EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Multi-modal image matching based on local frequency information

Xiaochun Liu[1,2*], Zhihui Lei[1,2], Qifeng Yu[1,2], Xiaohu Zhang[1,2], Yang Shang[1,2] and Wang Hou[1,2]

Abstract

This paper challenges the issue of matching between multi-modal images with similar physical structures but different appearances. To emphasize the common structural information while suppressing the illumination and sensor-dependent information between multi-modal images, two image representations namely Mean Local Phase Angle (MLPA) and Frequency Spread Phase Congruency (FSPC) are proposed by using local frequency information in Log-Gabor wavelet transformation space. A confidence-aided similarity (CAS) that consists of a confidence component and a similarity component is designed to establish the correspondence between multi-modal images. The two representations are both invariant to contrast reversal and non-homogeneous illumination variation, and without any derivative or thresholding operation. The CAS that integrates MLPA with FSPC tightly instead of treating them separately can more weight the common structures emphasized by FSPC, and therefore further eliminate the influence of different sensor properties. We demonstrate the accuracy and robustness of our method by comparing it with those popular methods of multi-modal image matching. Experimental results show that our method improves the traditional multi-modal image matching, and can work robustly even in quite challenging situations (e.g. SAR & optical image).

Keywords: Multi-modal image, Image matching, Image representation, Local frequency information, Wavelet transformation, Similarity measure

1. Introduction

Image matching that aims to find the corresponding features or image patches between two images of the same scene is often a fundamental issue in computer vision. It has been widely used in vision navigation [1], target recognition and tracking [2], super-resolution [3], 3-D reconstruction [4], pattern recognition [5], medical image processing [6], etc.. In this paper, we focus on the issue of matching for multi-modal (or multi-sensor) images that differ in relation to the type of visual sensor. There are many important issues that make multi-modal image matching a very challenging problem [7]. First, multi-modal images are captured using different visual sensors (e.g. SAR, optical, infrared, etc.) at different time. Second, images with different modalities are normally mapped to different intensity values. This makes it difficult to measure similarity based on their intensity values since the same content may be represented by different intensity values. The problem is further complicated by the fact that various intrinsic and extrinsic sensing conditions may lead to image non-homogeneity. Finally, the disparity between the intensity values of multi-modal images can lead to coincidental local intensity matches between non-corresponding content, which may make the algorithm difficult to search the correct solution. Hence, the focuses of multi-modal image matching reside in illumination (contrast and brightness) invariant representations, common structure extraction from varying conditions and robust similarity measure.

The existing approaches for multi-modal image matching can be generally classified as feature-based and region-based. Feature-based matching utilizes extracted features to establish correspondence. Interest points [8,9], edges [10], etc. are often used as the local features because of their robustness in extraction and matching. In [8], Scale Invariant Feature Transform (SIFT) and cluster reward algorithm (CRA) [11] are used to match multi-modal remote sensing images. The SIFT operator is first adopted

* Correspondence: lxc1448@gmail.com
[1]College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China
[2]Hunan Key Laboratory of Videometrics and Vision Navigation, Changsha 410073, China

to extract feature points and perform coarse match, and then the CRA similarity measure is used to achieve accurate correspondence. In [10], Yong *et al.* propose the algorithm for multi-source image matching based on information entropy which comprehensively considers of the intensity information and the edge direction information. For feature-based methods two requirements must be satisfied: (i) features are extracted robustly and (ii) feature correspondences are established reliably. Failure to meet either of them will cause this type of method to fail. In contrast to feature-based methods, region-based methods make use of the whole image content to establish correspondence. While most approaches use features for image matching, there is also a significant amount of work on region-based matching. In [12], local phase-coherence representation is constructed for multi-modal image matching. This representation has some merits that make it a promising candidate for handling situations where non-homogeneous image contrast exists: (i) it is relatively insensitive to the level of signal energy; (ii) it depends on the structures in the image and can emphasize the edges and ridges at the same time; and (iii) it has a good localization in the spatial domain. In [13], M. Irani *et al.* present an energy-image representation based on directional-derivative filters. A set of filters, oriented in the horizontal, vertical, and the two diagonal directions, are applied to the raw image, and then the derivative image is squared to get an "energy" image. Thus, the directional information is preserved in this energy representation. This approach, however, requires explicit directional filters and explicit filtering with Gaussian functions to create a pyramid. In addition, mutual information that has been commonly used and showed great promise in medical image processing is often adopted as the similarity measure for multi-modal image matching since it is insensitive to variation of intensities and doesn't require knowledge of the relationship (joint intensity distribution) of the two different modalities [14,15]. The main merit of region-based method is their ability of resistance against noise and texture distortions since abundant information can be adopted by using a relatively large template, and thus providing a high matching accuracy.

In this paper, we bring forward a local frequency information-based matching frame for multi-modal images. It takes advantage of the merits of both MLPA and FSPC by using the CAS, and can be used to match images captured by similar as well as different types of sensors at different time.

2. Image representations via local frequency information

The visual system of human can reliably recognize the same object/scene under widely varying conditions. If the illumination of a scene is changed by several orders of magnitude, our interpretation for it can keep unchanged largely. Thus, in the image matching the main form of invariance is invariance to illumination, this is particularly important for multi-modal images where non-homogeneous contrast and brightness variation frequently occur. In this work, the local frequency information is used to construct image representations namely FSPC and MLPA, which are both dimensionless and invariant to non-homogeneous illumination variation and contrast reversal, for multi-modal image matching.

2.1. Log-Gabor function

To preserve phase information, linear-phase filters that are nonorthogonal and in symmetric/anti-symmetric quadrature pairs should be used. In [16], J. Liu *et al.* use Gabor filters that can be tuned to any desired frequency or orientation and offer simultaneous localization of spatial and frequency information to construct local-frequency representation for multi-modal images. However, Gabor function cannot maintain a zero DC component for bandwidths over one octave. Log-Gabor filters have all the merits of Gabor filters and additionally allow constructing arbitrarily large bandwidth filters while still maintaining a zero DC component in the even-symmetric filter. Hence, in this work we prefer to use Log-Gabor filters that have a Gaussian transfer function when viewed on the logarithmic frequency scale, instead of Gabor filters, as the basis of our local frequency creation [17].

Due to the singularity of Log function at the origin, the 2D Log-Gabor filter needs to construct in the frequency domain. In polar coordinates system, the Log-Gabor function can be divided into two components: a radial component and an angular component. The radial component has a frequency response described by

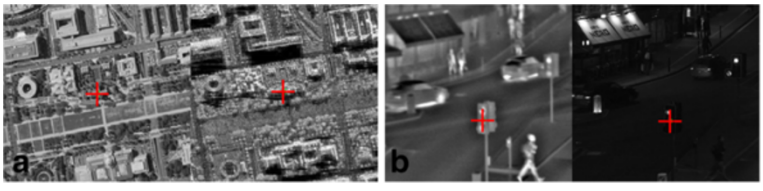

Figure 1 Matching result using different multi-modal image pair. (**a**) Optical and SAR image; (**b**) infrared and optical image.

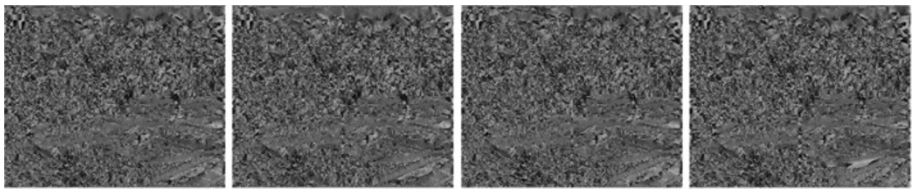

Figure 2 MLPAs corresponding to the images of Figure 4.

$$G_r(r) = \exp\left(-\frac{[\log(r/f_0)]^2}{2\sigma_r^2}\right) \quad (1)$$

And the angular component has a frequency response described by

$$G_\theta(\theta) = \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_\theta^2}\right) \quad (2)$$

The two components are multiplied together to construct the overall Log-Gabor filter which has the transfer function as

$$G(r,\theta) = G_r(r)G_\theta(\theta) \quad (3)$$

where (r, θ) represents the polar coordinates. As we can see from the definition formulas, the Log-Gabor filter is primarily determined by four parameters: f_0, θ_0, σ_r and σ_θ, where f_0 and θ_0 correspond to the center frequency and orientation angle, σ_r and σ_θ determine the scale and angular bandwidth respectively. The filter bank needs to make the transfer function of each filter overlap sufficiently with its neighbors so that the sum of all the transfer function forms a relatively uniform coverage of the spectrum.

2.2. Local frequency representations

In the search for invariant quantities in multi-modal images, the proposed approach is to take advantage of information from the frequency domain, rather than spatial domain. Let I denote the signal, and $LG^e_{n,\theta}$ and $LG^o_{n,\theta}$ denote the even-symmetric and odd-symmetric component of Log-Gabor function at the scale n and orientation θ. The response vector formed by the responses of each quadrature pair of filters can be expressed as

$$\left[e_{n,\theta}(x),\ o_{n,\theta}(x)\right] = \left[I(x) * LG^e_{n,\theta},\ I(x) * LG^o_{n,\theta}\right] \quad (4)$$

The values $e_{n,\theta}(x)$ and $o_{n,\theta}(x)$ can be regarded as real and imaginary parts of complex valued frequency component. The amplitude of the response vector at the scale n and orientation θ is given by

$$A_{n,\theta}(x) = \sqrt{e_{n,\theta}(x)^2 + o_{n,\theta}(x)^2} \quad (5)$$

and the phase is given by

$$\phi_{n,\theta}(x) = a\tan2\left(o_{n,\theta}(x), e_{n,\theta}(x)\right) \quad (6)$$

At each location x of a signal, we will have an array of these response vectors (each vector corresponds to one scale and orientation of filter). The response vectors form the basis of the proposed representations. The MLPA can be calculated as follow:

$$MLPA(x) = \begin{cases} \dfrac{a\tan2(F(x),H(x))}{\pi} \times 255, \\ \dfrac{\pi + a\tan2(F(x),H(x))}{\pi} \times 255, \end{cases} \quad (7)$$

$$if\ \ a\tan2(F(x),H(x)) \geq 0;$$
$$if\ \ a\tan2(F(x),H(x)) < 0.$$

where $F(x)$ and $H(x)$ can be calculated by summing the even and odd filter convolutions:

$$F(x) = \sum_\theta \sum_n e_{n,\theta}(x) \quad (8)$$

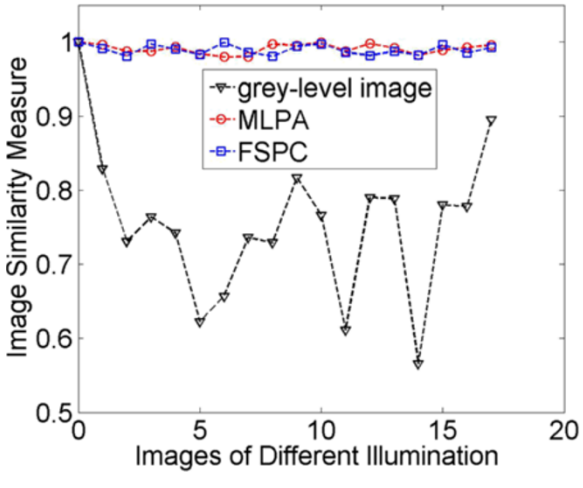

Figure 3 The illumination-invariant property of the proposed image representations.

Table 1 The values for the method parameters used in the experiments

Number of wavelet scales	Number of wavelet orientations	Wavelength of smallest scale filter	Scaling factor between successive filters	Cut-off value "c"	Gain factor "λ"
4	9	3	2.1	0.55	10

$$H(x) = \sum_{\theta} \sum_{n} o_{n,\theta}(x) \tag{9}$$

Contrast-reversal that may occur between the multi-modal images (e.g. Figure 1b) is eliminated by transferring the orientation of the mean local frequency vector [*F(x), H(x)*] that locates at the third/fourth quadrant (where $\alpha\tan2(F(x),H(x))<0$) to the first/second quadrant (where $\alpha\tan2(F(x),H(x))\geq 0$). Each value of MLPA, which is independent of the overall energy of the signal, is a measure of mean local phase angle. Hence, all MLPA maps have the same units, and are invariant to both scale and offset illumination changes (e.g. Figures 2 and 3). The main goal of MLPA is to eliminate the variation of intensity values between corresponding pixels of multi-modal image pair by using the phase information of local frequency. For a sophisticated matching algorithm, an outlier rejection mechanism is normally necessary since in many situations there are more "outliers" (non-common scene) than "inliers" (common scene) between multi-modal images. However, only by MLPA one cannot identify those inliers and eliminate the influence of the outliers. Hence, in this work the FSPC that aims to capture the common scene information while suppressing the illumination- and sensor-dependent information is developed by using the amplitude information of local frequency.

For multi-modal images, the signals are correlated primarily in high-frequency information, while correlation between the signals tends to degrade with the reduction of high-frequency information [13]. This is because high-frequency information (e.g. edge, contour, corner, junction, etc.) normally corresponds to the physical structure that is common to images with different modalities. On the other hand, low-frequency information depends heavily on the illumination and the photometric and physical imaging properties of sensors, and these are substantially different in multi-modal images. To capture the common physical structure, the high-pass filters (e.g. Sobel, Prewitt, Laplacian, etc.) that are working in spatial domain are reasonably adopted [10,13]. Those methods are straightforward and quite fast to compute. However, they normally depend on the intensity gradient information which highly relates with local image contrast, and therefore the non-homogeneous variation of contrast may degrade the performance of algorithm.

Working in frequency domain, phase congruency theory postulates that the structural information can be perceived at points where the local frequency components are maximally in phase, rather than assumes it is a point of maximal intensity gradient. The measure of phase congruency at a point x in a signal proposed by Morrone *et al.* in [18,19] can be expressed as

$$PC_1 = \frac{E(x)}{\sum_{\theta}\sum_{n} A_{n,\theta}} = \frac{\sqrt{F^2(x) + H^2(x)}}{\sum_{\theta}\sum_{n} A_{n,\theta}} \tag{10}$$

where $E(x)$ denotes the energy that is the magnitude of a vector sum. As we can see in the definition formula, phase congruency is the ratio of the energy $E(x)$ to the overall length taken by the local frequency components in reaching the end point. If all the local frequency components are in phase, all the response vectors would be aligned and the value of phase congruency, PC_1, would be a maximum of 1. If there is no coherent of phase, the value of PC_1 falls to a minimum of 0. Phase congruency is a quantity that is independent of the overall magnitude of the signal making it invariant to variation of image brightness and contrast.

Clearly, phase congruency is only significant if it occurs over a wide range of frequencies (phase congruency over many spectrums is more significant than phase congruency over narrow spectrums). Thus, as a measure of feature significance, phase congruency should be weighted by the frequency spread. To address the problem of the conventional phase congruency [18,19], we present a novel FSPC by using a weighing function that devalues the phase congruency at locations where the spread of filter responses is narrow. A measure of frequency spread can be defined as

Table 2 Comparisons of accuracy rates obtained from different methods

Image pair	Accuracy of Our method (%)	Accuracy of PC (%)	Accuracy of LFR (%)	Accuracy of FDDEI (%)	Accuracy of LSS (%)	Accuracy of MI (%)
Images of Figure 8	89.42	74.04	67.31	47.12	72.12	51.65
Images of Figure 9	97.25	81.32	79.12	76.92	80.77	74.73
Images of scene matching	96.63	85.78	82.69	76.73	83.24	74.89

Figure 4 Gray-level images with non-homogeneous illumination variation. The 1st image is the raw infrared image, and the rest are the synthetic images.

$$s(x) = \frac{1}{\sqrt{N}} \left(\frac{\sum_{\theta} \sum_{n} A_{n,\theta}(x)}{\sqrt{\sum_{\theta} \sum_{n} A_{n,\theta}^2(x)} + \varepsilon} \right) \tag{11}$$

where N denotes the total number of filters, and ε is used for avoiding division by zero and discounting the result when both $\sum_{\theta} \sum_{n} A_{n,\theta}(x)$ and $\sqrt{\sum_{\theta} \sum_{n} A_{n,\theta}^2(x)}$ are very small. The value of spread function, *s(x)*, varies between 0 and 1. If the distribution of filter responses is uniform over all spectrums, *s(x)*, reaches its maximum value of 1. The frequency spread weighing function can be constructed by applying a hyperbolic tangent function to the filter response spread value,

$$W(x) = \frac{1}{2} \left(1 + \tanh\left(\frac{\lambda}{2} (s(x) - c) \right) \right) \tag{12}$$

where c is the "cut-off" value, below which the value of phase congruency will be penalized, and λ is a gain factor that controls the sharpness of c. Thus, the definition of FSPC can be given as

$$FSPC(x) = \frac{W(x)E(x)}{\sum_{\theta} \sum_{n} A_{n,\theta}(x) + \varepsilon} \times 255 \tag{13}$$

Weighting by frequency spread has benefit of reducing those ill-conditioned responses that have the low frequency spread, as well as improving the localization accuracy of features, especially the smoothed features whose responses are normally uniform [20]. In addition, the noise resistance is also improved to some extent since the responses of noise are normally skewed to the high frequency end, and therefore have the relatively narrow frequency spectrums.

3. Matching using local frequency representations

Having obtained the local frequency representations, we then use them to perform matching operations. As we can see from the definitions of local frequency representations, MLPA primarily represents the phase information of local frequency, whereas FSPC mainly utilizes the amplitude information, which means MLPA and FSPC can be compensated each other to some extent since information independence. Hence, by using some proper fusion scheme that makes best use of the merits of MLPA and FSPC, one can achieve better matching performance. For example, the only use of MLPA may induce errors particularly in the texture-less image regions where FSPC normally has quite small value since the lack of significant features. In addition, it may be difficult to distinguish between two search windows that have similar MLPA but different FSPC.

In this work, we propose a novel confidence-aided similarity (CAS) measure to combine the MLPA and FSPC for improving matching robustness. CAS consists of two components: a similarity component and a confidence component. Let $FSPC_1$, $FSPC_2$, $MLPA_1$ and $MLPA_2$ denote a pair of values of FSPC and MLPA to be compared respectively, and the definition of CAS for a single signal can be expressed as

$$CAS_0 = d + c \tag{14}$$

where $d = -|MLPA_1 - MLPA_2|$, $c = \frac{1}{2}(FSPC_1 + FSPC_2)$. d is the similarity component that reflects how well the two signals resemble each other, and c is the confidence component that reflects the confidence that a match is correct.

MLPA with low FSPC is normally less reliable than those with high FSPC. Therefore, it is important to give

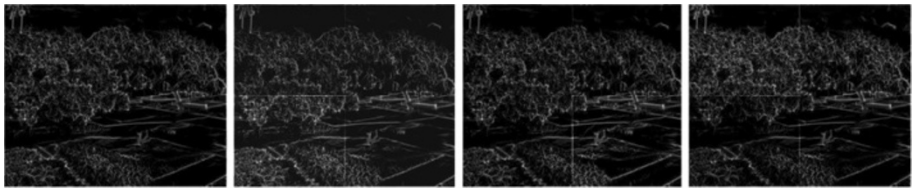

Figure 5 FSPCs corresponding to the images of Figure 4.

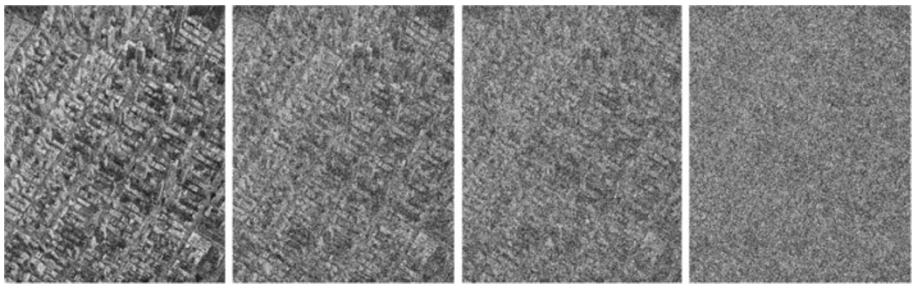

Figure 6 Noisy images with the SNR of 5.1728, 2.0026, 1.0864 and -0.47 respectively.

more confidence to the higher FSPC. In fact, the confidence component is the mean value of the two FSPCs, so the confidence highly relates with the significance of signals and will be given a larger value when both signals are significant. Hence, CAS_0 is normally given a relatively large value when the two pixels are both similar and significant and a relatively small value when they are not.

The CAS between two windows with a size of $(2n+1)$x $(2m+1)$ centered at (x, y) and *(u, v)* is given by

$$CAS_1(x,y;u,v) = C/2 - D \tag{15}$$

where:

$$C = \sum_{i=-n}^{n}\sum_{j=-m}^{m}((FSPC_1(x+i,y+j) + FSPC_2(u+i,v+j)) \tag{16}$$

$$D = \sum_{i=-n}^{n}\sum_{j=-m}^{m}|(MLPA_1(x+i,y+j) - MLPA_2(u+i,v+j)| \tag{17}$$

CAS_1 can be normalized so that its maximum value is equal to 1:

$$\begin{aligned} CAS_2(x,y;u,v) &= (C/2 - D)/(C/2) \\ &= 1 - 2D/C \end{aligned} \tag{18}$$

The equality can be further simplified as

$$CAS(x,y;u,v) = D/C \tag{19}$$

This measure returns 0 when the matching windows are identical. The denominator, C, is in fact related to the confidence component. For a same value of similarity D, the definition of CAS indicates a similarity is larger as the associated confidence components are high. It is apparent that CAS is invariant for the global linear illumination transformations: I→αI+b.

4. Implementation and experiments

The primary procedures for the proposed approach can be stated as follows. (1) Calculate the local frequency information by applying Log-Gabor wavelet transformation to raw multi-modal images; (2) Construct the local frequency representations—MLPA and FSPC based on the local frequency information; (3) Search the correspondence by minimizing the CAS between the template and the searching window. The values for the primary parameters used in the experiments are given in Table 1. These values are evaluated in a heuristic manner, and there is no need to change them for adapting different multi-modal scenes during the image matching. In the experiments, we notice that, for all parameters, a good value can be chosen across a relatively wide range of values. Actually, for the given wavelength of smallest scale filter, scaling factor between successive filters, cut-off value and gain factor, more wavelet scales and orientations can bring better experimental performance, but increase the computational time inevitably. Hence, we choose 4 and 9 as the number of wavelet scales and orientations respectively to compromise between performance and efficiency. To evaluate the performance of our method, we conduct numerous experiments using both synthetic and real images, and compare the experimental results with the state of the art methods, including four-directional derivative-energy image (FDDEI) [13], local frequency representation (LFR) [21], phase congruence (PC) [22], local symmetry score (LSS) [23], and mutual information (MI) [24]. In the experiments, joint histograms for calculating the MI are generated with 32×32 bins as suggested in

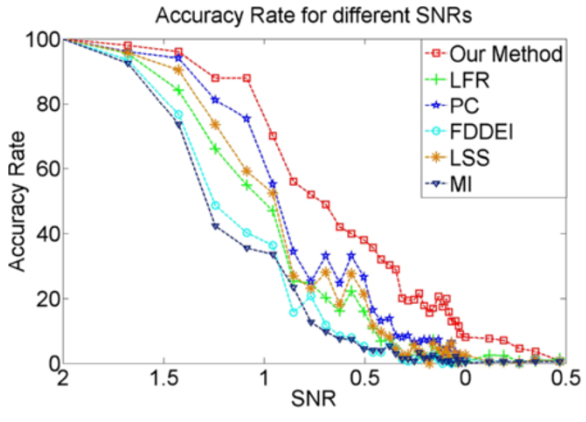

Figure 7 Accuracy Rates for different SNRs.

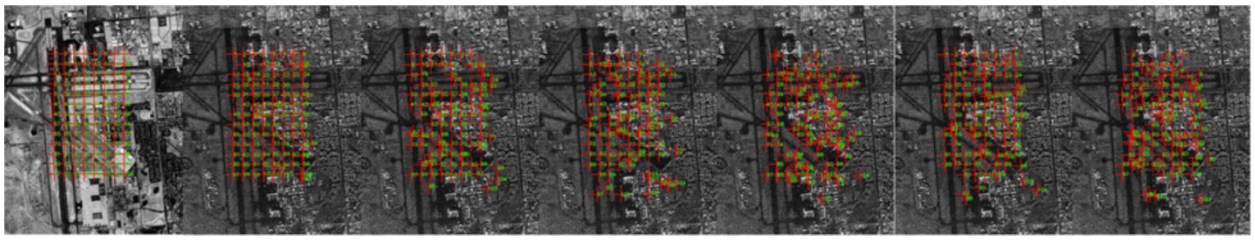

Figure 8 Matching results using the optical image and SAR. From left: Optical image (template center is labeled with the cross.); Our method; PC; LFR; FDDE; LSS; MI.

[11]. Local symmetry score is included due to its illumination invariant property and robustness to a range of dramatic variations. We adopt the product of the horizontal and vertical symmetry scores, which are based on a histogram of local gradient orientations and more stable to photometric changes, as the image representation, and the zero mean normalized cross correlation (ZNCC) as the similarity measure.

4.1. Illumination invariant property

Many visual and numerical experiments are first conducted to evaluate the illumination invariant property of the proposed MLPA and FSPC. The non-homogeneous illumination variation is synthesized by dividing an image into four equal parts and multiplying each part by a random scale factor to simulate the contrast variation and then adding a random constant factor to simulate the brightness variation. In Figure 4, we show a set of synthetic images with non-homogeneous illumination variation. We can observe the obvious contrast and brightness variation between different parts and different images. In Figures 2 and 5, we show the images of MLPA and FSPC corresponding to the images of Figure 4. As we can see, the illumination variation almost cannot be observed with unaided eye. At the boundary of each non-homogeneous illumination region, we can observe some straight line edges since the distribution of intensity values in the neighborhood of boundary is similar to that in the neighborhood of step edge.

To perform the numerical evaluation, we employ the normalized cross-correlation (NCC) to measure the similarity between the raw image and the synthetic image with non-homogeneous illumination variation. The definition of NCC can be expressed as

$$NCC = \frac{\sum_i \sum_j f(i,j)g(i,j)}{\sqrt{\sum_i \sum_j f^2(i,j) \sum_i \sum_j g^2(i,j)}} \quad (20)$$

where f and g denote the raw and synthetic images respectively. From Eq. 20, we can see that the value of NCC is highly related with the degree of non-homogeneous illumination variation. If there does not exist any non-homogeneous illumination variation, NCC will be given a maximum value of 1. The image of Figure 3 shows the results of numerical evaluation for gray-level images of Figure 4, MLPAs of Figure 2, and FSPCs of Figure 5. As we can see, the NCC values of MLPA and FSPC almost keep invariant to the non-homogeneous illumination variation, although the NCC values of gray-scale images are fluctuant with the varying degree of non-homogeneous illumination variation. The homogeneous illumination variation that can be considered as a type of non-homogeneous illumination variation is not particularly validated in this work. From the visual and numerical validation, we can clearly achieve the conclusion that both MLPA and FSPC can well keep invariant to non-homogeneous illumination validation.

4.2. Evaluation using synthetic images

We evaluate the matching accuracy and noise resistance using the synthetic images generated by adding the gaussian white noise generated using the imnoise function of Matlab 2010b to the raw images. The mean of

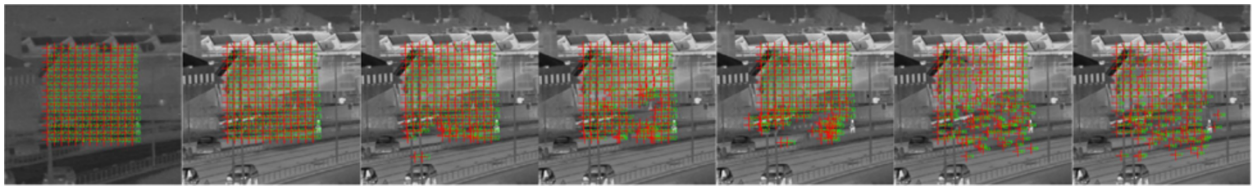

Figure 9 Matching results using the optical image and infrared image. From left: Optical image; Our method; PC; LFR; FDDE; LSS; MI.

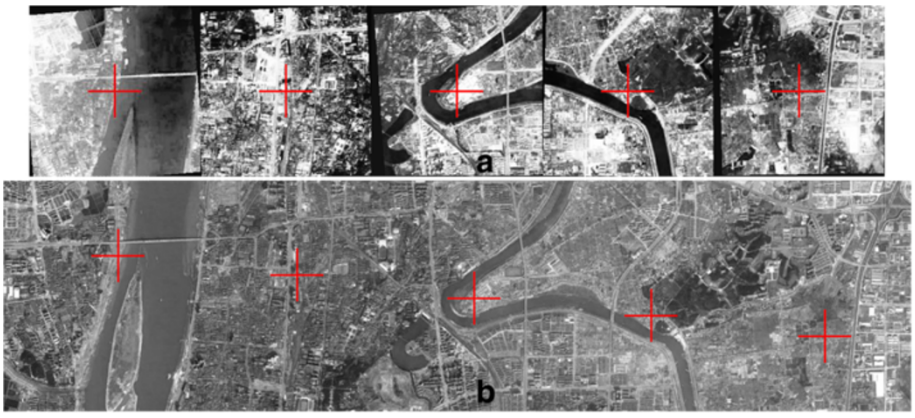

Figure 10 Matching results of the proposed method. (**a**) Aerial images of the intensifier charge coupled device (ICCD); (**b**) Reference image.

noise is given a same value of 0, and the variance is ascending from 0.1 to 3.5 gradually. Without loss of generality, we employ Signal to Noise Ratio (SNR) to describe the degree of noise. The definition of SNR is given as

$$SNR = 10 \times \log_{10}\left(\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(v(i,j))^2}{\sum_{i=1}^{M}\sum_{j=1}^{N}(u(i,j)-v(i,j))^2}\right) \quad (21)$$

where M and N denote the height and width of image, $v(i, j)$ and u(i, j) denote the intensity value of a pixel without and with noise respectively. The evaluation is performed as follows: (1) select a set of templates at 10-pixel intervals within the raw image; (2) search the corresponding points for the template centers in the noisy images using different methods. The raw image used for synthetic evaluation, whose content is composed of architecture, roads, vegetation, etc., is an optical satellite image with almost ideal imaging conditions. The sizes of raw image, template and search area are 1600×1200 (pixels), 101×101 (pixels) and 201×201 (pixels) respectively, and the total matching number is 26,825. The sizes of search area and template keep same to all methods for comparison equity. If the Euclid distance between the matching result and the ground truth is less than 2 pixels, we identify the matching result as correct. The experimental images with different degrees of noise are shown in Figure 6. As we can see, the image becomes more and more blurred as SNR decreases, and when the SNR decreases to -0.47, the image content almost cannot be identified with unaided eye. The accuracy rates obtained from different methods for different SNRs are shown in Figure 7. When SNR is larger than 2, all methods are not influenced since the smoothing effect of the relatively large template. And then the accuracy rates begin decreasing with the increase of

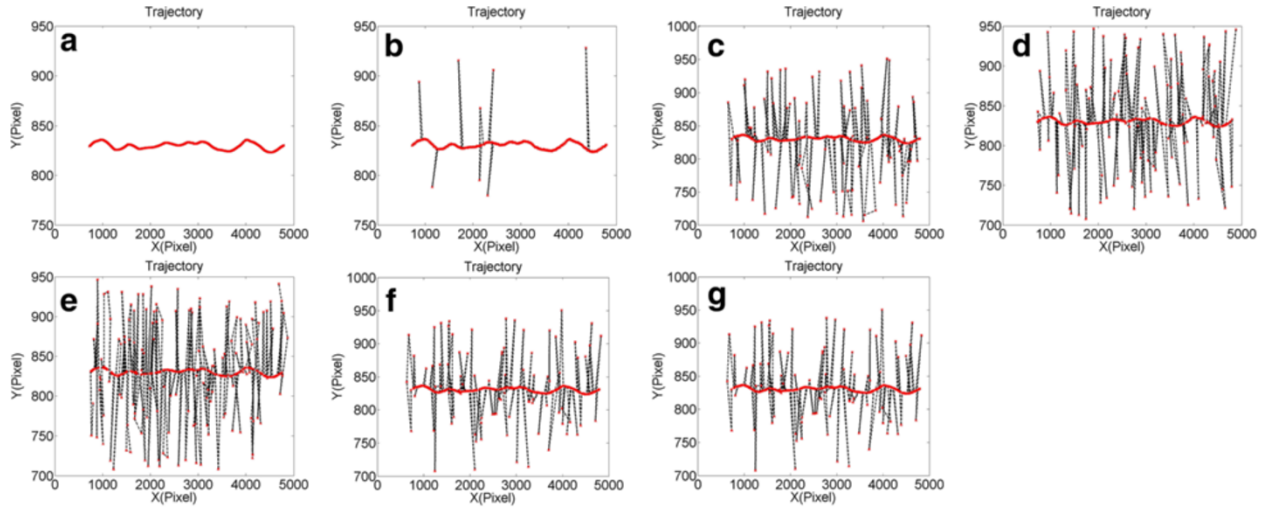

Figure 11 Comparisons of flight trajectories obtained from different methods. (**a**) GPS; (**b**) Our method; (**c**) PC; (**d**) LFR; (**e**) FDDE; (**f**) LSS; (**g**) MI.

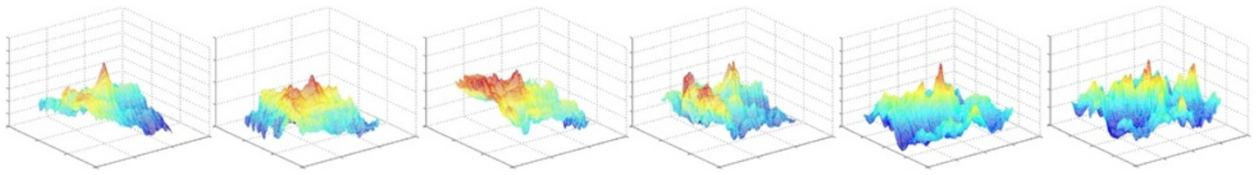

Figure 12 Correlation surfaces for the images of Figure 1a. From left: our method; PC; LFR; FDDE; LSS; MI.

noise degree, but the accuracy rates of conventional methods decrease more quickly than our method.

4.3. Matching accuracy evaluation using real images

To evaluate matching accuracy comprehensively and objectively, we perform numerous experiments using many real multi-modal image pairs. The database used for image matching includes 254 pairs of infrared and optical images and 52 pairs of SAR and optical images with a wide range of illumination and significant appearance changes caused by photometric and physical imaging properties of different sensors. The images of Figures 8 and 9 show two sets of matching results obtained from different methods, and in Table 2 we give the accuracy rates corresponding to the image pairs of Figures 8 and 9. As we can see in Figures 8, 9 and 10, the non-homogeneous contrast and brightness variation occurs frequently between the multi-modal image pairs, but the structural information still keeps common and reliable, for example, the edges of airport runway in Figure 8, the contours of cars, architecture, lampposts, persons and etc. in Figures 1 and 9. For the optical and infrared image pair of Figure 9, the conventional methods work well since the relatively significant feature and contrast variation, but the performance of conventional methods is degraded dramatically while handling the optical and SAR image pair of Figure 8 that has harsh speckle noise and contrast variation. The proposed method works reasonably well for those two situations.

We have applied the proposed method to scene matching used for Unmanned Aerial Vehicle (UAV) positioning, and conducted a total of 10,080 scene matching experiments using aerial images obtained by electro-optic pods, including 6,180 infrared images and 3,900 ICCD images. The ground scene consists of architecture, river, vegetation, farmland, highland, etc. The imaging time-range is day-and-night, and the imaging altitude ranges from 150 meters to 2000 meters. Since the reference image, obtained from space-borne optical sensor with a spatial resolution of 4 meters, normally has a relatively slow update rate, the aerial images and reference images are normally several years apart, and with dramatic differences in ground scenes (e.g. appearance/disappearance of architectures, growth/witherer of vegetation, drought/waterlogging of rivers, etc.), as well as changes caused by different sensors. In Figure 10, we show a set of results of scene matching. The geometric distortion caused by imaging attitude and altitude is eliminated by using the information of INS and altimeter. The truth values of scene matching are provided by GPS, which generally has a positioning accuracy better than 1 meter. If the difference between the scene matching result and GPS is less than 8 meters (2 pixels), we identify the result as correct; otherwise it fails. According to the criterion, the accuracy rate of our scene matching is 96.63%, which is well within the requirements for engineering, whereas the accuracy rates of PC, LFR, FDDE, LSS and MI are 85.78%, 82.69%, 76.73%, 83.24% and 74.89% respectively. In Figure 11, we show a set of flight trajectories measured by different methods. As we can see, the results of our method are more coincident with GPS. Very few false matches existing in our results can be effectively eliminated by the filtering operation (e.g. Kalman filter).

The shape of correlation surface is related to the confidence of matching result. We examine numerous correlation surfaces computed from multi-modal image pairs randomly chosen from the database of image matching and scene matching. In Figure 1a,b, we show two matching results for the optical & SAR images and the infrared & optical images, respectively, and in Figures 12 and 13, we show the correlation surfaces obtained from different methods for the matching results of Figure 1a,b, respectively. For the correlation matrix whose optimum

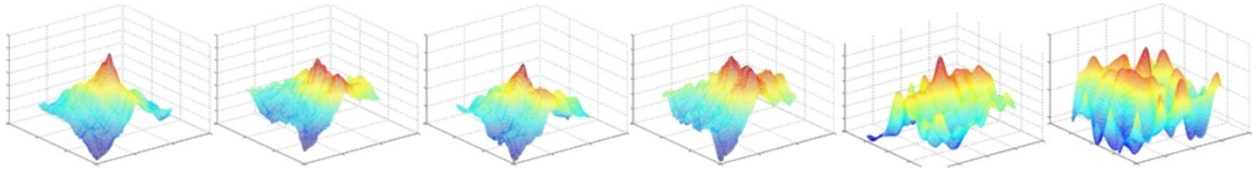

Figure 13 Correlation surfaces for the images of Figure 1b. From left: our method; PC; LFR; FDDE; LSS; MI.

corresponds to the minimum value, we reverse the correlation surface using the transformation: $M(i, j) \rightarrow M_{\max} - M(i, j)$ to transfer the optimum to the maximum value. For all correlation matrixes obtained from different methods, we use the transformation: $M(i, j) \rightarrow M(i, j)/M'_{\max}$ to transfer the maximum value to 1. Obviously, we can see that the surface of our method has fewer peaks and more distinct maximum. The conventional methods give a maximum peak not very dominant unlike the surface of our method for which the maximum stands out from the rest of the surface. In addition, the maximum peak of our method is narrower, and therefore can provide better localization ability.

It should be noted that the proposed method performs better than MI. The underlying assumption of MI is that the statistical relationship between the matching images is homogeneous over the whole image domain. It is normally true when intensities mapping between matching images is global and highly correlated or when structures with different intensities in one image have similar intensities in the other image, e.g. bond and background in CT and MR. However, the statistical relationships of intensities between multi-modal image pairs are normally not global and non-homogeneous as discussed above, which are quite different from the medical images. Therefore, MI may not be sufficient for matching multi-modal images. In addition, the absence of local spatial information in MI also weakens the matching robustness to some extent.

Since symmetries are a potentially robust and stable feature of many man-made and natural scenes, which makes it suitable to represent multi-modal images, LSS designed for scoring local symmetries whose performance is almost compatible with PC works reasonably well in our experiments, although its primary goal is to extract local features from images of architectural scenes.

From the evaluation using synthetic and real images, we can achieve the conclusion: since the considerations of noise resistance, illumination adaptability and common structure extraction and weighting, the proposed method can achieve higher accuracy rate, better matching confidence than the conventional methods for the test images used.

5. Conclusion

To achieve robust multi-modal image match, we first present two image representations—FSPC and MLPA based on the Log-Gabor wavelet transformation, and then design the CAS that combines confidence and similarity by using the information of FSPC and MLPA to find the correspondence. The proposed method has three main merits: (1) both MLPA and FSPC keep invariant for non-homogeneous illumination (contrast, brightness) variation and contrast reversal that frequently occur between multi-modal images; (2) FSPC can effectively capture the common scene structural information while suppressing the non-common sensor-dependent properties; (3) As the confidence factor, the structural information extracted by FSPC can be allocated more weighting softly by CAS. In addition, the proposed method is threshold-free, and therefore can retain as much image detail information as possible to resist noise influence and scene distortions between images. Experiments using numerous real and synthetic images demonstrate that our method can match multi-modal images robustly. Through comparison experiments, we also demonstrate the advantage of our method over the conventional methods. In the future, we plan to introduce the geometric transformation into our matching frame, and extend our method to image alignment.

Competing interests

The authors declare that they have no competing interests.

Acknowledgment

This work was partly supported by Oulu University, Finland. The authors would like to thank Prof. Janne Heikkila, Dr. Jie Chen and Guoying Zhao for their contributions. The authors also want to express their gratitude to the anonymous reviewers whose thoughtful comments and suggestions improved the quality of the article.

Received: 18 September 2012 Accepted: 18 December 2012
Published: 8 January 2013

References

1. G Conte, P Doherty, Vision-based unmanned aerial vehicle navigation using Geo-referenced information. EURASIP J. Adv. Sig. Process. **2009**, 1–18 (2009)
2. Z Kalal, K Mikolajczyk, J Matas, Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intel. **6**(1), 1–14 (2010)
3. P Vandewalle, S Susstrunk, M Vetterli, A frequency domain approach to registration of aliased images with application to super-resolution. EURASIP J. Adv. Sig. Process. **2006**, 1–14 (2006)
4. M Brown, D Lowe, "Unsupervised 3D object recognition and reconstruction in unordered datasets," in proc. Int. Conf. 3-D digit. Imag. Model, 56–63 (2005)
5. D Yingzi, B Craig, Z Zhi, "Scale invariant Gabor descriptor-based noncooperative iris recognition". EURASIP J. Adv. Sig. Process. **2010**, 1–13 (2010)
6. Y Yang, P Dong Sun, H Shuying, R Nini, "Medical image fusion via an EffectiveWavelet-based approach". EURASIP J. Adv. Sig. Process. **2010**, 1–13 (2010)
7. A Wong, J Orchard, Robust multi-modal registration using local phase-coherence representations. J. Sign. Process. Syst. **54**, 89–100 (2009)
8. WU Yingdan, MING Yang, "A multi-sensor remote sensing image matching method based on SIFT operator and CRA similarity measure". Proceedings of 2011 International Conference on Intelligence Science and Information Engineering, 115–118 (2011)
9. W Sasa, Z Zhenbing, Y Ping, G Zejing, "Infrared and visible image matching algorithm based on NSCT and DAISY". Proceedings of 2011 4th International Congress on Image and Signal Processing **4**, 2072–2075 (2011)
10. S Yong, H Jae, B Jong, Multi-sensor image registration based on intensity and edge orientation information. Pattern Recognition **41**, 3356–3365 (2008)
11. J Inglada, "Similarity measures for multi-sensor remote sensing images". Proceedings of Geoscience and Remote Sensing Symposium, Toulouse **5236**, 182–189 (2001)
12. P Kovesi, "Image correlation from local frequency information". Proceedings of the Australian Pattern Recognition Society Conference **1995**, 336–341 (1995)
13. PAM Irani, "Robust multi-sensor image alignment". Proceedings of the 6th International Conference on Computer Vision, 959–966 (1998)
14. PW Josien, JB Pluim, M Antoine, MA Viergever, "Mutual-information-based registration of medical images: a survey". IEEE Trans. Med. Imag. **22**(8), 986–1004 (2003)

15. PA Estévez, M Tesmer, CA Perez, JM Zurada, Normalized mutual information feature selection. IEEE Trans. Neural Netw. **20**(2), 189–201 (2009)
16. J Liu, BC Vemuri, F Bova, Efficient multi-modal image registration using local-frequency maps. Mach. Vis. Appl. **13**, 149–163 (2002)
17. J Morlet, G Arens, E Fourgeau, D Giard, Wave propagation and sampling theory-part II: sampling theory and complex waves. Geophysics **47**(2), 222–236 (1982)
18. MC Morrone, JR Ross, DC Burr, RA Owens, Mach bands are phase dependent. Nature **324**(6094), 250–253 (1986)
19. MC Morrone, RA Owens, Feature detection from local energy. Pattern. Recognit. Lett. **6**, 303–313 (1987)
20. P Kovesi, Phase congruency: a low-level image invariant. Psychol. Res. **64**, 136–148 (2000)
21. MI Elbakary, MK Sundareshan, Multi-modal image registration using local frequency representation and computer-aided design (CAD) models. Image Vis. Comput. **25**, 663–670 (2007)
22. L Zheng, L`r Robert, "Phase congruence measurement for image similarity assessment". Pattern. Recognit. Lett. **28**, 166–172 (2007)
23. H Daniel Cabrini, S Noah, "Image matching using local symmetry features". Proc. CVPR, 206–213 (2012)
24. P Viola, WM Wells, "Alignment by maximization of mutual information". Proc. ICCV, 16–23 (1995)

doi:10.1186/1687-6180-2013-3
Cite this article as: Liu *et al.*: **Multi-modal image matching based on local frequency information.** *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:3.