electronics

MDPI

Article

Infrared and Visible Image Registration Based on Automatic Robust Algorithm

Jingyu Ji [1], Yuhua Zhang [1], Zhilong Lin [1], Yongke Li [1], Changlong Wang [1,*], Yongjiang Hu [1,*] and Jiangyi Yao [2]

[1] Department of UAV, Army Engineering University, Shijiazhuang 050003, China; jijingyu@aeu.edu.cn (J.J.); ycenlgd@aeu.edu.cn (Y.Z.); zdy133127@aeu.edu.cn (Z.L.); lxwys@aeu.edu.cn (Y.L.)
[2] Equipment Simulation Training Center, Army Engineering University, Shijiazhuang 050003, China; yaojiangyi@aeu.edu.cn
* Correspondence: lwrit321@aeu.edu.cn (C.W.); mathzhz@aeu.edu.cn (Y.H.)

Abstract: Image registration is the base of subsequent image processing and has been widely utilized in computer vision. Aiming at the differences in the resolution, spectrum, and viewpoint of infrared and visible images, and in order to accurately register infrared and visible images, an automatic robust infrared and visible image registration algorithm, based on a deep convolutional network, was proposed. In order to precisely search and locate the feature points, a deep convolutional network is introduced, which solves the problem that a large number of feature points can still be extracted when the pixels of the infrared image are not clear. Then, in order to achieve accurate feature point matching, a rough-to-fine matching algorithm is designed. The rough matching is obtained by location orientation scale transform Euclidean distance, and then, the fine matching is performed based on the update global optimization, and finally, the image registration is realized. Experimental results show that the proposed algorithm has better robustness and accuracy than several advanced registration algorithms.

Keywords: image extraction; image matching; deep convolutional network; infrared and visible image; image registration

Citation: Ji, J.; Zhang, Y.; Lin, Z.; Li, Y.; Wang, C.; Hu, Y.; Yao, J. Infrared and Visible Image Registration Based on Automatic Robust Algorithm. *Electronics* **2022**, *11*, 1674. https://doi.org/10.3390/electronics11111674

Academic Editors: Miin-shen Yang and Cocianu Catalina-Lucia

Received: 19 April 2022
Accepted: 20 May 2022
Published: 25 May 2022

1. Introduction

In recent years, Unmanned Aerial Vehicles (UAV) have played an increasingly important role in many fields due to their high flexibility, low cost, and easy operation [1]. In the military, they are often utilized to perform reconnaissance, battlefield situation monitoring, and other tasks. Since infrared images have thermal radiation properties and visible images have light reflection properties, if the two are accurately registered and fused, the result not only preserves the clear details and edges of the visible image but also preserves the brightness information in the infrared image, making the target easier to identify [2]. As these two sensors are more common, at present, small UAVs are usually equipped with infrared and visible sensors for target detection or tracking. Using the above principles can make it easier for small UAVs to lock the target [3]. However, due to differences in the time period, distance, shooting angle, etc., of UAV aerial photography, the images obtained by multi-source sensors may not be strictly aligned due to the existence of translation, rotation, scaling, and other spatial transformation relationships. Therefore, registration needs to be carried out before fusion [4]. However, the information has a low overlap in the reconnaissance area, so the inconsistence of the information amplitude and resolution, large differences in viewing angles, etc., due to the constraints of terrain, time, climate, UAV flight trajectory, and other conditions, as well as the mutual constraints between the various loads when the infrared and visible sensors work simultaneously, mean that image registration is still a difficult task.

Existing image registration technology can be roughly divided into three categories: calibration-based registration technology [5], region-based registration technology [6],

and feature-based registration technology [7]. Calibration-based methods are fixed at a specific distance, so they are not suitable for UAV image registration applications. Region-based methods usually use the whole image information to establish a similarity measure of two images. They directly judge the intensity differences between images without extracting salient features. Under the condition that the image has a certain deformation in the local area or the heterogeneous image, the region-based registration methods usually fail to match successfully, so they are not suitable for the research in this paper. The feature-based methods are the most commonly used registration methods at present, and they are also a class of methods that have been widely improved by scholars. The algorithm in this paper is also an image registration algorithm based on feature point design. The detailed research process of image registration will be described in Section 2.

The above methods have been widely discussed in the past; however, registration of infrared and visible images for UAVs is still a difficult task. First, the spectra of infrared images is different from visible images, resulting in large differences in the corresponding regions and difficult registration implementation. Second, the resolutions of infrared images are different from visible images, resulting in a point in an infrared image that may not have matching areas in a visible image. Third, the low resolution of infrared images will lead to difficulty in feature point extraction. Fourth, there are viewpoint differences between infrared images and visible images, which will lead to difficulty in image matching.

In response to the above problems, a new feature point-based deep convolution automatic robust registration algorithm, named the deep convolutional network–rough to fine registration algorithm, was proposed for infrared and visible images. First, the feature points are fully extracted by the improved convolutional neural network, which solves the problem of feature point location between images; secondly, the problem of resolution difference is solved by the rough to fine feature matching method, which obtains an accurate image transformation matrix and solves the problem of inaccurate matching. The proposed registration algorithm fully utilizes image features and combines deep convolutional networks and the R2F method, which achieves accurate registration of infrared and visible images on UAVs. The main contributions of this paper are:

(1) A multi-scale feature descriptor is generated by utilizing a pre-trained deep convolutional network to obtain the feature points of the images. The advantages of convolutional neural network are effectively utilized in feature extraction to obtain accurately positioned feature points.

(2) A rough-to-fine feature point matching method is designed, which introduces the concept of location orientation scale transform Euclidean distance and fine matching, based on update global optimization, to obtain high-precision registration images.

(3) The feature point extraction comparison experiment and the image registration comparison experiment, respectively, prove that the proposed algorithm has a good performance in the feature point extraction and the overall image registration effect, which verifies the effectiveness and robustness of the proposed algorithm.

The paper is organized as follows. Section 2 outlines the research progress of image registration algorithms. Section 3 introduces the proposed registration algorithm, which mainly consists of deep convolutional feature extraction and the rough to fine method. The experiment results and analysis are presented in Section 4. Finally, the conclusion is given in Section 5.

2. Related Works

Image registration technology was developed in the 1970s, and it was generally used in military fields, such as missile guidance and aircraft navigation systems, in the early stage of development. With the advancement of software and hardware technology, image registration technology is applied in more and more fields. After years of development and accumulation, many excellent research results have been produced in the field of image registration. Initially, image registration algorithms were mostly based on image grayscale. Barnea et al. [8], in 1972, proposed the use of similarity metric functions to match images.

In 1995, Viola et al. [9] introduced the concept of mutual information, which opened a new door for image registration, and scholars have performed a great deal of research on this basis. For example, the multi-source medical image matching method, based on mutual information proposed by Maes et al. [10] in 1997, still inspires the processing of medical images. In 2005, Fan et al. [11] combined the advantages of wavelet transform and the mutual information method to propose an image registration method, which had gained extensive attention in the field of multi-source remote sensing registration images. The advantage of the grayscale-based image registration algorithm is that it is relatively simple to implement and has strong robustness; the disadvantage is that it requires significant computation and requires a strong grayscale correlation between images. It is difficult to obtain the best registration effect for grayscale information.

With the continuous development of technology, a large number of new image registration schemes continue to emerge, multi-source image registration research has been vigorously developed, and the image registration accuracy has also been greatly improved. At present, the research of image registration technology can be roughly divided into three categories: calibration-based image registration technology, region-based image registration technology, and feature-based image registration technology.

Calibration-based methods rely on calibrated cameras and can simply align images taken at the same time and view. The registration error of these methods is fixed at a certain distance, so it is not suitable for registration applications of UAV images. Unlike calibration-based methods, region-based registration methods and feature-based registration methods are automatic. However, region-based methods usually use the whole image information to establish a similarity measure of two images, directly compare and match the intensity difference between the images, and do not extract salient features. There are generally three region-based methods: the cross-correlation methods, the mutual information methods, and the Fourier methods. The registration accuracy of such methods can reach the pixel level, but there are occlusion and affine transformations. The region-based registration methods usually fail to match successfully under the conditions of local deformation of the images or with heterogeneous images.

Feature-based image registration methods are important tools for solving image registration problems due to their good invariant properties. In 1981, Moravec [12] proposed a method to detect the corners of image contours. In 1988, Harris [13] was inspired by it and developed a corner detection method that is not affected by image rotation, which is known as the famous Harris operator. Then, Lindeberg et al. [14] proposed scale space theory to solve the problem of scale invariance, and they designed operators, such as Hessian-Laplace and Harris-Affine, to perform affine transformation on images. After summarizing the previous research results and learning from each other's strengths, Lowe proposed the SIFT operator [15], with trans-epoch significance, in 2004. This operator can solve the problem of image registration in most complex situations, and it has been influential to this day. At the ECCV conference in 2006, Bay first proposed another famous operator: the SURF operator [16]. The SURF operator is an improvement of the SIFT algorithm, which greatly reduces the computational complexity in the feature extraction process, and it has higher robustness. After years of development, scholars have also proposed many excellent algorithms, such as BRISK [17], ORB [18], and multiple phase congruency directional patterns [19], which make the feature-based image registration methods more widely used.

With the widespread development of deep learning, convolutional neural networks (CNNs) have been utilized for image registration tasks [20]. In 2017, Ma et al. [21] proposed a feature registration method for full image representation based on CNN features. The features of the CNN were used to find keyframes with a similar appearance from the topological map. Then, the geometric features were checked by the consistency of the vector field to obtain the most similar key features and achieve the matching performance. DeTone et al. [22] used an end-to-end neural network to learn the homography between images, showing the superiority that is difficult to achieve by traditional image registration. The algorithm learned the homography between network parameters and images at the

same time, and it obtained the homography between images by outputting the offset of four coordinate points. Japkowicz et al. [23] proposed a dual convolutional neural network for image registration. Four of the convolutional layers are used to process two images simultaneously, and the other four are used to concatenate the feature maps to generate homography estimates, resulting in superior accuracy. However, because the network processes a pair of images in parallel and is trained hierarchically, the amount of parameters and computation of the network is greatly increased. Ty Nguyen et al. [24] proposed an unsupervised learning neural network, which achieved more robustness by computing the loss by exploiting the similarity between images. Li et al. [25] proposed a multiple vector (VLAD) encoding method, with local classification descriptors and CNN features, for image classification. Wen et al. [26] proposed a depth-guided color-coarse-to-fine image processing method based on convolutional neural networks, which solves the phenomenon of texture duplication and can effectively reserve the edge details of super-resolution images.

In summary, the image registration algorithm based on feature points is suitable for the research in this paper. Through the analysis of the literature, it is found that the above algorithm still has some shortcomings for the registration of infrared and visible images of UAVs. In this paper, the feature point-based image registration algorithm is used, and the feature point extraction is realized by using the advantages of the convolutional neural network for feature point location. Then, the traditional algorithm is used for subsequent matching and transformation.

3. Materials and Methods

In order to improve the problem of poor registration performance, a new image registration model is framed, as shown in the Figure 1. Different from the traditional image registration algorithm, in order to better extract feature points and perform correct matching, firstly, the deep convolution feature extraction network model is utilized to extract feature points from the image (i.e., step 1 in Figure 1), and then, the concept of location orientation scale transform Euclidean distance is introduced. The extracted feature points are roughly matched (i.e., step 2 in Figure 1); finally, a fine matching, based on update global optimization, is introduced to reduce the positional deviation of the feature points (i.e., step 3 in Figure 1). After the above operations, all the feature points corresponding to the infrared and visible images are correctly matched, and the final registered image pair is obtained by corresponding to the images.

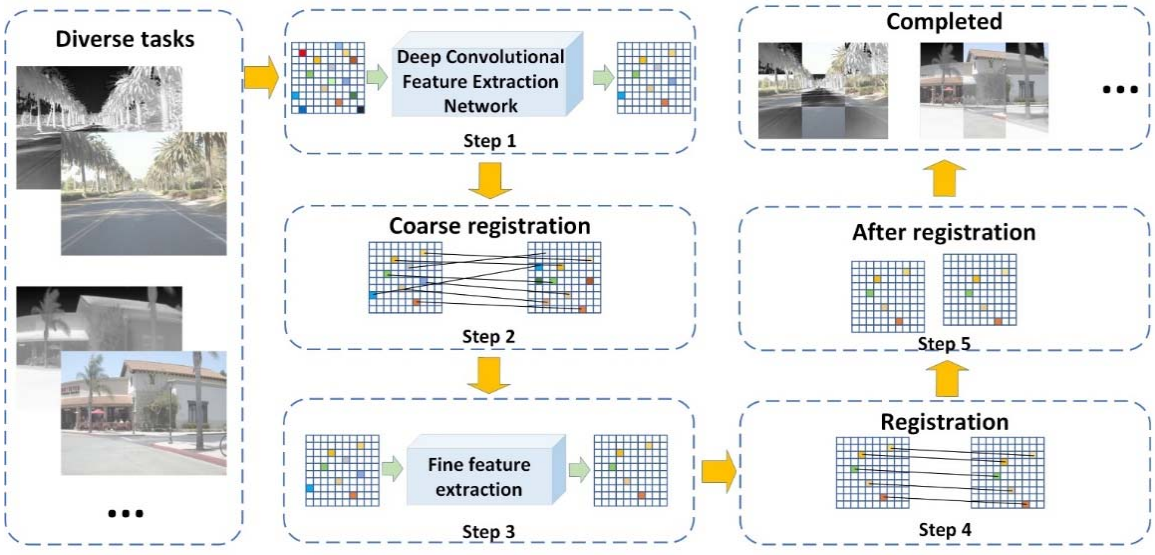

Figure 1. The algorithm framework of the article.

3.1. Feature Point Extraction by Deep Convolutional Network

The VGG-16 [27] network is an image classification network that can classify a large number of categories. It is often utilized in various computer vision feature extraction links. Its advantages are: (1) it has excellent image resolution ability; (2) it relies on connecting convolutional layers, pooling layers, and fully connected layers to build a network model. The structure is very simple and concise, so the network can be used for a variety of image processing tasks. (3) Its network structure is deep and can be trained with a large amount of diverse image data. Therefore, partial convolutional layers of VGG-16 were adopted, in this paper, to construct a deep convolutional network for feature descriptor extraction. The VGG-16 network is frequently utilized for feature detection in the field of image processing, such as the automatic detection of corn kernels from UAV images using the VGG-16 network [28] and Super-Resolution Generative Adversarial Networks (SRGAN) [29].

In order to obtain a single feature descriptor output, while taking into account the universality of the convolution filter and the size of the receptive field, multiple network layers are selected to construct the experimental model. The size of the input image can be arbitrarily set to a multiple of 32, but it may affect the computational efficiency, make the receptive fields of each feature point different, and even influence the performance of the network construction. Therefore, in order to maximize the network performance and be computationally efficient enough, the size of the input image is set to 224 × 224. The maximum pooling layer *pool*5_1, added after the three output layers *pool*3, *pool*4, and *pool*5, is utilized to construct the deep convolutional output network in this paper. Compared with the original VGG-16 model, our model removes the fully connected layer and adds a *pool*5_1 layer after the *pool*5 layer that can detect more general features.

As shown in Figure 2, the network structure model of this paper contains five convolution blocks, the first two blocks contain two convolutional layers, the third and fourth convolutional blocks contain three convolutional layers, and the last convolutional block contains one convolutional layer, each with a max-pooling layer at the end of them. A *Pool*5_1 layer, as a max pooling layer, is added to the end. A 28 × 28 grid is utilized to divide the input image into blocks, where each block corresponds to a 256-dimensional vector in the output of *pool*3, which is also the feature descriptor of *pool*3, and a central feature descriptor is produced by each 8 × 8 square. The feature map M_1 is directly obtained from the output of *pool*3. Different from how the *pool*3 output layer is processed, the feature map M_2 output by the *pool*4 layer (with a size of 14 × 14 × 512) is shared by four feature points in each 16 × 16 region, obtained by Kronecker (denoted by ⊗):

$$M_2 = OUT_{pool4} \otimes I_{2\times2\times1} \tag{1}$$

where OUT_{pool4} represents the output of the *pool*4 layer, I represents the tensor subscript shape and fills it with 1s.

The feature map M_3 of the *pool*5_1 layer is possessed by 16 feature points, which are represented as:

$$M_3 = OUT_{pool5\_1} \otimes I_{4\times4\times1} \tag{2}$$

Figure 3 shows the distribution of the above key points-descriptors. The gray circles represent M_1, which are produced in each 8 × 8 region. The blue circles represent M_2, which are generated in each 16 × 16 region. The yellow circle represents M_3, which are generated in the 32 × 32 region. Additionally, the shared relationship between them can be clearly seen in the figure.

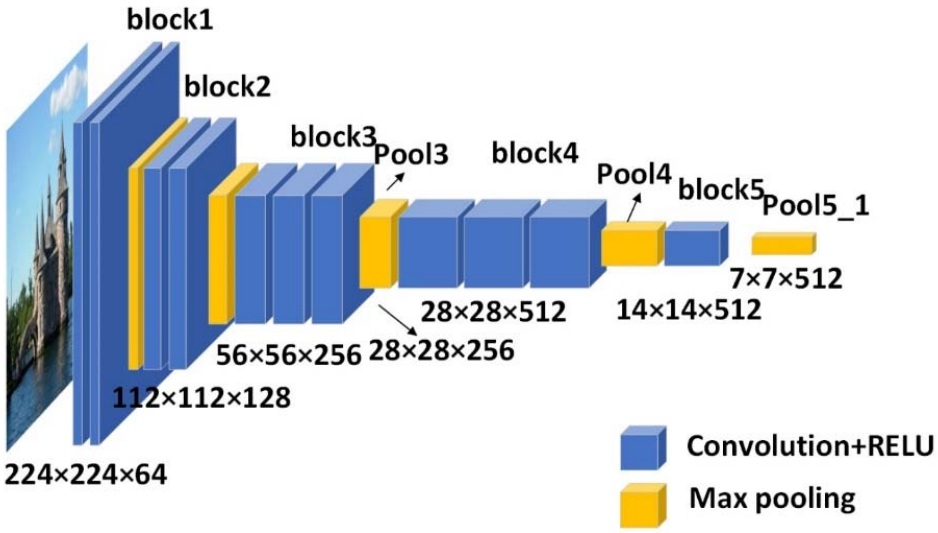

Figure 2. The structure of the deep convolutional network in this paper.

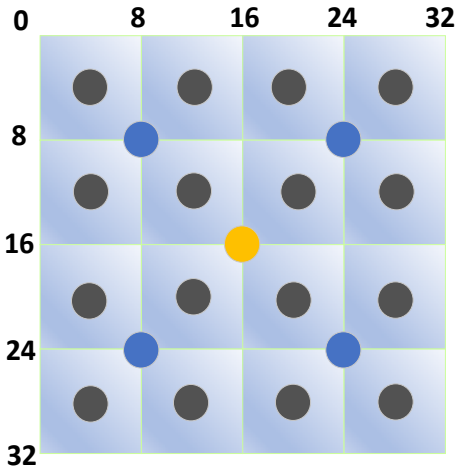

Figure 3. The distribution of feature descriptors.

After obtaining M_1, M_2, and M_3, the feature maps are normalized to unit variance.

$$M_i \leftarrow \frac{M_i}{\sigma(M_i)},\ i = 1, 2, 3 \tag{3}$$

where $\sigma(\cdot)$ represents the standard deviation of each element in the matrix. The descriptors, *pool*3, *pool*4, and *pool*5_1, of point a are represented by $F_1(a)$, $F_2(a)$, and $F_3(a)$, respectively.

3.2. Rough-to-Fine Feature Point Matching

3.2.1. Rough Matching

The spatial transformation of images usually includes translation, rotation, and scaling. Under the spatial transformation model, the correct feature point matching has the same position, main orientation, and scale in most cases. Therefore, correct matching can be performed by judging the spatial transformation information of each feature point in the two images.

Two feature point sets, $F = f_1, f_2, \cdots, f_M$ and $F' = f'_1, f'_2, \cdots, f'_N$, are extracted from the visible and infrared reference images, respectively. (x_i, y_i), α_i, and r_i, respectively, represent the position, main direction, and zoom scale of the key point f_i in the visible reference image. (x'_j, y'_j), α'_j, and r'_j, respectively, represent the position, main direction, and

scale of the key point f'_j in the infrared images. The position transformation error ε_p of f_i and f'_j, of the corresponding feature points, is represented:

$$\varepsilon_p\left(f_i, f'_j\right) = \|(x_i, y_i) - S(\left(x'_j, y'_j\right), \gamma)\| \tag{4}$$

where $S(\left(x'_j, y'_j\right), \gamma)$ is the spatial transformation model, and γ is the transformation variable. The main direction transformation error and scaling transformation error of feature points are expressed as

$$\varepsilon_\alpha\left(f_i, f'_j\right) = abs(\Delta\alpha_{i,j} - \Delta\alpha^*), \;\; \varepsilon_r\left(f_i, f'_j\right) = \left|1 - (s^*)\frac{r'_j}{r_i}\right| \tag{5}$$

where $\Delta\alpha^*$ and s^* represent the main direction difference and scaling transformation between the visible and infrared reference images, respectively, and $\Delta\alpha_{i,j} = \alpha_i - \alpha'_j$ represents the difference of main direction between f_i and f'_j. These parameters can be obtained from the histograms. In addition, r_i and r'_j represent the scale of the key point f_i and scale of the key point f'_j, respectively, and they can be obtained from the algorithm commands. Next, a robust connection distance called location orientation scale transform Euclidean distance (LOSTED) is defined as:

$$\text{LOSTED}\left(f_i, f'_j\right) = \left(1 + \varepsilon_p\left(f_i, f'_j\right)\right)\left(1 + \varepsilon_\alpha\left(f_i, f'_j\right)\right)\left(1 + \varepsilon_r\left(f_i, f'_j\right)\right)ED\left(f_i, f'_j\right) \tag{6}$$

where $ED\left(f_i, f'_j\right)$ represents the Euclidean distance of the descriptors in the feature points f_i and f'_j. Additionally, $\varepsilon_p\left(f_i, f'_j\right)$, $\varepsilon_\alpha\left(f_i, f'_j\right)$, and $\varepsilon_r\left(f_i, f'_j\right)$ represent the position transformation error, main direction transformation error, and scaling transformation error of the corresponding feature points mentioned above, respectively. LOSTED is minimal in most scenarios, as point pairs are matched accurately. The rough matching process in this paper is as follows:

(1) Initial matching: A ratio threshold is set as T, and the ratio of the nearest neighbor Euclidean distance to the next nearest neighbor Euclidean distance of the corresponding feature point is calculated. Then, we compare the calculated ratio with T and match the feature points that meet the threshold to obtain the key point pair set FF'. To build histograms of horizontal displacement, vertical displacement, scaling scale, and principal direction difference, the image transformations Δx^*, Δy^*, $\Delta\alpha^*$, and s^* are obtained from the histograms, as shown in Figure 4. According to the description in [30], the FSC algorithm can find the largest consistent sample set from the extracted sample set by setting the threshold relationship, and then finding the corresponding relationship through the transformation error, so the transformation parameters can be obtained. Therefore, we use this algorithm to calculate the initial transformation parameter γ from the feature point pair set FF'.

(2) Rematch: Since the circumferential angles at -180° and 180° are not continuous [see Figure 5], there are two main situations of the main direction difference. Actually, there should only be one main modal for the rotation angle. Once one of the two models is known, the other can be calculated as:

$$\Delta\alpha' = \begin{cases} \Delta\alpha + 360^\circ, \; \Delta\alpha \in [360^\circ, 0^\circ) \\ \Delta\alpha - 360^\circ, \; \Delta\alpha \in [0^\circ, 360^\circ) \end{cases} \tag{7}$$

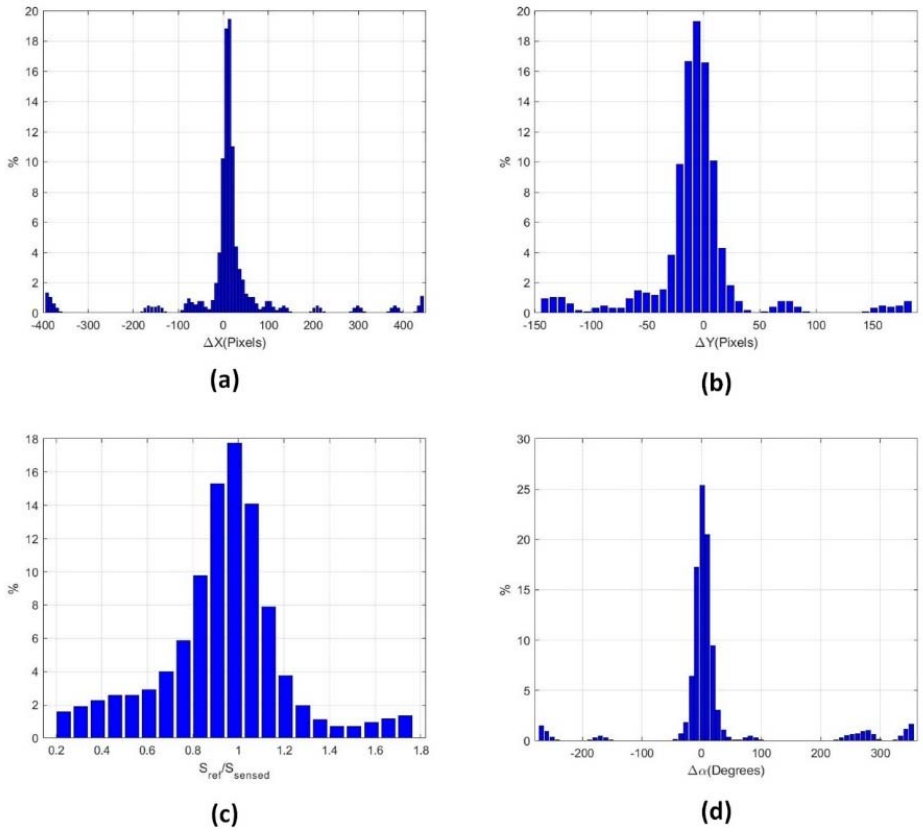

Figure 4. (**a**–**d**) Histograms of horizontal shifts, vertical shifts, scale ratio, and main direction difference.

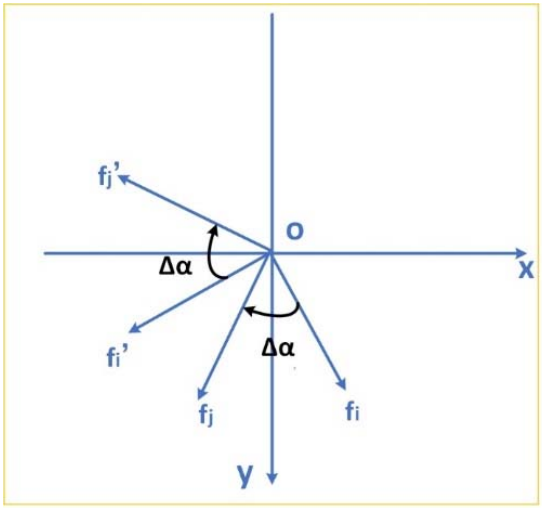

Figure 5. Main direction modal diagram.

$\Delta\alpha$ and $\Delta\alpha'$ represent the two angles in the main direction difference histogram, respectively. Figure 5 presents the two angles of the principal direction difference and the single mode of the rotation angle.

Based on the above, there are two combinations of Δx^*, Δy^*, $\Delta\alpha^*$, and s^*. For each combination, the distance metric is performed by LOSTED, and keypoints are matched by the ratio of the nearest Euclidean neighbor distance to the next nearest Euclidean neighbor distance. The ratio threshold is set as T'. Since two matchings are done, the feature points of one image may be matched more than once in the other image, so the point pair with the

smallest LOSTED is selected as the candidate matching pair to obtain a relatively accurate key point pair set FF_1'.

3.2.2. Fine Matching

After rough matching, the set of key point pairs $FF_1' = \{F_{vi}, F_{ir}\}$ is obtained, but due to the difference in resolution and viewpoint between infrared and visible images, there are still positional deviations in many matches. Aiming at this problem, a fine matching, based on update global optimization, is introduced to lower the positional deviation of key points.

The projection matrix is calculated by the least squares method:

$$\min f = \|F_{vi}M - F_{ir}\|_2 \tag{8}$$

The least squares solution of M is called:

$$M = \left(F_{vi}^T F_{vi}\right) F_{vi}^T F_{ir} \tag{9}$$

Then, the fitting point corresponding to F_{ir} is expressed as:

$$F_{irfit} = F_{vi}M \tag{10}$$

The residuals of F_{irfit} and F_{ir} are expressed as:

$$\sigma(i) = \|F_{irfit}(i) - F_{ir}(i)\|_2, i = 1, \ldots, N_c \tag{11}$$

where N_c is correct matching numbers.

Extensive experiments show that the point closest to the correct position is usually the fitted point with a large residual. Therefore, $\{\sigma(i)\}_{i=1}^{N_c}$ is used to find the $1/4$ element of the whole, with a large residual, by descending order. Then, replace the $1/4$ element in F_{ir} with the fitted points.

$$F_{ir} = \left\{F_{ir}(1), \ldots, F_{irfit}(i), \ldots, F_{irfit}(j), \ldots, F_{ir}(n)\right\} \tag{12}$$

where $F_{irfit}(i)$ and $F_{irfit}(j)$ are the points with large residuals.

Then, all points in F_{ir} are updated by repeating (9)–(12) until the residual summation equals 0 (considering the storage of floating-point numbers in the computer, 0 is replaced by a threshold of $\sigma_p = 10^{-4}$).

After updating the position of F_{ir} through global optimization, the matching items with obvious position errors are corrected, and finally, the final image registration is obtained by calculating the position transformation matrix of infrared and visible images.

To summarize, the flow of the proposed registration method is shown in Figure 6. First, the infrared images and visible images are initialized to 224×224, respectively, and then, the trained VGG-16 model parameters are called. The images are passed through the modified deep convolution feature extraction network in this paper, and the obtained output goes through the process of rough matching feature points and fine matching feature points. Finally, the final image registration is obtained by calculating the position transformation matrix of infrared and visible images.

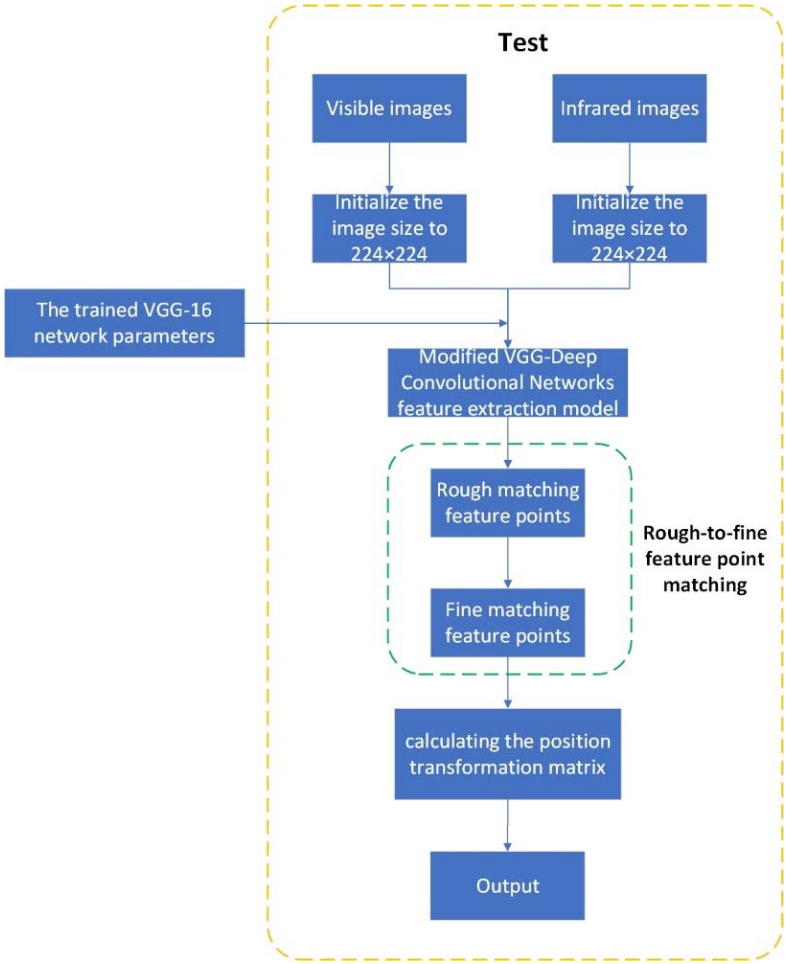

Figure 6. The flow of our algorithm.

4. Experiments and Discussion

In order to verify the validity and robustness of the proposed registration algorithm, the algorithm parameters and data sets are first configured. Then, the feature extraction part of the algorithm and the whole algorithm are experimentally verified, respectively, and the results are analyzed through qualitative and quantitative evaluation.

4.1. Experimental Setting

The training dataset is the classic ImageNet dataset at https://image-net.org/ (accessed on 13 February 2022), which can obtain more general feature extraction capabilities. In addition, our experiment is tested on visible and infrared image datasets, which consist of pictures taken by ourselves and a dataset created by Ma et al. at https://github.com/jiayi-ma/RoadScene (accessed on 2 March 2022). In our dataset, the buildings around the laboratory were selected by infrared and visible sensors for different time periods and different angles, with a total of 500 pairs of images. The dataset of Ma et al. is an infrared and visible dataset obtained by photographing some roads, vehicles, and pedestrians, with a total of 221 pairs of images. As shown in Figure 7, several groups of representative pictures were selected to display the results. The algorithm was built in Tensorflow and based on the PyCharm platform. The experimental platform, PyCharm and MATLAB 2018a, is adopted on a PC with a twelve Intel (R) Core (TM) i7-8700 CPU @ 3.2 GHz and NVIDIA GeForce RTX 2060. In the feature matching step, the ratio threshold T is automatically generated by the relatively reliable 256 pairs of key points. Analogously, T' is obtained by 128 pairs of key points. The algorithm in this paper is compared with several state-of-the-art algorithms,

including SI-PIIFD-LPM [31], HOSM [32], RIFT [33], and CNN [34]. Root mean square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM),mean absolute distance (MAE), and matching points accuracy (MPA) were selected to quantitatively evaluate the experimental results. Among them, the RMSE and MAE can be expressed as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left[\left(x_1^q - x_2^q\right)^2 + \left(y_1^q - y_2^q\right)^2\right]} \tag{13}$$

$$\mathrm{MAE} = \frac{1}{Q}\sum_{q=1}^{Q}\left|x_1^q - x_2^q\right| + \left|y_1^q - y_2^q\right| \tag{14}$$

where $\left(x_1^q, x_2^q\right)$ and $\left(y_1^q, y_2^q\right)$ are the qth pair of matching points of the visible images and infrared images, respectively, and Q is the number of matching point pairs.

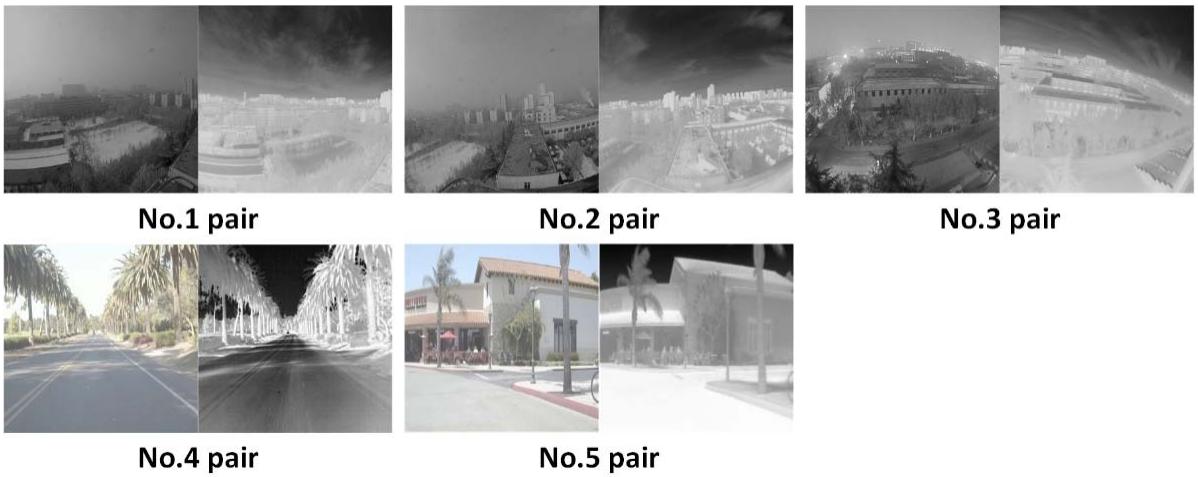

Figure 7. Input image pairs.

The matching points accuracy (MPA) can be expressed as:

$$\mathrm{MPA} = \frac{accuracy\ numbers}{total\ numbers} \tag{15}$$

where *accuracy numbers* represents the number of correct matching point pairs, and *total numbers* represents the total number of matching point pairs.

4.2. Comparison Results of Feature Point Extraction

To verify the superiority of the proposed algorithm, for extracting feature points with deep convolutional networks, the method is compared with SIFT, SUFT, and phase congruency-based methods, and the performance of the proposed algorithm is evaluated by computing repeatability [35]. Experiments illustrate the feature extraction capabilities of various algorithms by changing viewpoints, scales, and orientations of four pairs of images. The algorithm proposed in this paper ranks close to second among all tested methods, as shown in Figure 8. However, we believe that relying on repeatability results alone is not comprehensive, as the number of feature points changes with image rotation. For example, although the algorithm based on phase congruency always ranks first, after the image is rotated, it detects more invalid feature points, resulting in a sharp increase in the number of feature points, as shown in Figure 9. Although SIFT and SUFT do not have feature point explosion after image rotation, their performance is not as good as our algorithm. Since the convolutional feature extraction algorithm is more robust to changes in appearance and can better extract and retain important contours and details, we can conclude that the

proposed algorithm is more robust to feature point extraction from infrared and visible images, as shown in Figure 10.

4.3. Comparison Results of Image Registration

On the two datasets, the registration algorithm, in this paper, is compared with the HOSM, SI-PIIFD-LPM, RIFT, and CNN algorithms to test the overall performance of the proposed algorithm. The verification results are shown in Figures 11 and 12, and Figure 13 shows the quantitative evaluation results obtained by four evaluation indicators: RMSE, PSNR, SSIM, and MAE.

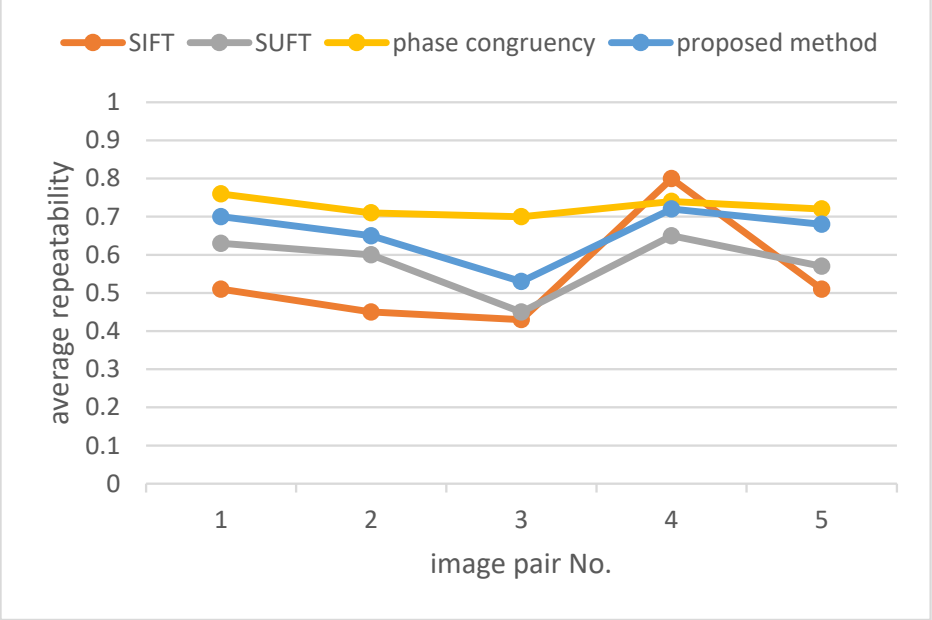

Figure 8. Average repeatability of several feature point extraction methods.

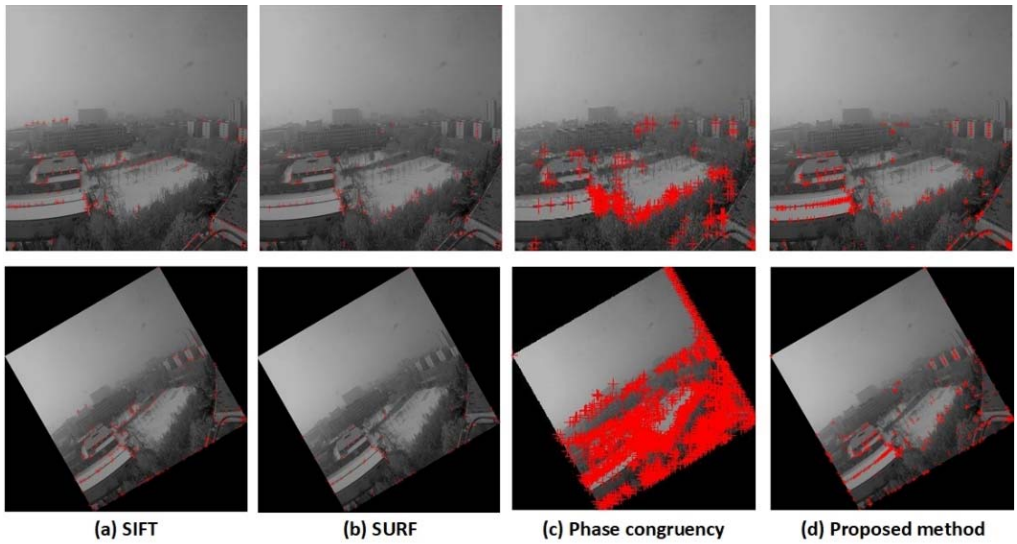

Figure 9. The result of the feature point extraction algorithm, in this paper, in the rotated image. (**a**) represents the result of the SIFT algorithm, (**b**) represents the result of the SURF algorithm, (**c**) represents the result of the Phase congruency algorithm, (**d**) represents the result of the proposed algorithm.

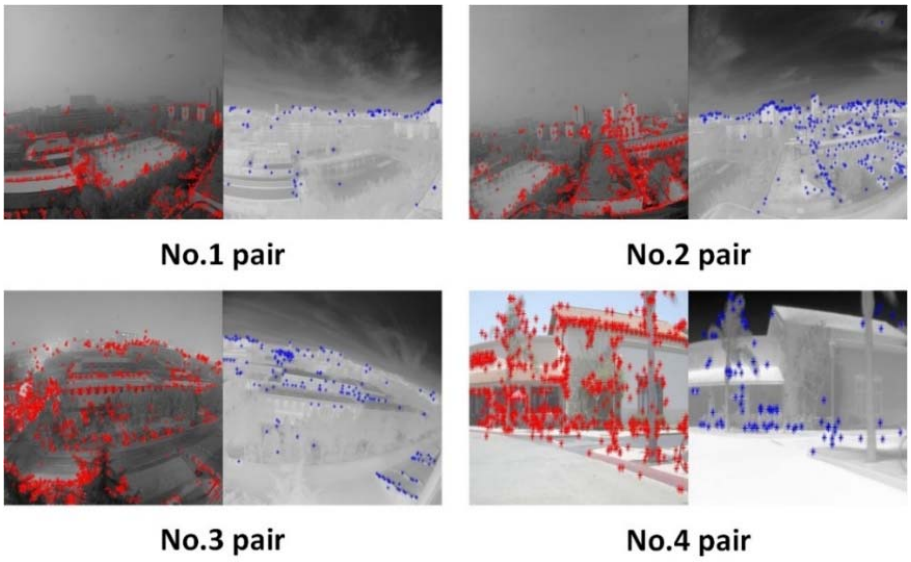

Figure 10. Feature extraction results, of the algorithm in this paper, in four pairs of images.

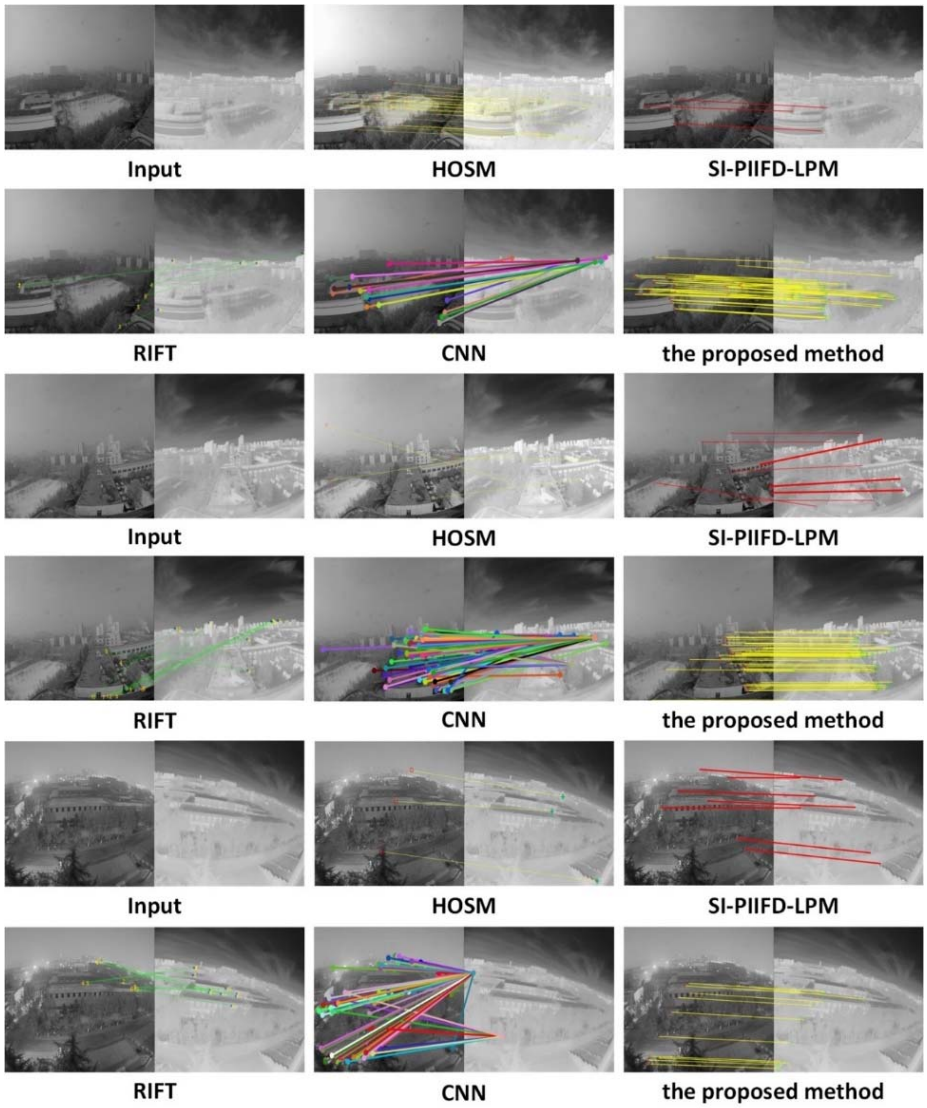

Figure 11. Registration results on our own dataset.

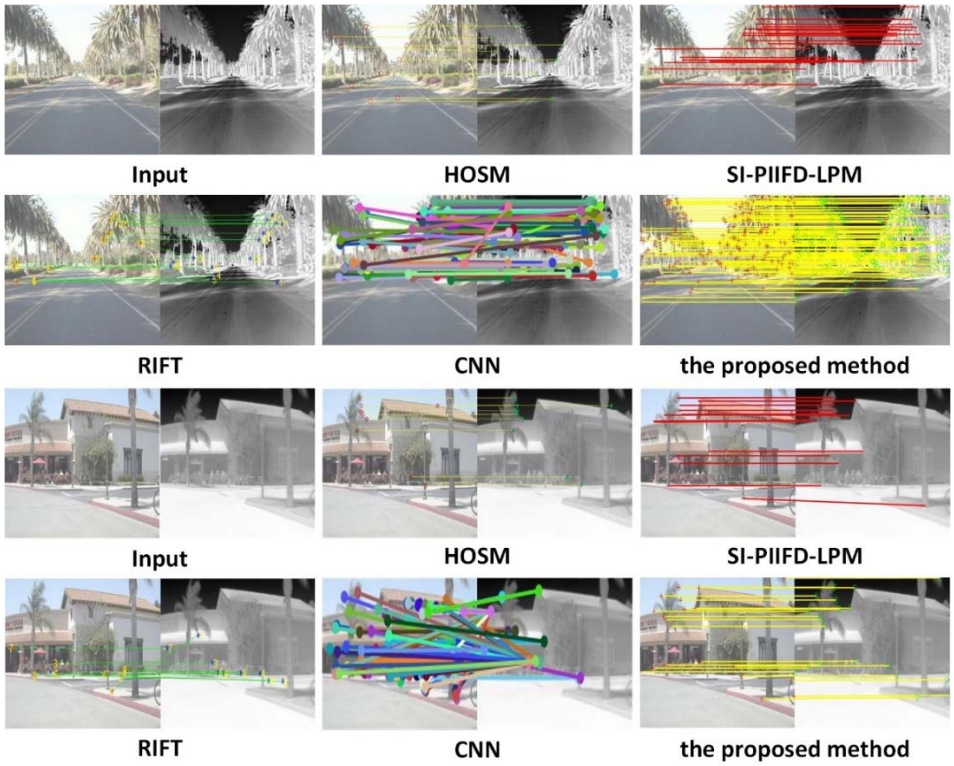

Figure 12. Registration results on the dataset of Ma et al.

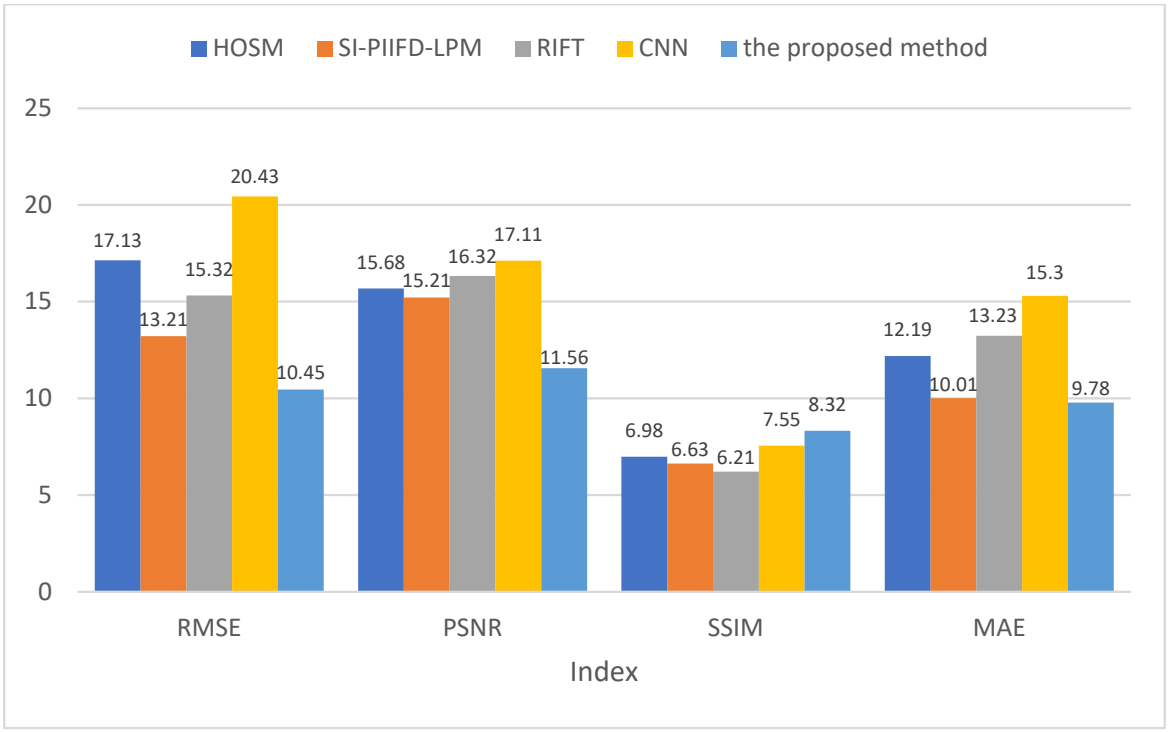

Figure 13. Quantitative evaluation index results of several algorithms.

In order to test the registration performance of the algorithm in the pictures taken at night, we selected some night picture pairs for testing. Due to the existence of image distortion and parallax interference, most of the algorithms performed in general. It can be seen, from Figures 11 and 12, that the registration performance of HOSM and SI-PIIFD-LPM

is better than that of the other two comparison methods, but the feature points are not extracted enough. Although the CNN method performs better in feature point extraction, its feature point matching ability is poor, resulting in poor registration results, and its RMSE, PSNR, and MAE indicators are also the largest among these methods. It can be seen, from Figure 13, that the results of the algorithm in this paper are all better in the four indicators. The values of RMSE, PSNR, and MAE are all the smallest, and the value of SSIM is the largest. Among them, the RMSE and PSNR indicators have smaller values than other algorithms, and their performance is more prominent. In contrast, the indicators of the SI-PIIFD-LPM algorithm are also lower, which corresponds to its visual performance. From the quantitative indicators, it can be seen that the registration effects of the other algorithms have shortcomings to varying degrees. According to Table 1, it can be seen that the MPA value of the proposed algorithm is much higher than that of the other four algorithms, which also verifies the registration performance of the proposed algorithm from another perspective. In short, the proposed algorithm performs well, on the whole, in both subjective visual performance and the value of evaluation indicators. The experimental results show that the proposed algorithm can effectively eliminate the interference of image distortion and parallaxing. At the same time, the method has good robustness under different scenes and different camera relations.

Table 1. The average matching points accuracy (MPA) of five methods.

Method	HOSM	SI-PIIFD-LPM	RIFT	CNN	The Proposed Method
MPA/%	82.79	88.64	85.79	85.21	91.13

4.4. Computational Efficiency Comparison

In order to verify the efficiency of the proposed algorithm, the proposed algorithm and the four comparison algorithms are tested on the same test dataset. The differences in the registration efficiency of the five algorithms were compared by calculating the average time required to register each pair of images. The experimental results are shown in Table 2. It can be seen that the efficiency of the proposed algorithm ranks second among the five algorithms in the test. Although it is not optimal, its efficiency is also remarkable. In the next step, it is also one of our research directions to further improve the algorithm structure and improve the overall efficiency of the algorithm.

Table 2. Average computational efficiency comparison of five methods.

Method	HOSM	SI-PIIFD-LPM	RIFT	CNN	The Proposed Method
Time/s	0.98	1.34	2.15	1.21	1.03

5. Conclusions

An automatic robust infrared and visible image registration algorithm, based on a deep convolutional network, was proposed in this paper. It makes full use of the feature extraction performance of the deep convolutional network to accurately locate feature points, and a rough-to-fine feature matching method is introduced. The initial screening is carried out by location orientation scale transform Euclidean distance, and the final accurate matching is achieved by optimizing the position of the global matching point, thereby solving the registration problem of infrared and visible images on UAVs. Image registration tested on two datasets shows that, compared with four advanced registration algorithms, the proposed algorithm can achieve good registration results by overcoming the shortcomings of obvious pixel differences between infrared and visible images, as well as blurred infrared image feature points. In the future, we will continue research in improving the efficiency and generalizability of our algorithm to improve its performance.

Author Contributions: Methodology, Y.L.; software, J.Y.; validation, C.W.; investigation, Y.Z.; resources, Y.H.; data curation, Z.L.; writing original draft preparation, J.J.; writing—review and editing, J.J., C.W., Y.Z., Y.L., Y.H., Z.L. and J.Y.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62171467 and the Natural Science Foundation of Hebei Province, grant number F2021506004.

Acknowledgments: The authors are grateful to Fuyu Huang for his help with the fundings and the preparation of figures in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ji, J.; Zhang, Y.; Lin, Z.; Li, Y.; Wang, C.; Hu, Y.; Yao, J. Infrared and Visible Image Fusion Based on Iterative Control of Anisotropic Diffusion and Regional Gradient Structure. *J. Sens.* **2022**, *2022*, 7144991. [CrossRef]
2. Cocianu, C.L.; Uscatu, C.R. Multi-Scale Memetic Image Registration. *Electronics* **2022**, *11*, 278. [CrossRef]
3. Wang, X.-H.; Huang, W.; Ouyang, J.H. Real-time image registration of the multi-detectors mosaic imaging system. *Chin. Opt.* **2015**, *8*, 211–219. [CrossRef]
4. Artola, L.; Hubert, G.; Gilard, O.; Ducret, S.; Perrier, F.; Boutillier, M.; Garcia, P.; Vignon, G.; Baradat, B.; Ricard, N. Single Event Upset Sensitivity of D-Flip Flop of Infrared Image Sensors for Low Temperature Applications Down to 77 K. *IEEE Trans. Nucl. Sci.* **2015**, *62*, 2979–2987. [CrossRef]
5. Di, K.; Zhao, Q.; Wan, W.; Wang, Y.; Gao, Y. RGB-D SLAM Based on Extended Bundle Adjustment with 2D and 3D Information. *Sensors* **2016**, *16*, 1285. [CrossRef] [PubMed]
6. Öfverstedt, J.; Lindblad, J.; Sladoje, N. Fast and robust symmetric image registration based on distances combining intensity and spatial information. *IEEE Trans. Image Process.* **2019**, *28*, 3584–3597. [CrossRef]
7. Chang, H.-H.; Wu, G.-L.; Chiang, M.-H. Remote Sensing Image Registration Based on Modified SIFT and Feature Slope Grouping. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1363–1367. [CrossRef]
8. Barnea, D.I.; Silverman, H.F. A class of algorithms for fast digital image registration. *IEEE Trans. Comput.* **1972**, *100*, 179–186. [CrossRef]
9. Viola, P.; Wells, W.M., III. Alignment by maximization of mutual information. *Int. J. Comput. Vis.* **1997**, *24*, 137–154. [CrossRef]
10. Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **1997**, *16*, 187–198. [CrossRef]
11. Fan, X.; Rhody, H.; Saber, E. Automatic Registration of Multi-Sensor Airborne Imagery. In Proceedings of the 34th Applied Imagery and Pattern Recognition Workshop (AIPR'05), Washington, DC, USA, 1 December 2005; pp. 81–86.
12. Moravec, H.P. Rover Visual Obstacle Avoidance. In Proceedings of the IJCAI, Vancouver, BC, Canada, 24–28 August 1981; Volume 81, pp. 785–790.
13. Harris, C.; Stephens, M. A combined corner and edge detector. In Proceedings of the Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988; Volume 15, pp. 10–5244.
14. Lindeberg, T. Scale-space theory: A basic tool for analyzing structures at different scales. *J. Appl. Stat.* **1994**, *21*, 225–270. [CrossRef]
15. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
16. Bay, H.; Tuytelaars, T.; Gool, L.V. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
17. Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary robust invariant scalable keypoints. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2548–2555.
18. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
19. Zhang, S.; Huang, F.; Liu, B.; Li, G.; Chen, Y.; Sun, L.; Zhang, Y. Robust registration for ultra-field infrared and visible binocular images. *Opt. Express* **2020**, *28*, 21766–21782. [CrossRef] [PubMed]
20. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
21. Ma, J.; Zhao, J. Robust topological navigation via convolutional neural network feature and sharpness measure. *IEEE Access* **2017**, *5*, 20707–20715. [CrossRef]
22. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Deep image homography estimation. *arXiv* **2016**, arXiv:1606.03798.
23. Nowruzi, F.E.; Laganiere, R.; Japkowicz, N. Homography estimation from image pairs with hierarchical convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 913–920.
24. Nguyen, T.; Chen, S.W.; Shivakumar, S.S.; Taylor, C.J.; Kumar, V. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2346–2353. [CrossRef]

25. Li, Q.; Peng, Q.; Yan, C. Multiple VLAD encoding of CNNs for image classification. *Comput. Sci. Eng.* **2018**, *20*, 52–63. [CrossRef]
26. Wen, Y.; Sheng, B.; Li, P.; Lin, W.; Feng, D.D. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Trans. Image Process.* **2018**, *28*, 994–1006. [CrossRef]
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Zan, X.; Zhang, X.; Xing, Z.; Liu, W.; Zhang, X.; Su, W.; Liu, Z.; Zhao, Y.; Li, S. Automatic detection of maize tassels from UAV images by combining random forest classifier and VGG16. *Remote Sens.* **2020**, *12*, 3049. [CrossRef]
29. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
30. Wu, Y.; Ma, W.; Gong, M.; Su, L.; Jiao, L. A novel point-matching algorithm based on fast sample consensus for image registration. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 43–47. [CrossRef]
31. Du, Q.; Fan, A.; Ma, Y.; Fan, F.; Huang, J.; Mei, X. Infrared and visible image registration based on scale-invariant piifd feature and locality preserving matching. *IEEE Access* **2018**, *6*, 64107–64121. [CrossRef]
32. Ma, T.; Ma, J.; Yu, K. A local feature descriptor based on oriented structure maps with guided filtering for multispectral remote sensing image matching. *Remote Sens.* **2019**, *11*, 951. [CrossRef]
33. Yu, K.; Ma, J.; Hu, F.; Ma, T.; Quan, S.; Fang, B. A grayscale weight with window algorithm for infrared and visible image registration. *Infrared Phys. Technol.* **2019**, *99*, 178–186. [CrossRef]
34. Yang, Z.; Dan, T.; Yang, Y. Multi-temporal remote sensing image registration using deep convolutional features. *IEEE Access* **2018**, *6*, 38544–38555. [CrossRef]
35. Schmid, C.; Mohr, R.; Bauckhage, C. Evaluation of Interest Point Detectors. *Int. J. Comput. Vis.* **2000**, *37*, 151–172. [CrossRef]