Machine Vision and Applications (2020) 31:10
https://doi.org/10.1007/s00138-020-01057-6

ORIGINAL PAPER

Cross-spectral registration of natural images with SIPCFE

Amir Hossein Farzaneh[1] · Xiaojun Qi[1]

Received: 22 January 2018 / Revised: 14 April 2019 / Accepted: 8 January 2020 / Published online: 6 February 2020

Abstract
Image registration is a viable task in the field of computer vision with many applications. When images are captured under different spectrum conditions, a challenge is imposed on the task of registration. Researchers carefully handcraft a local module insensitive to illumination changes across cross-spectral image pairs to tackle this challenge. We, in this paper, develop an optimized feature-based approach Single Instance Phase Congruency Feature Extractor (SIPCFE) to tackle the problem of natural cross-spectral image registration. SIPCFE uses the phase information of an image pair to quickly identify and describe reliable keypoints that are insensitive to illumination. It then employs a sequence of outlier removal processes to find the matching feature points accurately and the Direct Linear Transformation to estimate the geometric transformation to align the image pair. We extensively study the proposed approach for every module in the system to give more insights into the challenges. We benchmark our proposed method and other state-of-the-art feature-based methods developed for cross-spectral imagery on three datasets with various settings and image contents. The comprehensive analysis of cross-spectral registration results of natural images demonstrates that SIPCFE achieves up to 47.24%, 14.29%, and 12.45% accuracy improvement on the first, second, and third dataset, respectively, over the second best registration method in the benchmark.

Keywords Cross-spectral registration · Phase congruency · Near infrared · Feature-based image registration

1 Introduction

A cross-spectral image pair is a pair of two corresponding images captured in different imaging configurations such as different camera exposures, different camera positions, and different sensors. These different configurations make the images in one pair not perfectly aligned; hence, registering them is a challenging task in computer vision applications. When registering two images, the aim is to find a geometric transformation between a pair of corresponding images to compensate for the rotation, translation, and scaling differences. The transformation is then used to spatially align, superimpose, or match the images in a pair. With two registered images, it is easier to fuse information or describe the differences between them. Cross-spectral image registration has broad applications in remote sensing, object detection, noise reduction, 3D image reconstruction, image fusion, video surveillance, medical image analysis, and image mosaicking.

In this paper, we focus on registering the RGB spectrum and near-infrared (NIR) spectrum image pairs. The intensity variation presented in this type of cross-spectral images imposes an additional challenge in the task of registration. Figure 1 represents a pair of RGB-NIR cross-spectral images. As illustrated, the viewpoint for the NIR image shown on the right is slightly moved to the right. This difference in viewpoint is regarded as a translation between the pairs. A pair might also have differences in scale or rotation in the viewpoint. Additionally, because different sensors capture different color spectrums, each corresponding pixel between two images has a different range of values, which is regarded as intensity variation in this application. Registration methods are categorized into two classes, i.e., similarity measure-based global methods and feature-based local methods.

Methods relying on similarity measures are mainly built on global statistical dependencies between images. Mutual information (MI), which was initially introduced by Maes et al. [26] and Viola and Wells III [43], is a widely used similarity measure capturing the global structure of an image.

✉ Amir Hossein Farzaneh
farzaneh@aggiemail.usu.edu

Xiaojun Qi
xiaojun.qi@usu.edu

[1] Department of Computer Science, Utah State University, 4205 Old Main Hill, Logan, UT 84322-4205, USA

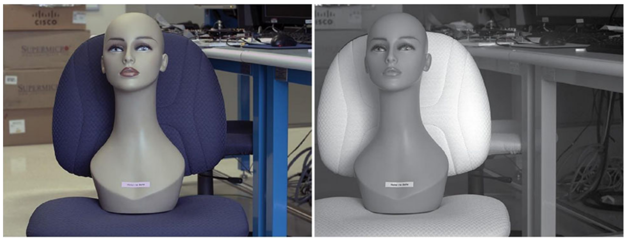


Fig. 1 Illustration of an RGB-NIR cross-spectral image pair

Therefore, MI is not capable of describing local structures and differentiating local intensity variations. These deficiencies compelled researchers to develop optimized MI-based registration methods to grasp local information. Pluim et al. [31] address the shortcomings of the original MI by combining the MI and gradient information to align the locations with a large gradient magnitude. This new measure also sets the orientation of the gradient at those locations to be similar. Rueckert et al. [37] develop a voxel similarity measure based on higher-order MI to address non-rigid registration. However, some misalignments are still present around the contours in rigid transformations. Studholme et al. [41] propose Regional MI (RMI) to utilize an entropy-based MI to introduce an extra channel to the joint intensity histogram. This approach shows robustness toward local contrast variations in medical images. Loeckx et al. [23] introduce conditional MI (CMI) to incorporate both intensity and spatial information of the image to be registered. However, the uncertainty of spatial distribution affects the performance of the method. Rivaz et al. [34] propose a self-similarity α-MI (SeSaMI) method using the gradient information to make it invariant to rotation and local affine intensity distortions. Although these proposed approaches inject some form of local representation in a global-based method, they are computationally complex, sensitive to noise, and time-consuming [16]. Ikena et al. [18] address the runtime issue using CUDA GPU programming, which needs special hardware and cannot be ported to all platforms. MI-based approaches [32] have been developed and optimized for medical images, which mostly are in grayscale. Therefore, they cannot handle natural images with richer details and higher intensity variation.

Locally solving the problem of cross-spectral registration has been tackled mostly with feature-based approaches. An end-to-end feature-based method finds a correspondence between the matching keypoints and estimates a transformation from one spectrum to another. It usually consists of three modules, namely keypoint extraction, feature extraction, and outlier removal. Firmenichy et al. [12] use classic keypoint detectors such as Harris corner detector [14] and a Gradient Direction Invariant version of Scale-Invariant Feature Transform (GDISIFT) to extract and match keypoints between RGB and NIR images. Hrkac et al. [17] use SUSAN and Harris corner detectors [14,40] and the Hausdorff distance to find the corresponding points. Han et al. [13] design a hybrid approach using both straight lines derived from edge pixels and keypoints as features to register IR and RGB images. They estimate an initial global transformation based on the straight lines and then use it to adapt the transformation on keypoints in small patches locally. Similarly, Zhao et al. [48] propose a hybrid approach and combine edges and keypoints extracted by phase congruency (PC) [21] as features. These features are further described by a multi-modality Robust Line Segment Descriptor (MRLSD), and a bidirectional matching method is utilized to find corresponding points. Qin et al. [33], inspired by the idea of encoding the gradient orientations in the image, develop a new descriptor called Histogram of Collinear Gradient-Enhanced Coding (HCGEC) to register Long Wave Infrared (LWIR) and RGB images. The Gixel Array Descriptor (GAD) proposed by Pang and Neumann [30] introduces the term Gixel to collect edge information extracted with the canny edge detector around a keypoint extracted by SURF. Several Gixels in a circular array construct the Gixel descriptor. GAD works well for both medical and natural images, but its running time is not optimized. Aguilera et al. [2] extract keypoints using the FAST detector [36] and describe features around them using Log-Gabor Histogram Descriptor (LGHD), an extended version of the Edge Histogram Descriptor (EHD) [1]. Kim et al. [19] develop a Dense Adaptive Self-Correlation (DASC) descriptor by taking advantages of an adaptive self-correlation measure and a randomized receptive field pooling learned by the linear discriminative learning. The disadvantage of feature-based methods is that finding repeatable and robust features between different spectrums and different image content is often a challenging task.

Other local-based methods [7,27,44], which do not fall in the category of feature-based approaches, have also been introduced. However, they are either sensitive to noise or can only tolerate a minimal amount of noise. Deep learning-based approaches have also been explored. For example, Large Deformation Diffeomorphic Metric Mapping (LDDMM) [46] is utilized to develop a 3D Convolutional Neural Network (CNN) architecture called Quicksilver to register two un-aligned medical images. LDDMM predicts the deformation parameters using the predicted initial momenta of the input. However, Quicksilver is optimized for medical images represented in 3D voxels. Additionally, deep learning methods require a sizeable pre-aligned dataset to train a network, which is not always easy to craft.

This paper proposes a fast, reliable, and robust image registration method to align the RGB and NIR image pair under different illumination conditions. We refer to the proposed method as Single Instance Phase Congruency Feature Extractor (SIPCFE). The contributions of the proposed method are as follows: (1) employing the PC method and its adaptive noise variant to extract the keypoints that are invariant to intensity changes; (2) incorporating the intermediate results from keypoint extraction, namely the Log-Gabor filter responses, in the feature description step to represent each keypoint using the histogram of oriented Log-Gabor filters; (3) designing a sequence of outlier removal processes to match corresponding keypoints accurately between the RGB and NIR image pair, which performs well regardless whether non-rigid or rigid correspondences are present in the data; and (4) utilizing the Direct Linear Transformation (DLT), a projective transformation, to estimate the geometric transformation for registering all the RGB points in the NIR domain. Finally, we conduct an extensive study of the image registration results on three sample datasets. Utah Water Research Laboratory (UWRL) and Computer Vision Lab at École Polytechnique Fédérale De Lausanne (EPFL) provide the first and second dataset, respectively. In addition, we choose a subset of the EPFL dataset to create image pairs with random rotation differences to evaluate the performance of the image registration methods under rotations. SIPCFE is evaluated on the registration results on these three datasets in terms of the Root Mean Square Error (RMSE). We show that the proposed method outperforms other state-of-the-art methods regarding accuracy. SIPCFE is faster on the second dataset and has comparable runtime on the first dataset compared to the second most accurate method. To the best of our knowledge, there is not a fully comprehensive study of the registration task in the literature, whereas the primary focus is on the evaluation of the keypoint extraction and keypoint description. This work is the first attempt to evaluate an end-to-end registration system from the perspectives of the performance of critical modules in the system and their impact on the whole system.

We organize the remainder of the paper as follows. Section 2 presents the proposed method. In Sect. 3, we present the evaluation method and the experimental results on the chosen datasets. An ablation study is pursued to develop a comprehensive guideline for future research. Finally, we deduce conclusions in Sect. 4.

2 Proposed SIPCFE method

In our application, we intend to register the RGB image onto the NIR image. Hence, the NIR and RGB images are referred to as the reference image (i.e., I_r) and the moving image (i.e., I_m), respectively. All the pixels in the reference image (e.g., NIR image) are kept static, and all the pixels in the moving image (e.g., RGB image) are transformed to the reference image plane via the geometric transformation found in the process. The algorithm overview of the proposed feature-based registration method, SIPCFE, is summarized in Algorithm 1.

Algorithm 1 The proposed feature-based algorithm (SIPCFE)

```
Input: original image pair $I = \{I_m, I_r\}$, number of scales $S$, number of orientations $O$.
Output: registered image pair $I' = \{I'_m, I_r\}$
for each $I_i$ in $I$ $(i \in \{m, r\})$ do
    for $s = 1$ to $S$ do
        for $o = 1$ to $O$ do
            Generate a Log-Gabor filter $lgf$ using $s$ and $o$
            Compute phase congruency $PC_i\{s, o\}$ by $I_i \circledast lgf$
        end for
    end for
    Compute the moment map $m_i$ from $PC_i$
    Extract keypoints $P_i^*$ from $m_i$
    Extract features $f_i$ around each keypoint $P_i^*$
end for
Find the pairs of putative matching keypoints $(P_m, P_r)$ in $I$ using an exhaustive search method
Apply Vector Field Consensus (VFC) on the feature pairs $(f_m, f_r)$ of $(P_m, P_r)$ to remove outliers and obtain robust matching keypoint pairs $(P'_m, P'_r)$
Find the geometric transformation $H$ based on $(P'_m, P'_r)$ using the Direct Linear Transformation (DLT) algorithm
Transform $I_m$ to $I'_m$ using $H$
```

Figure 2 presents the illustrative block diagram of the proposed image registration method, which consists of five components including keypoint extraction, keypoint feature description, keypoint feature matching, transformation estimation, and image registration. The aims of these five components are as follows:

- Keypoint extraction: Extracting distinct reliable and repeatable points in both the reference and the moving image.
- Keypoint feature description: Representing the keypoints in a compact but rich feature vector, which captures local information and is insensitive to intensity variation.
- Keypoint feature matching: Finding the corresponding matching keypoints between the RGB and NIR images and removing the outliers.
- Transformation estimation: Finding a geometric relationship between the matching keypoints in the form of a transformation matrix.
- Image registration: Aligning or superimposing the registered RGB image onto the NIR image.

In the following subsections, we explain each component in detail.

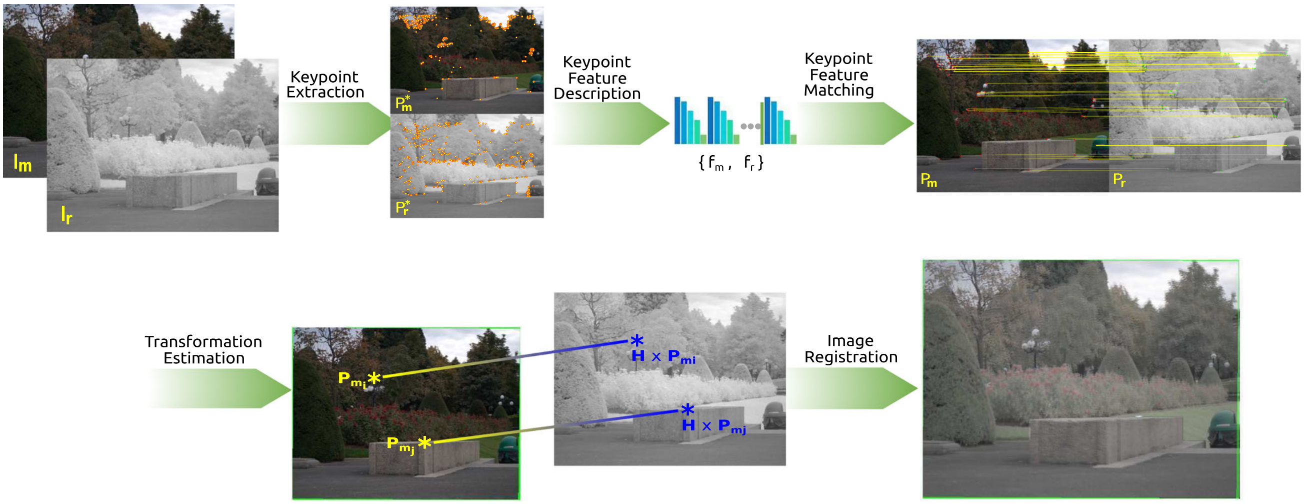


Fig. 2 The illustrative block diagram of the proposed image registration method

2.1 Keypoint extraction

A keypoint is a distinct spatial location representing what stands out in an image based on local information around the selected locations such as the corners. Therefore, unlike global measures, keypoints have to be insensitive to image rotation, translation, scale change, occlusions, and background clutter. Conventional keypoint extraction methods such as Harris [14], MinEigen [39], BRISK [22], HOG [8], SURF [4], MM-SURF [48], FAST [35] and its variants [36], and SIFT [24] extract local information based on statistical measures of the gradient. Gradient-based keypoint detectors degrade the performance of cross-spectral image registration when a large intensity variation exists between images. On the other hand, the phase congruency (PC) operator uses the local amplitude and phase of a signal at different locations as intermediate information and then performs principal moment analysis to extract keypoints. Since the collected information is highly localized with filter responses invariant to intensity changes, PC results in a keypoint extraction module, which is robust to varying illuminations usually presenting in image pairs captured in cross-spectral applications.

2.1.1 Computing PC for a 1D signal

Kovesi suggests PC [20,21] as a phase-based feature extraction, which is invariant to intensity changes between images and consequently makes it a dimensionless operator. This characteristic makes this feature detector robust in extracting features in cross-spectral images. The PC operator utilizes a Local Energy Model (LEM) [28] to extract features in an arbitrary image. In the LEM, features are described as signal locations that are in the most coherence state in the phase domain. Figure 3 illustrates that all the Fourier components are in phase at the step point of a one-dimensional square wave signal, which is highlighted in the green circular area.

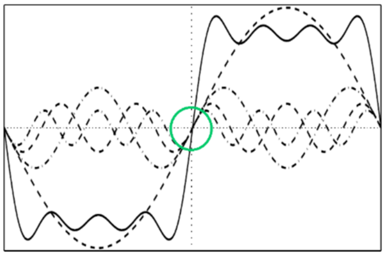


Fig. 3 Fourier components of a square wave signal, where the solid wave is the summation of the four dashed waves

The original PC by Morrone et al. [28] at location x in a one-dimensional signal is computed as the ratio of local energy to the sum of Fourier components' amplitudes at x. That is:

$$PC(x) = \frac{|E(x)|}{\sum_n A_n(x)} \tag{1}$$

where $E(x)$ is the local energy and $A_n(x)$ is the amplitude of the nth Fourier component at location x. Local energy can be redefined as a function of the cosine of the phase deviation, namely the difference between nth local phase component at location x (i.e., $\phi_n(x)$) and the mean of all local components at location x. As a result, (1) is rewritten as follows:

$$PC(x) = \frac{\sum_n A_n\left(\cos\left(\phi_n(x) - \overline{\phi_n}(x)\right)\right)}{\sum_n A_n(x)} \tag{2}$$

Kovesi [20] derives a modified local energy measure by subtracting the magnitude of the sine of the phase deviation from the original local energy. To further eliminate spurious response to noise and the fluctuations around a noisy step, an estimated noise energy value T is subtracted from the modified local energy measure before normalizing it by the sum of the Fourier response amplitudes. A small constant ϵ (e.g., 0.0001) is also added to the denominator to prevent the value from becoming unstable as the term $\sum_n A_n(x)$ becomes very small. Finally, to obtain a better-localized response, a frequency spread weighting factor $W(x)$ is applied to control the contribution of its corresponding modified local energy. This modified PC measure is as follows:

$$PC(x) = \frac{\sum_n W(x) \left\lfloor A_n(x) \left(\cos\left(\phi_n(x) - \overline{\phi_n}(x)\right) - \left| \sin\left(\phi_n(x) - \overline{\phi_n}(x)\right) \right| \right) - T \right\rfloor}{\sum_n A_n(x) + \epsilon} \tag{3}$$

where the rectifier $\lfloor\ \rfloor$ returns the enclosed value as it is if the value is positive; otherwise, the rectifier returns 0. The new modified formulation of PC not only provides better localization but also compensates the noise with an empirically determined optimal value T.

2.1.2 Computing PC for a 2D grayscale image

When computing PC for a $2D$ grayscale image, researchers use a bank of Log-Gabor filters with different scales and orientations to generate the responses of the image. Specifically, the filter response is used to replace the local energy information $E(x)$ and the magnitude of the filter response is used to replace the amplitude $A_n(x)$. In the proposed method, we empirically use 4 scales and 8 orientations to construct a bank of 32 Log-Gabor filters, which capture enough scale and orientation information.

In practice, the input image is convolved with a bank of Log-Gabor filters to obtain their responses (local energy) and the magnitude of the responses (amplitude). Finally, the two-dimensional PC at location x in a grayscale image is calculated by:

$$PC(x) = \frac{\sum_s W_o(x) \lfloor E_o(x) - T \rfloor}{\sum_o \sum\limits_s A_{so}(x) + \epsilon} \tag{4}$$

where $E_o(x)$ represents local energy at orientation o, $A_{so}(x)$ represents amplitude of the filter response at location x in the response image of scale s and orientation o, $W_o(x)$ is the frequency spread weighting factor at orientation o, and s and o represent the index of Log-Gabor filters' scale and orientation, respectively. For 2D PC, $W_o(x)$ penalizes narrow frequency distributions at the oth orientation and is defined by a sigmoidal function:

$$W_o(x) = \left(1 + \exp\left(g_o\left(c_o - \frac{1}{N_o}\left(\frac{\sum_s A_{so}(x)}{A_{\max}(x) + \epsilon}\right)\right)\right)\right)^{-1} \tag{5}$$

where c_o is the cutoff value, g_o is the gain factor that controls the cutoff sharpness, N_o is the total number of scales at orientation o, and $A_{\max}(x)$ is the maximum magnitude at x in the filters of all scales at all orientations.

2.1.3 Adaptive noise energy estimation

The noise energy value T in (4) is calculated based on the statistics of the filter response to the images. It is assumed that the noise is Gaussian; therefore, the response of Log-Gabor filters to noise forms Rayleigh distribution. By finding the Rayleigh distribution mode, we can easily calculate the mean and standard deviation, which can be further used to compute the noise energy (i.e., T). By default, the Median Absolute Deviation (MAD) criterion estimates the Rayleigh distribution mode, which serves as a simple estimator to estimate the noise statistics of an image with less complicated noise contamination. However, this estimated Rayleigh distribution mode does not always lead to accurate noise energy estimation. To address this issue, we devise a method to automatically choose an appropriate mode estimator and then compute a better estimation of T in (4). To this end, we adopt the idea of a wavelet-based de-noising algorithm [38] to use the joint statistics of the wavelet coefficients of natural images for estimating the noise. Specifically, we estimate the noise variance from the wavelet coefficients using a robust median estimator [10] by:

$$\sigma_n^2 = \frac{median(|Y|)}{0.6745} \tag{6}$$

where Y represents the wavelet coefficients at three first-level detail subbands including the horizontal subband LH_1, the vertical subband HL_1, and the diagonal subband HH_1. Here, $Y = \{y_1, y_2, y_3\}$ with y_1 being the wavelet coefficients at LH_1, y_2 being the wavelet coefficients at HL_1, and y_3 being the wavelet coefficients at HH_1. $|\ |$ gets the absolute value for each coefficient in Y. Using all of highest frequency subbands [5] makes the noise estimation adaptive to different subband characteristics. In our experiments, we use the 'db2' from the Daubechies wavelet family for a faster estimation. We classify input images into three types of noise severity based on the σ_n^2 value. Table 1 lists the empirically determined threshold values for σ_n^2 to determine the level of noise severity for an image.

For images with low noise severity, we do not need to estimate the noise statistics and directly set the noise energy value T as 0; for images with medium noise severity, we use the MAD criterion to estimate the Rayleigh distribution

Table 1 Classifying noise severity in images based on the variance value

σ_n^2 value	Noise severity
$\sigma_n^2 < 2$	Low
$2 \leq \sigma_n^2 < 5.5$	Medium
$\sigma_n^2 \geq 5.5$	High

Algorithm 2 Estimating the noise energy value T for PC

```
Input: grayscale image $gI$.
Output: the estimated noise energy value $T$.
Apply a 1-level 'db2' wavelet decomposition on $gI$
Extract the first-level subbands $LL_1$, $HL_1$, $LH_1$, and $HH_1$
Estimate the noise variance $\sigma_n^2$ using (6)
if $\sigma_n^2 < 2$ then $T \leftarrow 0$
if $2 \leq \sigma_n^2 < 5.5$ then use Median Absolute Deviation (MAD) mode estimator to calculate $T$
if $\sigma_n^2 \geq 5.5$ then use histogram mode estimator to calculate $T$
```

mode; for images with high noise severity, we use the histogram of the image to estimate the Rayleigh distribution mode. The histogram method serves as a more accurate way of estimating the noise statistics for more complex noise contamination. We summarize our noise estimation method in Algorithm 2.

2.1.4 Extracting image keypoints

To identify the keypoints in an image, we construct a moment map in which the larger moments encapsulate the corner strength information. In other words, the larger the moments, the higher strength of the corners. To this end, PC proceeds with a moment analysis as calculated in (7), (8), and (9).

$$a(x) = \sum_o \big(PC_o(x)\cos(\theta_o)\big)^2 \tag{7}$$

$$b(x) = 2\sum_o \big(PC_o(x)\cos(\theta_o)\big) \times \sum \big(PC(\theta_o)\sin(\theta_o)\big) \tag{8}$$

$$c(x) = \sum_o \big(PC_o(x)\sin(\theta_o)\big) \tag{9}$$

where $PC_o(x)$ is the phase congruency value at orientation o, θ_o is the axis angle of oth orientation, and $a(x)$, $b(x)$, and $c(x)$ are three kinds of second moments at location x. The PC operator utilizes these moments to construct the moment value at location x in a map m (i.e., $m(x)$):

$$m(x) = \frac{1}{2}\big(c(x) + a(x) - \sqrt{b(x)^2 + (a(x) - c(x))^2}\big) \tag{10}$$

In the proposed method, we exploit m to extract the corners in both RGB and NIR images. Specifically, the 1200 corners with the most strength (i.e., the largest 1200 values) in the map are chosen as the keypoints among the candidate corners for each image in a pair and are passed to the next module.

2.2 Keypoint feature description

We use a descriptor insensitive to illumination changes to represent features at keypoints due to the nonlinear intensity variation between the cross-spectral RGB-NIR image pair. We choose the Log-Gabor Histogram Descriptor (LGHD) [2], which is a distribution-based descriptor relying on high-frequency components to describe features around each keypoint. LGHD is a more robust candidate for our desired application than other state-of-the-art descriptors such as SIFT and PCEHD. LGHD uses the Log-Gabor filters in different scales and orientations to build a histogram as summarized in Algorithm 3.

Algorithm 3 LGHD feature descriptor

```
Input: grayscale image $gI$, its keypoints $P^*_{gI}$, its filtered phase congruency result $PC_{gI}$, number of scales $S$, number of orientations $O$, and patch size $M$
Output: LGHD features $f_{gI}$
for every keypoint $p_i^*$ in $P^*_{gI}$ ($i \in \{1, \ldots, n\}$) do
    Locate a patch $R_i$ of size $M$ around $p_i^*$
    Divide $R_i$ into 16 smaller subregions $sR_i$
    for each $sR_{i,j}$ in $sR_i$ ($j \in \{1, \ldots, 16\}$) do
        Calculate its corresponding oriented histogram
        $h_{i,j}$ at $O$ orientations in $PC_{gI}$
    end for
    Concatenate $h_{i,j}$ in all $S$ scales to obtain a feature
    vector $f_i$ of size $S \times O \times 16$
end for
Organize $\{f_1, f_2, \ldots f_n\}$ in a set to form LGHD features $f_{gI}$
```

Since PC itself uses LGHD, we combine the keypoint extraction and feature description module into a Single Instance Phase Congruency Feature Extraction module (SIPCFE) as one of our contributions. In other words, the Log-Gabor filter bank responses saved in the previous section (e.g., PC_{gI}), which are the results obtained by lines 1-7 of Algorithm 1 using 4 scales and 8 orientations, are directly used to extract LGHD features.

We use a patch of a pre-determined size around each keypoint to compute its histogram of oriented Log-Gabor filters in PC_{gI}. To this end, we divide each patch into 16 smaller subregions. For each subregion, we compute its corresponding magnitudes in PC_{gI} and determine the dominant orientation at each of its locations based on the maximum magnitude at all orientations. We then concatenate the histogram of the dominant orientations in each subregion in all scales to obtain a feature vector of size ($S \times O \times 16$). We observe that larger patches would allow us to consider possibly more informative descriptors, but at the same time, they would be more susceptible to occlusions and slower to compute. As a result, we empirically choose the patch size as 50×50 (i.e., $M = 50$) in our proposed method. At the end of this step, the RGB image has a set of feature vectors denoted as f_{RGB} to represent the characteristics of each keypoint and the NIR image has another set of feature vectors denoted as f_{NIR} to represent the characteristics of each keypoint.

2.3 Keypoint feature matching

As discussed in the previous section, we represent each keypoint by a feature vector of values. The corresponding feature points between the RGB-NIR pair are identified using an exhaustive matching method. Two keypoints match if their sum of absolute differences in their feature descriptor in all 512 dimensions is less than a certain threshold. This exhaustive matching method ensures that all potentially matching keypoints are uniquely identified and saved in a set of putative matching points. This exhaustive method, however, leaves us with outliers, which need to be removed to make our transformation estimation more accurate. Maintaining a robust set of corresponding points from a putative set of matched points is an essential step in the registration task. Before estimating a geometric transformation, one has to make sure that the estimation is done on a clean set of matched keypoints without outliers or a set of matched keypoints with a few outliers. Classic Sample Consensus (SAC) algorithms such as RANSAC or MSAC are highly sensitive to the proportion of outliers.

Moreover, they cannot handle non-rigid (non-parametric) correspondences. For our task, we adopt the idea of VFC algorithm [25] to represent the matching points by motion field samples and take advantage of the Expectation Maximization (EM) algorithm [9] to detect inliers and remove outliers. If the observed 2D sets of matched points are $\mathbf{P_m} = (x_m, y_m)^T$ and $\mathbf{P_r} = (x_r, y_r)^T$ with $\mathbf{P_m}$ representing the set of keypoints in the moving image and $\mathbf{P_r}$ representing the set of keypoints in the reference image, the motion field vector for each pair of matched points is:

$$\mathbf{v} = (s_n, t_n), \quad s_n = \mathbf{P_m}, \quad t_n = \mathbf{P_r} - \mathbf{P_m} \tag{11}$$

where s_n is the vector's starting point and t_n is the vector's terminal point. Next, we define the motion field set as:

$$S = \{(s_n, t_n) : n \in \mathbb{N}\} \tag{12}$$

The goal is to fit a mapping field function $\mathbf{f}$ so that:

$$t_n = f(s_n) \tag{13}$$

The robust estimation of $\mathbf{f}$ is obtained when there are no outliers present in the data. By assuming a Gaussian noise with zero mean, an arbitrary uniform standard deviation for the inliers, and a uniform distribution for the outliers, VFC employs the EM algorithm to estimate a set of parameters containing the distribution parameters θ and the mapping field function $\mathbf{f}$, where a slow-and-smooth [45] prior, which is generalized to a broad range of phenomena, on $\mathbf{f}$ is assumed. Maximum A Posterior (MAP) then estimates the optimal solution for θ by minimizing the energy. The EM algorithm estimates a posterior probability for each vector by updating the distribution parameters until convergence (i.e., reaching the desired minimum energy). At each iteration of EM, the solution for $\mathbf{f}$ is obtained via Tikhonov regularization [42] in a vector-valued Reproducing Kernel Hilbert Space (RKHS) [3]. The final solution enforces closeness of $\mathbf{f}$ to the inliers and maintains smoothness on the vector field of $\mathbf{f}$. Vectors with the posterior probability lower than a certain threshold (e.g., 0.75) are considered to be outliers.

2.4 Transformation estimation

Given a reference image r and a moving image m, the goal of image registration is to find a transformation function, $\mathbf{H} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, which maps all the pixels in the moving image to their corresponding pixels in the reference image. Here, d denotes the dimension of the data, which in our case is 2D (i.e., the x and y coordinates of matched keypoints). At this step, we aim to find the optimal projective transformation matrix to map all the matched points in the moving image to their corresponding matched points in the reference image. The projective transformation is suitable for our application since it does not preserve parallelism, length, and angle compared to the affine transformation. Specifically, we utilize the DLT algorithm [15] to estimate the projective transformation $\mathbf{H}$. We also constrain DLT to require at least 8 matched points instead of 4 matched points to increase its robustness to estimate $\mathbf{H}$. Hence, the task of registration is tagged as a *failure* if the keypoint feature matching method cannot identify at least 8 robust matched points. We denote the 2D inlier set of matched keypoints in the moving image as $\mathbf{P_m} = (x_m, y_m, 1)^T$, where $m = 1, 2, \ldots, N$, and N is the number of matched points with $N \geq 8$. Similarly, we denote the 2D inlier set of matched keypoints in the reference NIR image as $\mathbf{P_r} = (x_r, y_r, 1)^T$, where $r = 1, 2, \ldots, N$. The transformation equation is denoted as $\mathbf{P_r} = \mathbf{HP_m}$. The right-hand side of this equation is written as:

$$\mathbf{HP_m} = \begin{bmatrix} \mathbf{h}^{1T}\mathbf{P_m} \\ \mathbf{h}^{2T}\mathbf{P_m} \\ \mathbf{h}^{3T}\mathbf{P_m} \end{bmatrix} \tag{14}$$

where $\mathbf{h}^{jT}$ is the j-th row of the matrix $\mathbf{H}$. If we rewrite the equation $\mathbf{P_r} = \mathbf{HP_m}$ in the form of the vector cross product as $\mathbf{P_r} \times \mathbf{HP_m} = 0$, then:

$$\mathbf{P_r} \times \mathbf{HP_m} = \begin{bmatrix} y_r \mathbf{h}^{3T} x_m - \mathbf{h}^{2T} x_m \\ \mathbf{h}^{1T} x_m - x_r \mathbf{h}^{3T} x_m \\ x_r \mathbf{h}^{2T} x_m - y_r \mathbf{h}^{1T} x_m \end{bmatrix} \tag{15}$$

Since $\mathbf{h}^{jT} x_m = x_m^T \mathbf{h}^j$ for $j = 1, 2, 3$, we have three sets of equations in the form:

$$\begin{bmatrix} 0^T & -x_m^T & y_r x_m^T \\ x_m^T & 0^T & -x_r x_m^T \\ -y_r x_m^T & x_r x_m^T & 0^T \end{bmatrix} \times \begin{bmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{bmatrix} = 0 \tag{16}$$

which can be rewritten in the form of $\mathbf{A_i h} = 0$. To solve this, we need the first two rows since they are the only two independent linear equations. The n 2×9 matrices $\mathbf{A_i}$ are assembled into a single $2n \times 9$ matrix $\mathbf{A}$. The Singular Value Decomposition (SVD) of $\mathbf{A}$ is obtained so that $\mathbf{A} = \mathbf{UDV}^T$. If $\mathbf{h}$ denotes the last column of $\mathbf{V}$, it is a 9-value vector consisting of the entries of the matrix $\mathbf{H}$. In other words,

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \tag{17}$$

Using all pairs of the matched points between the RGB and NIR images, DLT estimates all 9 elements of the $\mathbf{H}$ matrix, which is further applied to all the pixels in the RGB image to register the RGB image onto the NIR image.

3 Experimental results

In this section, we discuss our experiments and the results to further evaluate the proposed image registration method and the state-of-the-art image registration methods. We test these image registration methods on three datasets. The first dataset contains 12 pairs of collected RGB-NIR remote sensing images with a resolution of 563 × 451 pixels, courtesy of Utah Water Research Laboratory (UWRL). This dataset is not aligned. To evaluate the registration methods, we have manually labeled at least 10 corresponding points between the RGB image and its NIR pair. Some sample images from this dataset are shown in Fig. 4.

The Computer Vision Lab at EPFL provides the second dataset [6], which includes 477 pairs of RGB-NIR images. The image pairs are categorized into nine different types of scenes: *country*, *field*, *forest*, *indoor*, *mountain*, *old building*, *street*, *urban*, and *water*. Each category contains at least 50 image pairs. To facilitate the evaluation, researchers have already aligned the EPFL dataset roughly for each pair. We show sample pairs from this dataset in Fig. 5.

We use the second dataset to create a subset of the EPFL dataset with rotated reference images (i.e., NIR images). Specifically, we randomly choose five images from each of nine categories of the EPFL dataset and apply random rotations in the angle range [1°, 45°] to the NIR images. Here, large angles are not employed because they are not commonly used in practice.

We have compared the proposed approach with different combinations of keypoint extractors and keypoint descriptors (denoted as keypoint extractor + cross-spectral descriptor), which are promised to deliver notable results under illumination varied applications. Two powerful keypoint extractors, namely, SIFT and PC, have been chosen for our benchmark. We describe the features at each keypoint using the four most commonly used cross-spectral descriptors such as LGHD, SIFT, Eight Local Directional Patterns (ELDP) [11], and Phase Congruency and Edge Histogram Descriptor (PCE-HD) [29]. Keypoint extractors such as FAST and MinEigen

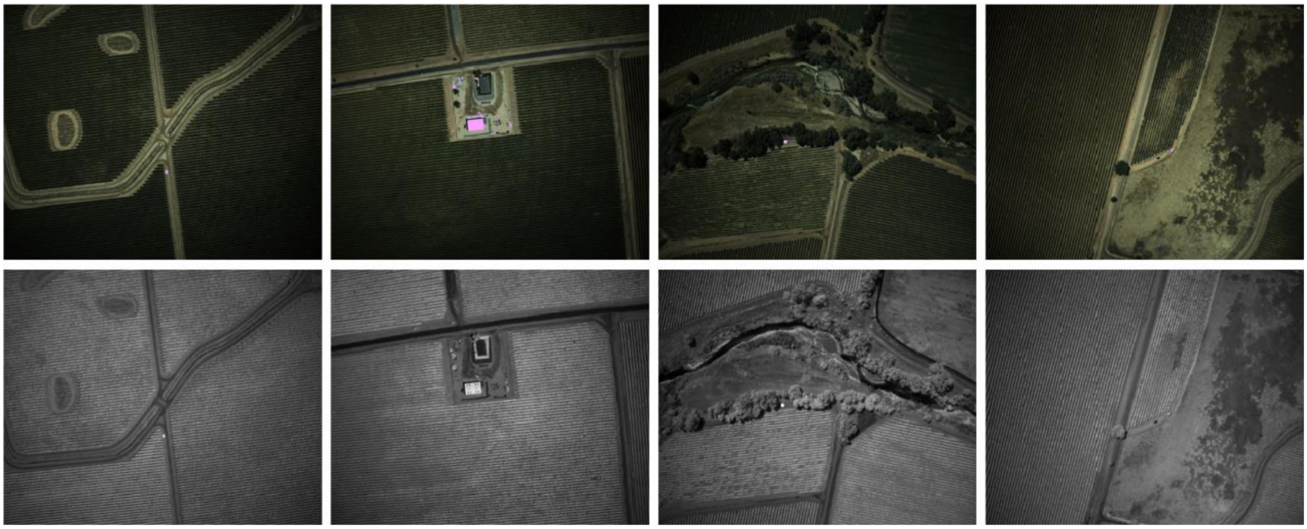


Fig. 4 Sample RGB-NIR image pairs from the UWRL dataset (top row: RGB images; bottom row: NIR images)

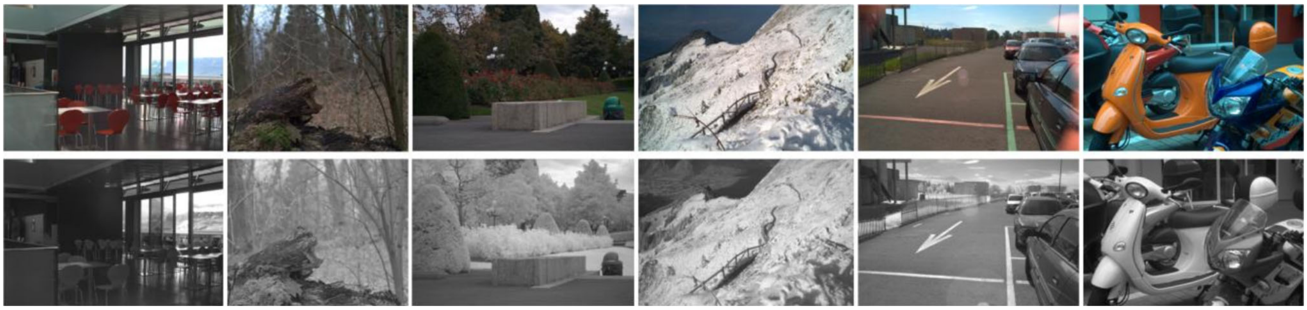


Fig. 5 Sample RGB-NIR image pairs from the EPFL dataset (top row: RGB images; bottom row: NIR images)

are powerful ones for broader types of images. However, they do not provide a sufficient number of robust and reliable keypoints on the three datasets. Hence, we exclude these two methods from our experiments. In addition, we extract the SIFT descriptors around each keypoint identified by SIFT since they can be easily extracted from the SIFT keypoint extraction process. However, extracting the SIFT descriptors for other keypoint extractors is difficult and we could not find any reliable online source code to do this. We use the following naming conventions {keypoint extractor + cross-spectral descriptors} to build the benchmark for six state-of-the-art image registration methods including {SIFT + LGHD}, {SIFT + SIFT}, {SIFT + ELDP}, {SIFT + PCEHD}, {PC + ELDP}, {PC + PCEHD}, and the proposed SIPCFE (i.e., an efficient version of {PC + LGHD}). It should be noted that the six state-of-the-art image registration methods use the same sequence of outlier removal process as proposed for SIPCFE (i.e., the exhaustive matching method followed by VFC) to find the reliably matched keypoints. The registration method {PC + SIFT} is not included in the comparison. We run our benchmarks on a 3.4GHz Intel Core i7 machine with 16GBs of RAM.

3.1 Evaluation measure

Let $H = (\mathbf{H}_x, \mathbf{H}_y)$ denote the transformation function to register the RGB image onto the NIR image, where H_x is the transformation at the x coordinate and $\mathbf{H}_y$ is the transformation at the y coordinate. Then, any point $\mathbf{p_m} = (x_m, y_m)$ in the moving image (RGB) has a relationship with its corresponding point $\mathbf{p_r} = (x_r, y_r)$ in the reference image (NIR) as follows:

$$\mathbf{p_r} = \mathbf{Hp_m} \tag{18}$$

Since the images are in different spectrums, we cannot use the intensity of registered points to evaluate the registration performance. Instead, we use the RMSE to evaluate the accuracy of the estimated transformation $\mathbf{H}$. Since the image pairs in the EPFL dataset are pre-aligned, we use $\mathbf{H}$ to register the inlier keypoints extracted from the RGB image to be aligned with their matching points in the NIR image. We then compute RMSE based on the number of pixels that the registered points shift away from their original locations in the NIR image. Specifically, if $\mathbf{P_i}$ is the set of matched points in the reference image and the set of their corresponding matched points in the moving image after employing the transformation is denoted by $\mathbf{P_j} = \mathbf{HP_i}$, RMSE is calculated by:

$$\text{RMSE} = \sqrt{||\mathbf{P_i} - \mathbf{P_j}||^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||\mathbf{p}_{r_i} - \mathbf{Hp}_{m_i}||^2} \tag{19}$$

where N is the number of matched points and i is the index of a pair of matched points in both the reference image and the moving image. Smaller values of RMSE represent a more accurate transformation from RGB to NIR. In literature, an RMSE of below 5 pixels is usually considered to be a fair error [47].

3.2 Results on the UWRL dataset

3.2.1 Effectiveness of the keypoints

To evaluate the accuracy of the proposed SIPCFE, we apply the estimated transformation on the manually labeled matched points in the RGB image and compute RMSE based on the registered RGB points and their corresponding labeled points in the NIR image. Figure 6 summarizes the RMSE performance on the UWRL dataset. As illustrated, SIPCFE outperforms the other methods with the smallest RMSE of 1.72 pixels. This is about 51.96% accuracy improvement compared to the best combination of SIFT-based registration method, namely, {SIFT + LGHD}, and 47.24% accuracy improvement compared to the overall second best method, namely, {PC + PCEHD}. Furthermore, we can see that only one variant of SIFT-based registration method yields an RMSE of 5 pixels or smaller. These results demonstrate that the PC keypoint extractor can extract more reliable and robust keypoints than the SIFT keypoint extractor since the PC keypoint extractor combined with any cross-spectral descriptor

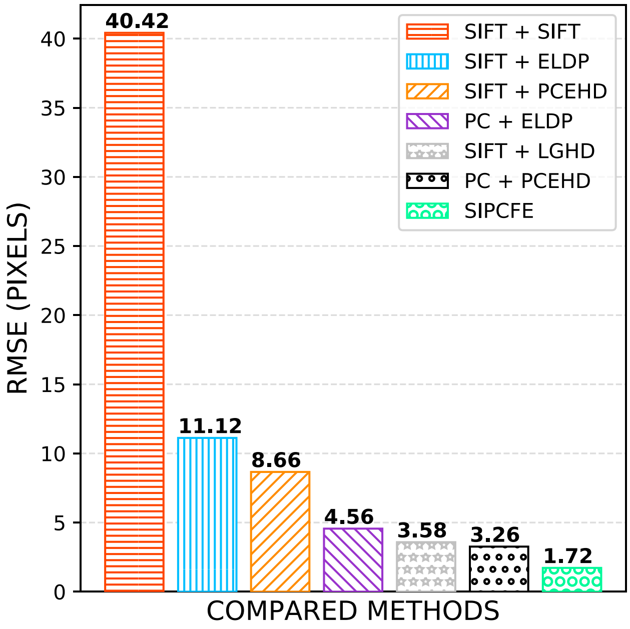


Fig. 6 Comparison of image registration results of different keypoint extraction and keypoint feature descriptors in terms of RMSE on the UWRL dataset

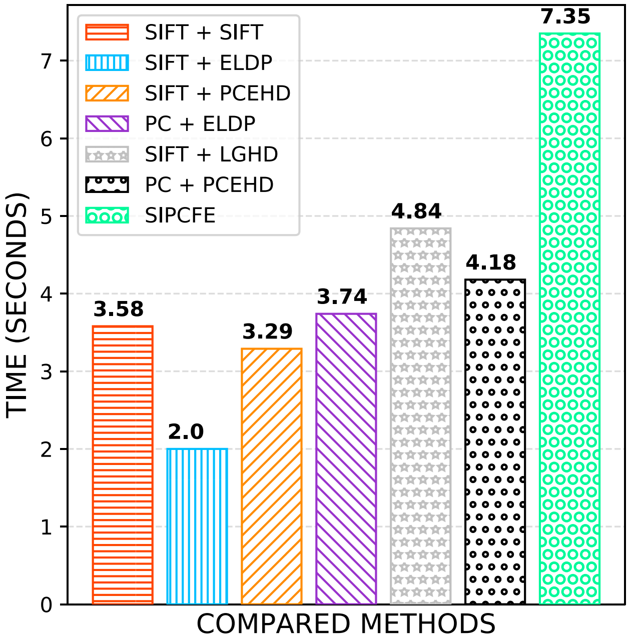


Fig. 7 Comparison of image registration results of different keypoint extraction and keypoint feature descriptors in terms of running times on the UWRL dataset

always achieves better registration results in terms of RMSE than the SIFT keypoint extractor combined with the same cross-spectral descriptor. In addition, the LGHD descriptor achieves the best discriminative power to generate more accurate matching pairs of keypoints since both PC and SIFT keypoint extractors combined with the LGHD descriptor achieves better registration results in terms of RMSE when compared with the same keypoint extractor with three other descriptors such as PCEHD, ELDP, and SIFT. In other words, our choice of the PC keypoint extractor and the LGHD cross-spectral descriptor is the best among the considered keypoint extractors and cross-spectral descriptors.

The average running times of the compared algorithms for 12 images in the UWRL dataset are summarized in Fig. 7. SIPCFE is not the fastest method compared to the other methods since the PC keypoint extractor extracts significantly more keypoints on this dataset than the SIFT keypoint extractor. For instance, SIPCFE extracts 960 keypoints on average on each image, while {SIFT + LGHD} extracts 447 keypoints on average. Evidently, SIPCFE processes more than twice as many keypoints as {SIFT + LGHD} does, which leads to the slower running time for PC-based image registration methods. However, SIPCFE (e.g., {PC + LGHD}) is only 1.5 times slower than the best SIFT-based registration method (e.g., {SIFT + LGHD}). This is mainly because VFC leaves SIPCFE with 131 inlier matched keypoints and leaves {SIFT + LGHD} with 142 inlier matched keypoints on average. This suggests that more outlier keypoints are removed from the set of putative matched points for SIPCFE. In general, SIPCFE tends to be slightly slower than other methods, but it delivers the least error in terms of RMSE by extracting more accurate inliers. Overall, SIPCFE performs the best on this benchmark dataset.

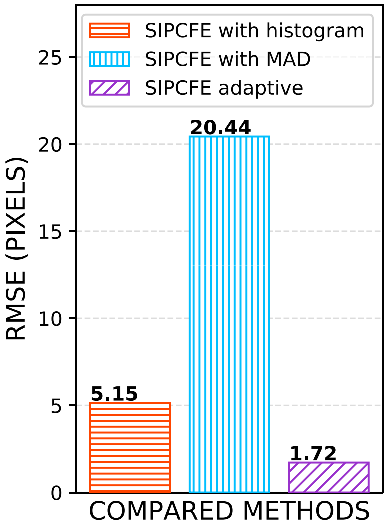


Fig. 8 Comparison of image registration results of the proposed SIPCFE and its variant systems using the two noise compensation methods in terms of RMSE on the UWRL dataset

3.2.2 Effectiveness of adaptive noise

We compare the performance of the proposed SIPCFE with its variant systems, where one of the three options of noise

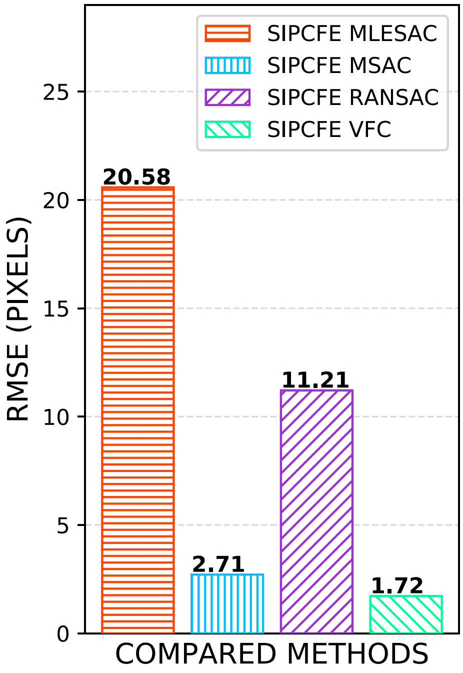


Fig. 9 Comparison of image registration results of the proposed SIPCFE and its variant methods using different outlier rejection methods in terms of RMSE on the UWRL dataset

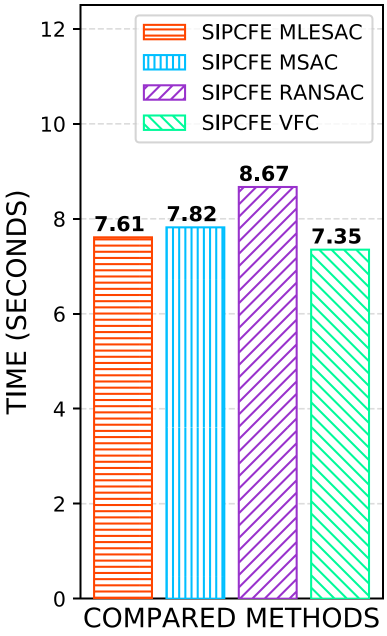


Fig. 10 Comparison of image registration results of different outlier rejection methods in terms of running time on UWRL dataset

compensation (e.g., MAD, histogram, and $T = 0$) is used in the PC function to alleviate noise effect. Figure 8 summarizes the performance of the algorithm on the UWRL dataset when using different noise compensation methods. It should be noted that the adaptive noise method treats all the images as clean (i.e., all the images are contaminated with a low level of noise). Therefore, for the UWRL dataset, the noise option of $T = 0$ achieves the same registration results as running the algorithm adaptively and we only present the results for running the proposed SIPCFE in Fig. 8. It is clear that the proposed SIPCFE outperforms its variant systems with the PC function with options of MAD and histogram. On the other hand, the variant systems with both options result in an average RMSE of over 5 pixels, which is undesirable. Specifically, the variant system with the option of MAD leads to a significant large RMSE value of 20.44 pixels. These results demonstrate the effectiveness of the proposed adaptive noise compensation method. They also show the importance of determining the correct noise level and employing the correct noise compensation value T when using PC to locate the reliable and robust keypoints.

3.2.3 Effectiveness of VFC

We compare the RMSE performance of the proposed SIPCFE method against its variant methods using Classic Sample Consensus (SAC) methods for outlier rejection. Specifically, we consider the widely used Random Sample Consensus (RANSAC) method and its derivations such as M-Estimator SAC (MSAC) and Maximum Likelihood SAC (MLESAC). These classic SAC methods generate a hypothesis (estimation) from random samples and verify it to the data for a certain number of iterations. At each iteration, a new random set of samples is chosen to update the estimation parameters. Verification is done by penalizing the outliers and estimating the best fitting function based on the inliers. Different SAC methods have different penalizing procedures. This mechanism makes these methods sensitive to the portion of outliers and any nonlinearity among inliers, which ultimately degrades the performance of the algorithm. Figure 9 compares the image registration results for the proposed SIPCFE and its variant methods using different SAC algorithms. It clearly shows that VFC outperforms the other classic outlier rejection methods and achieves the best performance for our application, in which nonlinearity (non-rigidness) might be present among inliers, or the portion of outliers might be significant. Among the SAC methods, MLESAC and RANSAC produce an undesirable RMSE of more than 10 pixels on average.

Figure 10 compares the running time of outlier rejection methods on the UWRL dataset. VFC is slightly faster than the other SAC methods. Specifically, on average, it takes the proposed SIPCFE around 7.35 seconds to register a pair of RGB-NIR images and its variant system using MLESAC as outlier removal around 7.61 seconds to register a pair of RGB-NIR images.

3.3 Results on the EPFL dataset

3.3.1 Effectiveness of the keypoints

The EPFL dataset contains nine categories of different natural scenes with at least 50 images in each category. To run our benchmark, we test each algorithm separately on each category. However, one blurry image in the *forest* category of the EPFL dataset makes SIPCFE and some other approaches either fail to register or yield extremely large RMSEs. Figure 11 (top row) shows this blurry RGB and NIR image pair. Figure 11 (middle row) shows the putative set of 102 matched points extracted by the exhaustive matching method proposed in SIPCFE. A lot of them are outliers, which accounts for 78% of the points in the set. Figure 11 (bottom row) shows the 29 matched points after removing outliers by VFC. It is clear that there are still some outliers remaining after the 2-step outlier removal process. These outliers are marked by red lines, as shown in the bottom row of Fig. 11. This is where both the proposed PC keypoint extractor, the LGHD feature descriptor, and VFC fall short as a combined module. When the sample set is small (i.e., the set of putative matched points is small), the performance of VFC's Tikhonov regularization degrades as the number of outliers increases. SIPCFE is not able to find enough matched points to feed into VFC. This situation also happens to all the methods that use PC as the keypoint extractor, namely {PC + ELDP} and {PC + PCEHD}.

Figure 12 illustrates the intermediate results of the two-stage keypoint feature matching for the same blurry image processed with {SIFT + LGHD}. The SIFT keypoint extractor is able to find a significant number of keypoints in both RGB and NIR images. Out of these keypoints, 1883 putative matched points as shown in the top row of Fig. 12 are found by the exhaustive matching method; since there are a lot of matched points, we have only highlighted a few of wrongly matched points. These matched points are further fed into VFC to remove 50 points as outliers. We show the resultant 1833 points in the bottom row of Fig. 12 formed a clean set of matched points to be fed into the DLT to find the transformation matrix for the registration task. To make our comparison benchmark feasible, we report the results for the *forest* category, with and without the blurry image.

Tables 2 and 3 summarize the RMSE and runtime performance of the proposed SIPCFE and six state-of-the-art registration methods on the EPFL dataset for each category, respectively. Since PC-based image registration methods are not able to find either enough keypoints or a robust set of matched keypoints for the blurry image pair as shown in the first row of Fig. 11, we deliberately remove this pair from the *forest* category to report the registration performance in terms of RMSE in pixels in Table 2. To make the comparison complete, we also present the RMSE results for all images in the nine categories of the EPFL database in Table 2 by showing the RMSE results in parentheses for the *forest* category and average RMSE results in parentheses for all images. Overall, the proposed SIPCFE method outperforms the other methods by an average of 2.28 pixels in RMSE with the smallest standard deviation of 1.84 for all the images except for the blurry image in the *forest* category. It should be noted that the smaller the standard deviation,

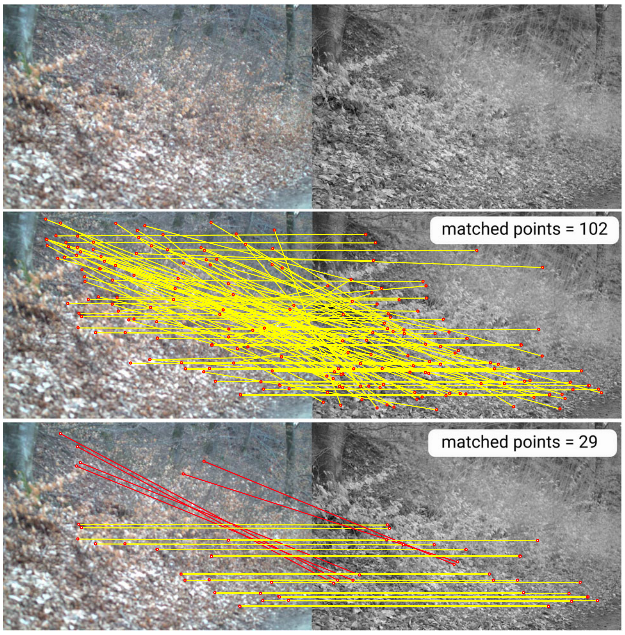


Fig. 11 The intermediate registration results for a sample blurry image from the *forest* category processed with SIPCFE. (top row: the sample blurry RGB-NIR image pair; middle row: the set of putative matched points using the exhaustive matching method of the SIPCFE; bottom row: the set of putative matched points cleaned with VFC of the SIPCFE)

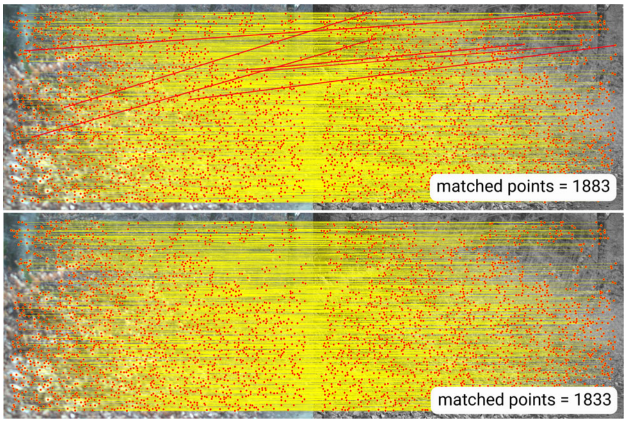


Fig. 12 The intermediate registration results for a sample blurry image from the *forest* category processed with {SIFT + LGHD}. (top row: the set of putative matched points extracted using the exhaustive matching method of the {SIFT + LGHD}; bottom row: the set of putative matched points cleaned with VFC of the {SIFT + LGHD})

the more robust the image registration is. As a result, we can safely say that SIPCFE is the most accurate and robust registration method among the compared methods. Specifically, this renders as a 14.29% accuracy improvement compared to the second best method {SIFT + LGHD}. Even with the blurry image included, SIPCFE slightly performs better than {SIFT + LGHD} with an accuracy improvement of 0.376%. Additionally, SIPCFE delivers an RMSE of below 5 pixels across all categories except the *water* category, while the second best method {SIFT + LGHD} delivers an RMSE of below 5 pixels across all categories except for the *mountain* category. Compared to {SIFT + LGHD}, SIPCFE significantly improves the registration performance for images in *mountain* and *old building* categories, slightly improves the registration performance for images in *indoor*, *street*, and *urban* categories, and achieves comparable registration performance for images in the other categories. It also achieves the best accuracy in terms of RMSE in *street* and *old building* categories among all the 7 compared methods. This suggests that SIPCFE performs the best in scenes with a lot of variety and corners, which are the features commonly seen in buildings and vehicles. The images in the *water* category seem to be the most challenging to register. This is mainly due to the homogeneous texture of water, which makes it hard for the keypoint extractors to find distinctive keypoints.

Table 3 shows that the proposed SIPCFE is on average 2 times faster than the second best registration method {SIFT + LGHD}; it runs faster than {SIFT + LGHD} for all nine categories and significantly improves the processing runtime by 2.6 times for the *forest* category, which takes the longest runtime for {SIFT + LGHD} to register. In addition, the runtime of SIPCFE has the third smallest standard deviation among all the compared methods and has a significantly smaller standard deviation compared to {SIFT + LGHD}. This proves that the running time of SIPCFE is scene-invariant and does not fluctuate too much. Our experiments also confirm that SIPCFE extracts almost the same number of features across all categories in EPFL. Other methods such as {SIFT + SIFT}, {SIFT + ELDP}, {PC + ELDP}, and {PC + PCEHD} are faster than our method. However, they do not deliver good accuracy across the categories. For example, the fastest registration method {SIFT + SIFT} achieves the second worst accuracy with an RMSE of 3.93 pixels. The second fast registration method {PC + ELDP} achieves the third worst accuracy with an RMSE of 3.84 pixels. This makes SIPCFE the best candidate from the perspectives of both accuracy and speed.

3.3.2 Effectiveness of the adaptive noise

Table 4 lists the RMSE on the EPFL dataset for the proposed SIPCFE and its variant methods, where one of the three options of noise compensation (e.g., MAD, histogram,

Table 2 The RMSE (in pixels) performance of the seven compared image registration methods on the EPFL dataset, where the RMSE results shown in the parenthesis of the *forest* category and the average listed in the last column are obtained by including the blurry image pair

METHOD	Category									Average ± std
	COUNTRY	FIELD	FOREST	INDOOR	MOUNTAIN	OLD BUILDING	STREET	URBAN	WATER	
SIFT+LGHD	**2.48**	**4.75**	0.94	0.68	5.69	3.13	1.78	0.50	**4.03**	2.66 ± 1.87
SIFT+SIFT	5.79	5.58	1.00	0.94	7.72	3.54	1.99	0.54	8.26	3.93 ± 3.00
SIFT+ELDP	3.74	5.50	0.90	1.02	5.69	3.12	1.62	0.54	11.96	3.79 ± 3.62
SIFT+PCEHD	3.84	5.66	**0.88**	0.87	3.92	2.65	1.96	0.48	8.46	3.19 ± 2.61
PC+ELDP	3.64	4.84	2.91 (failed)	0.68	3.63	1.18	2.28	0.37	15.06	3.84 ± 4.46 (failed)
PC+PCEHD	4.89	11.40	9.07 (failed)	**0.61**	**2.48**	1.25	2.81	**0.36**	9.57	4.72 ± 4.24 (failed)
SIPCFE	2.88	4.85	1.01 (4.36)	0.64	2.58	**1.16**	**1.53**	0.37	5.51	**2.28 ± 1.84 (2.65 ± 1.89)**
Average	3.89	6.08	2.38	0.77	4.53	2.29	1.99	0.45	8.97	

Bold-faced Category values are the top-performing methods with the lowest RMSEs in each category. Our proposed method and the best overall method are also bold-faced and $T = 0$) is used in the PC function. It clearly shows that the proposed SIPCFE leads to a better accuracy across all categories on average than its three variant systems. Specifically, the option of MAD mode estimator cannot handle all the images in the *water* category and is considered as a failed approach. The option of $T = 0$ delivers over 5 pixels RMSEs in *field* and *water* categories. The option of histogram mode estimator delivers a large 6.86 pixels RMSE in the *water* category, which is undesirable. The proposed SIPCFE has an average RMSE of over 5 pixels (e.g., 5.51 pixels) in the challenging *water* category.

3.3.3 Effectiveness of VFC

Table 5 lists the RMSE performance of the proposed SIPCFE method and its three variant methods, which use RANSAC, MSAC, and MLESAC to remove the outliers. For the convenience of comparison, we exclude the blurry image pair in the *forest* category from the experiments. It is clear that VFC achieves better performance than the other three SAC methods. With an average RMSE of 2.28 pixels, VFC improves the second best method (i.e., SIPCFE RANSAC) by 61.94%. In all categories except for the challenging *water* category, VFC achieves RMSEs of smaller than 5 pixels. The overall average RMSEs of the three variant methods are all over 5 pixels. The runtime performance of the proposed SIPCFE with different outlier rejection methods are tabulated in Table 6. VFC runs slightly slower than the classic SAC methods. From the perspectives of registration error and speed, the proposed SIPCFE with VFC as the outlier remover is the best candidate.

3.4 Results on the EPFL subset with rotated reference images

Table 7 summarizes the RMSE and runtime performance of the seven compared image registration methods on the rotated subset of the EPFL dataset, which contains 45 RGB-NIR (rotated) image pairs. The proposed SIPCFE method outperforms all other methods. It achieves the smallest RMSE of 2.32 pixels and the smallest standard deviation of 0.78 pixels. This is a 12.45% improvement in accuracy compared to the second best method (i.e., {SIFT + LGHD}). We can also observe that the two feature descriptors LGHD and SIFT offer comparable accuracy in the rotated subset as in the EPFL dataset when comparing the first two rows and the last row of Tables 7 and 2. The two other feature descriptors PCEHD and ELDP yield less accurate performance in the rotated subset. Therefore, we can safely say that SIFT and LGHD are more rotation-invariant in calculating robust features around keypoints. Lastly, the runtime performance follows almost the same trend as in Table 3, with SIPCFE being two times faster than the second best method.

Table 3 The runtime (in seconds) performance of the seven compared image registration methods on the EPFL dataset

METHOD	Category									Average ± std
	COUNTRY	FIELD	FOREST	INDOOR	MOUNTAIN	OLD BUILDING	STREET	URBAN	WATER	
SIFT+LGHD	20.33	19.99	23.44	13.12	19.34	16.95	17.76	15.68	18.06	18.30 ± 2.96
SIFT+SIFT	4.32	4.29	5.63	2.67	4.72	4.00	3.82	3.56	3.65	4.07 ± 0.82
SIFT+ELDP	8.31	8.12	9.84	5.27	8.11	7.16	7.26	6.63	7.03	7.53 ± 1.27
SIFT+PCEHD	11.87	11.81	13.60	8.33	11.95	10.34	10.48	9.28	10.21	10.87 ± 1.59
PC+ELDP	5.17	5.26	5.23	5.12	5.13	4.82	5.20	4.92	5.09	5.10 ± 0.14
PC+PCEHD	5.55	5.68	5.66	5.55	5.49	5.23	5.64	5.35	5.51	5.52 ± 0.15
SIPCFE	9.05	8.87	8.92	8.68	8.73	8.82	9.78	8.54	8.69	**8.90 ± 0.36**

Our proposed method and the corresponding run-time are bold-faced

Table 4 The RMSE (in pixels) performance of different noise compensation methods on the EPFL dataset

METHOD	Category									Average
	COUNTRY	FIELD	FOREST	INDOOR	MOUNTAIN	OLD BUILDING	STREET	URBAN	WATER	
SIPCFE with $T = 0$	2.92	5.12	1.13	0.64	2.61	1.16	1.44	0.37	5.37	2.29
SIPCFE with *MAD*	2.42	4.67	1.03 (4.39)	0.54	3.26	0.92	1.65	0.37	Failed	Failed
SIPCFE with *histogram*	2.4	4.91	1.01 (2.95)	0.53	2.15	0.95	1.63	0.37	6.86	2.31 (2.53)
SIPCFE adaptive	2.88	4.85	1.01 (4.36)	0.64	2.58	1.16	1.53	0.37	5.51	**2.28** (2.64)

Because the adaptive method is our proposed method, we have bold-faced it. Similarly the best method overall performance is bold-faced in the average column.

Table 5 The RMSE (in pixels) performance of different outlier rejection methods on the EPFL dataset

METHOD	Category									Average
	COUNTRY	FIELD	FOREST	INDOOR	MOUNTAIN	OLD BUILDING	STREET	URBAN	WATER	
SIPCFE RANSAC	7.52	13.53	2.62	1.91	6.42	3.45	3.51	1.26	13.72	5.99
SIPCFE MSAC	6.4	20.81	3.32	1.2	5.67	3.97	3.64	0.92	11.1	6.34
SIPCFE MLESAC	7.83	17.14	4.65	2.69	10.68	4.32	3.26	1.99	22.97	8.39
SIPCFE VFC	2.88	4.85	1.01	0.64	2.58	1.16	1.53	0.37	5.51	**2.28**

Table 6 The runtime (in seconds) performance of different outlier rejection methods on the EPFL dataset

METHOD	Category									Average
	COUNTRY	FIELD	FOREST	INDOOR	MOUNTAIN	OLD BUILDING	STREET	URBAN	WATER	
SIPCFE RANSAC	9.14	9.2	9.24	8.99	8.98	8.7	9.18	8.79	9.04	9.03
SIPCFE MSAC	8.97	9.01	9.07	8.82	8.81	8.48	8.99	8.58	8.87	8.84
SIPCFE MLESAC	8.95	9.01	9.07	8.8	8.9	8.78	9.3	8.89	9.24	8.99
SIPCFE VFC	9.46	9.27	9.7	9.16	9.44	9.04	9.44	9.04	9.27	9.31

Table 7 The RMSE (in pixels) and runtime (in seconds) performance of the seven compared image registration methods on the rotated EPFL dataset

METHOD	RMSE ± std	RUNTIME
SIFT+LGHD	2.65 ± 0.90	21.16
SIFT+SIFT	3.88 ± 2.40	2.34
SIFT+ELDP	4.15 ± 2.52	7.80
SIFT+PCEHD	3.22 ± 1.78	9.58
PC+ELDP	4.60 ± 2.85	4.42
PC+PCEHD	5.37 ± 4.60	4.81
SIPCFE	**2.32 ± 0.78**	10.13

4 Conclusions

In this paper, we propose an optimized feature-based approach called SIPCFE to quickly, reliably, and robustly register cross-spectral image pairs under different illumination conditions and rotations. Our major contributions include:

- Employing the PC method and its adaptive noise variant, which perform well under various illuminations, to identify reliable and robust keypoints that are invariant to intensity changes.
- Incorporating the Log-Gabor filter responses obtained from the keypoint extraction step to represent the characteristics around each keypoint using the histogram of the filter responses.
- Designing a sequence of outlier removal processes (i.e., exhaustive matching method followed by VFC) to find reliable matched keypoints accurately.
- Employing DLT to estimate the geometric transformation to align the image pair.
- Proposing the RMSE measure to evaluate the registration performance.

To evaluate the proposed method, we benchmark the proposed SIPCFE, its three variant methods incorporating different outlier removal algorithms, its three variant methods incorporating three different noise compensations, and six common feature-based approaches in the cross-spectral registration field on three datasets. For the remote sensing images in the first dataset from UWRL, SIPCFE achieves a 47.24% improvement in registration accuracy when comparing to the second best state-of-the-art method {PC + PCEHD}. The adaptive noise method proves to be working exceptionally better than any of the three noise compensation methods. For the second dataset from EPFL, SIPCFE achieves a 14.29% improvement in accuracy compared to the second best method in the benchmark (i.e., {SIFT + LGHD}). The adaptive noise method also proves to be a better approach for this dataset. VFC is the best candidate for the outlier rejection method for both datasets. Overall, SIPCFE outperforms other state-of-the-art feature-based methods that are developed for cross-spectral imagery from the perspectives of both accuracy and speed. However, SIPCFE cannot register a blurry image pair with good accuracy because the PC-based keypoint extractor cannot find either enough keypoints or a robust set of matched keypoints. For the third dataset, SIPCFE achieves a 12.45% improvement in accuracy compared to the second best method in the benchmark (i.e., $\{\text{SIFT} + \text{LGHD}\}$).

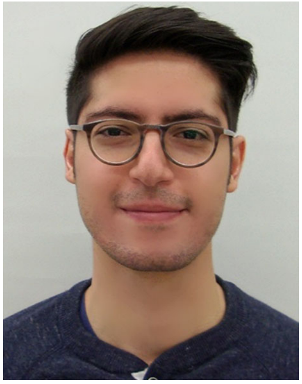
We will further improve the proposed method by exploring different parameters in the PC-based approach and considering more discriminative features to find better-matched keypoints.

References

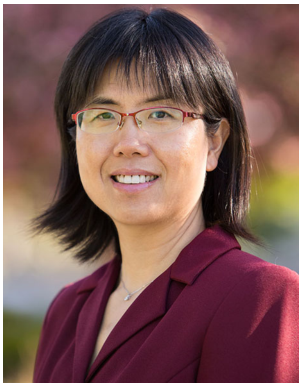
1. Aguilera, C., Barrera, F., Lumbreras, F., Sappa, A.D., Toledo, R.: Multispectral image feature points. Sensors **12**(9), 12661–12672 (2012)
2. Aguilera, C., Sappa, A.D., Toledo, R.: LGHD: A feature descriptor for matching across non-linear intensity variations. In: Proceedings of IEEE IEEE International Conference on Image Process (ICIP), pp. 178–181. IEEE (2015)
3. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**(3), 337–404 (1950)
4. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
5. Biswas, M., Om, H.: A new soft-thresholding image denoising method. Procedia Technol. **6**, 10–15 (2012)
6. Brown, M., Süsstrunk, S.: Multi-spectral SIFT for scene category recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 177–184. IEEE (2011)
7. Chen, M., Carass, A., Jog, A., Lee, J., Roy, S., Prince, J.L.: Cross contrast multi-channel image registration using image synthesis for MR brain images. Med. Image Anal. **36**, 2–14 (2017)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)
10. Donoho, D.L., Johnstone, J.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)
11. Faraji, M.R., Qi, X.: Face recognition under illumination variations based on eight local directional patterns. IET Biom. **4**(1), 10–17 (2015)
12. Firmenichy, D., Brown, M., Süsstrunk, S.: Multispectral interest points for RGB-NIR image registration. In: Proceedings of the IEEE International Conference on Image Processing, pp. 181–184. IEEE (2011)
13. Han, J., Pauwels, E.J., Zeeuw, P.D.: Visible and infrared image registration in man-made environments employing hybrid visual features. Pattern Recognit. Lett. **34**(1), 42–51 (2013)
14. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, pp. 10–5244. Manchester, UK (1988)
15. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)
16. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)
17. Hrkać, T., Kalafatić, Z., Krapac, J.: Infrared-visual image registration based on corners and hausdorff distance. In: Image Analysis, pp. 383–392 (2007)
18. Ikeda, K., Ino, F., Hagihara, K.: Efficient acceleration of mutual information computation for nonrigid registration using CUDA. IEEE J. Biomed. Health Inf. **18**(3), 956–968 (2014)
19. Kim, S., Min, D., Ham, B., Ryu, S., Do, M.N., Sohn, K.: DASC: dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2103–2112 (2015)
20. Kovesi, P.: Image features from phase congruency. Videre J. Comput. Vis. Res. **1**(3), 1–26 (1999)
21. Kovesi, P.: Phase congruency detects corners and edges. In: The Australian Pattern Recognition Society Conference, DICTA (2003)
22. Leutenegger, S., Chli, M., Siegwart, R.Y.: BRISK: Binary robust invariant scalable keypoints. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2548–2555. IEEE (2011)
23. Loeckx, D., Slagmolen, P., Maes, F., Vandermeulen, D., Suetens, P.: Nonrigid image registration using conditional mutual information. IEEE Trans. Med. Imaging **29**(1), 19–29 (2010)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
25. Ma, J., Zhao, J., Tian, J., Yuille, A.L., Tu, Z.: Robust point matching via vector field consensus. IEEE Trans. Image Proc. **23**(4), 1706–1721 (2014)
26. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging **16**(2), 187–198 (1997)
27. Mellor, M., Brady, M.: Phase mutual information as a similarity measure for registration. Med. Image Anal. **9**(4), 330–343 (2005)
28. Morrone, M.C., Ross, J., Burr, D.C., Owens, R.: Mach bands are phase dependent. Nature **324**(6094), 250–253 (1986)
29. Mouats, T., Aouf, N., Sappa, A.D., Aguilera, C., Toledo, R.: Multispectral stereo odometry. IEEE Trans. Intell. Transp. Syst. **16**(3), 1210–1224 (2015)
30. Pang, G., Neumann, U.: The Gixel array descriptor (GAD) for multimodal image matching. In: WACV, pp. 497–504 (2013)
31. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Image registration by maximization of combined mutual information and gradient information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 452–461. Springer (2000)
32. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. IEEE Trans. Med. Imaging **22**(8), 986–1004 (2003)
33. Qin, Y., Cao, Z., Zhuo, W., Yu, Z.: Robust key point descriptor for multi-spectral image matching. J. Syst. Eng. Electron. **25**(4), 681–687 (2014)
34. Rivaz, H., Karimaghaloo, Z., Collins, D.L.: Self-similarity weighted mutual information: a new nonrigid image registration metric. Med. Image Anal. **18**(2), 343–358 (2014)
35. Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. In: Proceedings of the 10th IEEE International Conference on Computer Vision, vol. 2, pp. 1508–1515. IEEE (2005)
36. Rosten, E., Porter, R., Drummond, T.: Faster and better: a machine learning approach to corner detection. IEEE Trans. Patt. Anal. Mach. Intell. **32**(1), 105–119 (2010)

37. Rueckert, D., Clarkson, M.J., Hill, D.L.G., Hawkes, D.J.: Non-rigid registration using higher-order mutual information. Proc. SPIE Med. Image Image Process **3979**, 439–447 (2000)
38. Sendur, L., Selesnick, I.W.: Bivariate shrinkage with local variance estimation. IEEE Signal Proc. Lett. **9**(12), 438–441 (2002)
39. Shi, J., Tomasi, C.: Good features to track. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600. IEEE (1994)
40. Smith, S.M., Brady, J.M.: SUSAN—a new approach to low level image processing. Int. J. Comput. Vis. **23**(1), 45–78 (1997)
41. Studholme, C., Drapaca, C., Iordanova, B., Cardenas, V.: Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change. IEEE Trans. Med. Imaging **25**(5), 626–639 (2006)
42. Tikhonov, A.N., Arsenin, V.Y., John, F.: Solutions of Ill-Posed Problems, vol. 14. Winston, Washington (1977)
43. Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. Int. J. Comput. Vis. **24**(2), 137–154 (1997)
44. Wachinger, C., Navab, N.: Entropy and Laplacian images: structural representations for multi-modal registration. Med. Image Anal. **16**(1), 1–17 (2012)
45. Weiss, Y., Adelson, E.H.: Slow and smooth: a Bayesian theory for the combination of local motion signals in human vision. MIT AI Lab Tech, Rep (1998)
46. Yang, X., Kwitt, R., Niethammer, M.: Quicksilver: fast predictive image registration—a deep learning approach. NeuroImage **158**, 378–396 (2017)
47. Zhao, C., Zhao, H., Lv, J., Sun, S., Li, B.: Multimodal image matching based on multimodality robust line segment descriptor. Neurocomputing **177**, 290–303 (2016)
48. Zhao, D., Yang, Y., Ji, Z., Hu, X.: Rapid multimodality registration based on MM-SURF. Neurocomputing **131**, 87–97 (2014)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Amir Hossein Farzaneh received the B.S. degree in Electrical Engineering from Shahid Beheshti University in 2013 and the M.S. degree in Electrical Engineering from Shahrood University of Technology in 2015. He is currently working toward the Ph.D. degree in Computer Science as a graduate research assistant in the Computer Science Department at Utah State University. He works as a research assistant at Computer Vision Laboratory of Utah State University. His research interests include computer vision, machine learning, and deep neural networks.



Xiaojun Qi received the B.S. degree in Computer Science from Donghua University in 1993, the M.S. degree in Computer Science from Shenyang Institute of Automation, the Chinese Academy of Sciences, in 1996, and the Ph.D. degree in Computer Science from Louisiana State University in 2001. In 2002, she joined the Department of Computer Science at Utah State University as a tenure-track Assistant Professor. In 2008, she received tenure and was promoted to Associate Professor. In 2015, she was promoted to Professor. She has been a senior IEEE member since 2010. She has expertise in artificial intelligence focusing on machine learning and computer vision. She has established and sustained a nationally recognized research program to solve challenging research problems. She has acquired competitive funding from reputable agencies and has published over 100 peer-reviewed scientifically rigorous and innovative papers in high-quality journals, book chapters, and conference proceedings. She has worked as a PI on more than 20 research projects and has supervised 82 students. She has also gained her professional leadership experience by serving on technical program committees, NSF panels, and as a reviewer.