Multi-Channel Defect Detection using Anisotropic Kernels

Maria Zontak

Electrical Engineering Department
Technion - Israel Institute of Technology

Supervised by Prof. Israel Cohen
Outline

1 Introduction
   - Problem Formulation
   - Recent Work
   - Objectives

2 Detection based on the Anisotropic Kernels
   - Detection using Anisotropic Kernels
   - Experimental Results

3 Multi-Channel Kernel-Based Detection
   - Motivation and Background
   - Joint Similarity Measure

4 Summary
Image Capture

SEM from Applied Materials, Rehovot, Israel

Maria Zontak
Technion - Israel Institute of Technology
Images

- The $External_1$ and $External_2$ detectors capture low energy secondary electrons. The acquired images contain the topography information of the sample (light and shadows), as if the light source pointed at the sample from top-right/top-left and the observation point remained above the sample.

- The $Internal$ detector captures high-energy backscattered electrons. The acquired image supplies the information about the edges and material of the sample.
Defect Variety and Unpredictability

There are no precise characteristics of the possible defects. Defects may

- include particles, open lines, shorts between lines, or other problems.
- be of various shapes, sizes, may belong to the wafer background or to its pattern.
- be predominant or scarcely noticeable.

The inspected wafer may contain many defects or no defects at all.
Reference-based Approach

Problem

The variety and unpredictability of defects makes it impossible to perform template matching based on some \textit{a priori} learned features and characterizations of detects.
Reference-based Approach

Problem

The variety and unpredictability of defects makes it impossible to perform template matching based on some *a priori* learned features and characterizations of detects.

Common Approach - Reference Image

- 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, verified to be clear of defects.
- The reference image is utilized to verify whether a pixel from the source image originates from the pattern clutter or not.
Simple Model

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:

\[ l_{\text{ref}}(s) = l_{\text{pat}}(s) + \delta_1(s) \quad \forall s \in \Omega \]
\[ l_{\text{src}}(s) = l_{\text{pat}}(s + r) + \delta_2(s) \quad \forall s \in \Omega , \]

where

\( l_{\text{src}}, l_{\text{pat}}, l_{\text{ref}} \) - source, noise-free pattern and reference images respectively;

\( \Omega \) - a set of indexes in the image domain;

\( \delta_1(s) \) and \( \delta_2(s) \) - independent white noise disturbances;

\( r \) - a translation vector, which is estimated by registration of the reference image to the source image.
The detection could be performed by simple thresholding of difference image $D(s) = I_{\text{ref}}(s) - I_{\text{src}}(s)$, receiving the following defect mask:

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

which is an empirical threshold.
Defect Detection Using Difference Image

- The detection could be performed by simple thresholding of difference image $D(s) = I_{ref}(s) - I_{src}(s)$, receiving the following defect mask:

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

- However, global threshold of pixel by pixel difference image yield high false alarm rate and is usually outperformed by more advanced statistical algorithms.

- For example, Single Hypothesis Test (SHT) could be applied on the difference image to detect defects.

  See Goldman and Cohen, 2004, for application of SHT for general anomaly detection.
Drawbacks of the Detection Based on the Difference Image

- The simple model handles only the translation differences between the images.
- Pattern variations differences could be as intense as the differences caused by defects ⇒ False detection may occur.
Drawbacks of the Detection Based on the Difference Image

- The simple model handles only the translation differences between the images.
- Pattern variations differences could be as intense as the differences caused by defects ⇒ False detection may occur.

Reference

Source

Difference

Onishi et al. 2002

- Use grayscale morphological dilation and calculate difference image according to the minimal distance between the reference and inspected images in the dilation range.
- Allows only slight misregistration and pattern variation.
Goals and Means

- To exploit the periodicity of patterned wafers and multi-channel information.
- To develop a more flexible similarity model that can significantly conceal the pattern variations disturbance and does not require precise registration.
Goals and Means

- To exploit the periodicity of patterned wafers and multi-channel information.
- To develop a more flexible similarity model that can significantly conceal the pattern variations disturbance and does not require precise registration.

**Lafon et al. 2005**

Studied the idea of anisotropic-kernels and diffusion maps.
Goals and Means

- To exploit the periodicity of patterned wafers and multi-channel information.
- To develop a more flexible similarity model that can significantly conceal the pattern variations disturbance and does not require precise registration.

Lafon et al. 2005
Studied the idea of anisotropic-kernels and diffusion maps.

Szlam et al. 2006
Incorporated the anisotropic-kernels approach for edges-preserving denoising application.
Source Image Reconstruction Using 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 $\xi_G$:

$$ s \rightarrow \xi_G(s) = \{ I_{src} \ast g_1(s), \ldots, I_{src} \ast g_d(s) \} $$

$$ s' \rightarrow \xi_G(s') = \{ I_{ref} \ast g_1(s'), \ldots, I_{ref} \ast g_d(s') \} ,$$

where $s, s' \in \Omega$ and $\Omega$ is a general set of indices in the image space.

Given $\xi_G(s')$ for all $s' \in \mathcal{N}_s = \{ s' \mid s' \in n_k(s) \}$, $\xi_G(s)$ is estimated by

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

where $n_k(s)$ is the set of $k$ nearest neighbors of $s$ in the image domain.

$$ W(s, s') = \exp^{-\rho(s, s')^2/\varepsilon} \text{ and } D(s, s') = \sum_{s' \in \mathcal{N}_s} W(s, s') ,$$

where $\rho$ is a metric in our feature space (usually Euclidean) and $\varepsilon$ is a similarity parameter.
Detection

A reconstructed source image is obtained by

\[
\hat{I}_{\text{src}}(s) = \frac{\sum_{s' \in N_s} \exp\left\{-\frac{\|\xi_G(s) - \xi_G(s')\|_2^2}{\varepsilon}\right\} I_{\text{ref}}(s')}{\sum_{s' \in N_s} \exp\left\{-\frac{\|\xi_G(s) - \xi_G(s')\|_2^2}{\varepsilon}\right\}}.
\]

- Under the null hypothesis \(H_0\), a source feature is reconstructible from similar features from the reference image.
- If the source feature arises from a defect, we assume that there are no similar features in the reference image, hence, the reconstructed pixel value would be close to zero:

\[
H_0 : \hat{I}_{\text{src}}(s) \rightarrow I_{\text{src}}(s) \Rightarrow s \notin A
\]

\[
H_1 : \hat{I}_{\text{src}}(s) \rightarrow 0 \Rightarrow s \in A,
\]

where \(A\) denotes a set of defect regions.
Non Linear Means (NL-means) Filters of Morel et al. 2005

- $g_{m,n}$ is an $[s_x \times s_y]$ matrix with one in $(m, n)$ position and zeros elsewhere.
- $\xi_G$ is the set of overlapping patches of the source and reference images embedded in $d = s_x \times s_y$ dimensions.
Filter Bank and Metric Choice

Non Linear Means (NL-means) Filters of Morel et al. 2005

- $g_{m,n}$ is an $[s_x \times s_y]$ matrix with one in $(m, n)$ position and zeros elsewhere.
- $\xi_G$ is the set of overlapping patches of the source and reference images embedded in $d = s_x \times s_y$ dimensions.

DCT-based Feature Space

- The embedding into $[s_x \times s_y]$ patches is the same (up to a rotation) as into $[s_x \times s_y]$ DCT coordinates.
  - The similarity weights are the same.
- Emphasizing only specific frequencies in the embedding coordinates with frequency weighting matrix leads to different representation.
Example of Reconstruction Using DCT with and without weighting matrix

Reference Image

Source Image

High Frequencies Attenuation, $\varepsilon = 0.05$

All Pass, $\varepsilon = 0.05$

High Frequencies Attenuation, $\varepsilon = 0.1$

All Pass, $\varepsilon = 0.1$
Example of Detection Using NL-means
Simulations on Brodatz Textures

Data Sets Construction

- The reference (reconstructing) dictionary set contains 5000 patches chosen randomly from one texture image.
- The FA-test data set contains additional different patches from that texture image.
- The PD-test anomalous set contains patches from additional texture.
Simulations on Brodatz Textures

Data Sets Construction
- The reference (reconstructing) dictionary set contains 5000 patches chosen randomly from one texture image.
- The FA-test data set contains additional different patches from that texture image.
- The PD-test anomalous set contains patches from additional texture.

ROC Curve Acquisition
- Non-reconstructible patches from the FA-test set \(\Rightarrow\) FAR.
- Non-reconstructible patches from the PD-test set \(\Rightarrow\) PD.
- ROC curve is obtained by varying the value of \(\varepsilon\) from low values (high PD and FAR) to high values (low PD and FAR).
The patch size is $5 \times 5$ pixels.

White Gaussian noise is added to either patches in the FA-data set or the reference dictionary.

⇒ The noise in the reference dictionary is less significant than the noise in the data.
Anomaly Size Influence

The patch size varies according to the anomaly size
⇒ As expected, bigger anomalous patches are more easily detected than smaller ones.
Patch Size Influence

- The anomaly size is constant: 5 \times 5 pixels.
- The patch size varies: 5 \times 5, 7 \times 7 and 9 \times 9 pixels. Empty area is filled with the reference pattern.

⇒ The patch should cover not only the anomaly but also its nearest surroundings, however it should not be too big.
Defect Detection

Source Image

Reconstructed Image

Difference Image

SHT on the Difference Image
Defect Detection

- Source Image
- Reconstructed Image
- Difference Image
- SHT on the Difference Image
A favorable case for NL-means is a periodic case.

A patch does not have to be identical to one reference patch but could be a combination of several patches.

Wide search region with repetitive pattern compensates for pattern variations and miss-registration.

Non-repetitive pattern reduces the achieved robustness.
Non-Periodic Pattern

Problem

- A favorable case for NL-means is a periodic case.
- A patch does not have to be identical to one reference patch but could be a combination of several patches.
- Wide search region with repetitive pattern compensates for pattern variations and miss-registration.
- Non-repetitive pattern reduces the achieved robustness.

*External*$_1$  *External*$_2$  *Internal*
Relation to the Parzen’s Non-Parametric Density Estimator

Parzen 1962

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

\[
p(x) \approx \hat{p}_\varepsilon(x) = \frac{1}{m} \sum_{i=1}^m b_\varepsilon(\|x - x_i\|)
\]

where \( b_\varepsilon \) is a normal density with zero mean and variance \( \varepsilon \).

Ruiz et al. 2001

- Empirical mapping of \( x \) on the set \( X = \{x_i\}_{i=1}^m \).

\[
x_X = (k(x_1, x), k(x_2, x), ..., k(x_m, x))^T
\]

- Gaussian kernel \( k_\varepsilon(x, y) = G_\varepsilon(x, y) \equiv e^{-\frac{1}{2}||x-y||_2^2/\varepsilon}} \)

\[
\Rightarrow \hat{p}_\varepsilon(x) = x_X \cdot \bar{1}_m = \frac{1}{m} \sum_{i=1}^m k_\varepsilon(x_i, x)
\]
Definition

- Single channel similarity measure
  \[
  \frac{1}{m} \sum_{i=1}^{m} k(x_i, x)
  \]

  relates to the probability of a patch \( x \) to arise from pattern statistics estimated using reference set of patches \( X \).

- We define joint similarity measure of the three channels to be
  \[
  J_{x,y,z} = \frac{1}{m} \sum_{i=1}^{m} k(x_i, x)k(y_i, y)k(z_i, z),
  \]

  where patches \( x, y \) and \( z \) relate to the \( \text{External}_1 \), \( \text{External}_2 \) and \( \text{Internal} \) channels.
Under the null hypothesis, spatial locations of the similar patches in the reference search regions should be consistent between the channels.

**Pattern**

**External\_1 Source**  
**External\_1 Reference**

**Pattern**

**External\_2 Source**  
**External\_2 Reference**

**Defect**

Maria Zontak  
Technion - Israel Institute of Technology
Statistical Foundation

Two Channels Similarity Measure \( \frac{1}{m} \sum_{i=1}^{m} k(x_i, x)k(y_i, y) \)

- Similarity consistency between the channels could be expressed by conditional probability:
  Instead of \( \hat{p}(x) = \sum_{i=1}^{m} \frac{1}{m} k(x_i, x) \)
Statistical Foundation

Two Channels Similarity Measure \[ \frac{1}{m} \sum_{i=1}^{m} k(x_i, x)k(y_i, y) \]

- Similarity consistency between the channels could be expressed by conditional probability:
  - We use
    \[ \hat{p}(x|y) = \sum_{i=1}^{m} \frac{k(y_i, y)}{\sum k(y_i, y)} k(x_i, x) . \]
Two Channels Similarity Measure \( \frac{1}{m} \sum_{i=1}^{m} k(x_i, x)k(y_i, y) \)

- Similarity consistency between the channels could be expressed by conditional probability:
  
  We use
  
  \[
  \hat{p}(x|y) = \sum_{i=1}^{m} \frac{k(y_i, y)}{\sum_{i=1}^{m} k(y_i, y)} k(x_i, x).
  \]

- Joint similarity checks the consistency of the similar patches locations and is determined according to:
  
  \[
  J_{x,y} = \hat{p}(x, y) = \hat{p}(x|y)\hat{p}(y) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, x)k(y_i, y).
  \]

Recall that \( \hat{p}(y) = \frac{1}{m} \sum k(y_i, y). \)
The joint detection compensates for non-periodic pattern and allows robustness for pattern variations.
Similarity Parameter Adjustment

Periodic pattern

The minimal value of $\varepsilon$ that allows good reconstruction of every reference patch from its neighboring patches excluding itself.

Non-Periodic Pattern

Pattern variations could differ from channel to channel. In every channel, $\varepsilon$ parameter should be scaled according to the similarity of the source and reference images, which could be measured according to the Bhattacharyya distance:

$$
\varepsilon_k = -\log \left( \sum_i \sqrt{R_k^i S_k^i} \right),
$$

where $R_i$ and $S_i$ are the frequency coded quantities in bin $i$ for the reference and source images' histograms, respectively.
Similarity Parameter Adjustment

### Periodic pattern

The minimal value of $\varepsilon$ that allows good reconstruction of every reference patch from its neighboring patches excluding itself.

### Non-Periodic Pattern

- Pattern variations could differ from channel to channel.
- In every channel $\varepsilon$ parameter should be scaled according to the similarity of the source and reference images, which could be measured according to the Bhattacharyya distance:

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

where $R_i$ and $S_i$ are the frequency coded quantities in bin $i$ for the reference and source images' histograms, respectively.
Conclusions

- In periodic patterned wafers:
  - A wide search region that covers at least one period of pattern, compensates for miss-registration.
  - Combination of several patches in the search region with more than one period, compensates for pattern variations.
Conclusions

- In periodic patterned wafers:
  - A wide search region that covers at least one period of pattern, compensates for miss-registration.
  - Combination of several patches in the search region with more than one period, compensates for pattern variations.
- Multi-channel constrained detection extends pattern variations robustness for the non-periodic patterns.
In periodic patterned wafers:
- A wide search region that covers at least one period of pattern, compensates for miss-registration.
- Combination of several patches in the search region with more than one period, compensates for pattern variations.

Multi-channel constrained detection extends pattern variations robustness for the non-periodic patterns.

Future Research:
- Depth map creation using External images will allow to reduce the dimensionality of the feature space.
- Local spatial adjustment of the similarity parameter $\varepsilon$ will improve detection of weakly-noticed defects in the smooth regions and robustness to pattern variations nearby edges.
- Kernel based similarity measure exploitation in other applications, for example texture segmentation.
Self-Reference Approach

**Guan et al. 2000, 2003**

A golden-block database is generated from the wafer image itself, then modified and refined when used in further inspections of the same pattern.
## Self-Reference Approach

**Guan et al. 2000, 2003**

A golden-block database is generated from the wafer image itself, then modified and refined when used in further inspections of the same pattern.

**Gleason et al. 2002**

Model self-similarities in the source image with fractal image encoding and detect defects without reference image.
Self-Reference Approach

**Guan et al. 2000,2003**
A golden-block database is generated from the wafer image itself, then modified and refined when used in further inspections of the same pattern.

**Gleason et al. 2002**
Model self-similarities in the source image with fractal image encoding and detect defects without reference image.

**Chang et al. 2005**
- A defect is detected according to the variance of gray level and sharp spatial irregularity, using unsupervised learning by a two-layer competitive Hopfield neural network.
- Reference image is not required.
Supplementary Information

Single Hypothesis Test (SHT) of the Difference Image

- A data set is constructed from overlapping patches formed around every pixel in the difference image $D(s)$.
- The expected vector $M$ and the covariance matrix $\Sigma$ of the data set are calculated, assuming that the defects are rare.
- The Mahalanobis distance of any vector $X$ from $M$ is obtained by
  \[ d^2(X) = (X - M)^T \Sigma^{-1} (X - M) , \]
  and the SHT is given by
  \[ d^2(X) \begin{cases} H_0 & \leq D^2 \text{,} \\ H_1 & > D^2 \end{cases} , \]
  where $D$ is a distance threshold.