Registration of Infrared Transmission Images Using Squared-Loss Mutual Information

Tomoya Sakai
Tokyo Institute of Technology
sakai@sg.cs.titech.ac.jp

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

Katsuichi Kitagawa
Toray Engineering Co., Ltd.
BXM02060@nifty.ne.jp

Kazuyoshi Suzuki
Toray Engineering Co., Ltd.
kazuyoshi_suzuki@toray-eng.co.jp

Abstract

Infrared light allows us to measure the inner structure of opaque samples such as a semi-conductor. In this paper, we propose a method of registering multiple infrared transmission images obtained from different layers of a sample for 3D reconstruction. Since an infrared transmission image obtained from one layer is contaminated with defocused images coming from other layers, registration with a standard similarity metric such as the squared error and the cross correlation does not perform well. To cope with this problem, we propose to use the *squared-loss mutual information* as an alternative similarity measure for registration, which is more robust against noise than ordinary mutual information. The practical usefulness of the proposed method is demonstrated in simulated and actual experiments.

Keywords

Infrared transmission image, Image registration, Squared-loss mutual information

1 Introduction

Non-destructive inspection of precision instruments such as semi-conductors is one of the most important manufacturing processes in precision industries, and measuring inner structure of opaque objects by *infrared light* is a promising means for this purpose. However, images obtained by infrared light from different layers of a sample often suffer misalignment because of asynchronous scanning of images from different layers. In this paper, we therefore consider the problem of registering multiple infrared transmission images obtained from different layers of a sample, and propose a new practical algorithm for 3D reconstruction. Our proposed method can be used, e.g., for identifying the position of defects in semi-conductor samples and thus can provide more precise information of the inner structure for inspection.

Related image registration problems have been explored, e.g., in photolithography processes for aligning circuit patterns and masks [11] and in pattern and photo-mask inspection for comparing target and reference images [12]. On the other hand, the image registration problem we are tackling in this paper is much more challenging than the previous works because an infrared transmission image obtained from one layer is contaminated with defocused images coming from other layers. For this reason, standard *linear* similarity metrics such as the *sum of squared differences* (SSD) and the *normalized cross-correlation* (NCC) [1, 2, 3, 4, 5, 6] are not suitable to registration of infrared transmission images.

Mutual information (MI) [7], which is a quantity of interest in the *information theory* community, allows us to capture non-linear variations between two random variables:

$$\mathrm{MI} := \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y, \tag{1}$$

where $p(x,y)$ is the joint probability density of x and y, and $p(x)$ and $p(y)$ are the marginal probability densities of x and y, respectively. MI is always non-negative, and takes zero if and only if x and y are statistically independent. Thus, MI measures the dependency between x and y, which describes more precise "relation" than linear correlations. For example, if $y = x^2$, x and y are uncorrelated but they are still statistically dependent and thus MI takes a strictly positive value. Note that MI is reduced to a linear correlation measure when x and y follow the centered Gaussian distributions. Thus, MI can be regarded as a generalization of linear correlations to non-Gaussian random variables.

Given this superior detectability of statistical dependency, MI should be regarded as a suitable similarity measure for registration of infrared transmission images. However, because of the log function included in the definition of MI, which is an extremely steep function near the origin, MI is highly sensitive to noise and outliers [8].

To overcome the excessive sensitivity of MI, a variant of MI called *squared-loss mutual information* (SMI) [9] was introduced:

$$\mathrm{SMI} := \frac{1}{2} \iint p(x)p(y) \left(\frac{p(x,y)}{p(x)p(y)} - 1 \right)^2 \mathrm{d}x\mathrm{d}y. \tag{2}$$

SMI is also always non-negative and takes zero if and only if x and y are statistically independent. Thus, SMI can be used as an alternative to MI for evaluating statistical dependency, without suffering the "log" problem. Furthermore, thanks to the simple squared-difference expression of SMI, it can be *analytically* approximated from samples in a statistically optimal way, and this analytic SMI approximator allows explicit computation of its derivative [10]. This is a highly useful property in image registration, because eventually we want to register images so that SMI is maximized with respect to some image transformation parameters. For these reasons, we propose to use SMI as our dependency measure for registration of infrared transmission images.

After reviewing an SMI approximator in Section 2, we describe our SMI-based image registration method in Section 3. Its performance is experimentally evaluated in Section 4, and we conclude in Section 5.

2 SMI Approximation

Since SMI defined by (2) contains unknown probability densities $p(x, y)$, $p(x)$, and $p(y)$, its value cannot be directly computed. In this section, we review how SMI is approximately computed from paired samples $\{(x_i, y_i)\}_{i=1}^n$ independently following $p(x, y)$.

A naive way to approximately compute SMI is to separately estimate the densities $p(x, y)$, $p(x)$, and $p(y)$ from samples $\{(x_i, y_i)\}_{i=1}^n$, and plug the estimated densities $\widehat{p}(x, y)$, $\widehat{p}(x)$, and $\widehat{p}(y)$ in the definition of SMI. However, such a plug-in approach is known to perform poorly, because the first step of estimating densities is performed without regards to the second step of plugging them in SMI. More specifically, dividing $\widehat{p}(x, y)$ by $\widehat{p}(x)$ and $\widehat{p}(y)$ can significantly magnify the estimation error of $\widehat{p}(x, y)$ when $\widehat{p}(x)$ and $\widehat{p}(y)$ take a small value [13].

To cope with this problem, the direct SMI approximator called *least-squares mutual information* (LSMI) [9, 8] was proposed. LSMI directly estimates the density ratio function,

$$r(x, y) := \frac{p(x, y)}{p(x)p(y)}, \tag{3}$$

without individually estimating each density. More specifically, the density ratio $r(x, y)$ is modeled by the following *multiplicative kernel model* [14]:

$$r_{\boldsymbol{\Theta}}(x, y) := \sum_{i,j=1}^{n} \Theta_{i,j} K(x, x_i) L(y, y_j), \tag{4}$$

where $K(x, x')$ and $L(y, y')$ are kernel functions for x and y, and $\boldsymbol{\Theta}$ is a parameter matrix whose (i, j)-element is $\Theta_{i,j}$. Below, we focus on the Gaussian kernel for $K(x, x')$ and

$L(y, y')$:

$$K(x, x') := \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right), \tag{5}$$

$$L(y, y') := \exp\left(-\frac{(y - y')^2}{2\sigma^2}\right), \tag{6}$$

where σ denotes the Gaussian bandwidth. When the sample size n is too large, we may reduce the number of kernel bases in the model (4) by, e.g., randomly selecting a subset. In our experiments, we randomly choose 100 kernel bases.

The parameter $\boldsymbol{\Theta}$ is determined so that the following squared error J is minimized:

$$\begin{aligned} J(\boldsymbol{\Theta}) &:= \frac{1}{2}\iint p(x)p(y)\Big(r_{\boldsymbol{\Theta}}(x, y) - r(x, y)\Big)^2 \mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2}\iint p(x)p(y)r_{\boldsymbol{\Theta}}(x, y)^2 \mathrm{d}x\mathrm{d}y - \iint p(x, y)r_{\boldsymbol{\Theta}}(x, y)\mathrm{d}x\mathrm{d}y + C, \end{aligned} \tag{7}$$

where C is a constant independent of $\boldsymbol{\Theta}$. By ignoring the irrelevant constant C, approximating the expectations with the empirical averages, and including the ℓ_2-regularizer for avoiding overfitting, the LSMI solution $\widehat{\boldsymbol{\Theta}}$ was shown to satisfy the following equation [14]:

$$\frac{1}{n^2}\boldsymbol{K}^2\widehat{\boldsymbol{\Theta}}\boldsymbol{L}^2 + \lambda\widehat{\boldsymbol{\Theta}} = \frac{1}{n}\boldsymbol{K}\boldsymbol{L}, \tag{8}$$

where $K_{i,j} = K(x_i, x_j)$, $L_{i,j} = L(y_i, y_j)$, and $\lambda \geq 0$ denotes the regularization parameter that is determined by *cross-validation* with respect to J [9]. The above equation is called the *discrete-time Sylvester equation* [15], and is known to be solved in $\mathcal{O}(n^3)$ time, e.g., via eigendecomposition.

Finally, based on another expression of SMI,

$$\mathrm{SMI} = \frac{1}{2}\iint p(x, y)r(x, y)\mathrm{d}x\mathrm{d}y - \frac{1}{2}, \tag{9}$$

the LSMI estimator is given as

$$\mathrm{LSMI} := \frac{1}{2n}\mathrm{tr}(\boldsymbol{K}\widehat{\boldsymbol{\Theta}}\boldsymbol{L}) - \frac{1}{2}. \tag{10}$$

3 Image Registration with LSMI

Let us consider the problem of registering images X and Y: we transform image X to $\widetilde{X}$ (for example, by rotation or translation), so that $\widetilde{X}$ "matches" Y as much as possible. As the matching score, we use LSMI between $\widetilde{X}$ and Y. We regard $\widetilde{x}$ as a pixel value of image $\widetilde{X}$ and y as a pixel value of image Y, and we generate paired samples $\{(\widetilde{x}_i, y_i)\}_{i=1}^n$ by coupling pixel values at corresponding points in the images.

More specifically, our goal is to maximize LSMI with respect to an image transformation parameter from X to $\widetilde{X}$. Here, we use a gradient-based method, such as a gradient ascent method and a quasi-Newton method, to find a (local) maximizer of LSMI. The gradient of LSMI is given by

$$\nabla \mathrm{LSMI} = \frac{1}{n^2\sigma^2}\mathrm{tr}(\widehat{\boldsymbol{\Theta}}^\top \widetilde{\boldsymbol{K}}\boldsymbol{Q}\widehat{\boldsymbol{\Theta}}\boldsymbol{L}^2) - \frac{1}{n\sigma^2}\mathrm{tr}(\boldsymbol{Q}\widehat{\boldsymbol{\Theta}}\boldsymbol{L}), \tag{11}$$

where

$$\widetilde{K}_{i,j} := K(\widetilde{x}_i, \widetilde{x}_j), \tag{12}$$

$$Q_{i,j} := \widetilde{K}_{i,j}(\widetilde{x}_i - \widetilde{x}_j)\nabla\widetilde{x}_i, \tag{13}$$

and $\nabla\widetilde{x}_i$ denotes the gradient of the pixel value at $\widetilde{x}_i$.

Below, for simplicity, we focus on translation as image transformation[1]. Then, the pixel values of the transformed image $\widetilde{X}$ are given by

$$\widetilde{x}(u,v) := x(u - w_u, v - w_v), \tag{14}$$

where (u, v) denotes the coordinates on the image $\widetilde{X}$, and $\boldsymbol{w} = (w_u, w_v)^\top$ denotes the amount of translation. The gradient $\nabla\widetilde{x}$ is expressed as

$$\nabla\widetilde{x} = -\begin{pmatrix} \dfrac{\partial\widetilde{x}(u,v)}{\partial u} \\ \dfrac{\partial\widetilde{x}(u,v)}{\partial v} \end{pmatrix}, \tag{15}$$

which is numerically computed by bilinear interpolation in our implementation.

The entire procedure of our LSMI-based image registration method is summarized as follows:

1. Initialize the image transformation parameter $\boldsymbol{w}$ to zero.

2. Determine σ and λ in LSMI by cross-validation.

3. Find a local maximizer $\widehat{\boldsymbol{w}}$ of LSMI by a quasi-Newton method.

4 Experiments

In this section, we experimentally demonstrate the performance of our proposed method.

We obtained infrared transmission images of size 2336 × 1632 using the surface profiler *SP500* by Toray Engineering Co., Ltd., equipped with an infrared light source (see Figure 1). We took 71 images from a semi-conductor sample, which were captured in the

[1] Note that our framework can handle any transformation as long as it is smooth with respect to transformation parameters.

Figure 1: The surface profiler SP500 equipped with an infrared light source.

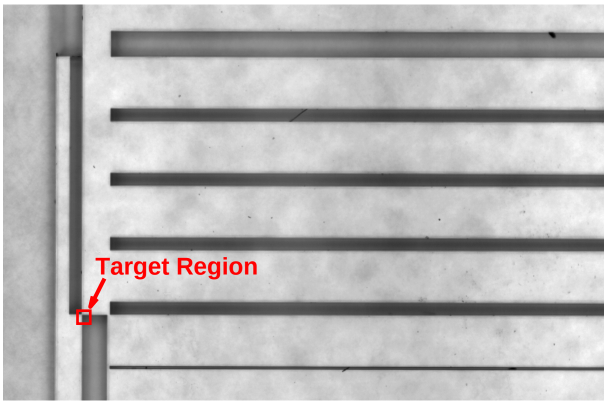


Figure 2: The target region used for registration.

layer-by-layer manner and small misalignments were incurred due to mechanical movements. We compared the performance of the proposed LSMI-based method with two conventional methods based on normalized cross-correlation (NCC) [3] and mutual information (MI) [16], where kernel density estimators for $p(x)$, $p(y)$, and $p(x, y)$ were used when computing the gradient of MI.

We selected the small region of size 50×50 pixels for registration (Figure 2). Examples of images extracted from the target region are shown in Figure 3, where the 1st layer is the closest to the surface and the 71st layer is the closest to the bottom.

4.1 Image Registration under Artificial Noise

First, we illustrate the behavior of the existing and proposed methods using images with artificial noise. More specifically, we first duplicated the original image in the 1st layer. Then we added the following noise to the images:

- Gaussian noise with mean zero and standard deviation $\tau = 0.01$, 0.02, and 0.03,

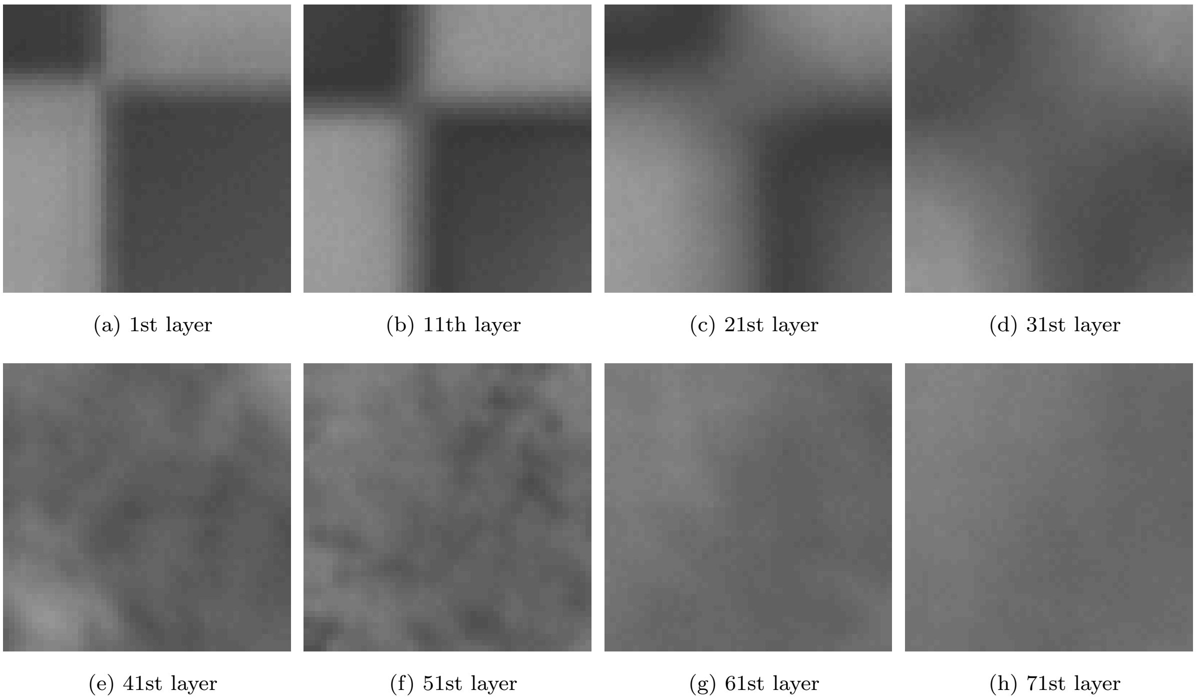

(a) 1st layer
(b) 11th layer
(c) 21st layer
(d) 31st layer
(e) 41st layer
(f) 51st layer
(g) 61st layer
(h) 71st layer

Figure 3: Examples of images extracted from the target region.

respectively.

- Gaussian blurring with standard deviation $\eta = 1$, 2, and 3, respectively.
- Impulse noise to relegate a pixel value to either black (the minimum value) or white (the maximum value) with probability $p = 0.001$, 0.002, and 0.003, respectively.

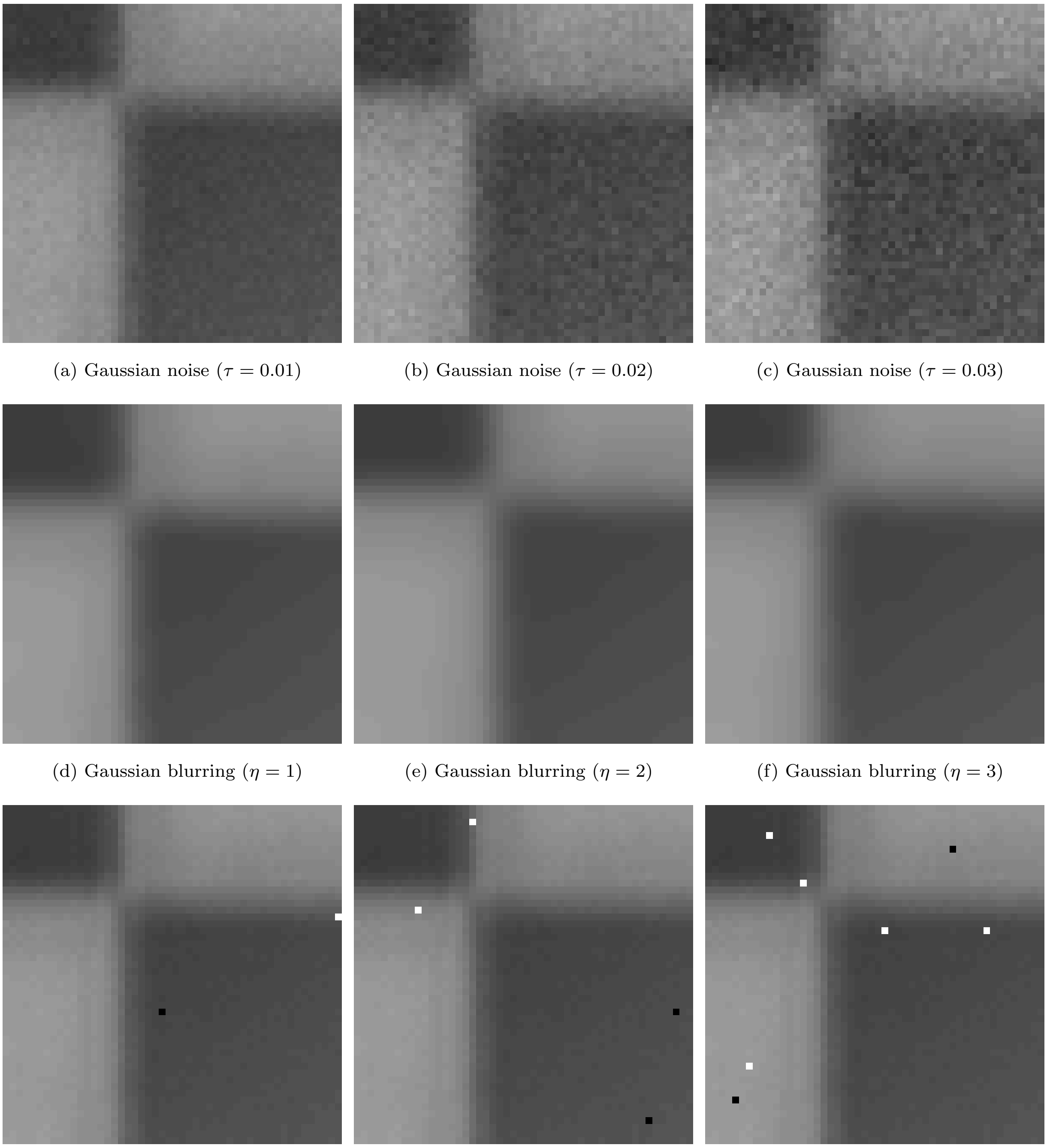
Note that the above Gaussian and impulse noise were added to both images, but Gaussian blurring was applied only to one of the images due to its deterministic nature. We then translated one image by $\boldsymbol{w}^* = (w_u^*, w_v^*)^\top$, where each element was drawn from the standard normal distribution. Figure 4 shows examples of images with artificial noise.

Table 1 summarizes the estimation error of registration:

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|, \tag{16}$$

where $\boldsymbol{w}^*$ is the ground-truth translation and $\widehat{\boldsymbol{w}}$ is its estimate by a registration method. This shows that the LSMI-based method outperforms the correlation-based and MI-based methods with statistical significance. The poor performance of the correlation-based method is due to local optima caused by intrinsic non-smoothness, while the MI-based method is heavily affected by the impulse noise. On the other hand, the proposed method seems not to suffer those problem too much, thanks to data-driven Gaussian width choice by cross-validation and its robustness property to noise and outliers.

Next, we investigate the effect of changing the number of kernel bases in the proposed method. Figure 5 depicts the means and the standard deviations of estimation error and



(a) Gaussian noise ($\tau = 0.01$)

(b) Gaussian noise ($\tau = 0.02$)

(c) Gaussian noise ($\tau = 0.03$)

(d) Gaussian blurring ($\eta = 1$)

(e) Gaussian blurring ($\eta = 2$)

(f) Gaussian blurring ($\eta = 3$)

(g) Impulse noise ($p = 0.001$)

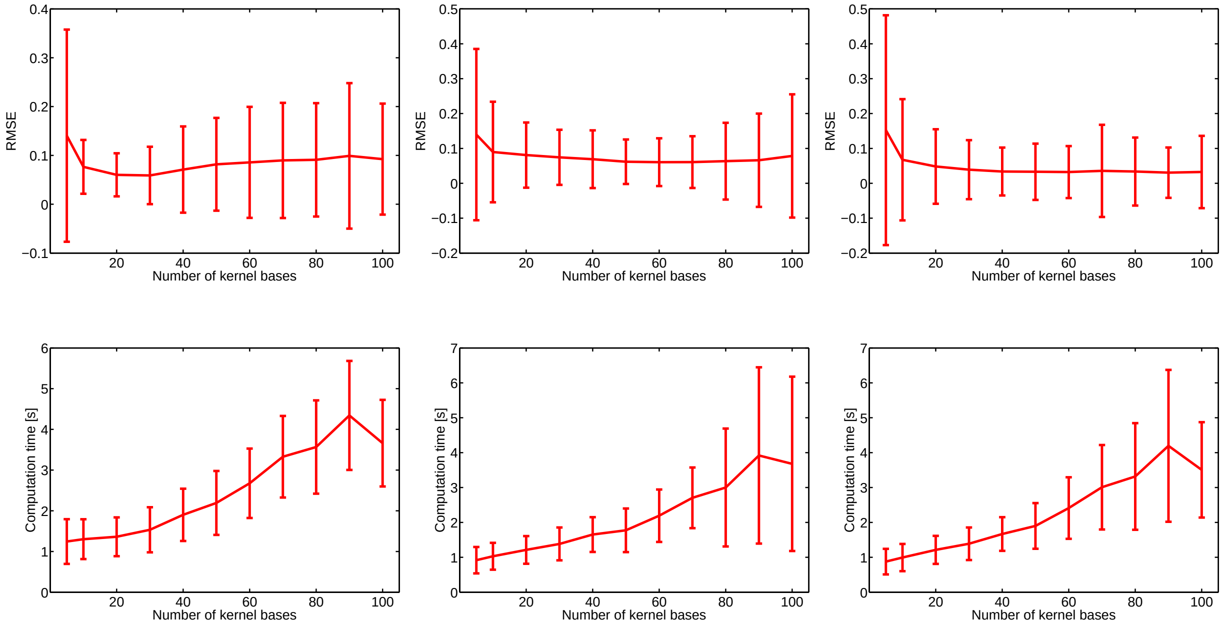
(h) Impulse noise ($p = 0.002$)

(i) Impulse noise ($p = 0.003$)

Figure 4: Examples of images with artificial noise. τ denotes the standard deviation of the Gaussian noise, η denotes the standard deviation of the Gaussian blurring, and p denotes the probability of having impulse noise.

Table 1: Means and standard deviations of estimation error over 100 runs. For each setup, the best method in terms of the mean error and comparable ones according to the t-test at the significance level 1% are specified by boldface.

Noise type	Noise level	Correlation	MI	LSMI
Gaussian noise	$\tau = 0.01$	0.89 (0.62)	0.27 (0.43)	**0.12** (**0.10**)
	$\tau = 0.02$	0.94 (0.61)	0.56 (0.57)	**0.16** (**0.11**)
	$\tau = 0.03$	0.89 (0.57)	0.78 (0.57)	**0.30** (**0.32**)
Gaussian blurring	$\eta = 1$	0.96 (0.55)	0.27 (0.48)	**0.08** (**0.12**)
	$\eta = 2$	1.07 (0.66)	0.42 (0.61)	**0.17** (**0.29**)
	$\eta = 3$	0.98 (0.58)	0.42 (0.60)	**0.12** (**0.10**)
Impulse noise	$p = 0.001$	1.13 (0.70)	0.85 (0.77)	**0.05** (**0.20**)
	$p = 0.002$	1.02 (0.71)	0.98 (0.71)	**0.05** (**0.18**)
	$p = 0.003$	0.96 (0.66)	1.01 (0.70)	**0.05** (**0.10**)



(a) Gaussian noise: $\tau = 0.01$

(b) Gaussian blurring: $\eta = 1$

(c) Impulse noise: $p = 0.001$

Figure 5: Means and standard deviations of estimation error (upper) and computation time (lower) over 1000 runs as functions of the number of basis functions in LSMI.

Table 2: Estimation error over all layers and the amount of translation between the 1st and 71st layers.

Methods	Estimation error	Translation	
		u-axis	v-axis
Ground truth	0	-10.9	-9.9
No registration	0.58	0	0
Correlation	0.45	-1.2	-7.4
MI	0.30	-6.9	-9.9
LSMI	0.23	-10.4	-13.4

computation time as functions of the number of kernel bases. The results show that the estimation error is not strongly affected by changing the number of kernel bases, but the computation time tends to grow as the number of kernel bases increases. This results show that, even if the number of kernel bases is reduced from 100 to 20–30, we can still obtain comparable estimation results only with about one third of computation time.

4.2 Multi-Layer Image Registration on Real Data

Finally, we apply our proposed image registration method to multi-layer image registration for all 71 images. More specifically, we align the 2nd image to the 1st image, then the 3rd image to the 2nd image, and this is continued until the 71st image is aligned to the 70th image.

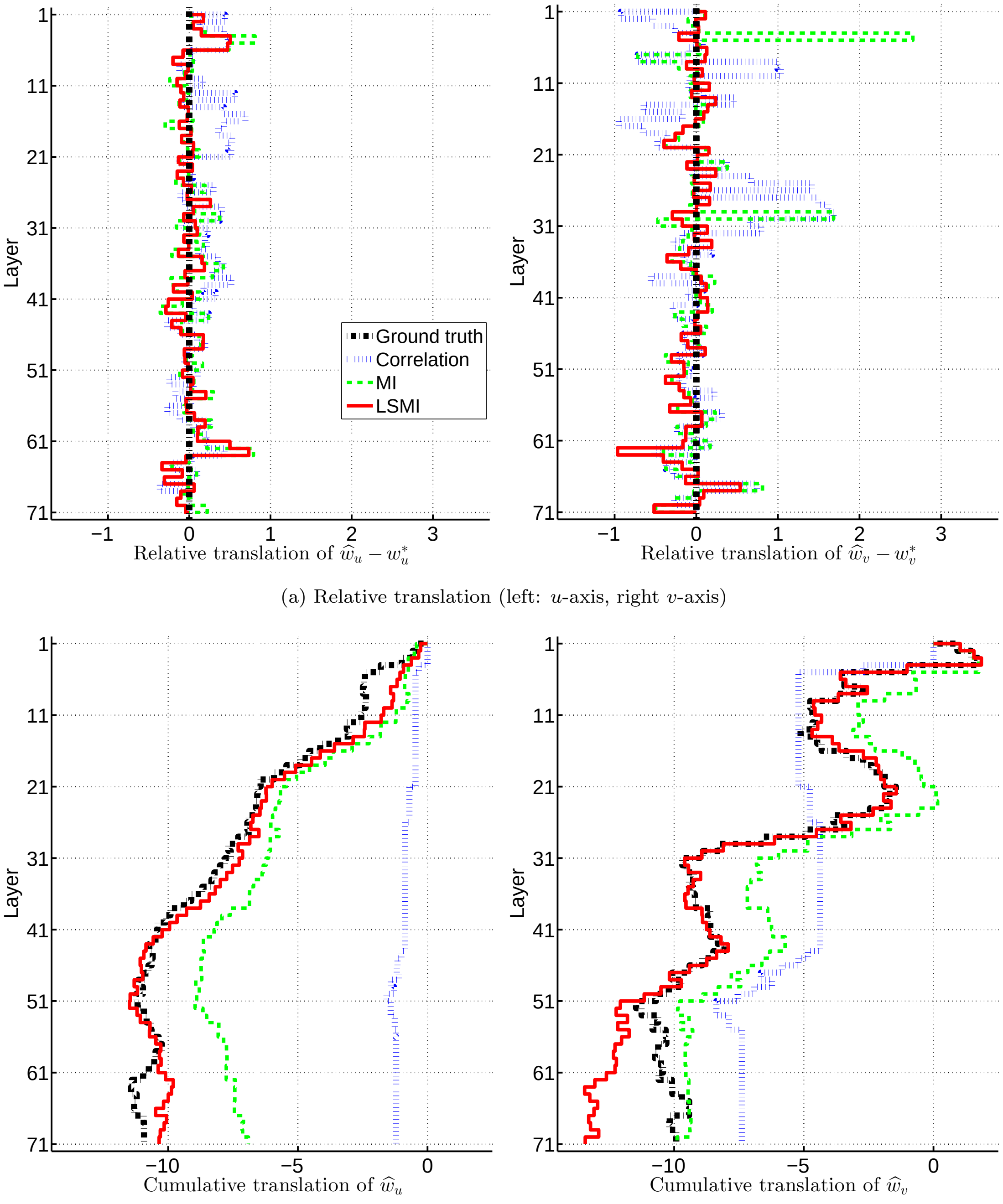
Figure 6 illustrates the registration results, and Table 2 describes the estimation error over all layers evaluated by

$$\frac{1}{70}\sum_{i=1}^{70}\|\widehat{\boldsymbol{w}}_{i,i+1}-\boldsymbol{w}^*_{i,i+1}\|, \tag{17}$$

where $\boldsymbol{w}^*_{i,i+1}$ is the manually annotated ground-truth of translation between the i-th and $(i+1)$-th layers, and $\widehat{\boldsymbol{w}}_{i,i+1}$ is its estimate by a registration method. From the results, we can confirm that the propose LSMI-based method is a promising alternative to the existing correlation-based and MI-based approaches.

5 Conclusion

In this paper, we proposed a novel image registration method that uses squared-loss mutual information (SMI) as a similarity measure. We experimentally demonstrated that the proposed method, in which an approximator of SMI called least-squares mutual information (LSMI) is used, outperforms the existing methods. We also investigated the effect of the number of kernel bases used in LSMI, and experimentally confirmed that even when the number of kernel bases is reduced, the proposed LSMI-based method still works



(a) Relative translation (left: u-axis, right v-axis)

(b) Cumulative translation (left: u-axis, right v-axis)

Figure 6: Results of multi-layer image registration.

well, with much less computation costs. In our future work, we will further investigate possible reduction in computation costs, e.g., by a coarse-to-fine approach.

In our current implementation for multi-layer image registration, images are registered one by one from top to bottom. Although this is simple, estimation error in the early layers can be cumulated in the latter layers. A potential advantage of the proposed method is that, unlike the cross correlation, SMI (and also its approximator LSMI) can be easily extended to more than two variables [17]: SMI for m variables $x^{(1)}, \ldots, x^{(m)}$ is defined as

$$\mathrm{SMI} := \frac{1}{2} \int \cdots \int p(x^{(1)}) \cdots p(x^{(m)}) \left(\frac{p(x^{(1)}, \ldots, x^{(m)})}{p(x^{(1)}) \cdots p(x^{(m)})} - 1 \right)^2 \mathrm{d}x^{(1)} \cdots \mathrm{d}x^{(m)}. \qquad (18)$$

Thus, in principle, it is possible to simultaneously handle similarities among all images. However, naive implementation of simultaneous registration of many images will be computationally too expensive. Our future challenge is to address this computational issue. Borrowing the idea from coordinate-wise ascent in optimization theory is expected to be promising.

References

[1] L. G. Brown, A survey of image registration techniques, ACM Computing Surveys 24 (4) (1992) 325–376.

[2] B. Zitová, J. Flusser, Image registration methods: A survey, Image and Vision Computing 21 (11) (2003) 977–1000.

[3] R. Szeliski, Image alignment and stitching: A tutorial, Tech. rep., MSR-TR-2004-92, Microsoft Research, 2004 (2005).

[4] M. V. Wyawahare, P. M. Patil, H. K. Abhyankar, Image registration techniques: An overview, International Journal of Signal Processing, Image Processing and Pattern Recognition 2 (3) (2009) 11–28.

[5] M. P. Deshmukh, U. Bhosle, A survery of image registration, International Journal of Image Processing 5 (3) (2011) 245–269.

[6] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, IEEE Transactions on Medial Imaging 32 (7) (2013) 1153–1190.

[7] C. Shannon, A mathematical theory of communication, Bell Systems Technical Journal 27 (1948) 379–423.

[8] M. Sugiyama, Machine learning with squared-loss mutual information, Entropy 15 (2013) 80–112.

[9] T. Suzuki, M. Sugiyama, T. Kanamori, J. Sese, Mutual information estimation reveals global associations between stimuli and biological processes, BMC Bioinformatics 10 (2009) S52:1–12.

[10] T. Suzuki, M. Sugiyama, Sufficient dimension reduction via squared-loss mutual information, Neural Computation 25 (3) (2013) 725–758.

[11] S. Slonaker, S. McNamara, K. Konno, R. Miller, N. Magome, T. Umatate, H. Tateno, Enhanced global alignment for production optical lithography, in: Proceedings of SPIE, Optical/Laser Microlithography, Vol. 0922, 1988, pp. 73–81.

[12] T. S. Newman, A. K. Jain, A survey of automated visual inspection, Computer Vision and Image Understanding 61 (2) (1995) 231 – 262.

[13] M. Sugiyama, T. Suzuki, T. Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, Cambridge, UK, 2012.

[14] T. Sakai, M. Sugiyama, Computationally efficient estimation of squared-loss mutual information with multiplicative kernel models, IEICE Transactions on Information and Systems E97-D (4) (2014) 968–971.

[15] V. Sima, Algorithms for Linear-Quadratic Optimization, Marcel Dekker, New York, NY, USA, 1996.

[16] P. Viola, W. M. Wells, III, Alignment by maximization of mutual information, International Journal of Computer Vision 24 (2) (1997) 137–154.

[17] T. Suzuki, M. Sugiyama, Least-squares independent component analysis, Neural Computation 23 (1) (2011) 284–301.