Journal of Process Control 62 (2018) 11–23

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Low-rank manifold optimization for overlay variations in lithography process

Zhichao Wang, Min Liu*, Mingyu Dong

Department of Automation, Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Received 4 January 2017
Received in revised form 7 November 2017
Accepted 10 November 2017

Keywords:
Semiconductor
Lithography process
Low-rank manifold
Riemannian optimization

ABSTRACT

Overlay variations occur frequently in lithography process, which should be controlled within the tolerance to guarantee the better pattern resolutions. The operational optimization of overlay aims to predict the unknown overlay variations and compensate them into the wafer production. Due to the uncertain yield, the overlay data for learning are usually incomplete, which makes the overlay optimization very challenging. This paper proposes a novel overlay optimization framework called low-rank manifold optimization (LRMO), which provides new insight to address incomplete overlay data via exploiting low-rank property. First, LRMO can use effectively the correlations from incomplete overlay data, which builds a low-rank model for overlay optimization. In addition, LRMO resorts to Riemannian optimization and designs an efficient algorithm for this low-rank model. The proposed LRMO algorithm analyzes the manifold structure of the overlay data and computes accurate overlay variations with a low computational complexity. The experiments validate that LRMO obtains satisfying performance on the operational optimization of overlay variations.

© 2017 Published by Elsevier Ltd.

1. Introduction

Lithography is an important and complex process for wafer production in semiconductor manufacturing [1–5]. Fig. 1 shows the lithography process that contains mainly two main steps. First, a pattern is created on the mask. Second, the light source passes through the lens repeatedly, which projects the pattern onto current layer of wafer. However, stepper errors and lens distortion may cause the misalignments between a current layer and a previous layer, known as overlay variations. Fig. 2 show eight overlay variations frequently considered in lithography, including *horizontal translation*, *vertical translation*, *horizontal expansion*, *vertical expansion*, *rotation*, *orthogonality*, *shot expansion* and *shot rotation*. These overlay variations must be accurately predicted and compensated into lithography process. Otherwise, the improper overlay variation will lead to the poor pattern resolutions and affect the wafer production.

Traditional overlay optimization approaches apply the regression model to approximate the unknown overlay variations based on collected sampling points. A multiple linear regression method is introduced in [6], which can solve a high-order overlay equation for assessing the unknown overlay variations. Subsequently, weighted least-squares regression considers the extra weights determined by the quality of sampling points [7]. Another weighted least-squares regression integrates high-order variables into linear overlay equation for improving performance [8,9]. In general, the effectiveness of these methods depends on the number of sampling points in a single wafer, and more sampling points will produce the more accurate overlay results [6]. However, a single wafer contains typically very limited points owing to sampling costs [10]. Thus, this implies that these methods may not have sufficient sampling points to get a good estimate of overlay variations.

To overcome aforementioned drawbacks, advanced process control (APC) performs batch-process control for improving process and device performance. Early APC methods [11–13] are generally implemented through exponentially weighted moving average (EWMA) based algorithms because of its simplicity and easy maintenance. Threaded run-to-run (R2R) [14–17] are the new paradigm of manufacturing with the goal of satisfying high-mix production environments. Threaded R2R divides historical data into different threads according to manufacturing contexts or characteristics, which reduces the sources of variation for one thread significantly. In particular, product-based EWMA (pb-EWMA) [5] and Group product-based EWMA (GP-EWMA) [16] are state-of-the art approaches improving the process performance. The idea of pb-EWMA was that it took an action based on the data from

* Corresponding author.
E-mail address: lium@tsinghua.edu.cn (M. Liu).

https://doi.org/10.1016/j.jprocont.2017.11.006
0959-1524/© 2017 Published by Elsevier Ltd.

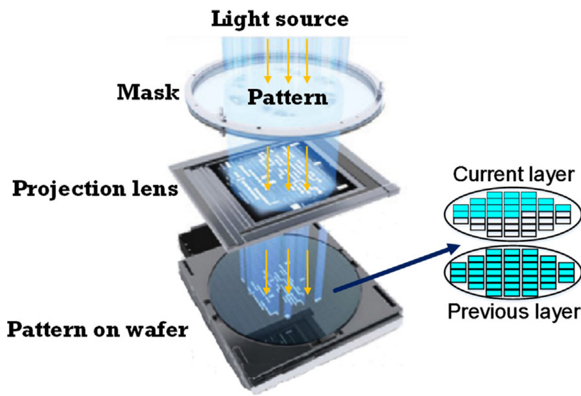

Fig. 1. Lithography process.

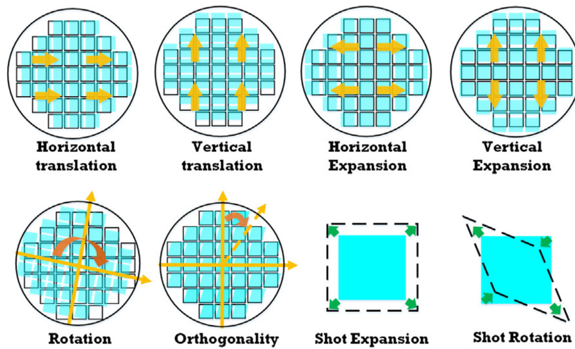

Fig. 2. Eight overlay variations of one wafer.

Table 1
Summary of overlay variations.

Overlay variations	Meaning
Horizontal translation	Translation degree on wafer in the directions of left/right
Vertical translation	Translation degree on a wafer in the directions of up/down
Horizontal expansion	Expansion degree on a wafer in the directions of left/right
Vertical expansion	Expansion degree on a wafer in the directions of up/down
Orthogonality	Orthogonal degree on a wafer with respect to the base axes
Rotation	Rotation degree on a wafer with respect to the base axes
Shot expansion	Expansion degree for the shots
Shot rotation	Rotation degree for the shots

the same product, ignoring the difference between tools. By using group products with similar characteristics, GP-EWMA adopts an adaptive k-means cluster algorithm to guarantee the quality of low-frequency products. However, the division of products by thread R2R depends on the frequency of products, which may fail to find the optimum weight and degrade the performance [18].

and Table 1

Alternatively, recent research resorts to train a model from multiple history wafers of production procedures. Neural network model [19] is employed to characterize the internal process dynamics and estimate the overlay variations based on history data analysis. Fusion method [20] utilizes a mixed training set with equipment parameters and overlay data to build a neural network model for predicting overlay variations. Virtual method [21,22] is presented to train an approximation function to improve the overlay optimization based on context information and sensor data. Pattern recognition technique [23] combines the available fingerprints with history overlay data for further improvements. In general, most of these methods must collect the overlay data of the k historical wafers as the training samples (i.e., $k = 3$ in [19]), where the history wafer means wafers with the same type. However, very often the collected sample comes with block-wise missing entries; for example, one sample may only have 2 or 1 history wafers that are less than k, making the overall data incomplete. In this case, samples with missing entries are discarded, resulting in a severe loss of available information. Moreover, it is known that there exist inherent correlations among multiple historical wafers, since they are produced by the same machine and belong to the same wafer type. In contrast, most existing training methods focus on training multiple historical wafers separately and thus cannot utilize the intrinsic useful correlation information. Identifying the correlation may contribute to performance improvements of overlay. Therefore, how to effectively exploit overlay data and construct a robust and accurate overlay model remains largely unexplored.

This paper aims to address these limitations by developing a novel framework termed low-rank manifold optimization (LRMO), which can obtain accurate overlay variations from incomplete and correlated overlay data effectively. The details of the proposed framework are described in Fig. 3. Our main contributions can be summarized as follows:

- By exploiting the underlying correlation nature of the overlay data, optimizing overlay variation can be formulated as a low-rank model. Since the overlay samples with the same wafer type are produced by the same machine, which thus promotes the underlying correlations and similarities among overlay variations with low-rank structure.
- Wafers usually contain incomplete overlay data during the lithography process. The proposed low-rank model can make full use of all incomplete overlay data that are usually discarded by existing overlay methods. The low-rank model can estimate the incomplete overlay data, which avoid to lose useful overlay information and can offer superior performance for overlay optimization.
- To solve this low-rank model, we design a Riemannian optimization algorithm over manifold to obtain optimal overlay variations. Particularly, the proposed algorithm exploits smooth geometries over the low-rank manifold and utilizes Riemannian gradient strategy for overlay optimization. This allows for computing accurate overlay variations with a lower complexity.

The remainder of the paper is organized as follows: Section 2 analyzes the challenges for overlay optimization. Section 3 gives the low-rank model and Section 4 presents low-rank algorithm for overlay optimization. Section 5 discusses how to compensate the results of our approach into the lithography process. Section 6 implements extensive experiments to verify the performance of our framework. Finally, we conclude the paper in Section 7.

2. Challenges

The goal of overlay optimization is to predict the unknown overlay variations and compensate them into the wafer production. Overlay variations refer to the displacement variations of an exposed layer relative to the previous exposed layer, which are caused mainly by stepper [8]. In this paper, we considered the following eight overlay variations as shown in Table 1. These eight overlay variations of a wafer can be formulated as the following vector form

$$[T_h, T_v, E_h, E_v, R, O, S_e, S_r] \tag{1}$$

Download English Version:

https://daneshyari.com/en/article/7104397

Download Persian Version:

https://daneshyari.com/article/7104397

Daneshyari.com