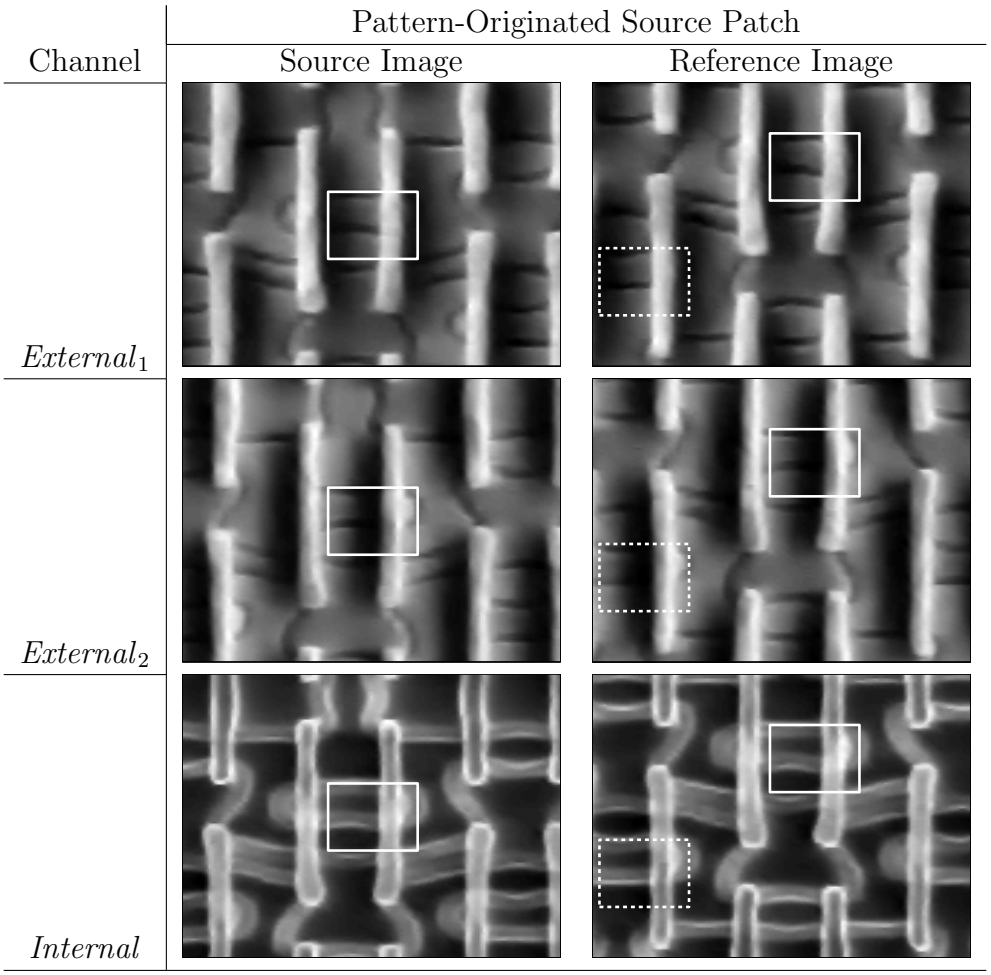


Figure 5.2: Example of similarity consistency between channels for a defect-free source patch. The figure shows four columns of $External_1$, $External_2$ and *Internal* images. In the left column, a source patch that is free of defects is delineated by a rectangular. The second column shows the corresponding reference images, and the patches that are most similar to the source patch. The locations of the most similar reference patches are the same in all channels (cf. second column).

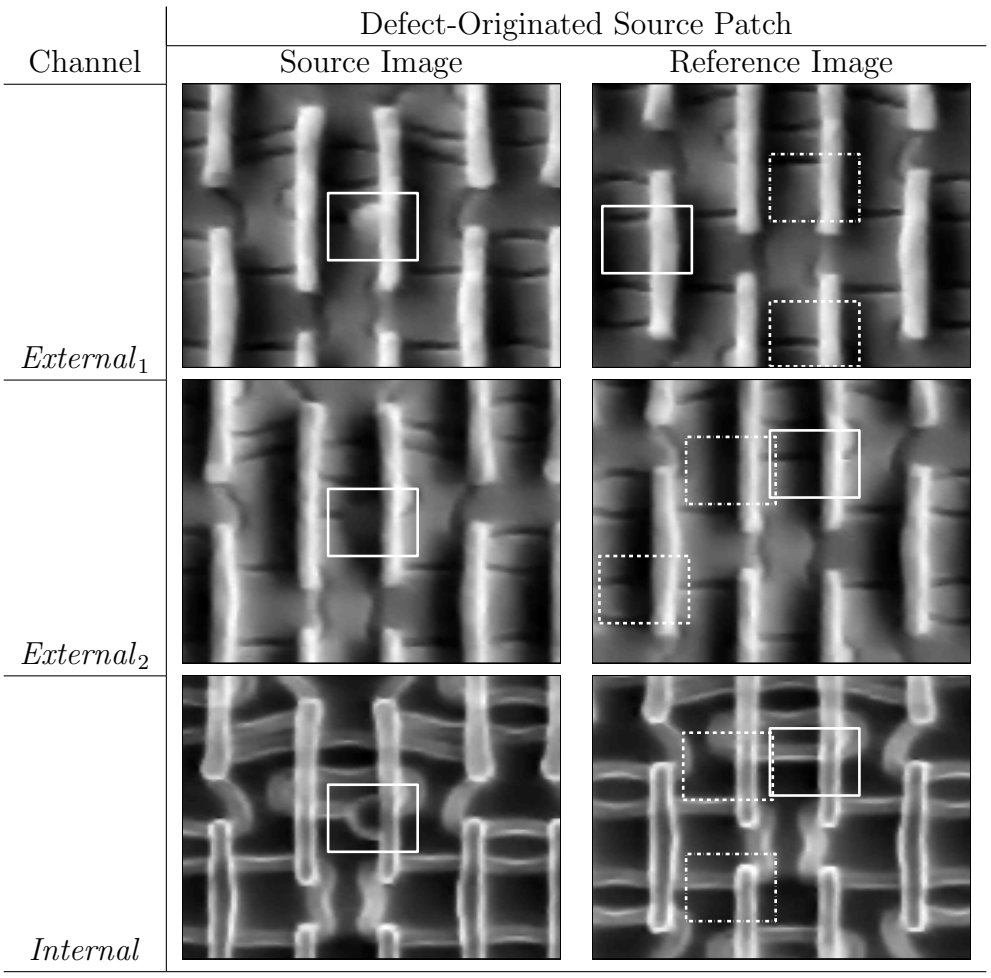


Figure 5.3: Example of similarity consistency between channels for a defect source patch. The figure shows two columns of $External_1$, $External_2$ and *Internal* images. In the left column, a source patch that contains a defect is delineated by a rectangular. The second column shows the corresponding reference images, and the most similar patches. The locations of the most similar reference patches may be different in each channel (cf. second column).

shows the corresponding reference images, and the most similar patches. It turns out that the locations of the most similar reference patches are the same for all channels in case the source patch is free of defects (cf. second column). However, in case the source patch contains a defect, the locations of the most similar reference patches may be different for each channel (cf. forth column). Consistent similarity is an additional characteristic that may be used to enhance multi-channel distinction between defective and pattern-originated regions.

5.4 Joint Similarity Measure

In this section we introduce the joint similarity measure, which constitutes a basis for the multichannel defect detection algorithm. For clarity of presentation, first we consider two-channel consistent similarity, and then extend the formulation to three channels. Based on a data sample $X = \{x_1, ..., x_m\}$, the general form of a kernel density estimator is given as $\hat{p}_{\varepsilon,\gamma}(x) = \sum_{i=1}^{m} \gamma_i b_\varepsilon(\|x - x_i\|)$, where γ_i are nonuniform weighting coefficients. According to the relation (eq. 3.18) between the Parzen's estimator and the similarity measure, the weighting coefficients γ_i could be regarded as *a priori* estimation of a similarity between a reference patch x_i and a source patch x. Without any *a priori* assumptions about these similarity relations, we set γ_i to

$$\gamma_i^{(0)} = \frac{1}{m}, \quad \forall i \ . \tag{5.4}$$

We refer to eq. (5.4) as a zero-order *a priori* estimation of a reference patch x_i to be similar to the source patch x. Given the *within-similarity map* of a patch y and the reference patches $\{y_i\}_{i=1}^{m}$ and using the assumption of consistent similarity between the channels, we propose the following first-order refinement of eq. (5.4)

$$\gamma_i^{(1)} = \frac{k(y_i, y)}{\sum_{i=1}^{m} k(y_i, y)} \ , \tag{5.5}$$

where x and y relate to the same pixel $\mathbf{s}$ in two different channels. Hence, we can write the conditional estimated probability of a patch x to arise from the pattern statistics, given a *within-similarity map* of a respective patch y from another channel, as follows:

$$\hat{p}(x|y) = \sum_{i=1}^{m} \gamma_i^{(1)} k(x_i, x) \ . \tag{5.6}$$

Note, that if $k(y_i, y) = K, \forall i$, then $\gamma_i^{(1)}$ reduces to $\gamma_i^{(0)}$ and eq. (5.6) reduces to eq. (3.18).

Under the assumption of consistent similarity between the channels and the proposed eq. (5.5) and eq. (5.6) we have

$$\begin{aligned} \hat{p}(x, y) = \hat{p}(x|y)\hat{p}(y) &= \left(\sum_{i=1}^{m} k(x_i, x) \frac{k(y_i, y)}{\sum k(y_i, y)} \right) \left(\frac{1}{m} \sum k(y_i, y) \right) \\ &= \frac{1}{m} x_X^T \cdot y_Y \ . \end{aligned} \tag{5.7}$$

Hence, we propose to define the scalar product of the two *within-similarity maps* as the joint similarity measure of two channels:

$$J_{\mathrm{xy}} = \frac{1}{m} x_X^T \cdot y_Y \ . \tag{5.8}$$

The single-channel similarity measure, given by eq. (5.2), quantifies the total similarity of a source patch x to its reference patches. Hence, it answers the question whether a source image has similar patches in its reference search region. The joint similarity measure answers this question and the additional one, which is whether the similar patches locations in the first channel are consistent with the similar patches locations in the second channel. Geometrically, it relates to the cosine of the angle between the *within similarity maps* x_X and y_Y (consistent similarity measure), multiplied by their L_1 norms (single-channel similarity measures).

Note that, the joint probability is symmetric

$$\begin{aligned}\hat{p}(y,x) &= \left(\sum_{i=1}^{m} k(y_i,y)\frac{k(x_i,x)}{\sum k(x_i,x)}\right)\left(\frac{1}{m}\sum k(x_i,x)\right)\\ &= \frac{1}{m}y_Y^T \cdot x_X = \frac{1}{m}x_X^T \cdot y_Y = \hat{p}(x,y)\ . \end{aligned} \tag{5.9}$$

However it disobeys the triangle inequality (*i.e.* it is non-metric distance), which often happens in case of distance functions that are robust to outliers and noise [31]. It can be shown that its negative natural logarithm is a probabilistic distance [16]. For the convenience of presentation we adopt its positive natural logarithm as a likelihood measure.

From eqs. (5.8) and (5.7) we can define a joint similarity measure for three channels by:

$$\begin{aligned}J_{\text{xyz}} &= \hat{p}(x,y,z) = \hat{p}(x,y|z)\hat{p}(z)\\ &= \left(\sum_{i=1}^{m} k(x_i,x)k(y_i,y)\frac{k(z_i,z)}{\sum k(z_i,z)}\right)\frac{1}{m}\sum_{i=1}^{m} k(z_i,z)\\ &= \frac{1}{m}\sum_{i=1}^{m} k(x_i,x)k(y_i,y)k(z_i,z)\ , \end{aligned} \tag{5.10}$$

and the corresponding joint likelihood to the pattern is measured by

$$L_{\text{xyz}} = \log(J_{\text{xyz}})\ . \tag{5.11}$$

5.5 Implementation of the Algorithm

Algorithm 2 summarizes the reconstruction and decision procedures for multi-channel defect detection. To verify whether a pixel from the source image belongs to a defect area according to eq. (4.8), we execute the following steps. Relative patches are constructed around every pixel in the source images of all the channels and transformed into vectors (similarly to the single-channel algorithm). The respective *within similarity maps* are calculated in

Algorithm 2 Multi-Channel Defect Detection Using NL-means Estimation

1: {$\mathbf{s}$ - pixel index in the source images, Ω - image support}
2: **for all** $\mathbf{s} \in \Omega$ **do**
3: ch $\Leftarrow 1$
4: **for all** CHANNELS **do**
5: $P_{\mathrm{ch},\mathbf{s}} \Leftarrow$ construct a patch of size $[s_x \times s_y]$ around pixel $\mathbf{s}$ in the source image of the respective channel
6: $i \Leftarrow 1$
7: {$\mathbf{s}'$ - pixel index in the reference image, $\mathcal{N}_{\mathbf{s}}$- search region neighborhood of $\mathbf{s}$}
8: **for all** $\mathbf{s}' \in \mathcal{N}_{\mathbf{s}}$ **do**
9: $P^i_{\mathrm{ch},\mathbf{s}'} \Leftarrow$ construct a patch of size $[s_x \times s_y]$ around pixel $\mathbf{s}'$ in the respective channel
10: $\mathcal{W}^i_{\mathrm{ch}} \Leftarrow \exp(-\frac{\rho(P_{\mathrm{ch},\mathbf{s}}, P^i_{\mathrm{ch},\mathbf{s}'})^2}{2\varepsilon_{\mathrm{ch}}})$ {ρ - a distance metric}
11: $i \Leftarrow i + 1$
12: **end for**
13: ch $\Leftarrow$ ch $+ 1$
14: **end for**
15: $J_{\mathbf{s}} = \Sigma_{\forall i} \Pi_{\forall \mathrm{ch}} \mathcal{W}^i_{\mathrm{ch}}$
16: **if** $\log(J_{\mathbf{s}}) < \tau$ **then**
17: $\mathbf{s} \in \mathcal{A}$ {$\mathcal{A}$ is a set of defect}
18: **end if**
19: **end for**

every channel for the determined search region via eq. (3.17), and the joint similarity measure (likelihood to the pattern) is calculated according to eq. (5.10). Finally, the detection is performed by thresholding the log likelihood values.

Pattern variations between the source and reference images diminish the similarity of the source feature to its reference features. Too small ε applies high similarity demand, which may cause false detection in a presence of pattern variations between the source and reference images. Often pattern variations are more disturbing in the *Internal* channel, what can be for example observed by comparing Figs. 1.2(b)-(d) and Fig. 1.4. It is important to diminish the influence of the channel with high pattern variations on the detection procedure, by adjusting the related ε to be relatively large. In the presented

examples we base our choice of the similarity parameter on the global similarity of the source and reference images in every channel, which is determined according to the Bhattacharyya distance between the gray levels histograms of the images.

The Bhattacharyya measure can be used to compare the similarity between two histograms as follows. Let R_i to be the frequency coded quantity in bin i for the reference image histogram and S_i a similar quantity for the source image histogram. The Bhattacharyya distance $-\log\left(\sum_i \sqrt{R_i S_i}\right)$ provides a measure of similarity between the two histograms and hence between the source and reference images. The successful applications of the Bhattacharyya measure for histogram matching and images similarity testing can be found in numerous applications [1, 58]. Thus, we propose to determine the similarity parameter for every channel according to

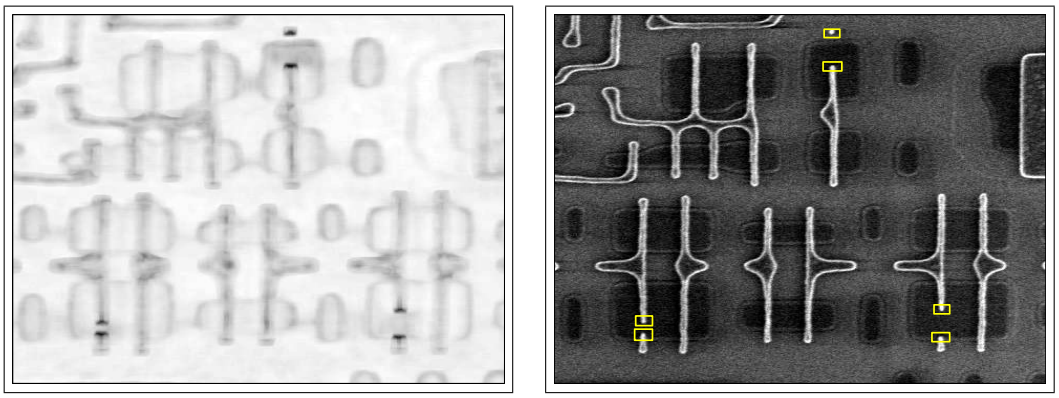
$$\varepsilon_k = -\log\left(\sum_i \sqrt{R_i^k S_i^k}\right) . \tag{5.12}$$

The above approach allows relative adjustment of ε between the channels, the absolute value remains to be determined by single-channel adjustment procedure, like self-reconstruction of the reference image (see Chapter 4, Implementation of the Algorithm).

5.6 Experimental Results

Figure. 5.4 shows detection results applying the joint similarity measure to the three images from Fig. 1.2. Thresholding low likelihood values reveals the defect edges that are marked on the *Internal* image.

In this example only the edges of the defects are detected, because the patch window is relatively small compared to the whole defect. In general, the predicted size of defects determines the choice of a patch size. However defects' sizes are not known *a priori* and may vary from defect to defect.



(a) (b)

Figure 5.4: Detection results using a joint similarity measure:(a) Joint similarity $\log(\hat{p}(x, y, z))$ (the gray level represents the likelihood that a patch is similar to the pattern). (b) *Internal* image with overlaid detection results (denoted by light rectangles).

Small size patches (11×11 in our example) will capture small defects and large defects edges, whereas large patches will miss smaller defects. Moreover, the computational complexity increases as the patch size becomes larger, hence smaller patches are generally preferable.

The proposed kernel-based algorithm was successfully applied to images with pattern variations, wherever the algorithm that was based on the difference image failed. The multi-channel algorithm was also compared with single-channel algorithm that was based on calculation of log-likelihood in every channel separately, without similarity consistency constraint. The joint multi-channel detection outperformed separate detections, in case of non-periodic patterns and defects that were not evident in all the channels. Tables 5.1 and 5.2 present several examples of the comparison, which involve different patterns. Table 5.1 presents detection results, which include the number of exact detections (D) and the number of false detections (FD). Table 5.2 present separation tolerance for the same cases, which was calculated according to the ratio between the range of threshold values that allow exact detection without false alarms and the range of the log-likelihood values. In cases 1–5 the source

Case	Detection Results							
	Single Channel Detection						Multi-Channel Detection	
	External$_1$		*External*$_2$		*Internal*			
	D	FD	D	FD	D	FD	D	FD
1	1(p)	0	1(p)	0	0	1	1	0
2	0	2	0	1	1	0	1	0
3	0	2	1	0	0	2	1	0
4	1	0	1	4	0	1	1	0
5	0	1	1	0	0	1	1	0
6	2 of 2	> 5	1(p) of 2	1	1(p) of 2	5	2 of 2	0
7	8 of 9	0	9 of 9	> 5	2 of 9	0	9 of 9	0

Table 5.1: Detection and false detection results obtained by using the multi-channel and single-channel algorithms.

Case	Separation Tolerance			
	Single Channel Detection			Multi-Channel
	External$_1$	*External*$_2$	*Internal*	Detection
1	0.3	0.3	0	0.5
2	0	0	0.4	0.3
3	0	0.35	0	0.3
4	0.1	0	0	0.2
5	0	0.1	0	0.1
6	0	0	0	0.1
7	0	0	0	0.1

Table 5.2: Separation tolerance of multi-channel and single-channel algorithms. It was calculated according to the ratio between the range of threshold values that allow exact detection without false alarms and the range of the log-likelihood values.

image contains only one defect, while in cases 6–7 the source image contains several defects (two and nine respectively) of different sizes and shapes. In all the tested cases, the threshold was set automatically for optimal outliers detection. The search region and the patch sizes were identical for all cases. In cases where the detection was partial, *i.e.* only part of a defect was detected, this was stated (p). The results show that performance of the single channel algorithm is not sufficient. The application of the single channel algorithm to the $External_1$, $External_2$ or *Internal* channel yields either low detection rate or high false detection rate. Furthermore, combing the single channel detection results by 'AND' or 'OR' operations over the three channels, may either increase the false detection rate or decrease the detection rate. Compared to the single channel algorithm, the multi-channel algorithm enables to improve the detection rate while decreasing the false detection rate.

Figures 5.5 and 5.6 present an example of the advantage of the multi-channel detection procedure over simple union or intersection of the detection results in different channels. In the presented example, the defect is less distinguished in the $External_1$ and *Internal* channels. Hence, the single-channel detections in these channels are characterized by false detections, as demonstrated in Figs. 5.5(d),(f). It is evident that the union of the results yields high false detection rate and could not be applicable. The defect is a bump, which appears as non-overlapping bright blobs in the *External* images. The detected regions in these channels are orthogonal, that is their intersection is empty. Hence, the intersection of the detection results in the three channels is empty and misses the defect. The joint detection, presented in Fig. 5.6(a) detects the defect precisely, as it is presented in Fig. 5.5(b)-(d). The joint detection allows to avoid misses and faults compared to the union, which may cause high false detection rate, and intersection, which may miss a defect especially when the defect is visible only in one channel.

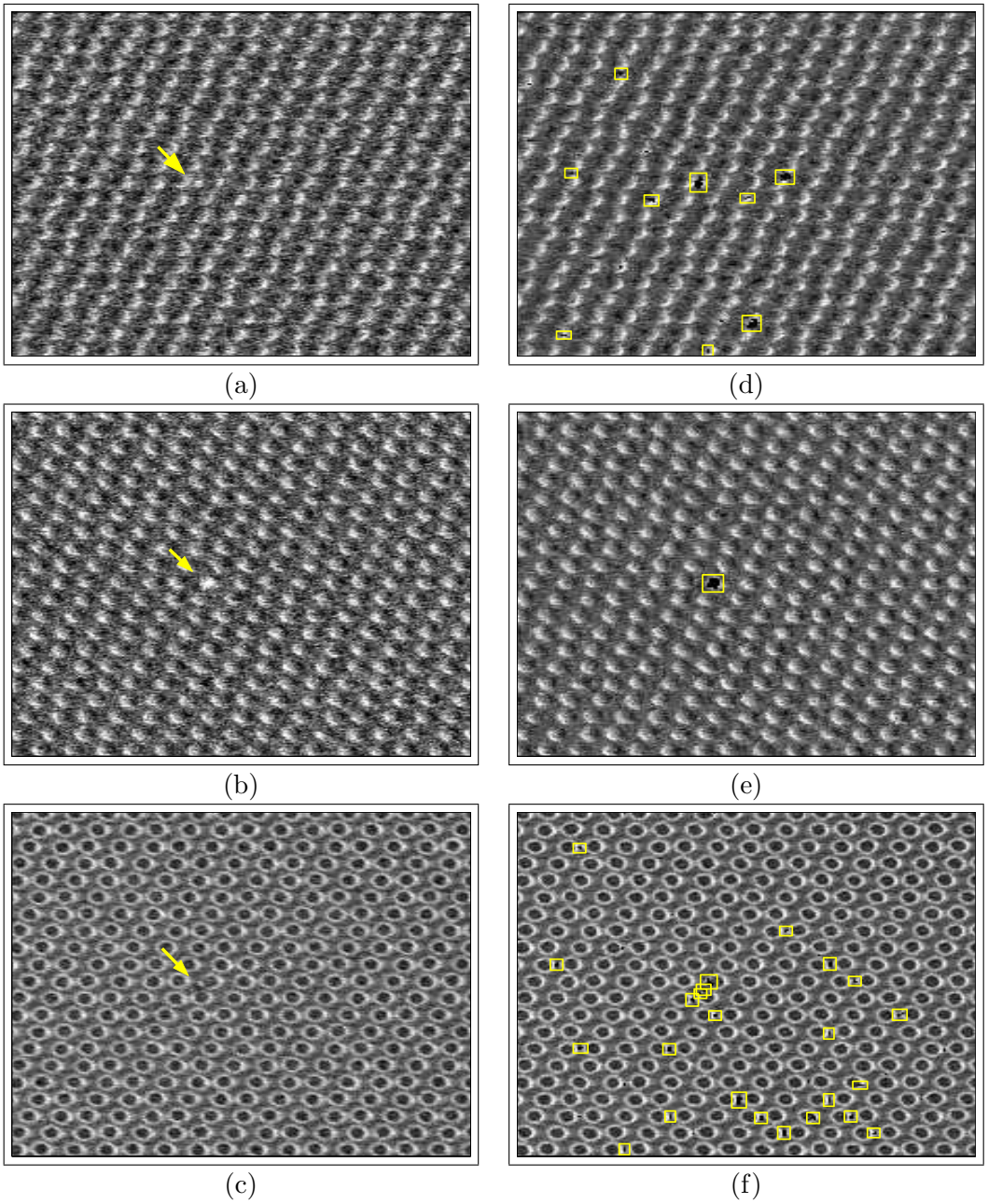


Figure 5.5: (a)-(c) Source images in the $External_1$, $External_2$ and *Internal* channels, respectively; (d)-(f) Single-channel reconstruction results in the three channels; Non-reconstructed (black) regions are detected as defects and marked with light rectangles (the regions are marked only if they contain more than two pixels).

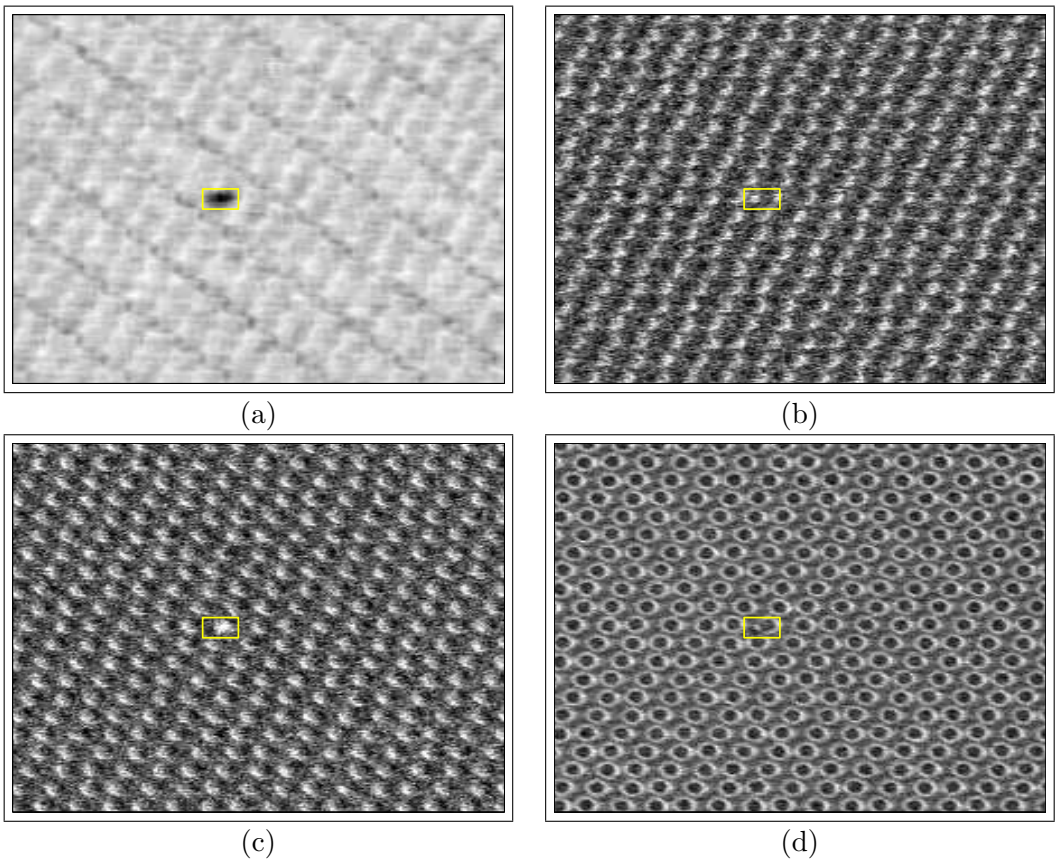


Figure 5.6: Detection results using a joint similarity measure:(a) Joint similarity $\log(\hat{p}(x, y, z))$ (the gray level represents the likelihood that a patch is similar to the pattern). (b)-(d) All channels images with overlaid detection results (denoted by light rectangles).

5.7 Summary

We have proposed an algorithm for automatic defect detection in wafers using three channels of SEM images. The kernel-based detection algorithm exploits the periodic nature of the wafer pattern and compensates for the pattern variations and misregistration. If the inspected pattern is not periodic, the proposed method exploits the multi-channel information to compensate for pattern variations. We have introduced a kernel-based similarity measure that quantifies similarity relations between the inspected patch and its reference patches under the assumption that the locations of similar patches in the search regions

are invariant for all the channels. Finally, we have demonstrated improved performances of the constrained multichannel detection comparing to the single channel detection in the case of non-periodic pattern. We have also demonstrated that the multi-channel detection is preferable over simple union or intersection of the single-channel detections.

Chapter 6

Conclusions

6.1 Summary

We have presented the problem of defect detection in patterned wafers. We have described the production of three-channels images using SEM tool. We have demonstrated that the variety and unpredictability of the defects make it impossible to base the detection on *a priori* characteristics of the defects. We have described a common reference based detection technique that compares an inspected image to a reference image, which was previously aligned to it. We have discussed the drawbacks of this technique in case of pattern variations nearby edges. We have shown that simple modeling, based on translation differences between the reference and source images doesn't consider pattern variations differences and yields high false detection rate. We have introduced the anisotropic kernels and reviewed their recent application in edge-preserving de-noising. We have proposed novel detection that is robust for pattern variations and misregistration. Our procedure is based on the reconstruction of the source image from the reference image using anisotropic kernel.

The source and reference images are mapped into a feature space, where every feature with origin in the source image is estimated by a weighted sum of neighboring features from the reference image. The set of neighboring features is determined according to the spatial neighborhood in the original image

space, and the weights are calculated from exponential distance similarity function. We assume that pattern-originated features of the source image can be reconstructed from features of the reference image due to similarity and periodicity of patterns in the source and reference images. When a source feature contains a defect, there are no similar features in the reference image, and therefore the feature cannot be reconstructed from features of the reference image, which indicates the presence of a defect.

We have discussed possible feature spaces and the reconstruction error. Using NL-means feature space, we have evaluated the proposed algorithm by analyzing the receiver operating characteristics (ROC), and demonstrated its improved defect detection performances compared to the recent anomaly detection technique, applied to the difference image. Finally, we have discussed the robustness of the proposed algorithm for pattern variations and misregistration. Registration of a reference image relatively to the source image is not required, as long as a pattern is periodic and a search region covers at least one pattern period. Furthermore, a source pixel doesn't have to be equal to only one pixel, but could be equal to a combination of several similar pixels, which allows to overcome the problem of pattern variations. For periodic patterns, the search region is taken often more than one period of the pattern, in order to increase the number of potentially similar reference patches.

For the non-periodic patterns, we have proposed to exploit the multichannel information, in order to compensate for pattern variations. We have introduced a kernel-based similarity measure that quantifies similarity relations between the inspected patch and its reference patches under the assumption that the locations of similar patches in the search regions are invariant for all the channels. We have demonstrated improved performances of the constrained multichannel detection compared to the single channel detection in case of non-periodic pattern. We also demonstrated the the presented multi-channel automatic detection outperforms simple union or intersection of single-channel

detection results, especially in case when the defect is not clearly distinguished in all the channels.

6.2 Future Research

The method we have proposed in this thesis opens a number of interesting topics for future study:

Performance Acceleration: The main disadvantage of the proposed algorithm is its relatively high computational complexity. As we have presented in Chapter 4 the computational complexity of the detection procedure is $O(n \cdot m \cdot d)$, where n is a number of the inspected pixels, m is a neighborhood dimension (search region size) and d is a feature space dimension (patch size). Especially in the case of periodic patterned wafers, it is quite desirable to expand the size of the search window as much as possible, and it is therefore useful to give a fast version. The future research could explore acceleration by a modification of the algorithm using a multi-scale implementation, similar to that proposed for image denoising applications [5]. The main idea is first to perform a search in the coarsest scale and to continue the search in finer scales only in regions that were found similar in the coarser scales. The procedure is applied to the pyramidal representation of the images. Starting from the coarsest scale, the pixels that could not be reconstructed are detected as defects and other pixels are reconstructed again in the next, finer scale. A search in the finer scale could be restricted only to the pixels, related to the k nearest neighbors from the coarser scale, as show in Fig. 6.1.

Additionally, the implementation of the proposed algorithm can be accelerated by calculating in parallel the log-likelihood in (5.11), for all pixels in the inspected imags. Consider a problem in three channels, where an image contains n pixels and a search region contains m reference features of size d.

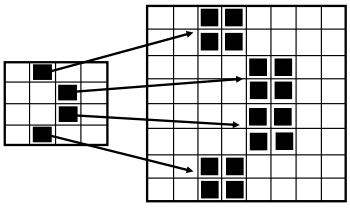


Figure 6.1: The left search region is the coarser scale of the right search region. Pixels (marked with black squares) are found to be the nearest neighbors at this scale. According to these nearest neighbors, the search in the finer scale could be restricted only to the related pixels, which are marked by arrows and black squares.

The number of operations for all the pixels is $n(4+m(3(3d+4)+2))$ (where the term $3d+4$ is due to the single weight calculation that involves d subtractions, d squares, $d-1$ sums, one division and one exponent, which requires four operations). For example, the computation for a 512-by-512 pixels image, with a search region of 51×51 pixels and patch size 11×11 pixels, requires 750G operations. The amount of required data is $3 \times 2 \times n \times 4 = 6$MB, where source and reference images from 3 channels with n pixels in float precision are considered. The current state-of-the-art Graphical Processing Unit (GPU) model GeForce 9800 GX2, [43], has 128GB/sec memory bandwidth and 76.8G/sec operations rate. Hence, an efficient GPU implementation could approach the theoretical 10 seconds limit, using a single GPU card.

Reduction of the Feature Space Dimension: The feature space dimension, which is determined according to the patch size, influences the computational complexity, and therefore reducing the dimension could be advantageous. However, by choosing a patch to be too much small, we could cause missed detection of the defects. Next, we present an issue for the future research, which will allow to reduce the multi-channel feature space dimension

without loss in the information of the original feature space. Using Euclidian distance, the joint similarity measure given in eq. (5.10) could be viewed as a single similarity measure with a combined feature:

$$\begin{aligned} L_{\text{xyz}} &= \frac{1}{m}\sum_{i=1}^{m} k_{\varepsilon_x}(x_i, x)\cdot k_{\varepsilon_y}(y_i, y)\cdot k_{\varepsilon_z}(z_i, z) \\ &= \frac{1}{m}\sum_{i=1}^{m} \exp\left(-\frac{1}{2}\frac{(\varepsilon_y\varepsilon_z\|x-x_i\|^2+\varepsilon_x\varepsilon_z\|y-y_i\|^2+\varepsilon_x\varepsilon_y\|z-z_i\|^2)}{\varepsilon_x\varepsilon_y\varepsilon_z}\right) . \end{aligned}$$

Where the combined feature is

$$\mathbf{s} \rightarrow v = \begin{bmatrix} \sqrt{\varepsilon_y\varepsilon_z}x \\ \sqrt{\varepsilon_x\varepsilon_z}y \\ \sqrt{\varepsilon_x\varepsilon_y}z \end{bmatrix} . \tag{6.1}$$

and the combined similarity parameter is $\varepsilon = \varepsilon_x\varepsilon_y\varepsilon_z$. The dimension of the expanded feature space is increased by factor 3, compared to the single channel feature space. This is a disadvantage due to the increase in the computational complexity. However, the new feature space is redundant, because both the $External_1$ and $External_2$ images supply the information about the topography of the wafer. Merging this information into a depth map and using the relative features, instead of the $External_1$ and $External_2$ features, will reduce the dimension without loss of information.

Local Spatial Adjustment of the Similarity Parameter: In our research we used global ε similarity parameter, to adjust the scale of similarity interest between the inspected patch and its reference patches. However, for many patterns, there is no one choice of scale, which is appropriate for each point. Smoother regions are characterized by the redundancy of the patches, hence smaller ε (higher similarity requirement) could be applied to achieve detection of scarcely noticeable defects. On contrary, to prevent pattern variations disturbance, ε should be larger (lower similarity requirement) nearby edges. Hence, local spatial adjustment of the similarity parameter will prevent

false detections nearby edges and maintain high detection rate in the smooth areas. The future research could examine possible methods to characterize the pattern structures, for example edge detection and clustering.

Texture Segmentation Application: The proposed similarity measure could serve in texture segmentation applications. For example, if possible texture components are known, reference dictionaries could be constructed for every possible component. To segment a pixel, we try to reconstruct the respective patch using the reference dictionaries. The segmentation will be performed by comparison of the likelihood (eq. 5.3) values of the different reconstructions. If the components are not known, but their number is given, a rough segmentation of the image could be performed using k-means clustering. According to the initial guess, the reference dictionaries could be constructed and the above presented procedure could be performed. In case of constructing the dictionaries from the image itself, it is important to prevent the reconstruction of a patch by itself. If no *a priori* information is given, an unsupervised segmentation could be performed by the analysis of similarity relations between the pixels. This implies a construction of a weight matrix, which represents the similarity relations between all the pixels in the image. The weight matrix allows to analyze the connectivity of a graph, which vertices are the image pixels. The connected vertices (pixels with high similarity relations) will constitute a cluster and the segmentation will be achieved. However, the size of such matrix will be $[n \times n]$, where n is the number of the pixels. This is disadvantageous from the computational point of view and requires high memory storage space. The future research could explore effective construction of such weight matrix.

Bibliography

[1] F.J. Aherne, N.A. Thacker, and P.I.Rockett. Optimal pairwise geometric histograms. In *Proc. 8th British Machine Vision Conference*, pages 480–490, Essex, UK, 1997.

[2] Ahmed Nabil Belbachir, Mario Lera, Alessandra Fanni, and Augusto Montisci. An automatic optical inspection system for the diagnosis of printed circuits based on neural networks. In *Proc. 40th IEEE-Industry Applications Society 2005 Annual Meeting*, pages 680–684, Hong-Kong, China, October 2005.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 13:1373–1397, 2003.

[4] A. Buades, B. Coll, and J. M. Morel. Denoising image sequences does not require motion estimation. In *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 70–74, Como, Italy, September 2005.

[5] A. Buades, B. Coll, and J. M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling and Simulation*, 4(2):490–530, 2005.

[6] G. A. Carpenter, M. A. Rubin, and W. W. Streilein. Artmap-fd: familiarity discrimination applied to radar target recognition. In *International Conference on Neural Networks*, volume 3, pages 1459–1464, Houston, TX, 1997.

[7] Chuan Yu Chang, Si Yan Lin, and MuDer Jeng. Using a two-layer competitive hopfield neural network for semiconductors wafer defect detection. In *Proc. IEEE International Conference on Automation Science and Engineering*, number 5, pages 301–306, Edmonton, Canada, August 2005.

[8] Long-Wen Chang, Ching-Yang Wang, and Shiuh-Ming Lee. Designing JPEG quantization tables based on human visual system. *Signal Processing: Image Comunication*, 16(5):501–506, January 2001.

[9] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006.

[10] Ronald R. Coifman, Mauro Maggioni, Yossi Keller, Arthur D. Szlam, Frederic J. Warner, and Steven W. Zucker. Geometries of sensor outputs, inference, and information processing. *Intelligent Integrated Microsystems. Proceedings of the SPIE*, 6232, 2006.

[11] Ronald R. Coifman, Mauro Maggioni, Steven W. Zucker, and Ioannis G. Kevrekidis. Geometric diffusions for the analysis of data from sensor networks. *Current Opinion in Neurobiology*, 15:576–584, 2005.

[12] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reaction coordinates and low dimensional representation of stochastic systems. in preperation.

[13] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[14] C.E. Costa and M. Petrou. Automatic registration of ceramic tiles for the purpose of fault detection. *Machine Vision and Applications*, 11(5):225–230, February 2000.

[15] X.L. Dai, M.A. Hunt, and M.A. Schulze. Automated image registration in the semiconductor industry: A case study in the direct to digital holography inspection system. In *Proc. SPIE, Machine Vision Applications in Industrial Inspection XI*, volume 5011, Santa Clara, CA, January 2003.

[16] Pierre Devijver and Josef Kittler. *Pattern Recognition: A Statistical Approach*. Prentice-Hall International, London, 1982.

[17] B.E. Dom and V. Brecher. Recent advances in the automatic inspection of integrated circits for pattern defetcs. *Machine Vision and Applications*, 8(1):5–19, January 1995.

[18] D.L. Donoho. De-noising by soft-thresholding. *IEEE Transaction on Information Theory*, 41(3):613–627, May 1995.

[19] D.L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100:5591–5596, 2003.

[20] Robert P. W. Duin. On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Transactions on Computer*, pages 1175–1179, November 1976.

[21] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, CA, USA, second edition, 1990.

[22] S. Gleason, R. Ferrell, T. Karnowski, and K. Tobin. Detection of semiconductor defects using a novel fractal encoding algorithm. In *Proc. SPIE. Process Integration, and Diagnostics in IC Manufacturing*, volume 4692, pages 61–71, March 2002.

[23] A. Goldman. Anomaly subspace detection based on a multi-scale markov random field model. Master's thesis, The Technion - Israel Institute of Technology, Haifa, Israel, 2004.

[24] A. Goldman and I. Cohen. Anomaly detection based on an iterative local statistics approach. *Signal Processing*, 84(7):1225–1229, July 2004.

[25] A. Goldman and I. Cohen. Anomaly subspace detection based on a multi-scale markov random field modeling. *Signal Processing*, 13(5):870–881, September 2005.

[26] Sheng-Uei Guan, Pin Xie, and Hong Li. A golden-block-based self-refining scheme for repetitive patterned wafer inspections. *Machine Vision and Applications*, 13(5-6):314–321, March 2003.

[27] J. Ham, DD. Lee, S. Mika, and B. Sch˙Technical report.

[28] Y. Hamamura, T. Kumazawa, K. Tsunokuni, A. Sugimoto, and H. Asakura. An advanced defect-monitoring test structure for electrical screening and defect localization. *IEEE Transactions on Semiconductor Manufacturing*, 17(2):104–110, May 2004.

[29] T. Hiroi, S. Maeda, H. Kubota, K. Watanabe, and Y. Nakagawa. Precise visual inspection for LSI wafer patterns using subpixel image alignment. In *Proc. 2nd IEEE Workshop on Applications of Computer Vision*, pages 26–34, Sarasota, Florida, USA, December 1994.

[30] T. Hiroi, C. Shishido, and M. Watanabe. Pattern alignment method based on consistency among local registration candidates for LSI wafer pattern inspection. In *Proc. 6th IEEE Workshop on Applications of Computer Vision*, pages 257–263, Orlando, Florida, USA, 2002.

[31] David Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Class representation and image retrieval with non-metric distances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583–600, June 2000.

[32] R. A. Johnson and D. W. Wichem. *Applied Multi-variate Statistical Analysis*. 1998.

[33] S. M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory.* Upper Saddle River, NJ: Prentice-Hall, 1993.

[34] E. J. Kelly. Adaptive detection and parameter estimation for multidimensional signal models. Technical Report 848, MIT Lincoln Laboratory, Lexington, MA, April 1989.

[35] S. Lafon. *Diffusion Maps and Geometric Harmonics.* PhD thesis, Yale University, 2004.

[36] Brodatz like textures. http://www-dbv.cs.uni-bonn.de/image/texture.tar.gz.

[37] M. Fischer M. Lindenbaum and A. Bruckstein. On gabor contribution to image enhancement. *Pattern Recognition*, 27:1–8, 1994.

[38] C. Manikopoulos and S. Papavassiliou. Network intrusion and fault detection: a statistical anomaly approach. *IEEE Communications Magazine*, 40:76–82, October 2002.

[39] D. Manolakis and G. Shaw. Detection algorithms for hyperspectral imaging applications. *IEEE Signal Processing Magazine*, pages 29–43, January 2003.

[40] G. Manson, G. Pierce, and K. Worden. On the long term stability of normal condition for damage detection in a composite panel. In *4th international conference on damage assessment of Structures (DAMAS)*, pages 359–369, Cardiff,UK, June 2001.

[41] J. S. Marron. A comparison of cross-validation techniques in density estimation. *Annals of Statistics*, 15(1):152–162, 1987.

[42] G. McLachlan and D. Peel. *Finite Mixture Models.* John Wiley and Sons, 2000.

[43] NVIDIA. http://www.nvidia.com.

[44] H. Onishi, Y. Sasa, K. Nagai, and S. Tatsumi. A pattern defect inspection method by parallel grayscale image comparison without precise image alignment. In *Proc. 28th IEEE Annual Conference of the Industrial Electronics Society*, volume 3, pages 2208 – 2213, Santa Clara, CA, November 2002.

[45] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[46] H. V. Poor. *An Introduction to Signal Detection and Estimation.* New York: Springer-Verlag, 1994.

[47] Richard J. Radke, Srinivas Andra, Omar Al-Kofahi, and Member Badrinath Roysam. Image change detection algorithms: A systematic survey. *IEEE Transcations on Image Processing*, 14(3):294–307, March 2005.

[48] K. Rao and P. Yip. *Discrete Cosine Transform.* Academic Press, New York, 1990.

[49] I. S. Reed and X. Yu. Adaptive multiple-band cFAR detection of an optical pattern with unknown spectral distribution. *IEEE Trans. Acoustics, Speech, Signal Processing*, 38(10):1760–1770, October 1990.

[50] P. Rosin. Thresholding for change detection. *Computer Vision and Image Understanding*, 86(2):79–95, May 2002.

[51] P. Rosin and E. Ioannidis. Evaluation of global image thresholding for change detection. *Pattern Recognition Letters*, 24(14):2345–2356, October 2003.

[52] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[53] Alberto Ruiz and Pedro E. Lopez de Teruel. Nonlinear kernel-based statistical pattern analysis. *IEEE Transactions on Neural Networks*, 12(1):16–32, January 2001.

[54] B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. Smola. Input space vs. feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10:1000–1017, September 1999.

[55] L. Shadhan and I. Cohen. Detection of anomlaies in texture images using multi-resolution random field modeling. *Signal Processing*, 87(12):3045–3062, December 2007.

[56] N. Shankar and Z. Zhong. Defect detection on semiconductor wafer surfaces. *Microelectronic Engineering*, 77:337–346, April 2005.

[57] A. Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, July 2006.

[58] S.Liapis and G.Tziritas. Color and texture image retrieval using chromaticity histograms and wavelet frames. *IEEE Transactions on Multimedia*, 6(5):676–686, October 2004.

[59] A. Szlam. *Non-Stationary Analysis on Datasets and Applications*. PhD thesis, Yale University, New Haven, Connecticut, USA, May 2006.

[60] L. Tarassenko. Novelty detection for the identification of masses in mammograms. In *IEEE 4th International Conference on Artificial Neural Networks*, volume 4, pages 442–447, Cambridge, UK, 1995.

[61] Du-Ming Tsai and Cheng-Hsiang Yang. An eigenvalue-based similarity measure and its application in defect detection. *Image and Vision Computing*, 23(12):1094–1101, November 2005.

[62] Du-Ming Tsai and Cheng-Hsiang Yang. A quantile-quantile plot based pattern matching for defect detection. *Pattern Recognition Letters*, 26(13):1948–1962, October 2005.

[63] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, New York, 1982.

[64] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[65] V. N. Vapnik. *Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communications and Control).* Wiley, 1998.

[66] A. R. Webb. *Statistical Pattern Recognition.* John Wiley and Sons, second edition, July 1990.

[67] A. R. Webb. *Statistical Pattern Recognition.* John Wiley and Sons, July 2002.

[68] Tang Xianghong, Xie Shuqin, and Li Qiliang. Watermarking for the digital images based on model of human perception. In *Proc. IEEE International Conference of Neural Netweoks & Signal Processing*, pages 14 – 17, Nanjing, China, December 2003.

[69] P. Xie and S. Guan. A golden-template self-generating method for patterned wafer inspection. *Machine Vision and Applications*, 12(3):149–156, October 2000.

[70] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignement. Technical Report CSE-02-019, Pennsylvania State University, 2002.