Li *et al. EURASIP Journal on Advances in Signal Processing* [redacted]
DOI 10.1186/s13634-015-0240-z

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Robustly building keypoint mappings with global information on multispectral images

Yong Li*, Hongbin Jin, Wei Qiao, Jing Jing and Hang Yu

Abstract

This paper proposes an approach to robustly build keypoint mappings on multispectral images. The distinctiveness and repeatability of descriptors often decrease significantly on multispectral images and thus give unreliable keypoint mappings. To complement this decrease, global information over entire images is induced in this work to evaluate keypoint mappings. Initial keypoint mappings are established by utilizing descriptors. A pair of keypoint mappings determines a similarity transformation T, and then it is evaluated with the induced global information that is defined to be the similarity metric between the reference image and the transformed image by T. A process is utilized that iteratively considers the pairs of keypoint mappings and searches the best reference matched keypoint for every test keypoint. Experimental results show that the proposed approach can provide more reliable keypoint mappings than SIFT, ORB, FREAK, and ISS on multispectral images.

Keywords: Multispectral imaging; Keypoint mappings; Global information

1 Introduction

Multispectral imaging has been widely applied in a variety of applications such as monitoring of natural disaster and battlefield surveillance. The fusion of images taken by different spectral light can often provide more information about objects of interest and scenes than a single-spectrum light. A satisfying fusion usually requires image registration as the building block, and the registration performance has a great effect on the fusion quality.

1.1 Related work

Registering multispectral images has been a challenging problem due to the lack of explicit or implicit relationship between the values of corresponding pixels. In literature, there are two categories of registration methods, registration based on image features and registration based on image intensity [1]. Among intensity-based methods are mutual information [2], MIND [3], and maximum likelihood (ML) [4]. Let $I_r(x,y)$ and $I_t(x,y)$ denote the reference and test image. Intensity-based methods typically construct an objective/registration function $f(I_r(x,y), I_t^T(x,y))$ of the transformation parameter T between images. Then, the task of aligning $I_r(x,y)$ and $I_t(x,y)$ amounts to searching for the T at which $f(I_r(x,y), I_t^T(x,y))$ achieves the extremum.

The problem with intensity-based methods is that any optimization technique may fail to find the ground truth transformation parameters [5]. To improve the convergence of an optimization algorithm, the misalignment is often assumed to be small, e.g., several pixels. This assumption is equivalent to the following: an estimate $\tilde{T}$ of the ground truth can be obtained falling into the converging basin of $f(I_r(x,y), I_t^T(x,y))$, allowing for the optimization algorithm to achieve the global extremum. When the misalignment is relatively large, any optimization algorithm may easily be trapped in local extrema, ending with an unsuccessful registration.

Another category of intensity-based methods is Fourier methods. The translation of two images in spatial domain corresponds to the peak of the inverse Fourier transform of the product of two Fourier transformations. Tzimiropoulos et al. [6] propose a FFT-based approach to aligning scale-invariant images in which the log-polar Fourier is used to estimate the scaling and rotation. Pan et al. [7] propose multilayer fractional Fourier transform (MLFFT) to improve the accuracy of registering images with respect to both rotation and scaling. The problem with the Fourier methods lies in the difficulty that

*Correspondence: yli@bupt.edu.cn
School of Electronic Engineering, Beijing University of Posts and Telecommunications, Xitucheng Road, 100876, Beijing, China

translation, rotation, and scaling can not be dealt with simultaneously generally.

Other intensity-based techniques include region-based confidence weighted M-estimators [8] that deal with image sets with arbitrarily shaped local illumination variations caused by changes and movement of light sources. Zosso et al. [9] propose geodesic active fields that couple the registration term and regularization term. The energy of the deformation field is measured with the Polyakov energy weighted by a suitable image distance. Xing and Qiu [10] propose the using of nonparametric local smoothing to determine the underlying transformation, which does not need to assume that the mapping transformation has a certain type of parametric form. Liu et al. [11] propose mean local phase angle (MLPA) and frequency spread phase congruency (FSPC) using local frequency information to emphasize the common structural information while suppressing the sensor-dependent information.

Feature-based registration methods firstly build feature mappings and then compute the transformation parameters without resorting to any optimization techniques. In the past, a variety of image features such as keypoints have been proposed. Among commonly used features are keypoints and descriptors. Lowe [12] proposed the scale invariant feature transform (SIFT) detecting keypoints and descriptors invariant to scale and rotation. A main orientation is assigned to a keypoint, and the local gradient pattern with respect to the main orientation is computed as its descriptor. Bay et al. [13] proposed Speeded-Up Robust Features (SURF). SURF has the same repeatability and distinctiveness as SIFT but is computed faster than SIFT by employing integral images. Alahi et al. [14] propose Fast Retina Keypoint (FREAK). FREAK is a cascade of binary strings computed by comparing image intensities over a retinal sampling pattern. Ambai and Yoshida [15] propose compact and real-time descriptors (CARD). CARD can be computed rapidly by utilizing lookup tables to extract histograms of oriented gradients.

SIFT, SURF, FREAK, and CARD are suitable for monomodal images. To utilize descriptors for building keypoint mappings on multispectral images, partial intensity invariant feature descriptor (PIIFD) was proposed that adapted the gradient pattern to gradient and region reverse [16]. Saleem and Sablatnig [17] proposed using normalized gradients for computing descriptors to achieve robustness against intensity changes between multispectral images. Wang et al. [18] proposed modified sift feature extraction algorithm with shape-context descriptor (MSSCD). MSSCD computes a 3D histogram of edge point locations and orientations around a keypoint as its shape context descriptor.

1.2 The proposed approach

Although MSSCD and PIIFD improve the matching ability of these descriptors on multispectral images, they still generate a high ratio of incorrect mappings since the amount of common information decreases on them. Our previous work [19] considered affine transformations and utilized global information to evaluate triplets of keypoint mappings. To obtain the best matched reference keypoint for a test keypoint, an iterative process is employed that exhausts all triplets of possible keypoint mappings, and the computational complexity of the iterative process is large. For many multispectral images however, a translation [20] or a similarity transformation [21] may be enough to account for the misalignment. Observing this, this paper proposes utilizing global information and descriptors to establish keypoint mappings on two images between which the misalignment can be accounted for by similarity transformations. Since two keypoint mappings are required for calculating a similarity transformation, the computational complexity of exhausting pairs of keypoint mappings is greatly reduced compared with exhausting triplets of keypoint mappings.

The contribution of this paper is to utilize global information to build keypoint mappings. The proposed method has a much lower computational cost than exhausting triplets of keypoint mappings, but can still robustly build keypoint mappings on multispectral images. The matching ability of descriptors decreases on multispectral images, and hence the ratio of correct keypoint mappings is not so high as on monomodal images. Due to this, other information must be employed to help build robust keypoint mappings. One option is to increase the size of the local window for computing descriptor, allowing for more information to be encoded by descriptors. In most existing descriptors on single-spectrum images, one main orientation suffices to characterize the (local) geometric mistransformation since in sufficiently small regions any transformations reduce to rotation, translation, and scaling. However, for a window of a larger size on multispectral images, correctly assigning a main orientation is itself a challenging task [22].

To enhance the matching ability of descriptors, this work proposes utilizing information over entire images. Two keypoint mappings are needed to determine a similarity transformation that comprises scaling and rotation. The determined rotation and scaling in effect serve as a main orientation for the entire image when used as computing descriptors. The proposed method is similar to RANSAC in that both methods sample the combinations of keypoint mappings and then evaluate the sampled combinations. However, it differs essentially from RANSAC in that RANSAC only utilizes keypoint positions, i.e., in RANSAC, the sampled combinations are assessed with

the number of correct mappings in the rest. Due to the low ratio of correct mappings on multispectral images, the correct/good combinations are often mis-assessed to be incorrect/bad. While the proposed method utilizes the global information (encoded by the similarity metric) so that good mappings "conform to" the content of entire images, and thus the keypoint mappings of high similarity metric are more likely to be correct than built with RANSAC.

The rest of this paper is organized as follows, Section 2 discusses the proposed method, Section 3 analyzes the complexity of the proposed algorithm, Section 4 presents the experimental results, and Section 5 concludes this paper.

2 Proposed approach

This section presents the registration approach to aligning multispectral images. The misalignment is assumed to be small (i.e., not wide-baseline) and can be accounted for by a similarity transformation. For a test keypoint, the distance constraint is applied to narrow the space of its mapping candidates. Given a pair of keypoint mappings, a similarity transformation T is determined, then the similairty metric between $I_r(x,y)$ and $I_t^T(x,y)$ is calculated over entire images. Intuitively, the greater the similarity metric, the better the pair "conforms to" the entire image content. The insight of this paper comes from the following observation. Descriptors around keypoints encode the local information, and two keypoints are matched if the local information around them have the most common information/structure. However, multispectral images contain less common information than the-same-band (monomodal) images. Therefore, the local information around keypoints, i.e., descriptors, can not provide so many correct keypoint mappings, especially when the spectral difference is large. Intuitively, the local information on multispectral images becomes insufficient to decide whether a keypoint mapping is correct. Consequently, other complementary information to descriptors is necessitated for building reliable keypoint mappings.

To build keypoint mappings with descriptors on multispectral images, global information is utilized in this paper to evaluate keypoint mappings. A keypoint mapping is decided to be correct if its resulting T yields a large similarity metric between $I_r(x,y)$ and $I_t^T(x,y)$. This paper deals with similarity transformations, which require at least two keypoint mappings (i.e., a pair) for determining the misalignment. Since there are multiple such pairs of keypoint mappings, an iterative process is employed to search the best matched reference keypoint for every test keypoint. Calculating T uses the information over entire images, and we call it global information.

2.1 Distance constraints

This section follows the notations used in the previous work [19]. A similarity transformation is a simplified affine transformation $T = (A, \mathbf{t})$, and it transforms a point (x,y) to (u,v) by

$$\begin{aligned}\begin{pmatrix} u \\ v \end{pmatrix} &= A \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \mathbf{t} \\ &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}, \end{aligned} \tag{1}$$

where

$$\begin{aligned}\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \cdot \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} \\ &= \begin{pmatrix} s \cdot \cos(\theta) & -s \cdot \sin(\theta) \\ s \cdot \sin(\theta) & s \cdot \cos(\theta) \end{pmatrix}. \end{aligned} \tag{2}$$

Set a to $s \cdot \cos(\theta)$ and b to $s \cdot \sin(\theta)$, then A can be written as,

$$A = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

Thus, Equation 1 can be rewritten as,

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}. \tag{3}$$

Note, although A comprises four entries a_{11}, a_{12}, a_{21}, a_{22} in Equation 1, there are only four unknown variables in Equation 3, a, b, t_x, t_y, to be determined. This is the reason that a similarity transformation needs only two keypoint mappings, as compared with an affine transformation that needs three mappings.

When the misalignment is relatively small, the spatial distance of a test keypoint $\mathbf{p_t}$ in $I_t(x,y)$ to its corresponding point $\mathbf{p_r}$ in $I_r(x,y)$ is small [19, 23].

Formally, $\|\mathbf{p_t} - \mathbf{p_r}\|_2 \leq \sqrt{2}\|A - I\|_\infty \cdot \|\mathbf{p_t}\|_2 + \|\mathbf{t}\|_2 < T_{trd}$. T_{trd} is a threshold to be set. In this work, T_{trd} is set to the 1/4 the maximum of the height and width of images to be aligned, i.e., $T_{trd} = 1/4 \cdot \max\{H, W\}$, where H (W) is the height (width) of images. $1/4 \cdot \max\{H, W\}$ is used here since a point $\mathbf{p_t}$ will not move farther than it given that the unknown misalignment is relatively small. Under this assumption, the distance constraint can easily rule out a large number of wrong keypoint mappings.

A wrong mapping here is referred to two matched keypoints that are spatially far away from each other.

2.2 Building initial keypoint mappings

This section discusses building initial keypoint mappings. SURF [13] is used to detect keypoints and descriptors. Let $K_t^i, i = 1, \ldots, N_t$, denote the ith keypoint on $I_t(x,y)$, and $K_r^j, j = 1, \ldots, N_r$, denote the jth keypoint on $I_r(x,y)$. Follow the notation in [16], let $f_t^i, i = 1, \ldots, N_t$, denote the descriptor associated with K_t^i, and $f_r^j, j = 1, \ldots, N_r$, denote

the descriptor associated with K_r^j. $K_r^{j_0}$ and $K_t^{i_0}$ are said to be matched if

$$D\left(f_t^{i_0}, f_r^{j_0}\right) < 0.8 \cdot D\left(f_t^{i_0}, f_r^{j_1}\right),$$

where $f_r^{j_1}$ is the second closest neighbor to $f_t^{i_0}$.

Due to the gradient reversal and region reversal, the repeatability and distinctiveness decrease significantly on multispectral images, and hence the initial keypoint mappings contain a high ratio of incorrect ones [24]. The set of initially built keypoint mappings are used in Section 2.4 for searching the best matched reference keypoint for every test keypoint.

2.3 Evaluating a pair of keypoint mappings

Given a keypoint mapping $\left(K_t^{i_1} \sim K_r^{j_1}\right)$, $K_r^{j_1}$ is the best matched keypoint to $K_t^{i_1}$. "Best matched" means the local region around $K_r^{j_1}$ is more similar to $K_t^{i_1}$ than other keypoints on $I_r(x, y)$. Further evaluation of this mapping is often accomplished by applying "consistence check" to the set of initial mappings. RANSAC [25] is a commonly used technique to separate out correct mappings. When the ratio of wrong mappings is high, it often fails to work. Observing this, this paper proposes utilizing global information to compensate the decrease of the matching ability of descriptors.

Consider a pair of keypoint mappings, $\left(K_t^{i_1} \sim K_r^{j_1}\right)$, and $\left(K_t^{i_2} \sim K_r^{j_2}\right)$. Let (u_k, v_k) denote the location of $K_r^{j_k}, k = 1, 2$, and (x_k, y_k) denote the location of $K_t^{i_k}, k = 1, 2$, then by Equation 3 the two keypoint mappings give

$$\begin{pmatrix} x_1 & -y_1 & 1 & 0 \\ y_1 & x_1 & 0 & 1 \\ x_2 & -y_2 & 1 & 0 \\ y_2 & x_2 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ t_x \\ t_y \end{pmatrix} = \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{pmatrix}. \quad (4)$$

Once a, b, t_x, t_y are determined, $I_t(x, y)$ is transformed by T to obtain $I_t^T(x, y)$ with Equation 3. The similarity metric $S(I_r(x, y), I_t^T(x, y))$ between $I_r(x, y)$ and $I_t^T(x, y)$ is computed. The greater the similarity metric, the closer the two keypoint mappings to be correct.

The similarity metric measures the similarity between $I_r(x, y)$ and $I_t^T(x, y)$, and thus it characterizes the closeness of T to the ground truth. This work applies the number of overlapped edge pixels (NOEP) as the similarity metric,

$$S\left(I_r(x, y), I_t^T(x, y)\right) := \text{NOEP}\left(E_r(x, y), E_t^T(x, y)\right), \quad (5)$$

where $E_r(x, y)$ and $E_t^T(x, y)$ are edge maps of $I_r(x, y)$ and $I_t^T(x, y)$, respectively. $S\left(I_r(x, y), I_t^T(x, y)\right)$ simply counts the number of overlapped edge pixels.

The NOEP represents the similarity of two edge maps. It serves as a descriptor in the sense that it encodes the distribution of edge points and hence characterizes the content structure. Alternatively, NOEP can be treated as a simplified version of edge of histogram (EOH) [26] calculated on an entire image instead of a local window. Due to the gradient reversal [27], gradient orientation is unreliable, but the position of edges tends to be stable and has been used for computing similarity metric on multispectral images [20]. Note the superscript T of $E_t^T(x, y)$ in (5) is a generalized version of main orientation. The main orientation of a keypoint accounts for the local geometric view difference with rotation. For an entire image (a larger window), a more complex transformation is required other than only rotation to align the entire image (a larger descriptor).

2.4 Searching for the best matched keypoint

Due to the multimodality, some test keypoints may not have any corresponding reference keypoints on $I_r(x, y)$. To rule out incorrect mappings from the initially built keypoint mappings, this section computes for every test keypoint $K_t^i, i = 1, \ldots, N_t$, the maximum similarity metric $S_{\max}^t(i)$ it can yield. Then, the vector $S_{\max}^t(i), i = 1, \ldots, N_t$, is ordered, and the test keypoints ranked top 15 % is preserved to calculate the final transformation parameters. In short, this section includes two steps, the first is to compute the maximum similarity metric for every test keypoint, and the second is to choose test keypoints for computing the transformation parameters.

There are N_t keypoints on $I_t(x, y)$, $K_t^i, i = 1, \ldots, N_t$, so the number of pairs of test keypoints is $\binom{N_t}{2}$. For a test keypoint $K_t^{i_0}$, it appears in $N_t - 1$ pairs $(K_t^{i_0}, K_t^i), i \neq i_0$. Thus, for the test keypoint $K_t^{i_0}$, $N_t - 1$ NOEPs can be calculated with the pair comprising $K_t^{i_0}$ and K_t^i by Equation 5. The maximum similarity metric for $K_t^{i_0}$ is achieved by considering all $N_t - 1$ such pairs. Formally,

$$S_{\max}^t(i_0) = \max_{i, i \neq i_0} \text{NOEP}\left(I_r(x, y), I_t^{T_{i_0,i}}(x, y)\right), \quad (6)$$

where $T_{i_0,i}$ is determined by $(K_t^{i_0}, K_t^i)$ and their initial mapping reference keypoints.

To compute the maximum similarity metric for every test keypoint, an iterative process is employed that exhausts all pairs of keypoint mappings.

The iterative process picks a pair of test keypoints $\left(K_t^{i_1}, K_t^{i_2}\right)$ and their reference mapped keypoints $\left(K_r^{k_{i_1}}, K_r^{k_{i_2}}\right)$, the distance constraint in Section 2.1 is applied to $\left(K_t^{i_1}, K_r^{k_{i_1}}\right)$ and $\left(K_t^{i_2}, K_r^{k_{i_2}}\right)$, to remove the keypoint mappings with a greater distance than the threshold T_{trd}. Additionally, we require the distance between two test keypoints in a pair be greater than a threshold T_{ttd}, as a pair consisting of smaller-distance keypoints often provides unreliable transformation. In this work, $T_{ttd} = 10$.

The similarity transformation T is determined with $\left(K_t^{i_1}, K_t^{i_2}\right) \sim \left(K_r^{k_{i_1}}, K_r^{k_{i_2}}\right)$, and the similarity metric is calculated. The iterative process considers all pairs of keypoint mappings and stores the maximum similarity metric for every test keypoint. It is summarized in Algorithm 1.

Algorithm 1: Iteratively processing pairs of keypoint mappings

input : $I_r(x,y)$, $I_t(x,y)$.
output: $S^t_{\max}$.

1 **Extract image features**:
- Detect keypoints K_r^i and descriptors $f_r^i, i \in [1, N_r]$, from $I_r(x,y)$, and K_t^i and $f_t^i, i \in [1, N_t]$, from $I_t(x,y)$.
- Generate edge maps $E_r(x,y)$ and $E_t(x,y)$ from $I_r(x,y)$ and $I_t(x,y)$. They are used for computing similarity metric.

Precompute:

- The spatial distance between test/reference keypoints, $d\left(K_t^i, K_t^j\right), \forall i,j \in [1, N_t]$, $d\left(K_r^i, K_r^j\right), \forall i,j \in [1, N_r]$
- The spatial distance between test and reference keypoints, $d\left(K_t^i, K_r^j\right), \forall i \in [1, N_t], j \in [1, N_r]$

for $i_1, i_2 \in [1, N_t]$ **do**

1. Require $i_1 < i_2$.
2. **if** $d\left(K_t^{i_1}, K_t^{i_2}\right) < T_{ttd}$ **then** Continue; **end if**
3. Find the matched reference keypoint to $K_t^{i_1}$, $K_r^{k_{i_1}}$, and the matched reference keypoint to $K_t^{i_2}$, $K_r^{k_{i_2}}$.
4. Require $k_{i_1} \neq k_{i_2}$.
5. Require $\left(P_t^{i_1}, P_t^{i_2}\right) \sim \left(P_r^{k_{i_1}}, P_r^{k_{i_2}}\right)$ satisfying the geometrical constraint in Section 2.1.
6. Determine T between $\left(P_t^{i_1}, P_t^{i_2}\right)$ and $\left(P_r^{k_{i_1}}, P_r^{k_{i_2}}\right)$ by Equation 4.
7. Transform edge points of $I_t(x,y)$ by the determined T.
8. Compute similarity metric $S\left(I_r(x,y), I_t^T(x,y)\right)$ by Equation 5.
9. Update $S^t_{\max}(i_1)$ and $S^t_{\max}(i_2)$ by Equation 6.

end for

3 Complexity analysis

This section analyzes the computational complexity of the proposed method. Firstly, we discuss the computational cost when the distance constraints are not applied, and then give the real running time when the constraints are applied. Since there are N_t test keypoints, the number of combinations of two test keypoints is $\binom{N_t}{2}$. If we are dealing with affine transformations, at least three keypoint mappings are needed to determine an affine transformation. Three keypoint mappings form a triplet, and there are totally $\binom{N_t}{3}$ such triplets. Consequently, the number of triplets of keypoint mappings is roughly N_t times that of the pairs of keypoint mappings.

On multispectral images, the closest reference keypoint may not be the correct one, so multiple mapping candidates are assigned to a test keypoint [28]. If N_c mapping candidates are assigned to every test keypoint like ref. [19], then the computational cost of the proposed method is $\binom{N_t}{2} \cdot N_c^2$, and the computational cost of the approach to dealing with affine transformations in [19] is $\binom{N_t}{3} \cdot N_c^3$, which is about $N_t \cdot N_c$ times that of the presented algorithm. The similarity transformation used in this paper is sufficient to account for a wide variety of images, e.g., the remote sensing images and slices of medical images. When the misalignment does not involve a lot of skewing, the computational cost of the presented method is roughly $N_t \cdot N_c$ times less than the affine transformation model.

Next, we analyze the real running time. Table 1 gives the running time of the proposed method and the time required by affine transformation models on different datasets. See Section 4.1 for the explanation on the datasets utilized in this work. From Table 1, it can be seen that the computational cost of the presented algorithm is about at least 10∼20 times less than the algorithm applying affine transformation models, which is much smaller than $N_t \cdot N_c$ as analyzed above. The reason is that a relatively large percent of triplets (pairs) of keypoint mappings are ruled out by the distance constraints discussed

Table 1 Mean and standard deviation of the running time in seconds of the proposed method and the time required by affine transformation models

Dataset	Proposed method		Affine transformations	
	μ_t (s)	σ_t (s)	μ_t (s)	σ_t (s)
EOIR	6.23	10.29	63.35	480.56
Visible_nir	5.02	4.59	155.49	228.39
Country	43.27	31.97	109.82	141.35
Field	55.92	62.10	946.60	2540.73
Forest	80.78	47.24	938.17	1181.63
Indoor	46.57	48.84	875.28	1274.22
Mountain	44.01	43.74	1316.91	2601.72
Oldbuilding	72.93	74.19	1954.95	3306.19
Street	43.38	38.65	954.92	3242.05
Urban	94.24	67.59	5585.69	6886.24
Water	26.08	25.32	259.07	444.64

Processor: Intel i7-3770 3.40 GHz, RAM: 8 GHz

in Section 2.1. This also verifies the usefulness of the distance constraints in reducing the computational cost.

However, even with the distance constraints, the mean running time for affine transformation models is still about 10∼20 times more than the proposed method. For the image pairs between which the misalignment contains little skew, a similarity transformation is sufficient to account for the misalignment.

In Table 1, we can also see that the proposed method does not provide a real-time registration on any dataset. Two factors contribute to the computational cost, the number of pairs of keypoint mappings, and the complexity of calculating similarity metric in (5) between two images. To improve the running the speed, there are two aspects accordingly. The first is to reduce the number of pairs of keypoint mappings. For this, improving the matching ability of descriptors is a direction as otherwise the ratio of correct keypoint mappings is low and hence we would need to consider sufficiently many pairs of keypoint mappings.

The second is to substitute (5) for a simpler similarity metric of lower complexity. An image feature that effectively represents the entire image for use in registration, and a fast-running similarity metric can improve the running speed. Additionally, a multiresolution technique is an option to lower the computational expense of the presented method. A heuristic calculation of similarity metric will reduce some computation, like the search of extremum points in SIFT [12]. The edge points are assigned to different priorities, and those of a high priority will firstly be used for computing similarity metric. If they do not contribute much to the similarity metric then the calculation stops.

4 Experimental results

This section presents experimental results. The proposed method is compared with the SIFT [12], FREAK [14], improved symmetric-SIFT (ISS) [27], and ORB [29]. The SIFT and ORB are mostly designed for single-mode images, and they are expected to perform well on single-mode images, e.g., visible images. The FREAK is (partly) designed for multispectral images, and the ISS is completely designed for multispectral images.

4.1 Datasets used to test the performance

Three datasets are used to test the performance of the proposed approach. In general, the larger the spectral difference, the stronger the multimodality, and consequently the less repeatable the keypoint and local gradient pattern [30]. We investigate the performance of the proposed method on multispectral images of varying spectral difference.

Dataset 1 (EOIR) includes 101 image pairs acquired by ourselves, one image taken with the visible camera and the other taken with the mid-wave infrared camera (3–5 μm).

Dataset 2 (Visible_nir) includes real-world hyperspectral image (RWHI) from [31] containing 50 scenes. The images in this dataset were acquired by sequentially tuning a filter through a series of 31 narrow wavelength bands, each with approximately 10-nm bandwidth and centered at steps of 10 nm from 420 to 720 nm (refer [31] for details). We use the 50 image pairs of 420 and 720 nm. Dataset 3 is from [32] including 477 images in 9 categories of scenes: Country, Field, Forest, Indoor, Mountain, Old-building, Street, Urban, Water. The image pairs in dataset 3 are taken with visible camera (RGB) and near infrared (NIR) camera. Since the image pairs in dataset EOIR have a larger spectral distance than the image pairs in dataset Visible_nir and the 9 categories in dataset 3, they contain less common information, and hence the repeatability of descriptors decreases more on dataset EOIR.

The texture information in images is important for establishing keypoint mappings. The image content in dataset EOIR covers 2D indoor scenes, 2D outdoor scenes (e.g., wall of buildings), 3D outdoor scenes, and Landsat images. The 9 categories in dataset 3 cover different scenes as well, on which the performance of keypoint mappings can be effectively evaluated.

4.2 Results of keypoint mappings

This section analyzes the performance of the keypoint mappings built with the SIFT [12], FREAK [14], ISS [27], ORB [29], and the presented method. We first present the visual results of keypoint mappings on the image pairs of dataset EOIR, since this dataset is the most challenging. And then a quantitative analysis is conducted on the performance of keypoints built with different methods.

Figure 1 gives the keypoint mappings on an image pair taken with the visible camera and mid-wave infrared camera. Due to the significant decrease of the repeatability and distinctiveness on multispectral images, the keypoint mappings built with the SIFT, ISS, ORB, and FREAK contain many incorrect mappings as shown in Fig. 1b–e. The global information applied in the proposed method helps establish reliable mappings as shown in Fig. 1a, since it effectively compensates the insufficiency of the distinctiveness of the descriptors. Additionally, the repeating structures in the image content cause mismatches since the local patches at these structures are similar to each other even if the repeatability of the descriptors does not decrease. In such image pairs, it is very difficult to establish keypoint mappings of a high correct rate relying solely on the information encoded by local descriptors.

Figure 2 shows the keypoint mappings on another image pair taken with the visible camera and mid-wave infrared camera. Figure 2b gives the keypoint mappings built with the SIFT. The local patch near the tail light of a car in the

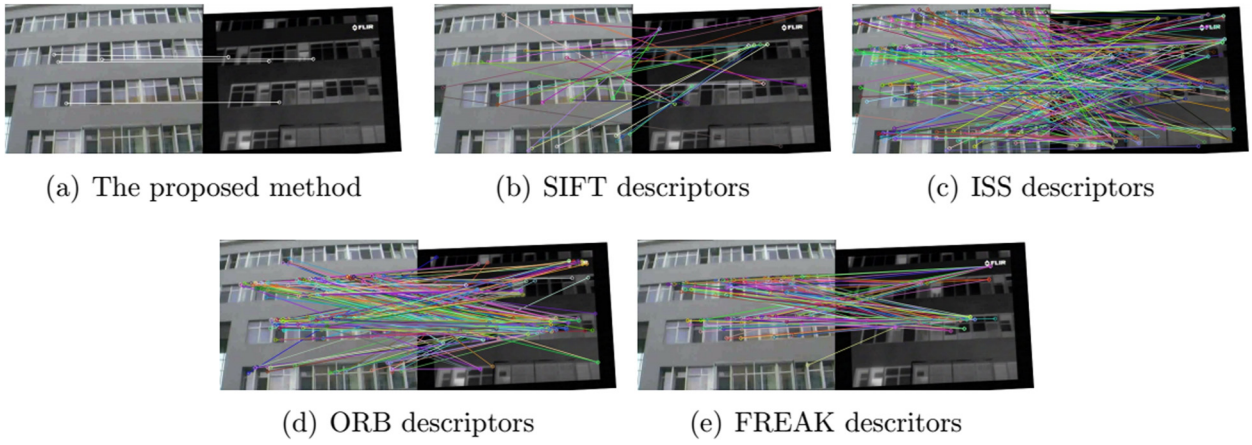

Fig. 1 Comparison of the keypoint mappings built by different methods on an image pair of an outdoor scene in dataset 2. **a** The proposed method, **b** SIFT descriptors [33], **c** ISS descriptors [27], **d** ORB descriptors [29], and **e** FREAK descriptors [14]

visible image is similar to the patch near a window in the infrared image, so the two keypoints are matched. A similar phenomenon can be observed on the result of FREAK as shown in Fig. 2e. The ISS and ORB shown in Fig. 2c, d provide more keypoint mappings than the SIFT and the FREAK, but the ratio of correct mappings is not markedly higher than the SIFT and FREAK.

Figure 3 shows the keypoint mappings on a Landsat multispectral image pair built with different methods. On this image pair, the SIFT, ISS, ORB, and FREAK perform much better than on the image pair shown in Figs. 1 and 2. The reason is that the characteristic of the infrared image is close to that of the visible image, i.e., the dark regions (pixels) in the visible image also correspond to the dark ones in the infrared image. Thus, the similar local gradient patterns for computing descriptors are also close to each other and consequently keypoint mappings can be robustly established. In terms of a local gradient pattern, this image pair can be partly viewed as single-mode images.

Next, a quantitative analysis is conducted on the performance of keypoint mappings. Specifically, the number of correct mappings is calculated for each method on every dataset. Assume that (K_t^i, K_r^j) is a keypoint mapping, then it will be viewed as correct if $d(T(K_t^i), K_r^j) < d_M$, where d_M is a threshold to be set. In literature, different thresholds on the distance between mapped keypoints have been used to determine whether keypoint mappings are correct or not. These thresholds include 2, 3, 4, 5, etc. To eliminate the effect of thresholds on the performance evaluation of different methods, this work employs the histogram of the distance between mapped keypoints. d_M is set to multiple

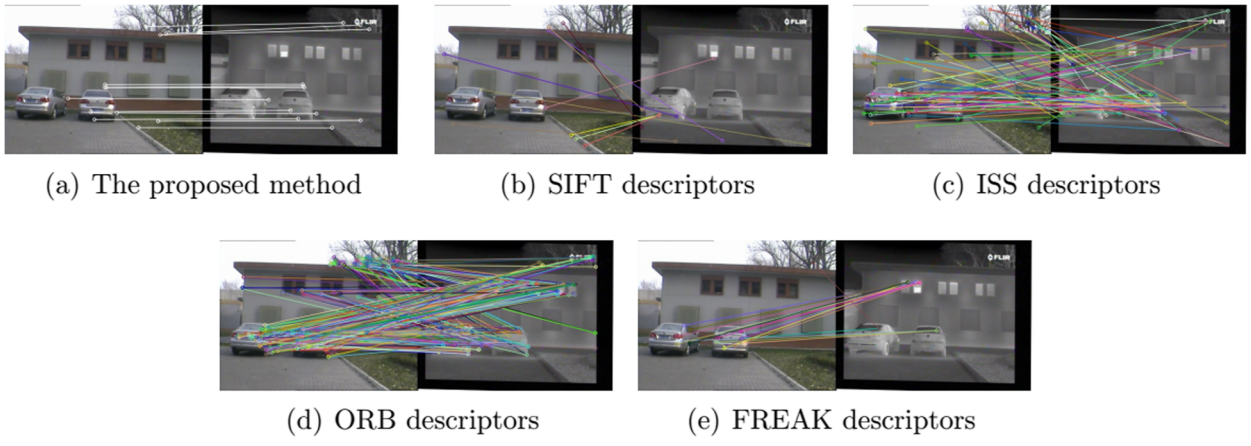

Fig. 2 Comparison of the keypoint mappings on an image pair that contains similar local intensity/gradient patterns. Due to this, in (**e**), the keypoints near the tail lights of the two cars in left image are mapped to keypoints near a window in the right image. **a** The proposed method, **b** SIFT descriptors [33], **c** ISS descriptors [27], **d** ORB descriptors [29], and **e** FREAK descriptors [14]

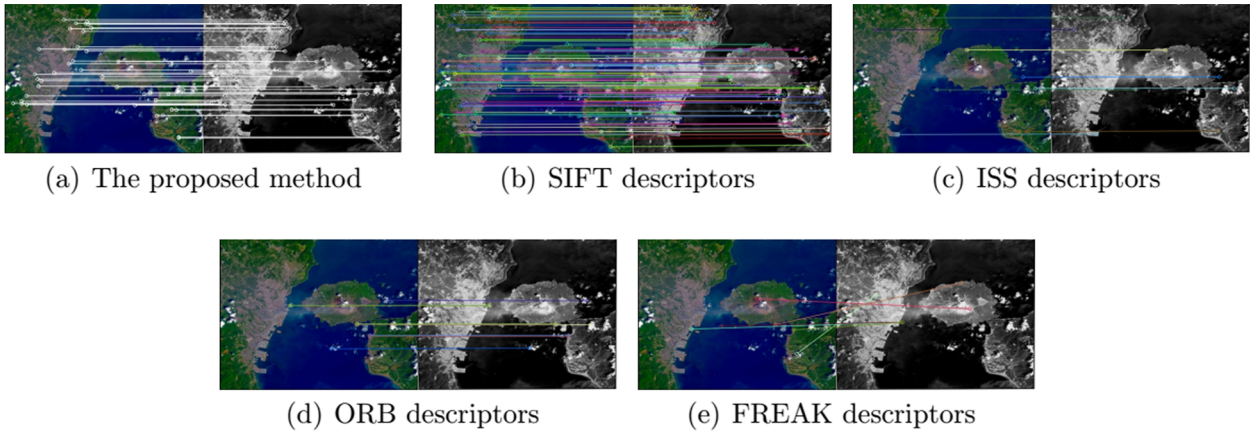

Fig. 3 Keypoint mappings on a Landsat multispectral image pair. The characteristic of the visible image is close to that of the infrared image. From this aspect, the two images can be partly viewed as single-mode, and so the SIFT, ISS, ORB, and FREAK perform better. **a** The proposed method, **b** SIFT descriptors [33], **c** ISS descriptors [27], **d** ORB descriptors [29], and **e** FREAK descriptors [14]

values, and we count the number of keypoint mappings for which the distance between the two keypoints is smaller than d_M.

The histogram of the distances between mapped keypoints is generated as follows. The bins are set to $[0, 2]$, $[2, 5]$, $[5, 10]$, $[10, 20]$, and $[20, \infty]$. For example, the bin $[2, 5]$ counts the number of keypoint mappings with the distance less than 5 but greater than 2, and the bin $[10, 20]$ counts the number of mappings with the distance greater than 20. Note, other setups for the bins can be used here if a better comparison can be achieved for different methods.

Table 2 gives the histogram of distances between mapped keypoints for different methods. On all datasets, the presented method performs better than other methods. There are two aspects showing the advantage of the presented method over other methods. The first is that the presented method provides a higher ratio of keypoint mappings that have a relatively small distance. For example, on the dataset "Field," the presented method yields 2480 mappings that have a distance falling in $[0, 2]$, while SIFT does 691, ISS does 246, ORB does 295, and FREAK does 11. The second is that the presented method provides a lower ratio of keypoint mappings that have a distance greater than 20. For example, on the dataset "EOIR," the SIFT yields 187 keypoint mappings of distance greater than 20, ISS does 295, ORB does 4616, and FREAK does 358.

One observation on the comparison result shown in Table 2 is that the dataset "EOIR" is the most challenging. In most cases, all methods including the proposed one perform worse on "EOIR" than other datasets. Take the ORB method for an example, and we consider the number of keypoint mappings that have a distance greater than 20 (the worst case). It provides 4616 mappings on dataset "EOIR" that have a distance falling in ">20", only 3 mappings on dataset Visible_nir, 6745 mappings on Country, 2805 mappings on Field, 373 mappings on Forest, 598 mappings on Indoor, 359 mappings on Mountain, 27 mappings on Oldbuilding, 731 mappings on Street, 8 mappings on Urban, and 2354 mappings on Water. The reason is that as aforementioned, the multimodality of the image pairs in dataset EOIR is greater than that in other datasets, and hence the matching performance of descriptors decrease on EOIR.

Another observation is that ISS does not perform better than SIFT, although ISS is designed to adapt the descriptor of SIFT to multispectral images. On dataset EOIR, the SIFT performs slightly better or comparable to ISS, and on other, datasets the SIFT performs evidently better than ISS since the multimodality decreases on them. The underlying mechanism causing this phenomenon needs further investigation on more types of multispectral images.

Table 2 clearly shows that the matching performance of descriptors decreases when the multimodality of image data increases. On single-mode images, the SIFT and ORB perform fairly well for coping with integer-pixel alignment. On multispectral images, the common information around keypoints may be not enough for robustly establishing keypoint mappings. Either more common and distinctive information near keypoints can be encoded to enhance the matching ability of descriptors, or complementary information at regions not-too-near keypoints are desired to correctly build keypoint mappings. This will be the future work.

5 Conclusions

This work presents a registration approach on multispectral images. A similarity transformation is considered for

Table 2 The distribution of the distances between matched keypoints

	[0–2]	[2–5]	[5–10]	[10–20]	> 20	[0–2]	[2–5]	[5–10]	[10–20]	> 20	[0–2]	[2–5]	[5–10]	[10–20]	> 20
			EOIR					Visible_nir					Country		
Proposed	234	102	54	244	0	5732	0	0	0	0	875	35	9	2	0
SIFT	30	7	9	6	187	4204	44	1	0	21	379	87	47	14	224
ISS	31	16	5	13	295	1879	27	8	7	106	213	54	18	12	854
ORB	328	87	50	122	4616	12,946	563	3	0	3	195	95	59	56	6745
FREAK	11	11	5	10	358	9	2	4	11	189	12	10	15	35	9517
			Field					Forest					Indoor		
Proposed	2480	99	40	60	0	2814	24	0	0	0	4803	0	0	0	0
SIFT	691	116	28	15	183	6045	1829	328	9	199	464	23	7	13	108
ISS	246	73	37	10	747	6	2	0	4	1229	299	26	16	7	155
ORB	295	168	65	49	2805	2316	763	22	6	373	391	57	17	19	598
FREAK	11	9	11	25	3946	0	0	2	5	876	69	36	36	103	5351
			Mountain					Oldbuilding					Street		
Proposed	4676	20	1	38	0	6783	31	0	0	0	3369	70	5	0	0
SIFT	742	258	31	14	65	696	81	17	0	11	356	94	29	1	13
ISS	179	98	25	5	170	281	32	10	2	46	179	54	18	9	273
ORB	269	186	17	3	359	304	103	30	1	27	209	144	49	10	731
FREAK	97	13	28	58	5112	43	24	20	65	6779	37	6	20	82	6193
			Urban					Water							
Proposed	12,685	0	0	0	0	1714	30	6	66	0					
SIFT	735	8	4	3	16	425	44	17	11	101					
ISS	389	14	0	7	67	263	35	11	16	515					
ORB	366	17	10	0	8	300	122	69	78	2354					
FREAK	111	34	35	115	9629	32	20	26	75	6831					

accounting for the misalignment between two images. Global information over entire images is induced to help evaluate the quality of keypoint mappings. Compared with the methods that solely use descriptors for building keypoint mappings, the proposed approach effectively compensates the insufficiency of the repeatability and distinctiveness of descriptors and hence provides more correct mappings.

Several future research directions can be done to further improve the performance. The matching ability of descriptors can be researched by analyzing, extracting, and encoding the common information between multispectral images. Although it is not the focus of this work, the matching ability can improve the overall registration accuracy. Another direction is on the similarity metric that has been used for evaluating the quality of keypoint mappings. It carries the global information of entire images and its effective characterization will be expected to bring more precise keypoint mappings.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the National Natural Science Foundation of China (Grants No., NSFC- 61170176), Fund for the Doctoral Program of Higher Education of China (Grants No., 20120005110002), Fund for Beijing University of Posts and Telecommunications (Grants No., 2013XD-04, 2013XZ10), Fund for National Great Science Specific Project (Grants No. 2014ZX03002002-004).

Received: 16 January 2015 Accepted: 12 May 2015
Published online: 01 July 2015

References

1. LG Brown, A survey of image registration techniques. ACM Comput. Surv. **24**(4), 325–376 (1992)
2. P Viola, III, WMW, in *Proceedings of the Fifth International Conference on Computer Vision*. Alignment by Maximization of Mutual Information (IEEE, Cambridge, MA, 1995), pp. 16–23
3. MP Heinrich, M Jenkinson, M Bhushan, T Matin, FV Gleeson, SM Brady, JA Schnabel, Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)
4. S Chen, Q Guo, H Leung, Bossé, A maximum likelihood approach to joint image registration and fusion. IEEE Trans. Image Process. **20**(5), 1363–1372 (2011)
5. P Thévenaz, M Unser, Optimization of mutual information for multiresolution image registration. IEEE Trans. Image Process. **9**(12), 2083–2099 (2000)
6. G Tzimiropoulos, V Argyriou, S Zafeiriou, T Stathaki, Robust fft-based scale-invariant image registration with image gradients. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1899–1906 (2010)
7. W Pan, K Qin, Y Chen, An adaptable-multilayer fractional fourier transform approach for image registration. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 400–413 (2009)
8. MM Fouad, RM Dansereau, AD Whitehead, Image registration under illumination variations using region-based confidence weighted m-estimators. IEEE Trans. Image Process. **21**(3), 1046–1060 (2012)
9. D Zosso, X Bresson, J-P Thiran, Geodesic active fieldsa geometric framework for image registration. IEEE Trans. Image Process. **20**(5), 1300–1312 (2011)
10. C Xing, P Qiu, Intensity-based image registration by nonparametric local smoothing. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 2081–2092 (2011)
11. X Liu, Z Lei, Q Yu, X Zhang, Y Shang, W Hou, Multi-modal image matching based on local frequency information. EURASIP J. Adv. Signal Process. **2013**(3), 1–11 (2013)
12. DG Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
13. H Bay, A Ess, T Tuytelaars, LV Gool, Speeded up robust features (surf). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
14. A Alahi, R Ortiz, P Vandergheynst, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. FREAK: Fast Retina Keypoint (IEEE Providence, RI, 2012), pp. 510–517
15. M Ambai, Y Yoshida, in *IEEE International Conference on Computer Vision*. CARD: Compact And Real-time Descriptors (IEEE, Barcelona, 2011), pp. 97–104
16. J Chen, J Tian, N Lee, J Zheng, RT Smith, AF Laine, A partial intensity invariant feature descriptor for multimodal retinal image registration. IEEE Trans. Biomed. Eng. **57**(7), 1707–1718 (2010)
17. S Saleem, R Sablatnig, A robust sift descriptor for multispectral images. IEEE Signal Process. Lett. **21**(4), 400–403 (2014)
18. W Bingjian, L Quan, L Yapeng, L Fan, B Liping, L Gang, L Rui, Image registration method for multimodal images. Appl. Opt. **50**(13), 1861–1867 (2011)
19. Y Li, R Stevenson, Incorporating global information in feature-based multimodal image registration. J. Electron. Imaging. **23**(2), 023013-1–023013-14 (2014)
20. KM Simonson Jr, S M D, FR Tanner, A statistics-based approach to binary image registration with uncertainty analysis. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 112–125 (2007)
21. G Yang, CV Stewart, M Sofka, C-L Tsai, Registration of challenging image pairs: Initialization, estimation, and decision. IEEE Trans. Pattern Anal. Mach. Intell. **29**(11), 1973–1989 (2007)
22. S Gauglitz, M Turk, T Höllerer, in *British Machine Vision Conference*. Improving Keypoint Orientation Assignment (BMVC Press, University of Dundee, 2011)
23. Y Wu, W Ma, M Gong, A novel point-matching algorithm based on fast sample consensus for image registration. IEEE Geosci. Remote Sens. Lett. **12**(1), 43–47 (2015)
24. MT Hossain, SW Teng, G Lu, in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. Achieving High Multi-Modal Registration Performance Using Simplified Hough-Transform with Improved Symmetric-SIFT (IEEE, Fremantle, WA, 2012), pp. 1–7
25. MA Fischler, RC Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
26. C Aguilera, F Barrera, F Lumbreras, AD Sappa, R Toledo, Multispectral image feature points. Sensors **12**, 12661–12672 (2012)
27. MT Hossain, G Lv, SW Teng, G Lu, M Lackmann, in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. Improved Symmetric-SIFT for Multi-modal Image Registration (IEEE, Noosa, QLD, 2011), pp. 197–202
28. Y Wu, W Ma, M Gong, L Su, L Jiao, A novel point-matching algorithm based on fast sample consensus for image registration. IEEE Geoscience Remote Sensing Lett. **12**(1), 43–47 (2015)
29. E Rublee, V Rabaud, K Konolige, G Bradski, in *IEEE Computer Vision (ICCV)*. ORB: An Efficient Alternative to SIFT Or SURF (IEEE, Barcelona, 2011), pp. 2564–2571
30. Z Ghassabi, J Shanbehzadeh, A Sedaghat, E Fatemizadeh, An efficient approach for robust multimodal retinal image registration based on ur-sift features and piifd descriptors. EURASIP J. Image Video Process. **2013**(25), 1–15 (2013)
31. A Chakrabarti, T Zickler, in *IEEE Conference on Computer Vision and Pattern Recognition*. Statistics of Real-World Hyperspectral Images (IEEE, Providence, RI, 2011), pp. 193–200
32. M Brown, S Süsstrunk, in *IEEE Conference on Computer Vision and Pattern Recognition*. Multi-Spectral SIFT for Scene Category Recognition (IEEE, Providence, RI, 2011), pp. 177–184
33. Y Ke, R Sukthankar, in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors (IEEE, Washington, DC, 2004), pp. 506–513