Machine Vision and Applications (2010) 21:129–141
DOI 10.1007/s00138-008-0146-y

ORIGINAL PAPER

Defect detection in patterned wafers using anisotropic kernels

Maria Zontak · Israel Cohen

Received: 4 April 2007 / Accepted: 2 May 2008 / Published online: 4 June 2008

Abstract Wafer defect detection often relies 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 paper, we propose a defect detection procedure, 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. The performance of the proposed algorithm is evaluated and its advantage is demonstrated compared to using an anomaly detection algorithm.

Keywords Semiconductor defect detection · Anomaly detection · Anisotropic kernels · Image reconstruction · Similarity measure · NL-means

This research was supported by Applied Materials Inc., Rehovot, Israel.

M. Zontak · I. Cohen (✉)
Department of Electrical Engineering,
Technion—Israel Institute of Technology,
Technion City, Haifa 32000, Israel
e-mail: icohen@ee.technion.ac.il

M. Zontak
e-mail: zontakm@tx.technion.ac.il

1 Introduction

Defect detection in wafers is a critical component of the 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 [1–5]. 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, and 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 [6–8]. Moreover, 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.

Xie and Guan [9] and Guan et al. [10] proposed to generate a golden-block database from the wafer image itself, and then modify and refine its content when used in further inspections of the same pattern. Gleason et al. [11] modeled self-similarities in the source image with fractal image encoding and detected defects without image registration. Onishi et al. [12] proposed a reference-based method that does not require exact registration. Grayscale morphological dilation

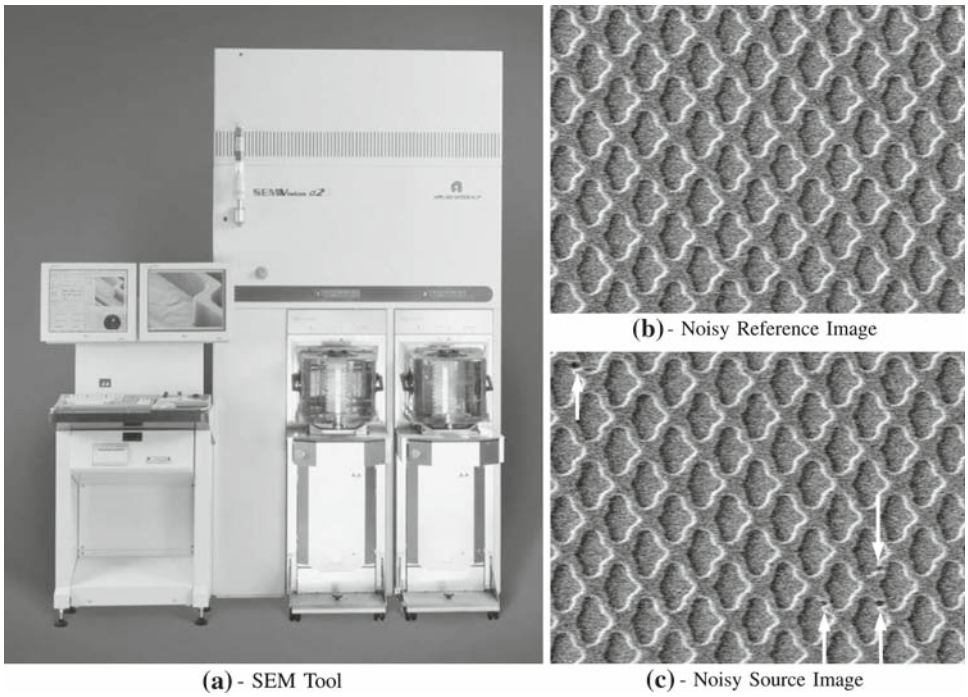


Fig. 1 SEM tool and example images it produces: **a** Defect review scanning electron microscope designed in Applied Materials Israel; **b** Reference image (clean of defects) and **c** Source image. Defects are indicated by *arrows*

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. Chang et al. [13] used an unsupervised learning by a two-layer competitive Hopfield neural network for defect detection. Their method does not require a reference image and enables detection based on the variance of gray level and sharp spatial irregularity.

In this paper, we propose a reference-based method for defect detection, which avoids image registration and is robust to pattern variations. The proposed procedure involves anisotropic kernel reconstruction of the source image using a reference image. The idea of anisotropic kernels was studied by Lafon and Coifman [14,15] and successfully applied to image denoising applications by Szlam [16]. Here, we exploit anisotropic kernels for the application of wafer defect detection. 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 weights are calculated by using an exponential distance similarity function for a set of features in the spatial neighborhood of the source feature in the original image space. We show that features originating from defect regions are not reconstructible from the reference image and hence can be identified.

The paper is organized as follows. In Sect. 2, we present the motivation for developing the proposed algorithm. In Sect. 3, we provide a theoretical framework for source image reconstruction from the reference image using anisotropic kernel and discuss feature space selection and reconstruction error. In Sect. 4 we present the implementation of the proposed algorithm. In Sect. 5, we demonstrate the application of the proposed algorithm to wafer defect detection and show its improved performance compared to using a Single Hypothesis Test (SHT) [17]. Finally, in Sect. 6 we discuss the robustness of the proposed algorithm to pattern variations and misregistration and summarize in Sect. 7.

2 Motivation and goals

In this section, we describe a standard procedure for defect detection in patterned wafers, discuss its drawbacks and present our objectives.

2.1 Background

Figure 1a shows a wafer inspection tool, namely Defect Review Scanning Electron Microscope (SEM), which detects defects of different types during the wafer manufacturing process. The inspection of the wafer begins with imaging of its surface. The analyzed wafer is illuminated with electrons, which cause interactions on the wafer's surface. These interactions lead to subsequent emission of electrons that supplies the information about the edges and the material of the inspected wafer. This information is rendered into two-dimensional intensity distribution that can be stored as a digital image and analyzed for defect detection (see Fig. 1b, c).

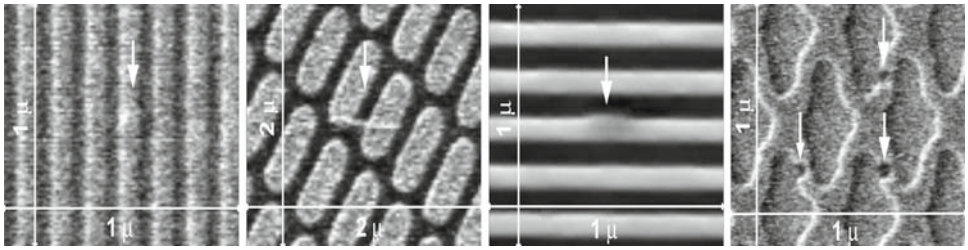


Fig. 2 Examples of defects. Defects are of various shapes and sizes and may belong to the wafer background or to its pattern

There are no precise characteristics of the possible defects. Defects may include particles, open lines, shorts between lines or other problems. Figure 2 demonstrates that they may be of various shapes, sizes, and 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 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.

Pattern to pattern comparison is the most suitable technique for an SEM-based inspection system. This comparison could be performed using a reference image captured from another wafer's die that was preliminarily checked to be clean of defects [2–5,7,18] or using self-reference approach based on the golden template construction from the repeating cells in the image [9,10]. 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. With this approach, image registration between the reference image and the source image is a major problem. Furthermore, in some cases only a single image that is under inspection is available. A self-reference technique avoids the above problems and does not require a reference image and registration. However, it cannot always detect the exact placement of the defects, but only their existence. The golden template is constructed from the analyzed image, so if the image contains defects, the template that is obtained by averaging the pattern blocks will also contain reduced defects. Hence, the reference-based technique, which relies on another defect-free die from the same wafer, is more popular. Next, we describe the general framework [4,5,7,18,19] of the comparison between reference and source images.

2.2 Pixel-based comparison

Using the reference image from Fig. 1b we would like to verify whether a pixel $\mathbf{s}$ from the source image in Fig. 1c originates from the pattern clutter or not. For this purpose, we denote the null hypothesis by

$$H_0 : \mathbf{s} \in \mathcal{P}, \tag{1}$$

which assumes that a pixel with coordinates $\mathbf{s} = (i, j)$ belongs to the pattern clutter $\mathcal{P}$. Under this hypothesis, a 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:

$$\begin{aligned} I_{\text{ref}}(\mathbf{s}) &= I_{\text{pat}}(\mathbf{s}) + \delta_1(\mathbf{s}) \quad \forall \mathbf{s} \in \Omega \\ I_{\text{src}}(\mathbf{s}) &= I_{\text{pat}}(\mathbf{s} + \mathbf{r}) + \delta_2(\mathbf{s}) \quad \forall \mathbf{s} \in \Omega, \end{aligned} \tag{2}$$

where I_{src}, I_{pat}, I_{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 $\delta_2(\mathbf{s})$ 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.

2.2.1 *Simple differencing*

The difference image $D(\mathbf{s}) = I_{\text{ref}}(\mathbf{s}) - I_{\text{src}}(\mathbf{s})$, calculated by subtracting the source image from the reference image, is used in several defect detection applications [7,9,20]. As a preprocessing step, denoising and registration of the source and reference images must be performed. It is important that the denoising procedure will preserve edges and will not blur the defects. For example, soft-threshold wavelet denoising [21] is applied to the images in Fig. 1b, c, and the results are shown in Fig. 3. The denoised images are usually registered using various techniques [7,8], mostly based on maximizing the correlation between blocks.

Figure 4a demonstrates the absolute difference image of the source and reference images shown in Fig. 3. Large differences are apparent in regions where the null hypothesis does not hold. Filtering and thresholding of the difference image can reveal the defective regions in the inspected wafer [9,20]. The defect mask is generated according to the following decision rule:

$$B(\mathbf{s}) = \begin{cases} 1, & \text{if } |D(\mathbf{s})| > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Often, the threshold τ is chosen empirically. Rosin [22,23] surveyed and reported experiments on many different criteria for choosing τ for general change detection applications. However, global threshold of pixel-by-pixel differencing

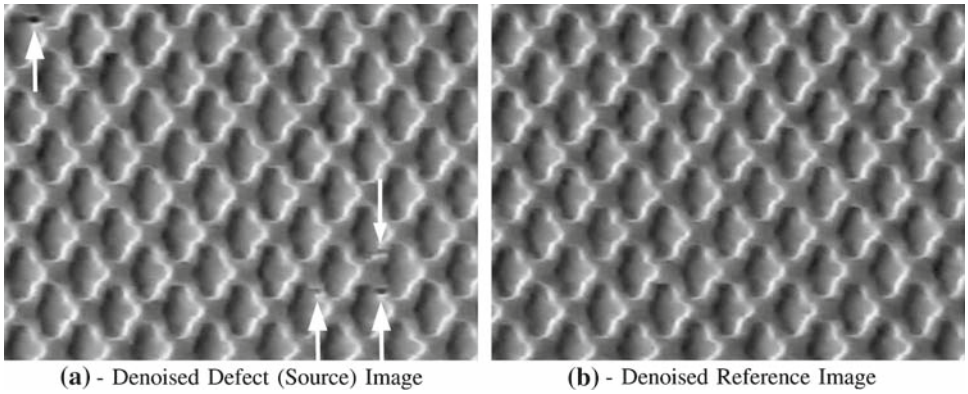


Fig. 3 **a** Image from Fig. 1b after performing stationary wavelet denoising, the *arrows* point at the defects; **b** Image from Fig. 1c after the same denoising procedure

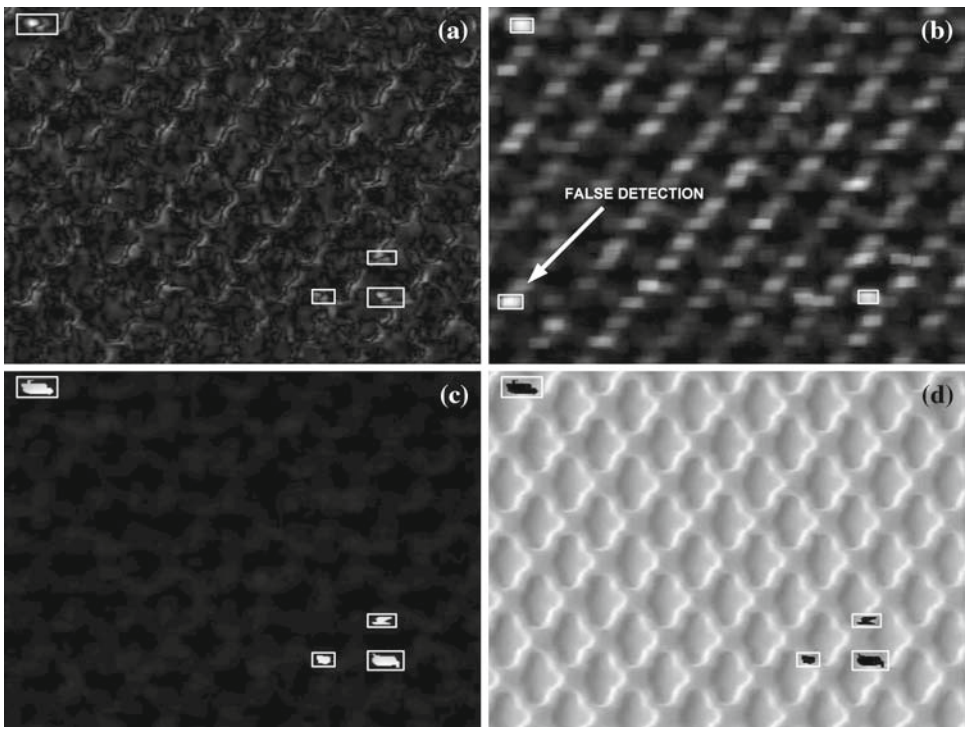


Fig. 4 **a** Difference image. Large differences may result from pattern variations and may obscure the defects (marked with *white frames*); **b** Detection based on thresholding of SHT of the difference image reveals only two of the four defects and one is falsly detected as defect. Thresholding with a lower threshold will detect the missed defects, but add more false detections; **c** Difference image between the reconstructed source image and the source image; **d** Detection based on unreconstructed regions of the reconstructed source image

yield high false alarm rate and is usually outperformed by more advanced statistical algorithms.

2.2.2 Single hypothesis test

The decision rule in many anomaly detection algorithms is cast as a statistical hypothesis test [24,25]. 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 (see Eq. 1) 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 a decision upon one of the hypotheses using the classical framework of hypothesis testing [26–28].

The variety and unpredictability of defects makes it impossible to characterize the H_1 hypothesis and to construct the respective PDF $p(I_{\text{ref}}, I_{\text{src}}|H_1)$. On the contrary, characterizing the null hypothesis 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. (2). An anomaly detection algorithm based on SHT [17,29] of the difference image $D(\mathbf{s})$ allows one to check null hypothesis fulfillment without any statistical knowledge about the defects.

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 $D(\mathbf{s})$. Given the expected vector

M and the covariance matrix Σ of the constructed data set, the Mahalanobis distance of any vector $\mathbf{X}$ from M is obtained by

$$d^2(\mathbf{X}) = (\mathbf{X} - M)^T \Sigma^{-1} (\mathbf{X} - M), \tag{4}$$

and the SHT is given by

$$d^2(\mathbf{X}) \underset{H_1}{\overset{H_0}{\lessgtr}} D^2, \tag{5}$$

where H_1 and H_0 represent hypotheses of anomaly presence and absence, respectively, and D 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. (2) is Gaussian. Under these assumptions, $d^2(\mathbf{X})$ is distributed $\chi^2_n(0)$ (central chi-squared distribution with n degrees of freedom), where n is the number of pixels in the patch constructed around pixel $\mathbf{s}$. The decision threshold D for a desired false alarm rate is calculated according to

$$P_{\mathrm{FA}} = 1 - p_{d^2}(\xi < D^2). \tag{6}$$

The model presented in Eq. (2) handles only the translational differences between the images. However, the source and reference image patterns are not identical and pattern variations may occur on nearby edges. These differences could be as intense as the differences caused by defects (see Fig. 4a), which may cause false detection. Figure 4b shows that applying SHT to the difference image of the images from Fig. 3 leads to false detections, due to the pattern variations differences that predominate over small defects differences. Pattern variation differences invalidate the assumption that the constructed feature vector $\mathbf{X}$ of the difference image is Gaussian distributed under the null hypothesis. Hence, the SHT threshold in the Eq. (5) could not be computed using Eq. (6).

Onishi et al. [12] tried to overcome the problems of pattern variations and misregistration by using a 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 variation, because it does not exploit the neighborhood replication of the periodic pattern.

2.3 Objectives

The aim of this work is to develop a more flexible similarity model that can significantly conceal the pattern variations disturbance and does not 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 from the source image we can find similar patches in the reference image. Therefore, 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:

$$\hat{I}_{\mathrm{src}}(\mathbf{s}) = \frac{1}{\sum W(\mathbf{s}, \mathbf{s}')} \sum_{\mathbf{s}' \in \mathcal{N}_{\mathbf{s}}} W(\mathbf{s}, \mathbf{s}') I_{\mathrm{ref}}(\mathbf{s}'), \tag{7}$$

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

$$\mathcal{N}_{\mathbf{s}} = \left\{ \mathbf{s}' \,|\, \mathbf{s}' \in n_k(\mathbf{s}) \right\}, \tag{8}$$

$n_k(\mathbf{s})$ is the set of k nearest neighbors of $\mathbf{s}$. 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. (7) 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. (2), 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. On the contrary, if a patch of the source image contains a defect, there are no similar patches in the reference image and the patch cannot be reconstructed from patches of the reference image. Hence, the detection is obtained by

$$\begin{aligned} H_0 &: W(\mathbf{s}, \mathbf{s}') \neq 0 \quad \exists \mathbf{s}' \in \mathcal{N}_{\mathbf{s}} \\ H_1 &: W(\mathbf{s}, \mathbf{s}') = 0 \quad \forall \mathbf{s}' \in \mathcal{N}_{\mathbf{s}}. \end{aligned} \tag{9}$$

Figure 4c demonstrates the improved difference image based on difference between the estimated source image (Fig. 4d) and the original source image (Fig. 3a). 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. 4a, b and detects all the four defects without false alarms.

3 Detection using anisotropic kernels

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 one proposed by Szlam [16] for image denoising applications.

3.1 Kernel representation of the source image using a reference image

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

$$\begin{aligned} \mathbf{s} &\to \xi_G(\mathbf{s}) = \{I_{\text{src}} * g_1(\mathbf{s}), \ldots, I_{\text{src}} * g_d(\mathbf{s})\}, \\ \mathbf{s}' &\to \xi_G(\mathbf{s}') = \{I_{\text{ref}} * g_1(\mathbf{s}'), \ldots, I_{\text{ref}} * g_d(\mathbf{s}')\}, \end{aligned} \tag{10}$$

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 $\mathbf{s}' \in \mathcal{N}_\mathbf{s}$, we estimate $\xi_G(\mathbf{s})$ by

$$\hat{\xi}_G(\mathbf{s}) = \frac{1}{D(\mathbf{s})} \sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} W(\mathbf{s}, \mathbf{s}') \cdot \xi_G(\mathbf{s}'), \tag{11}$$

where $\mathcal{N}_\mathbf{s}$ is denoted in Eq. (8). According to [16], we choose

$$W(\mathbf{s}, \mathbf{s}') = \exp^{-\rho(\mathbf{s},\mathbf{s}')^2/\epsilon}, \tag{12}$$

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

$$\rho^2(\mathbf{s}, \mathbf{s}') = \|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2. \tag{13}$$

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. Finally, returning to the image domain, a reconstructed source image is obtained by

$$\hat{I}_{\text{src}}(\mathbf{s}) = \frac{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2/\epsilon\} I_{\text{ref}}(\mathbf{s}')}{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|_2^2/\epsilon\}}. \tag{14}$$

Under the null hypothesis H_0, a patch from the source image is reconstructible from similar patches from the reference image. According to Eq. 9, if the source patch contains a defect there are no similar patches in the reference image and the reconstructed pixel is determined to be zero:

$$\begin{aligned} H_0 &: \hat{I}_{\text{src}}(\mathbf{s}) \to I_{\text{src}}(\mathbf{s}) \Rightarrow \mathbf{s} \notin \mathcal{A}, \\ H_1 &: \hat{I}_{\text{src}}(\mathbf{s}) \to 0 \Rightarrow \mathbf{s} \in \mathcal{A}, \end{aligned} \tag{15}$$

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

3.2 Filter banks

Szlam [16] construct G from non-local means filters (NL-means) of Buades et al. [30], where $g_{m,n}$ is an $[s_x \times s_y]$ matrix with one in (m, n) 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. For example, given two different orthonormal bases, $\{g_1, \ldots, g_n\}$ and $\{\tilde{g}_1, \ldots, \tilde{g}_n\}$, for a subspace $V \subset L^2$, because convolution is linear we have $m_G = O\, m_{\tilde{G}}$, 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 similarity weights constructed from these embeddings are the same [16]. 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 [31] and watermarking [32]. 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.

Figure 5a, b presents 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 [33], 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 Fig. 5c with d and Fig. 5e 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. 5c the reconstruction was possible even in the regions that were not reconstructed using NL-means feature space in Fig. 5d (the black regions). However, lower similarity requirement of the high frequencies coefficients reduces the reconstruction quality of the pattern details, because smooth regions and fine structures are not distinguished as well. The loss of detail in reconstruction with weighting matrix is especially evident in Fig. 5e, where higher ϵ is used. On the contrary, the NL-means (DCT with all-pass matrix) reconstruction in Fig. 5f succeeds in preserving more fine structures. To summarize, DCT coordinates feature space with frequency weighting matrix is

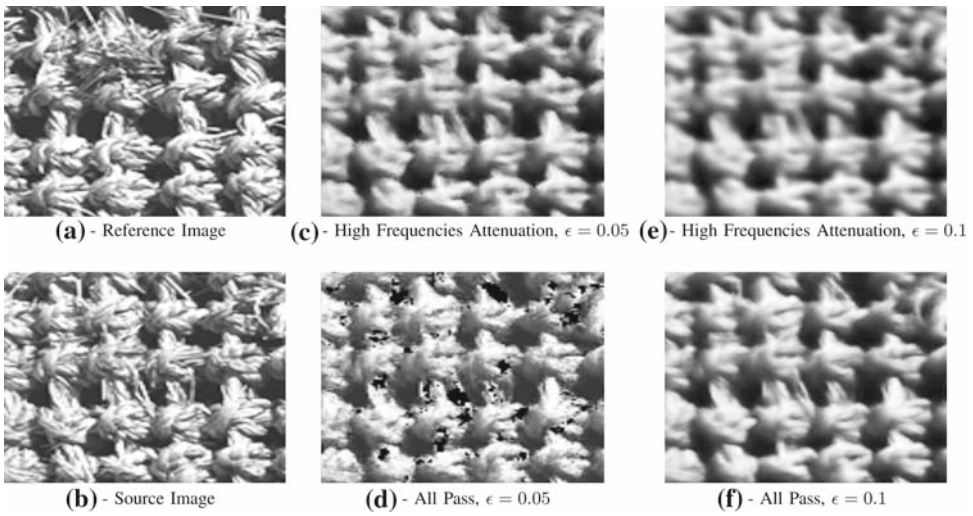


Fig. 5 DCT-based reconstruction: **a** Reference image; **b** source image; **c**, **e** images reconstructed using the [8 × 8] DCT transform with a frequency weighting matrix given in [33] (suppression of the high frequencies), $\epsilon = 0.05$ and 0.1, respectively; **d**, **f** images reconstructed using the [8 × 8] DCT transform with all pass matrix (all the entries equal 1), $\epsilon = 0.05$ and 0.1, 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

advantageous in the case of images with high spectral activity. These images are usually characterized by a 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.

3.3 Analysis of the representation error

To analyze the estimation error of Eq. (14), we construct a set of points $X = \{\xi_G(\mathbf{s}) \bigcup \xi_G(\mathbf{s}') | \forall \mathbf{s}' \in \mathcal{N}_\mathbf{s}\} = \{x_1, x_2, \ldots, x_N\}$, 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 $\mathcal{M}$, where $\mathcal{M} \subset \mathbb{R}^m$ is a d-dimensional compact Riemannian manifold [15, 34]. The last assumption is not trivial under finite set X and will be discussed below.

We assume that a smooth function, $f : \mathcal{M} \to \mathbb{R}$, exists 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

$$f(\xi_G(\mathbf{s})) = \langle \xi_G(\mathbf{s}), \chi_{\text{central}} \rangle = I_{\text{src}}(\mathbf{s}), \tag{16}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product on $\mathcal{M}$, which is a Riemannian manifold and by definition is associated with a smooth inner product. Therefore, the function proposed in Eq. (16) is smooth as required. The same function is valid for the reference image features construction.

Denoting the estimation operator by E, we can rewrite Eq. (14) as

$$(Ef)(\xi_G(\mathbf{s})) = \frac{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|^2/\epsilon\} f(\xi_G(\mathbf{s}'))}{\sum_{\mathbf{s}' \in \mathcal{N}_\mathbf{s}} \exp\{-\|\xi_G(\mathbf{s}) - \xi_G(\mathbf{s}')\|^2/\epsilon\}}. \tag{17}$$

Under the above assumptions, the estimation error is given by [34]

$$\hat{I}_{\text{src}}(\mathbf{s}) - I_{\text{src}}(\mathbf{s}) = (Ef)(\xi_G(\mathbf{s})) - f(\xi_G(\mathbf{s})) = \frac{\epsilon}{4} \Delta f(\xi_G(\mathbf{s})) + O(\epsilon^2). \tag{18}$$

Hence, for $\Delta f(\xi_G(\mathbf{s})) \ll 1/\epsilon$ the estimation error is negligible. Unfortunately, ϵ cannot be too small, because Eq. (18) holds only if the number of points N grows faster than $\epsilon^{-(\frac{d}{2}+1)}$ [15]. 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 $\eta : \mathcal{M} \to X$, with small norm, such that every point in X can be written as $x + \eta(x)$, for some $x \in \mathcal{M}$. It was shown in [15] that the approximation used for obtaining Eq. (18) is valid as long as the similarity parameter $\sqrt{\epsilon}$ remains larger than the size of the perturbation $\|\eta(x)\|$. In our case, we could rewrite any feature as

Algorithm 1 Defect Detection using NL-means estimation

{**s** - pixel index, f - source image, $\hat{f}$ - reconstructed source image}
for all $\mathbf{s} \in f$ **do**
  $P_{\mathbf{s}} \Leftarrow$ construct a patch of size $[s_x \times s_y]$ around pixel **s**
  $i \Leftarrow 1$
  {**r** - pixel index, f_{ref} - reference image, $\mathcal{N}_{\mathbf{s}}$- search region neighborhood of **s**}
  for all $\mathbf{r} \in \mathcal{N}_{\mathbf{s}}$ **do**
    $P^i_{\mathbf{r}} \Leftarrow$ construct a patch of size $[s_x \times s_y]$ around pixel **r**
    $\mathcal{W}^i \Leftarrow \exp(-\frac{\rho(P_{\mathbf{s}}, P^i_{\mathbf{r}})^2}{\epsilon})$ {ρ - a distance metric}
    $i \Leftarrow i + 1$
  $\mathcal{S}_{\mathcal{W}} \Leftarrow \Sigma_i \mathcal{W}^i$
  if $\mathcal{S}_{\mathcal{W}} = 0$ **then**
    for all i **do**
      $\mathcal{W}^i \Leftarrow 0$
  else
    for all i **do**
      $\mathcal{W}^i \Leftarrow \frac{\mathcal{W}^i}{\mathcal{S}_{\mathcal{W}}}$
  $\hat{P}_{\mathbf{s}} \Leftarrow \Sigma_{\forall i} \mathcal{W}^i \cdot P^i_{\mathbf{r}}$ {source image patch estimation using reference neighboring patches}
  $\mathcal{D}(\mathbf{s}) \Leftarrow \|\hat{P}_{\mathbf{s}} - P_{\mathbf{s}}\|_2$ {difference image value at pixel **s** calculation}
  $\hat{f}(\mathbf{s}) \Leftarrow \Sigma_{\forall i} \mathcal{W}^i \cdot f_{\text{ref}}(\mathbf{r}_i)$
  if $\hat{f}(\mathbf{s}) = 0$ **then**
    $\mathbf{s} \in \mathcal{A}$ {$\mathcal{A}$ is a set of defect regions}

$$\xi_G(\mathbf{s}) = x(\mathbf{s}) + r(\mathbf{s}) + n(\mathbf{s}), \tag{19}$$

where $x(\mathbf{s})$ is an ideal point that belongs to the manifold $\mathcal{M}$ and represents the original pattern, $n(\mathbf{s})$ denotes a noise term, and $r(\mathbf{s})$ denotes pattern variations under H_0 or a defect term under H_1. The same notations are valid for $\mathbf{s}'$. As a preprocessing step, the noise should be suppressed by applying a denoising operator to the images. Assuming that the remaining noise is negligible, the estimation error, presented in Eq. (18), is valid if

$$\|r(\mathbf{s})\| < \sqrt{\epsilon}. \tag{20}$$

For a defect-originated point, we assume that $\|r(\mathbf{s})\|$ is larger than $\sqrt{\epsilon}$. Hence, the reconstruction in Eq. (17) does not hold, which indicates the presence of a defect.

4 Implementation of the algorithm

Algorithm 1 summarizes the reconstruction and decision procedures for defect detection as described in Eqs. (10)–(15). To verify whether a pixel from the source image belongs to a defect area according to Eq. (15), we execute the following steps. A patch around every pixel in the source image is transformed into a vector in the feature space using Eq. (10), whose dimension is related to the defect size. A patch should be sufficiently big to contain the defect and its nearest surroundings, but not too big, 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. (11). 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. (12) that are calculated from Euclidian distances in the feature space [16]. The parameter ϵ in Eq. (12) 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., $1e-15$). 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. (11), and the tested pixel using Eq. (14). 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 (image size), m is neighborhood dimension (search region size) and d is a feature space dimension (patch size). Therefore from a computational point of view, m and d should be as small as possible. In our experiments we used a region that covered two to three pattern periods.

5 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 [17,35]. In all the presented experiments the reconstruction is performed using the NL-means feature space.

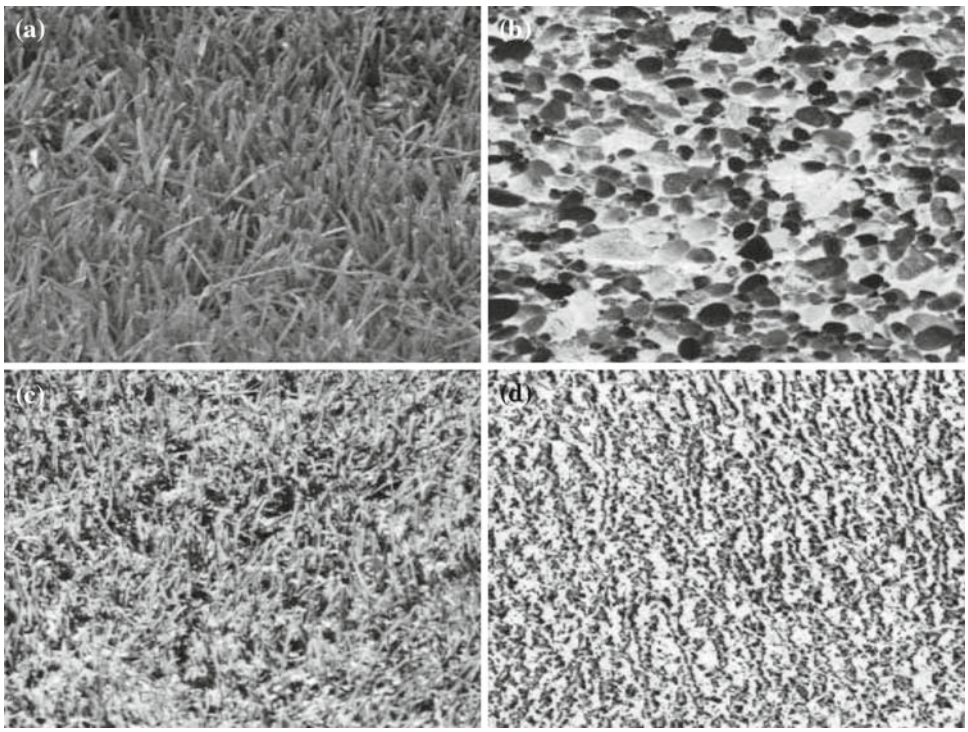
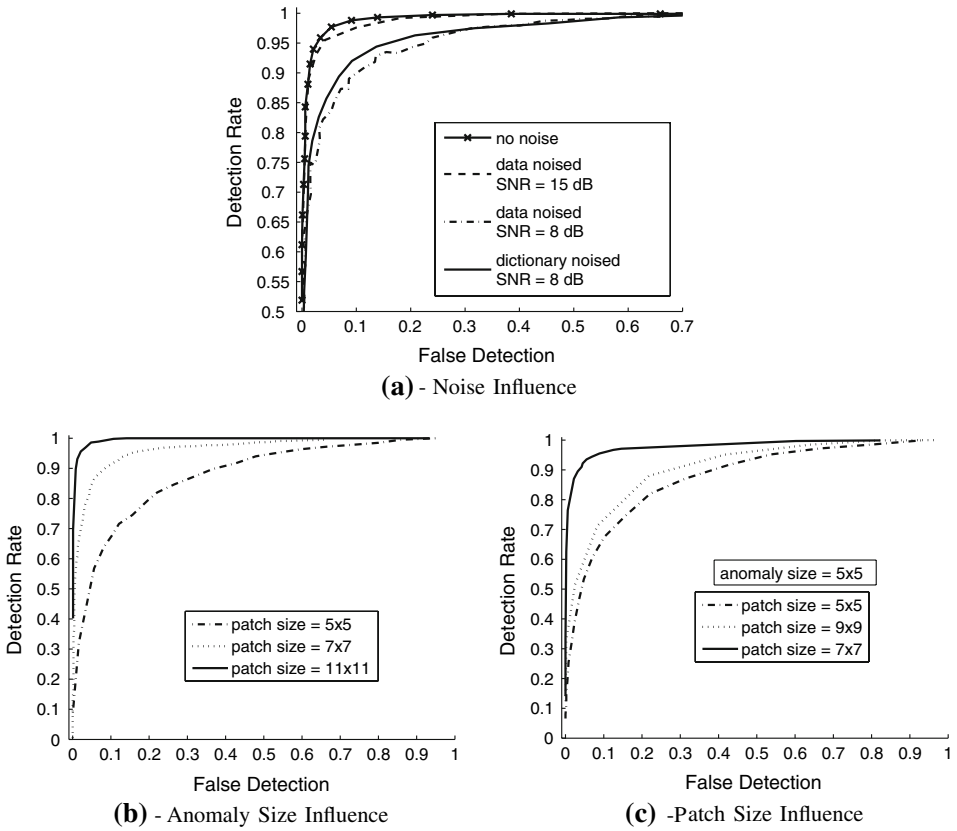


Fig. 6 Brodatz textures for constructing a reference dictionary and anomalous patches. **a** and **b** Textures for the reference dictionary and FAR-test data sets; **c** and **d** textures for generating anomalous patches

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 5,000 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 latter 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. (11), and if the patch is not reconstructible, it contributes to the 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 the detection 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. 6a for construction of the reference dictionary, and the texture shown in Fig. 6c 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 7a 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 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. 6b for construction of the reference dictionary, and the texture shown in Fig. 6d for the anomalous patches. The influence of the anomaly size on the detection performances was studied. Figure 7b 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 7c 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 does not contain an anomaly pattern is filled with a 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 (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



(a) - Noise Influence

(b) - Anomaly Size Influence

(c) -Patch Size Influence

Fig. 7 Performances dependence on: **a** additive white gaussian noise; **b** varying anomaly size (patches grow accordingly); **c** varying patch size with constant anomaly

requires calculation of Mahalanobis distance given in Eq. (4) in a feature space of the difference image and applying the SHT according to Eq. (5) to the result. The feature space is constructed from patches formed around every pixel in the difference image, and the size of the patches is the same as the size of the patches in the kernel-based detection algorithm. Figures 8 and 9 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 (Figs. 1, 3, 4) also demonstrates the advantage of the proposed algorithm in case of multiple defects. In the case of a single defect, the SHT threshold in Eq. (5) 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.

6 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, 8 and 9. Figure 10 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. 10a, does not have to be identical to one reference patch, but could be a combination of several marked patches from Fig. 10b. Moreover, Buades et al. [36] 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. 10b. Patches in Fig. 10b are marked as similar to the inspected patch from Fig. 10a, if their similarity measure according to Eq. (12) 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 compensation for pattern variation will be less effective and fusion of

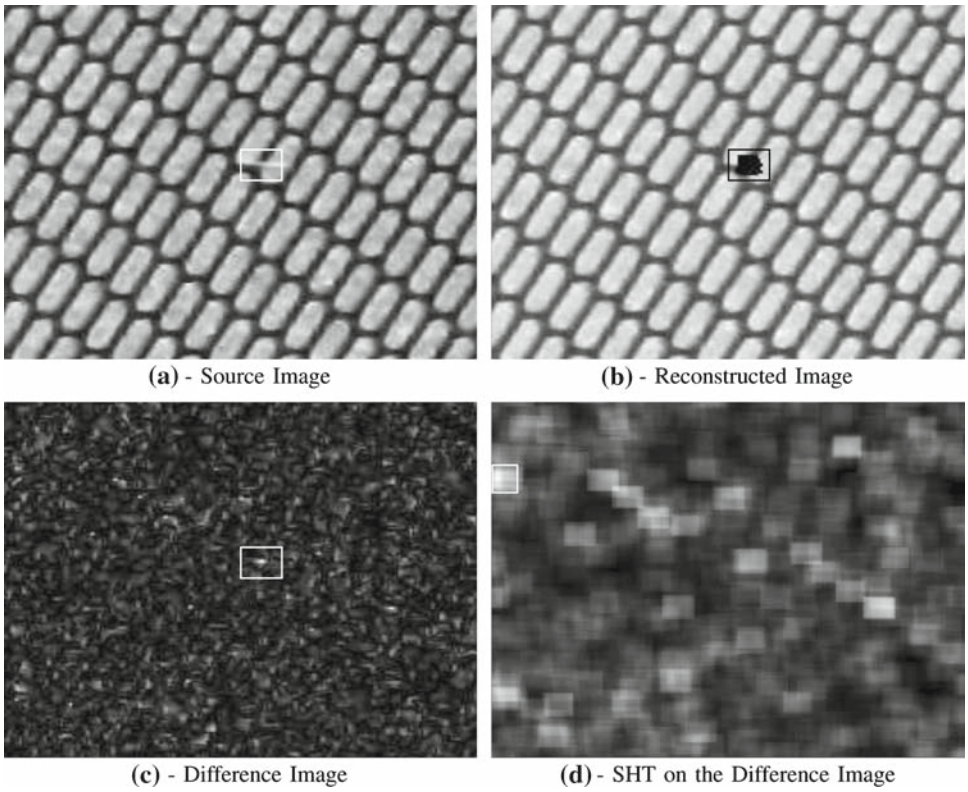


Fig. 8 Wafer defect detection. **a** Wafer image containing a defect (designated by *white frame*); **b** image reconstructed by Algorithm 1 ($\epsilon = 0.04$, $\mathcal{N}_{\mathbf{s}} = [81 \times 81]$, $s_x \times s_y = [13 \times 13]$); **c** difference image (the *white frame* is around the defect location); **d** SHT on the difference image yields false detections and misses

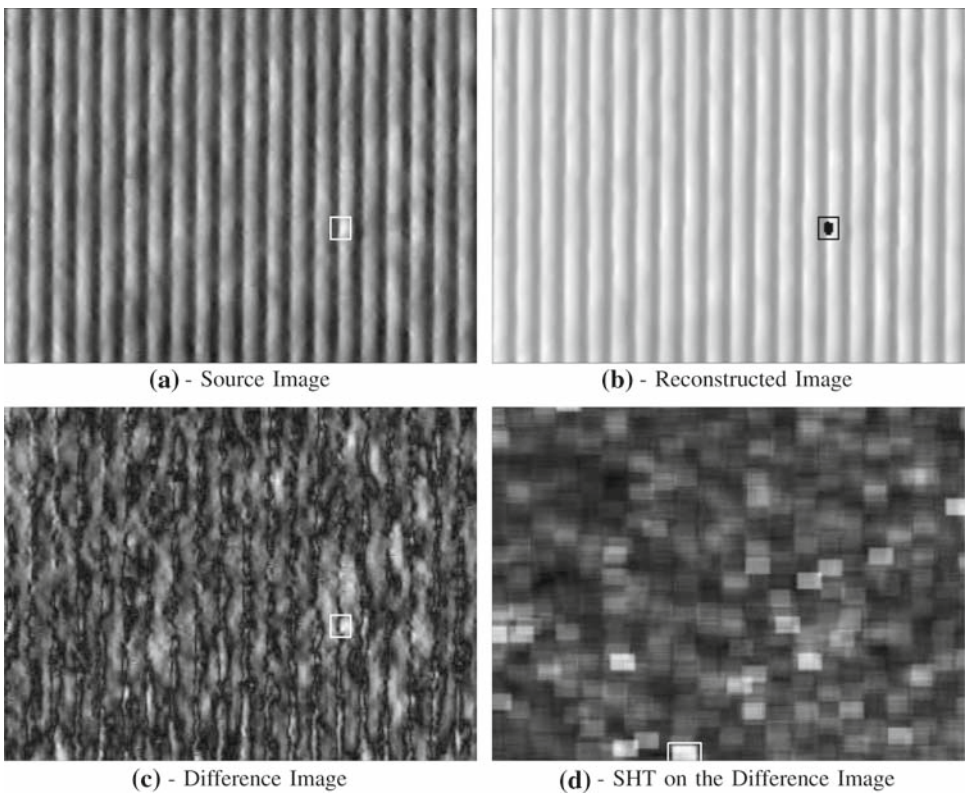


Fig. 9 Wafer defect detection. **a** Wafer image containing a defect (designated by *white frame*); **b** image reconstructed by Algorithm 1 ($\epsilon = 0.03$, $\mathcal{N}_{\mathbf{s}} = [151 \times 151]$, $s_x \times s_y = [13 \times 13]$); **c** difference image (the *white frame* is around the defect location); **d** SHT on the difference image yields false detections and misses

detections in different SEM channels (images with different perspective of the wafer) may be required to prevent false detections caused by pattern variations. The optimal fusion process remains an issue for a future research. The misregistration is tolerable only within the search region, hence either rough registration should be performed, or the search region

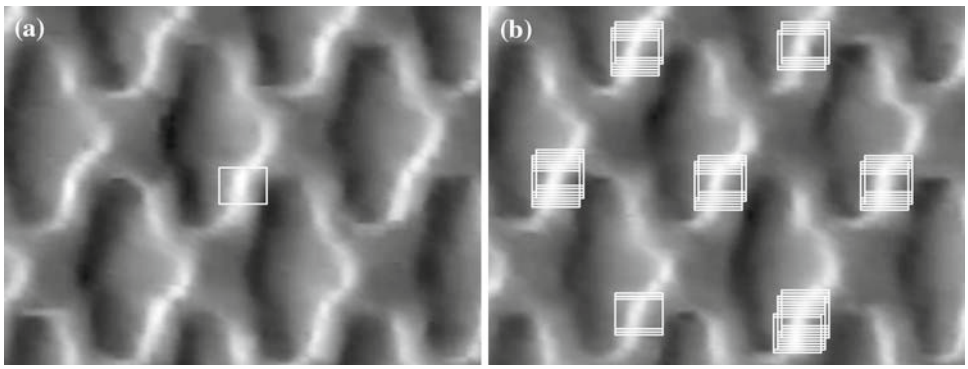


Fig. 10 Exploitation of periodicity of the pattern. **a** Region from a source image. A small patch is marked with a *white frame* in the center; **b** Aligned region from the reference image. Patches in **b** are marked as similar to the inspected patch from **a**, if their similarity measure according to Eq. (12) is above a determined threshold

should be chosen very large, which is disadvantageous due to the increased computational complexity.

7 Summary

We have presented a defect detection approach, which avoids image registration of the source and reference images and is robust to pattern variations. The proposed method relies on anisotropic reconstruction of the source image from the reference image. Although the computational complexity of the proposed algorithm is relatively high, further reduction in complexity can be achieved by some modifications. For example, a multi-scale implementation, similar to that proposed for image denoising applications [30], may be advantageous in our framework. 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 coarser scales. Additionally, the implementation of the proposed algorithm can be accelerated by calculating in parallel the similarity weights for all pixels in the reconstructed image. The proposed algorithm can also be combined with standard state-of-the-art wafer defect detection algorithms to reduce the false alarm rate without increasing the missed detection rate. Suspicious regions are first detected by conventional defect detection algorithms. Subsequently, the reconstruction procedure is applied only to patches around the suspicious pixels according to proposed algorithm, and regions that are not reconstructible are identified as defects.

Acknowledgments The authors thank Gil Sod-Moriah from Applied Materials Inc., Rehovot, Israel, for helpful discussions and constructive suggestions. They would also like to thank the anonymous reviewers for their insightful comments that helped to improve the quality of this work.

References

1. Shankar, N., Zhong, Z.: Defect detection on semiconductor wafer surfaces. Microelectronic Eng. **77**, 337–346 (2005)
2. Tsai, D.M., Yang, C.H.: A quantile-quantile plot based pattern matching for defect detection. Pattern Recognit. Lett. **26**(13), 1948–1962 (2005)
3. Tsai, D.-M., Yang, C.-H.: An eigenvalue-based similarity measure and its application in defect detection. Image Vision Comput. **23**(12), 1094–1101 (2005)
4. Dom, B., Brecher, V.: Recent advances in the automatic inspection of integrated circits for pattern defetcs. Machine Vision Appl. **8**(1), 5–19 (1995)
5. Hiroi, T., Maeda, S., Kubota, H., Watanabe, K., Nakagawa, Y.: Precise visual inspection for LSI wafer patterns using subpixel image alignment. In: Proc. 2nd IEEE Workshop on Applications of Computer Vision. Sarasota, Florida, USA, Dec. 1994, pp. 26–34
6. Dai, X., Hunt, M., Schulze, M.: 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, vol. 5011, Santa Clara, CA
7. Hiroi, T., Shishido, C., Watanabe, M.: 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. Orlando, Florida, USA, pp. 257–263 (2002)
8. Costa, C., Petrou, M.: Automatic registration of ceramic tiles for the purpose of fault detection. Machine Vision Appl. **11**(5), 225–230 (2000)
9. Xie, P., Guan, S.: A golden-template self-generating method for patterned wafer inspection. Machine Vision Appl. **12**(3), 149–156 (2000)
10. Guan, S.-U., Xie, P., Li, H.: A golden-block-based self-refining scheme for repetitive patterned wafer inspections. Machine Vision Appl. **13**(5-6), 314–321 (2003)
11. Gleason, S., Ferrell, R., Karnowski, T., Tobin, K.: Detection of semiconductor defects using a novel fractal encoding algorithm. In: Proc. SPIE. Process Integration, and Diagnostics in IC Manufacturing, vol. 4692, Mar. 2002, pp. 61–71
12. Onishi H., Sasa Y., Nagai K., Tatsumi S. (2002) 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, vol. 3. Santa Clara, CA, pp. 2208–2213
13. Chang, C.Y., Lin, S.Y., Jeng, M.: Using a two-layer competitive hopfield neural network for semiconductors wafer defect detection. In: Proc. IEEE International Conference on Automation Science and Engineering, no. 5, Edmonton, Canada, pp. 301–306 (2005)
14. Lafon, S.: Diffusion Maps and Geometric Harmonics. Ph.D. dissertation, Yale University, New Haven, Connecticut, USA, 2004
15. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmonic Anal. **21**(1), 5–30 (2006)

16. Szlam, A.: Non-stationary Analysis on Datasets and Applications. Ph.D. dissertation, Yale University, New Haven, Connecticut, USA, May 2006
17. Goldman, A., Cohen, I.: Anomaly detection based on an iterative local statistics approach. Signal Process. **84**(7), 1225–1229 (2004)
18. Hamamura, Y., Kumazawa, T., Tsunokuni, K., Sugimoto, A., Asakura, H.: An advanced defect-monitoring test structure for electrical screening and defect localization. IEEE Trans. Semiconductor Manufact. **17**(2), 104–110 (2004)
19. Radke, R.J., Andra, S., Al-Kofahi, O., Badrinath Roysam, M.: Image change detection algorithms: A systematic survey. IEEE Trans. Image Process **14**(3), 294–307 (2005)
20. Belbachir, A.N., Lera, M., Fanni, A., Montisci, A.: 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, Hong-Kong, China, Oct. 2005, pp. 680–684
21. Donoho, D.: De-noising by soft-thresholding. IEEE Trans. Inform. Theory **41**(3), 613–627 (1995)
22. Rosin, P.: Thresholding for change detection. Comput. Vision Image Understanding **86**(2), 79–95 (2002)
23. Rosin, P., Ioannidis, E.: Evaluation of global image thresholding for change detection. Pattern Recognit. Lett. **24**(14), 2345–2356 (2003)
24. Kelly, E.J.: Adaptive detection and parameter estimation for multidimensional signal models. MIT Lincoln Laboratory, Lexington, MA, Tech. Rep. 848, Apr. 1989
25. Reed, I.S., Yu, X.: Adaptive multiple-band cFAR detection of an optical pattern with unknown spectral distribution. IEEE Trans. Acoustics, Speech, Signal Process. **38**(10), 1760–1770 (1990)
26. Kay, S.M.: Fundamentals of Statistical Signal Processing: Detection Theory. Prentice-Hall, Upper Saddle River, NJ (1993)
27. Poor, H.V.: An Introduction to Signal Detection and Estimation. Springer, New York (1994)
28. Webb, A.R.: Statistical Pattern Recognition. Wiley, New York (2002)
29. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego, CA, USA (1990)
30. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simulat. **4**(2), 490–530 (2005)
31. Chang, L.-W., Wang, C.-Y., Lee, S.-M.: Designing JPEG quantization tables based on human visual system. Signal Process. Image Comun. **16**(5), 501–506 (2001)
32. Xianghong, T., Shuqin, X., Qiliang, L.: Watermarking for the digital images based on model of human perception. In: Proc. IEEE International Conference of Neural Netweoks & Signal Processing, Nanjing, China, Dec. 2003, pp. 14–17
33. Rao, K., Yip, P.: Discrete Cosine Transform. Academic Press, New York (1990)
34. Singer, A.: From graph to manifold laplacian: The convergence rate. Appl. Comput. Harmonic Anal. **21**(1), 128–134 (2006)
35. Goldman, A.: Anomaly subspace detection based on a multi-scale markov random field model. Master's thesis, The Technion—Israel Institute of Technology, Haifa, Israel
36. Buades, A., Coll, B., Morel, J.M.: Denoising image sequences does not require motion estimation. In: Proc. IEEE Conference on Advanced Video and Signal Based Surveillance, Como, Italy, Sept. 2005, pp. 70–74