Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Improvement of multi-scale image alignment techniques in computer vision based on numerical analysis methods

Huaijiang Teng[1], Zhenbo Zhang[1,✉]

[1] *Heilongjiang Open University, Harbin, Heilongjiang, 150080, China*

ABSTRACT
Image alignment is a fundamental problem in the field of computer vision and an important prerequisite for carrying out many other tasks. Firstly, the theoretical basis and realization method of image alignment as well as the process and the method of alignment are introduced to provide alignment ideas. Subsequently, an image alignment method based on the union of multi-scale features is proposed, and a new loss term is introduced to the small-scale features therein, which further improves the distinguishability of the small-scale feature descriptors while guaranteeing the invariance of the large-scale feature descriptor matching therein. Three common alignment algorithms (RIFT algorithm, HAPCG algorithm, and SAR-SIFT algorithm) are selected for stability assessment and quantitative evaluation on the dataset, and an image enhancement algorithm with histogram equalization is used to enhance the dataset. The results show that the feature stability of this paper's method is described as 99.1%, which is better than other algorithms. Meanwhile the desired effect is achieved on the dataset.

Keywords: image alignment, multi-scale alignment, histogram equalization, image recognition

1. Introduction

Computer vision converts the physical information of a real scene into a two-dimensional image, and simulates human's ability to understand and judge image information through techniques such as image enhancement, data mining and machine learning. Image alignment is an important precursor step in computer vision to align two or more images so that they can be subsequently processed and analyzed under the same spatial reference system [1-3]. The purpose of image alignment is to solve the problem of spatial mismatch between different images for applications such as multimodal image fusion, image stitching, and image correction [4]. Image alignment is needed to improve the accuracy and reliability of information extraction and analysis in the fields of medicine, remote sensing, industrial inspection, and automated driving [5-7]. In medical imaging, it can be used to align medical

✉ Corresponding author.
E-mail address: 15546202925@163.com (Z. Zhang).
Received 02 April 2024; Revised 15 July 2024; Accepted 10 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-047

images at different times or imaging modalities for applications such as lesion analysis, surgical navigation, and treatment monitoring [8-9]. For remote sensing, it can be used to synthesize high-resolution images and digital maps [10-11]. For industrial inspection, it can be used for precise localization and defect detection in machine vision [12-13].

In general, various factors such as image noise, image distortion, geometric transformation, and non-rigid deformation are the problems and research challenges faced by image alignment [14]. Among them, image noise and distortion are caused by the limitations of image acquisition equipment and environmental disturbances, while geometric transformations and non-rigid deformations are due to the shooting angle and the characteristics of deformed objects [15-18]. These factors lead to differences between images, making image alignment complex and difficult. Therefore, studying how to overcome these difficulties and improve the accuracy and robustness of the alignment algorithms has been a hot and difficult research topic in the field of image processing and computer vision.

Wang, Y. et al. proposed a moving direct linear transform (MDLT) method by generating the sample labels of the drawings to reflect their local features, inputting the feature data into a convolutional neural network for training in order to compute multiple pairs of local matches between the images to be aligned, and estimating the local one-shot matrices to achieve the image alignment by using the MDLT, the proposed improved image alignment algorithm achieves higher alignment accuracy and robustness [19]. Beroiz, M. et al. designed a Python module for graphical alignment that does not depend on WCS information, namely the Astroalign algorithm, which matches information such as diffusion functions at different points to align or differential image analysis of images to achieve accurate alignment of images in astronomy [20]. Brüstle, S. developed a multidimensional downhill simplex method for image alignment workflows and quantitatively examined the performance of this transform estimation method in improving the accuracy of geo-referenced image alignment, and experiments have shown that image alignment workflows based on this method are robust and feasible in terms of the differences in the appearance of the image to be measured and the reference image [21]. Song, J. et al. classified the centerline-based 3D-2D image alignment problem of blood vessels as an iterative n-point perspective (PnP) problem, and introduced a regenerative kernel Hilbert space algorithm and an iterative reweighted least squares algorithm, which effectively improved the real-time and accurate performance of vascular interventional robots in aligning non-rigidly shaped graphics [22]. Cocianu, C. L. et al. conducted a study on high-precision image alignment in the presence of geometric perturbations and proposed a memetic algorithm, which combines bio-inspired, evolutionary computation, and clustering search techniques to efficiently solve the problem of precocious convergence under different scale representations [23]. Sokooti, H. et al. designed a convolutional neural network architecture capable of training images to estimate displacement vector fields, integrating image content at multiple scales in order to make the network contextually informative, and efficiently solving the problem of non-rigid image alignment through a learning approach [24]. Liu, S. et al. designed a multiscale cross-feature network for medical image alignment task by introducing a feature crossover module to enhance the information flow in the hidden layer of the image and combining it with a three-bit convolutional attention module to improve the correlation between different modal features of the image [25].

The study firstly describes the methods and principles of image alignment, and then proposes an alignment method with the combination of multi-scale features, which utilizes the intermediate layer of small-scale features to maintain distinguishability, and utilizes the output layer of large-scale features to maintain invariance, and in addition, introduces a distinguishability constraint loss for the intermediate layer of small-scale features. Three datasets are selected to be enhanced with histogram equalization image enhancement algorithm. Different algorithms such as RIFT, SAR-SIFT, and HAPCG are used to conduct quantitative evaluation experiments and alignment performance comparisons to prove the effectiveness of the method in this paper.

2. Image alignment for multi-scale feature association in machine vision

2.1. Image Preprocessing

Image preprocessing generally includes: digitization of images, geometric correction, image restoration and image enhancement, etc. Here we only briefly introduce image enhancement and geometric correction [26]. Image enhancement is actually a gray scale transformation, and is important because differences in the gray scale of an image are a major obstacle to alignment. Geometric correction, on the other hand, is used to correct for systematic errors generated by the sensor during image acquisition and random errors generated by the sensor in position.

2.1.1. Image enhancement. Image enhancement techniques cover: histogram correction, gray scale transformation, image smoothing and color enhancement. According to the domain of action of image enhancement, it can be divided into: (1) spatial domain processing method; (2) frequency domain processing method. Image enhancement techniques in the spatial domain are said to actually refer to the plane itself, and the pixel processing of the image i.e.: the meaning of spatial domain image enhancement methods [27]. It can be defined as:

$$g(x,y) = T[f(x,y)] \tag{1}$$

where $f(x,y)$ is the input image, $g(x,y)$ is the processed image and T is an operation on f. In the neighborhood processing technology, the completion of the null domain filtering is done with the help of neighborhood operations under the image space template, which can be divided into two categories of image smoothing and image sharpening according to the different functions. Image smoothing: an image in its acquisition and transmission process, it is very likely to suffer from a variety of noise interference, which will make the quality of the image greatly reduced. The operation of processing these interferences is called image smoothing.

Image Sharpening: Sharpening of an image is the process of highlighting the edge or outline information of an image in image recognition. Frequency Domain Image Enhancement: Frequency domain or “transform domain”, the basic principle of enhancement is to make the image in a certain range of the transform domain is suppressed, while other components are not affected. Therefore, the frequency distribution of the output image can be changed to achieve the effect of enhancement. To achieve the above effect, the main steps are: calculate the Fourier transform of the image we want to enhance; convolve it with a (this should be set by yourself) transfer function; finally, carry out the Fourier inverse transform of the obtained result to get the enhanced image.

2.1.2. Geometric correction. The problem of geometric deformation is often encountered when acquiring the initial image, and there are many factors that contribute to the deformation, many of which cannot be solved by the imaging techniques we have developed so far. One of the more common geometric deformations is nonlinear deformation, which often makes image alignment much more difficult, and the image alignment does not reach our expected matching accuracy. Therefore, it is often possible to achieve a satisfactory accuracy only after precise geometric correction and then image alignment.

2.2. Principles and steps of image alignment

2.2.1. Basic principles of image alignment. In layman's terms, image alignment is the technique of aligning two or more images generated in the same state, or more precisely, to find the point-to-point relationship between N images. Image alignment can be summarized as a spatial and gray scale transformation between images. That is, the coordinates of one of the images $X(x,y)$ are first mapped to a newly defined coordinate system to produce new coordinates $x' = (x', y')$, and then its pixels are resampled. The premise of image alignment is that the images should be partially identical

to each other, i.e., a part of the image reflects the same target region of the original image, which is a basic condition in image alignment.

Here we assume that two images are represented by two-dimensional matrices I_1 and I_2, and $I_1(x,y)$ and $I_2(x,y)$ denote the gray values at the corresponding position (x,y), respectively, then the mapping relationship between the images can be expressed as:

$$I_2(x,y) = g(I_2(f(x,y))) \tag{2}$$

where f represents a transformation of a two-dimensional spatial coordinate:

$$(x',y') = f(x,y) \tag{3}$$

And g is a one-dimensional grayscale alignment problem is straightforward, that is, how to find the optimal spatial and grayscale transformations, the transformations so that the two images to achieve the maximum accuracy of the match. Usually the grayscale transformation is only used in applications related to sensor changes in optics, and is not needed in other cases. In most cases, the main idea of solving the alignment problem is still to find the spatial transformation or this geometric transformation, the above equation can be changed to:

$$I_2(x,y) = I_1(f(x,y)) \tag{4}$$

2.2.2. Weak classifiers. Manual image alignment requires people's experience and common sense accumulated over years of work and study to select suitable feature points in the image, then select the spatial transformation model of the two images and use interpolation to complete the matching of the two images. This method is subject to human intervention and has a lot of uncertainty, so the application range is narrow, and it may be seen in medical image alignment.

2.2.3. Automatic image alignment. After feature extraction and matching, only the establishment of a correct transform or mapping model can be followed by stitching and fusion, which is actually the most fundamental and important place of image alignment technology. According to the complexity of the transformation model, it can be divided into rigid body transformation, affine transformation, projection transformation and nonlinear transformation.

1) Rigid Body Transformation. We use rigid body transformation to transform any two points in the image to be aligned, which can make the distance between the two points unchanged after the transformation. Rigid-body transformation includes translation, rotation and inversion. The rigid body transforms point (x,y) to point $(x^{'},y^{'})$:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\psi & \pm\sin\psi \\ \sin\psi & \mp\cos\psi \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{5}$$

where $\begin{bmatrix} t_x \\ t_v \end{bmatrix}$ is the translation and ψ is the rotation angle.

2) Affine Transformation. The affine transformation adds a scaling operation to the rigid body transformation, that is, a proportional scaling factor is added to each dimension of the image. The affine transform matrix transformation is given. After affine transformation, the expression from point (x,y) to point $(x^{'},y^{'})$ is expressed as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{6}$$

where $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ is a real matrix.

3) Projective transformations. Projective transformations differ from affine transformations in that they do not guarantee parallelism while keeping the state constant. Projective transformations can be expressed as linear transformations, but only in higher dimensions. In two dimensions, the point (x, y) is transformed to (x', y') as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} + \begin{bmatrix} x \\ y \end{bmatrix} \tag{7}$$

where $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$ plays a similar role as in affine transformations.

4) Nonlinear transformation. The model is more complex, we do not introduce too much, simply put, it will make any line in the image after the transformation becomes no longer a straight line, but an arc. In two dimensions, the nonlinear transformation from point (x, y) to point (x', y') is:

$$(x', y') = F(x, y) \tag{8}$$

$$x' = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \cdots \tag{9}$$

$$y' = b_{00} + b_{10}x + b_{01}y + b_{20}x^2 + b_{11}xy + b_{02}y^2 + \cdots \tag{10}$$

2.3. Image alignment based on multi-scale feature association

2.3.1. Multi-scale feature networks:

1) Network structure. For deep learning based image descriptors such as L2-Net, HardNet, SOSNet, etc., there is often the problem of invariance and distinguishability contradiction. Different from the above methods, the image alignment method based on multi-scale feature union (MSD-Net SIFT) proposed in this paper, in addition to the large-scale features of the output layer, simultaneously utilizes the small-scale features that have better preservation of the detail information, so as to achieve the improvement of accuracy when matching.

The main body of the network structure of MSD-Net is divided into two parts, i.e., the small-scale feature description network (SSD-Net), and the large-scale feature description network (LSD-Net), which are connected in series, i.e., the output feature map of the SSD-Net is the input of the LSD-Net. For the convolutional layers in MSD-Net other than the output layer, a batch normalization (BN) layer as well as a ReLU activation function are followed. In addition, since the pooling layer has no learnable parameters, a convolutional layer with a sliding step size of 2 is used instead of the pooling layer for the downsampling operation, as in L2-Net and HardNet.

2) Loss function. Small-scale feature loss term: let the input image block of MSD-Net be denoted as P. Denote the parameters of SSD-Net and LSD-Net by Ω^s and Ω^t, respectively, and their respective outputs can be denoted as $F^s(\Omega^s, P)$ as well as $F^L(\Omega^L, F^S(\Omega^S, P))$. For the input sample $X = \{(A_i, P_i)\}$, denote the output of SSD-Net as $Y^s = \{(\mathrm{A}_i, \mathrm{P}_i)\}$, where:

$$\mathrm{A}_i = F^s(\Omega^s, A_i), \mathrm{P}_i = F^s(\Omega^s, P_i) \tag{11}$$

From the structure of SSD-Net network, it can be seen that its output feature $F^s(\Omega^s, P) \in \square^{16\times16\times128}$,

so $F^S(\Omega^S, P)_{8,8,:}$ that is, represents the small-scale features corresponding to the center point of the image block (the index starts from 0, and after that, this paper all follows this convention). Considering that the SSD-Net output feature scale (sensory field) is relatively small and retains more detailed information compared to the output features of LSD-Net, it is proposed to add distinguishability constraints on this basis to further improve its distinguishability. Similar to HardNe, distinguishability also adopts the idea of the most difficult negative sample mining, but focuses only on the near-neighbor region features. Define the distance difference:

$$\Delta d_i^{(1)} = \min_{p,q \in W} d((\mathrm{P}_i)_{p,q,:}, (\mathrm{A}_i)_{8,8,:}) - d((\mathrm{P}_i)_{8,8,:}, (\mathrm{A}_i)_{8,8,:}) \tag{12}$$

$$\Delta d_i^{(2)} = \min_{p,q \in W} d((\mathrm{A}_i)_{p,q,:}, (\mathrm{P}_i)_{8,8,:}) - d((\mathrm{A}_i)_{8,8,:}, (\mathrm{P}_i)_{8,8,:}) \tag{13}$$

where $W = \{x \in \square \mid 0 \le x \le 15 \wedge x \ne 8\}$, $d(u,v)$ denotes the Euclidean distance after normalization of the modulus of the vector, i.e:

$$d(u,v) = \left\| \frac{u}{\| u \|_2} - \frac{v}{\| v \|_2} \right\|_2 \tag{14}$$

Large-scale feature loss term: for the set of input samples $X = \{(A_i, P_i)\}$, the output of LSD-Net is represented as $Y^L = \{(a_i, p_i)\}$, where:

$$a_i = F^L l(\Omega^L, F^S(\Omega^S, A_i)r), p_i = F^L l(\Omega^L, F^S(\Omega^S, P_i)r) \tag{15}$$

Combining the structure of SSD-Net as well as LSD-Net, it can be seen that $a_i, p_i \in \square^{128}$. For large-scale features, it has a larger sensory field, and thus it is easier to maintain robustness in feature matching, so this paper adopts the HardNet loss function for constraints, as shown in Equation (16):

$$\mathrm{L}_L = \mathrm{L}_h(Y^L) \tag{16}$$

where L_h is defined according to equation (17). For the set of input samples $X = \{(A_i, P_i)\}$, the overall loss function throughout the training process is:

$$\mathrm{L}(X) = \mathrm{L}_S + \mathrm{L}_L \tag{17}$$

While in this paper, the network is trained with a set of input samples containing both $X^1 = \{(A_i, P_i^1)\}$ as well as $X^2 = \{(A_i, P_i^2)\}$, so the loss function for the entire MSD-Net training is:

$$\mathrm{L}_M = \frac{\mathrm{L}(X^1) + \mathrm{L}(X^2)}{2} \tag{18}$$

By establishing constraints on the small-scale features output from SSD-Net and large-scale features output from LSD-Net at the same time, the distinguishability and matchability of the small-scale features are enhanced, and the robustness of the large-scale features is ensured during matching.

2.3.2. Joint localization of multi-scale features. Considering that there are a large number of key points in an image, if for each key point, an image block of size 128×128 is cropped with its center, it will make a large number of overlapping areas between image blocks and image blocks, and if descriptors are computed independently for each of these image blocks, it will lead to computational redundancy. Therefore, this paper proposes a joint localization method based on multi-scale features to realize the localization from the feature points of the source image to the target image Let the source image be denoted as I_A, and the target image as I_B, and take the two as the inputs of MSD-Net respectively, and the outputs of SSD-Net and LSD-Net are denoted as $(\mathrm{A}^S, \mathrm{A}^L)$ and $(\mathrm{B}^S, \mathrm{B}^L)$ respectively. Let a key point in Fig. I_A have coordinates (x_A, y_A) on the original image, and make

sure that its distance to the image boundary is far enough, and by indexing, its small-scale features as well as large-scale features can be obtained:

$$f^S = \mathrm{A}^S_{y^S,x^S;}, f^L = \mathrm{A}^L_{y^L,x^L;} \tag{19}$$

Among other things:

$$x^S = \left\lfloor \frac{x_A + 4}{8} \right\rfloor, y^S = \left\lfloor \frac{y_A + 4}{8} \right\rfloor \tag{20}$$

$$x^L = \left\lfloor \frac{x_A - 48}{16} \right\rfloor, y^L = \left\lfloor \frac{y_A - 48}{16} \right\rfloor \tag{21}$$

On this basis, the small-scale feature distance map DM^S as well as the large-scale feature distance map DM^L are defined as follows:

$$DM^S_{ij} = d(f^S, B^S_{i,j,:}), DM^L_{ij} = d(f^L, B^L_{i,j,:}) \tag{22}$$

3. Results of numerical experiments and analysis

3.1. Experimental setup and treatment

3.1.1. Experimental setup. We trained and evaluated our method on three datasets, ADNI, RIRE, and UWF.The ADNI dataset is a combination of MRI and PET information of the brain from people of different ages collected for the study of the cause of Alzheimer's disease, and contains 200 pairs of MRI-PET images.The RIRE is designed to compare retrospective CT-MR and PET-MR used by multiple groups alignment techniques, which contained 109 pairs of CT-MR images. For each dataset, we cropped the paired images to 1024 × 1024.We compared the proposed method with traditional alignment methods including SIFT and Voxel, RSC, LapIRN, ProsRegNet, MICDIR.All experiments were performed on NVIDIA Tesla T4 GPUs.

3.1.2. Image enhancement. Design monotone nonlinear mapping function f using histogram equalization method to convert original image to equalized image. Assume that the histogram distributions of the original image A, and the equalized image are $H_A(D), H_B(D)$ respectively.

$$D_B = f(D_A) \Rightarrow D_B + \Delta D_B = f(D_A + \Delta D_A) \tag{23}$$

The gray scale D_A is converted to D_B by the mapping function:

$$\int_{D_A}^{D_A+\Delta D_A} H_A(D)\, dD = \int_{D_B}^{D_B+\Delta D_B} H_B(D)\, dD \tag{24}$$

In particular, there are $H_B(D)$, there:

$$\int_0^{D_A} H_A(D)\, dD = \int_0^{D_B} H_B(D)\, dD \tag{25}$$

Ideally, the histogram after equalization is uniformly distributed with $H_B(D) = \frac{A_0}{L}$. Where A_0 is the number of pixel points and L is the gray scale range with a value of 256 (0 to 255). There are:

$$\int_0^{D_A} H_A(D)\, \mathrm{d}D = \frac{A_0 D_B}{L} = \frac{A_0 f(D_A)}{L} \tag{26}$$

The mapping function is calculated:

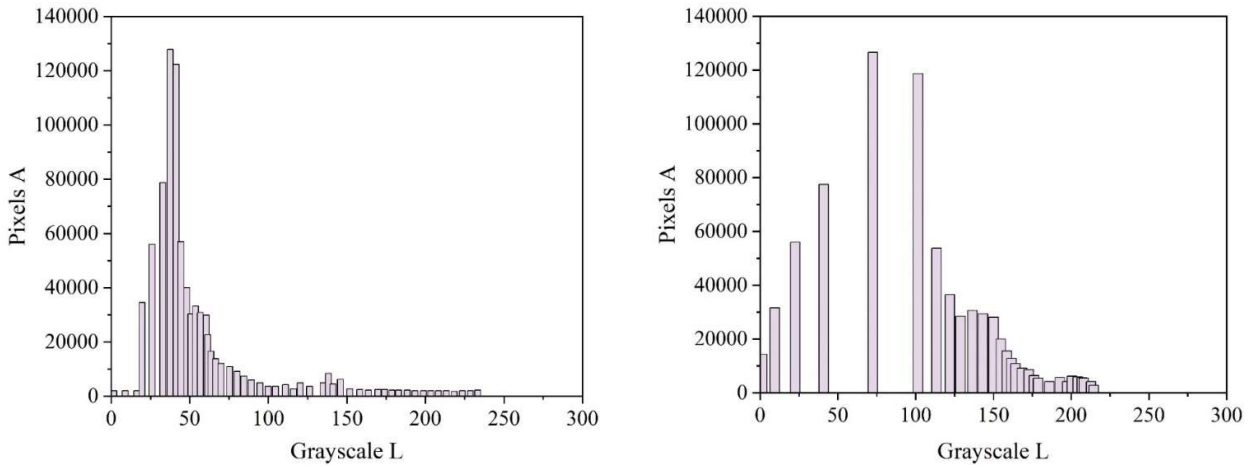
$$f(D_A) = \frac{L}{A_0} \int_0^{D_A} H_A(D)\, dD \tag{27}$$

Its discrete form is:

$$f(D_A) = \frac{L}{A_0} \sum_{u=0}^{D_A} H_A(u) \tag{28}$$

The discrete value is the pixel value of each point after histogram equalization.

The image information was acquired under dark, indoor conditions and the histogram equalization was performed in the Spyder software environment, and the results are shown in Fig. 1, from which it can be seen that the brightness of the original image is low, and its grayscale is concentrated at about 50. After the histogram equalization, the brightness of the image is significantly increased and the gray level is uniformly distributed from 0 to 227.



(a)Primary image (b)Equalizing image

Fig. 1. Histogram equalization treatment results

3.2. Quantitative analysis results

3.2.1. Evaluation Experiment of Gray Scale Aberration Resistance. The evaluation experiments on the resistance to gray-scale distortion used the four hyperspectral data from the Pavia Center dataset described earlier with four gray-scale distortions, and a total of 20 images including the original image were obtained.

In this step part of the experiment, this section is based on the original 4 hyperspectral data, FAST corner point detection is performed on these 4 images respectively, and based on the results of the detection, the positions with the highest number of repetitive occurrences under different bands are selected as the key points, and a total of 82 are selected to be featured on each of these 20 sets of data using an algorithm to characterize these 82 key positions, and a set of feature description vectors is obtained $D(i,j)$, where $i=1,2,3...M$, $j=1,2,3...N$ (M is the number of data used as comparison images, $M=20$ in this experiment, N is the number of selected key locations, $N=82$ in this experiment), and each $D(i,j)$ is a feature description vector.

It is assumed that the feature askance of each feature descriptor is not comparable between its different dimensional data, so in this paper, the simplest MSE method is used for the difference measure of two features. For 82 key positions, 82 variance calculations $S(j)$ can be obtained, where $j=1,2,3...N, S(j)$ the formula is as follows:

$$S(j)=\frac{\sum_{i=1}^{M}(D(i,j)-\bar{D}(j))^2}{M} \tag{29}$$

where $\bar{D}(j)$ is the average of the feature descriptors for the j nd position under 20 sets of data, i.e:

$$\bar{D}(j)=\frac{\sum_{i=1}^{M}D(i,j)}{M} \tag{30}$$

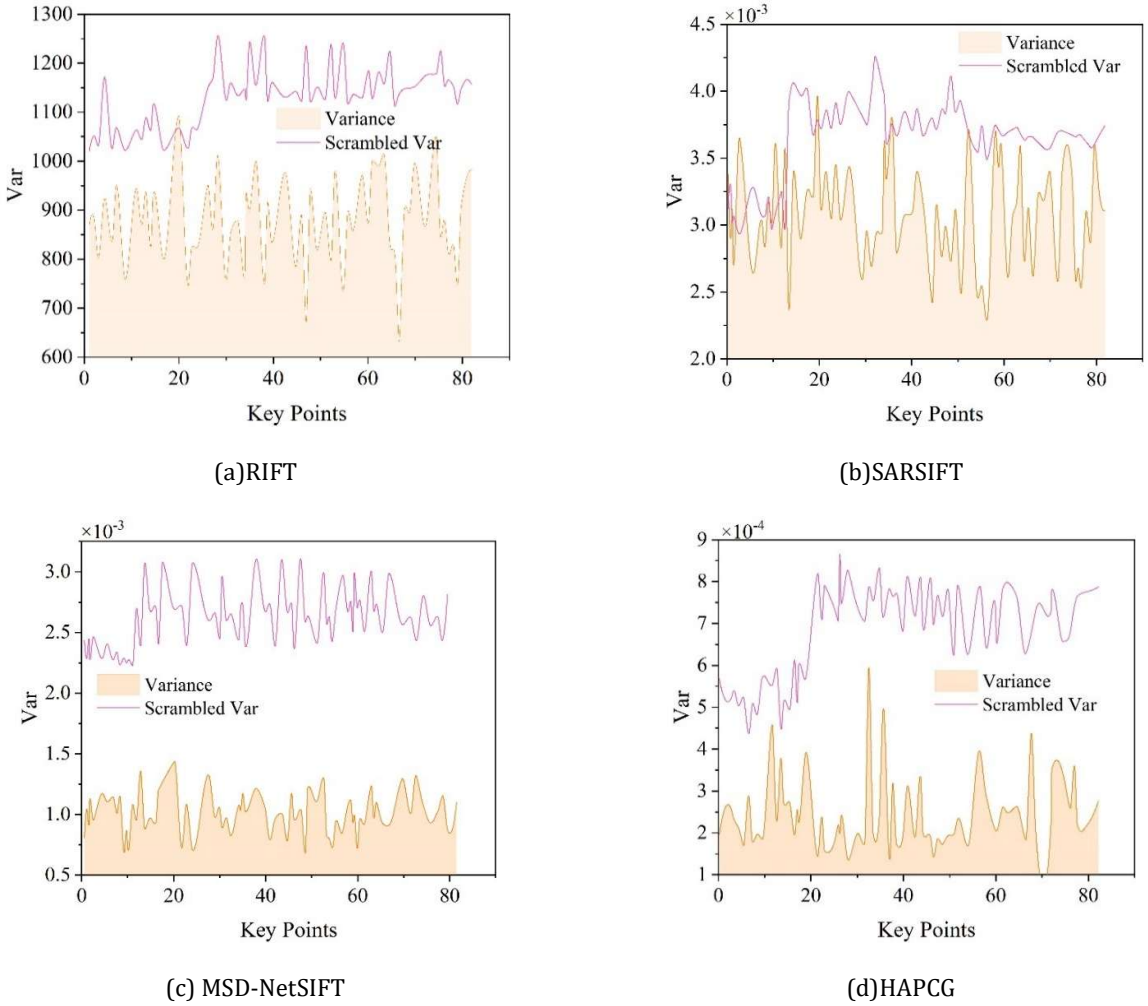
At the same time, the data in $D(i,j)$ are disrupted to obtain the variances computed in the above manner at different locations, and the values of the two sets of variances are observed. The gray-scale distortion experiment is shown in Fig. 2, and the variance of similar feature vectors (i.e., 20 sets of feature description vectors computed at the same location from 20 sets of gray-scale distorted data) is roughly smaller than the variance of dissimilar feature vectors (i.e., feature description vectors at

different locations). The minimum value of the variance after upsetting is used as a threshold value, and values above the threshold in $S(j)$ are regarded as failed values to be excluded, so that a pass rate, i.e., the stability R of the feature descriptors under that set of gray-scale distortion data, can be calculated for S the feature descriptors.

$$R = \frac{n}{N} \times 100\% \tag{31}$$

where N is the total number of selected key positions and n is the number of key positions with variance below the threshold.

The final calculated experimental results for evaluating the gray-scale distortion resistance of the feature descriptors of the four algorithms are shown in Fig. 2. From the figure, it can be seen that the variance of the similar feature vectors is overall smaller than that of the dissimilar feature vectors, so according to the calculations for the experimental data of this grayscale aberration, the stability of the RIFT feature descriptor is 94.01%, the stability of the SARSIFT feature descriptor is 35.03%, the stability of the MSD-Net SIFT feature descriptor is 99.1%, and the stability of the HAPCG feature descriptor is 96.11%. The stability of HAPCG feature descriptor is 96.11%.



(a)RIFT (b)SARSIFT

(c) MSD-NetSIFT (d)HAPCG

Fig. 2. Results of the grayscale distortion experiments

3.2.2. Scale invariance evaluation experiments. The data used in the scale invariance evaluation experiments are several sets of SAR image data from MiniSAR, based on which a total of 40 sets of data are transformed in the scale range of 0.5~1.7. In other words, the stability of the feature descriptors of the four algorithms in 40 different scales is calculated in this experiment.

By calculating the ratio of the number of points that remain stable to the total number of points in each scale of the feature descriptors of the four algorithms, the stability in that scale can be calculated. The results of the average scale invariance of the feature descriptors of the four algorithms obtained in this experiment with respect to the scale on five images are shown in Fig. 3.

From the figure, we can see that the MSD-Net SIFT feature descriptor has the best scale invariance, and still maintains superior stability at the scale of 0.8~1.3, and starts to decline after exceeding the scale, but still has the best performance among several descriptors; the scale invariance of the RIFT descriptor is only second to that of the SIFT descriptor, and the SAR-SIFT and HAPCG are the least.

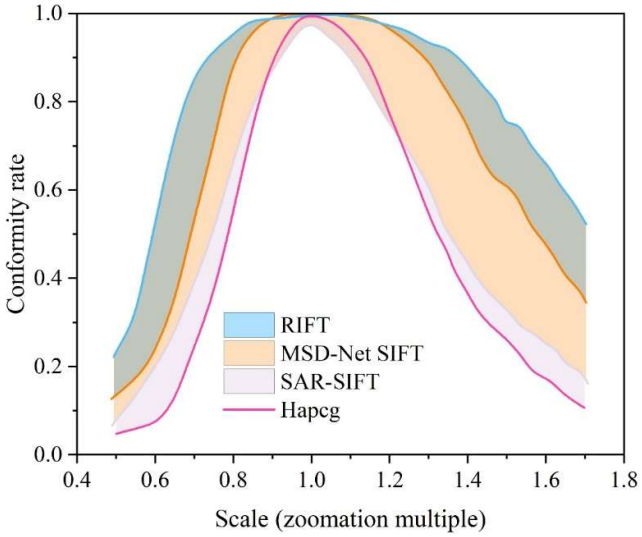


Fig. 3. Results of the scale-invariance experiments with the four algorithms

3.2.3. Noise Resistance Evaluation Experiments. There are many types of noise, and coherent spot noise is mainly dominant in SAR images, so this type of noise is used as the added variable in this experiment. In this experiment, the variance of added coherent spot noise is used as a variable, from 0 to 0.4, with each interval of 0.02 as a group, and a total of 21 sets of experiments are calculated to evaluate the average stability of the four algorithms in the face of several sets of data from the MiniSAR dataset with different degrees of noise, which is calculated in the same way as the previous experiments on the type of scale invariance and rotational invariance, and will not be repeated here. The results of the noise resistance evaluation experiments are shown in Fig. 4. As can be seen from the figure, the MSD-Net SIFT feature descriptor has the best performance in the face of noise, followed by the RIFT and HAPCG feature descriptors, and the SAR-SIFT feature descriptor has the worst performance.

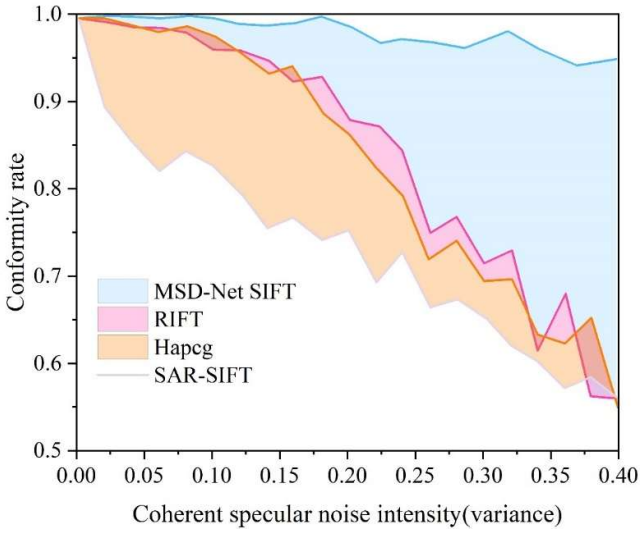


Fig. 4. Algorithm evaluation experiment

3.3. Experimental results and analysis

Our method was evaluated on three datasets using five evaluation metrics, including DSC, SSIM, MI, CC, and Time. DSC measures the degree of overlap between the ROIs in the post-alignment image and those in the baseline image, which takes the value range of [0,1], and the closer it is to 1, the higher the overlap is.SSIM measures the structural similarity between the post-alignment image and the baseline image. MI can measure the probability distribution of the intensity of the benchmark image and the image to be aligned, which is a great reference value for evaluating the effectiveness of cyclic consistency loss in our method.CC measures the degree of correlation between the aligned image and the benchmark image, which takes the value in the range of [-1,1], and the closer the value is to 1, the higher the correlation between the two images. Time can test the efficiency of our methods.

The quantitative results of all the compared methods on the ADNI dataset are shown in Table 1. As can be seen from the table, MSD-Net SIFT outperforms the other methods on all metrics on the ADNI dataset. For the SSIM metrics, MSD-Net SIFT is far ahead compared to the other methods because it has a bootstrap for structural consistency loss. For MI metrics, since MSD-Net SIFT has the bootstrap of cyclic consistency loss, it is highly similar to the image to be aligned both in terms of content and intensity distribution. For DSC metrics and CC metrics, the difference between MSD-Net SIFT and ProsRegNet method is 0.031 and 0.009, respectively, which is not a big difference, but it also has a lead. This is because ProsRegNet uses thin-plate spline to estimate the transformation of parameters, and the accuracy will be higher compared to other methods. However, the experiment shows that the estimation of thin plate spline is not as good as MSD-Net SIFT for refining the feedback parameters.

Table 1. Quantitative results of all comparison methods on the ADNT dataset

Methods	DSC	SSIM	MI	CC	Time
Affine	0.818(0.165)	0.839(0.109)	1.907(0.234)	0.869(0.055)	6s
Sift	0.822(0.137)	0.858(0.101)	1.933(0.171)	0.874(0.041)	8s
Voxel	0.851(0.132)	0.911(0.024)	1.951(0.158)	0.893(0.012)	5s
RSC	0.875(0.140)	0.913(0.011)	1.953(0.151)	0.907(0.034)	5s
LapIRN	0.871(0.034)	0.911(0.019)	1.958(0.093)	0.905(0.014)	4s
ProsRegNet	0.892(0.021)	0.908(0.015)	1.959(0.122)	0.903(0.019)	3s
MICDIR	0.904(0.013)	0.902(0.011)	1.954(0.113)	0.902(0.029)	3s
MSD-Net SIFT	0.923(0.013)	0.917(0.034)	1.972(0.128)	0.911(0.039)	2s

The quantitative results of all the compared methods on the UWF dataset are shown in Table 2. As can be seen from the table, for the UWF dataset, the correlation coefficient (CC) of MSD-Net SIFT is lower than that of LapIRN.This is because LapIRN also ensures good differential characteristics on the basis of multi-resolution, which makes it show good performance when facing complex vascular structures. However, its post-alignment image differs from the to-be-aligned image in terms of content, which can cause greater interference in practical clinical applications.

Table 2. Quantitative results of all comparison methods on the UWF dataset

Methods	DSC	SSIM	MI	CC	Time
Affine	0.674(0.167)	0.801(0.107)	1.895(0.235)	0.872(0.056)	6s
Sift	0.689(0.138)	0.803(0.101)	1.921(0.171)	0.877(0.042)	8s
Voxel	0.707(0.135)	0.806(0.024)	1.939(0.159)	0.896(0.012)	5s
RSC	0.713(0.140)	0.806(0.011)	1.938(0.151)	0.907(0.034)	5s
LapIRN	0.708(0.034)	0.807(0.018)	1.946(0.097)	0.908(0.014)	4s
ProsRegNet	0.709(0.022)	0.807(0.016)	1.947(0.124)	0.906(0.018)	3s
MICDIR	0.712(0.014)	0.806(0.013)	1.942(0.115)	0.905(0.029)	3s
MSD-Net SIFT	0.714(0.012)	0.808(0.035)	1.949(0.129)	0.902(0.038)	2s

The quantitative results of all the compared methods on the RIRE dataset are shown in Table 3.The DSC of MSD-Net SIFT is slightly lower than that of ProsRegNet with a difference of 0.0003 but higher than the other methods. This is because RIRE is a multimodal dataset and ProsRegNet is an algorithm that is specifically designed to align multimodal images, so ProsRegNet performs a little better on multimodal data. But MSD-Net SIFT is better than ProsRegNet in other metrics and in the alignment time, which shows that MSD-Net SIFT shows good performance on multimodal dataset.

Table 3. Quantitative results of all comparison methods on the RIRE dataset

Methods	DSC	SSIM	MI	CC	Time
Affine	0.798(0.019)	0.811(0.195)	1.803(0.039)	0.813(0.028)	5s
Sift	0.799(0.101)	0.804(0.034)	1.806(0.124)	0.815(0.044)	7s
Voxel	0.806(0.094)	0.822(0.153)	1.809(0.112)	0.814(0.012)	4s
RSC	0.807(0.247)	0.825(0.165)	1.809(0.257)	0.816(0.015)	4s
LapIRN	0.805(0.024)	0.828(0.108)	1.813(0.085)	0.815(0.022)	3s
ProsRegNet	0.811(0.201)	0.827(0.103)	1.811(0.158)	0.814(0.014)	2s
MICDIR	0.805(0.009)	0.828(0.148)	1.805(0.121)	0.815(0.027)	2s
MSD-Net SIFT	0.808(0.031)	0.835(0.138)	1.816(0.138)	0.817(0.018)	1s

As can be seen from the three tables, the MSD-Net SIFT alignment time is very short, indicating that the method is also very efficient. In terms of SSIM and MI indexes, MSD-Net SIFT performs optimally on all three datasets and achieves excellent results in all indexes, which proves that the method in this

paper is effective and reasonable.

4. Conclusion

In this paper, an image alignment method based on multi-scale feature union (MSD-Net SIFT) is proposed and the histogram equalization image enhancement algorithm is applied to the image data. Finally, the multi-scale feature union based image alignment method is experimented with four algorithms, namely SAR-SIFT, RIFT and HAPCG. Firstly, the anti-aliasing ability of the feature descriptors of several algorithms is analyzed, and experimental comparisons are made on three datasets to verify the effectiveness of the method in this paper. The results show that the MSD-Net SIFT feature descriptor stability is 99.1%, which is better than other algorithms. Meanwhile, MSD-Net SIFT shows great superiority in the alignment effect.

References

[1] Rigaud, B., Simon, A., Castelli, J., Lafond, C., Acosta, O., Haigron, P., ... & de Crevoisier, R. (2019). Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncologica, 58(9), 1225-1237.

[2] Lei, Y., Fu, Y., Wang, T., Liu, Y., Patel, P., Curran, W. J., ... & Yang, X. (2020). 4D-CT deformable image registration using multiscale unsupervised deep learning. Physics in Medicine & Biology, 65(8), 085003.

[3] Haskins, G., Kruger, U., & Yan, P. (2020). Deep learning in medical image registration: a survey. Machine Vision and Applications, 31(1), 8.

[4] Dhara, A., & Bagaini, C. (2020). Seismic image registration using multiscale convolutional neural networks. Geophysics, 85(6), V425-V441.

[5] Guislain, M., Digne, J., Chaine, R., & Monnier, G. (2017). Fine scale image registration in large-scale urban LIDAR point sets. Computer Vision and Image Understanding, 157, 90-102.

[6] Velesaca, H. O., Bastidas, G., Rouhani, M., & Sappa, A. D. (2024). Multimodal image registration techniques: a comprehensive survey. Multimedia Tools and Applications, 1-29.

[7] Nazir, I., Haq, I. U., AlQahtani, S. A., Jadoon, M. M., & Dahshan, M. (2023). Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion. Journal of Sensors, 2023(1), 6683438.

[8] Arar, M., Ginger, Y., Danon, D., Bermano, A. H., & Cohen-Or, D. (2020). Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13410-13419).

[9] Gopalakrishnan, V., Dey, N., & Golland, P. (2024). Intraoperative 2d/3d image registration via differentiable x-ray rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11662-11672).

[10] Paul, S., & Pati, U. C. (2021). A comprehensive review on remote sensing image registration. International Journal of Remote Sensing, 42(14), 5396-5432.

[11] Sui, X., Zheng, Y., Jiang, Y., Jiao, W., & Ding, Y. (2021). Deep multispectral image registration network. Computerized Medical Imaging and Graphics, 87, 101815.

[12] Gao, C., Li, W., Tao, R., & Du, Q. (2022). MS-HLMO: Multiscale histogram of local main orientation for remote sensing image registration. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

[13] Havel, J., Merciol, F., & Lefèvre, S. (2019). Efficient tree construction for multiscale image representation and processing. Journal of Real-Time Image Processing, 16, 1129-1146.

[14] Zhu, B., Zhou, L., Pu, S., Fan, J., & Ye, Y. (2023). Advances and challenges in multimodal remote sensing image registration. IEEE Journal on Miniaturization for Air and Space Systems, 4(2), 165-174.

[15] Kuppala, K., Banda, S., & Barige, T. R. (2020). An overview of deep learning methods for image registration with focus on feature-based approaches. International Journal of Image and Data Fusion, 11(2), 113-135.

[16] Zou, B., He, Z., Zhao, R., Zhu, C., Liao, W., & Li, S. (2020). Non-rigid retinal image registration using an unsupervised structure-driven regression network. Neurocomputing, 404, 14-25.

[17] Zampieri, A., Charpiat, G., Girard, N., & Tarabalka, Y. (2018). Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 657-673).

[18] Patel, R. K., & Kashyap, M. (2023). The study of various registration methods based on maximal stable extremal region and machine learning. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11(6), 2508-2515.

[19] Wang, Y., Yu, M., Jiang, G., Pan, Z., & Lin, J. (2020). Image registration algorithm based on convolutional neural network and local homography transformation. Applied Sciences, 10(3), 732.

[20] Beroiz, M., Cabral, J. B., & Sanchez, B. (2020). Astroalign: A Python module for astronomical image registration. Astronomy and Computing, 32, 100384.

[21] Brüstle, S. (2018, April). Evaluation of different image processing methods in the context of an image registration workflow. In Geospatial Informatics, Motion Imagery, and Network Analytics VIII (Vol. 10645, pp. 110-120). SPIE.

[22] Song, J., Yang, K., Zhang, Z., Li, M., Cao, T., & Ghaffari, M. (2024, May). Iterative PnP and its application in 3D-2D vascular image registration for robot navigation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 17560-17566). IEEE.

[23] Cocianu, C. L., & Uscatu, C. R. (2022). Multi-scale memetic image registration. Electronics, 11(2), 278.

[24] Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B. P., Išgum, I., & Staring, M. (2017). Nonrigid image registration using multi-scale 3D convolutional neural networks. In Medical Image Computing and Computer Assisted Intervention– MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20 (pp. 232-239). Springer International Publishing.

[25] Liu, S., Wei, G., Fan, Y., Chen, L., & Zhang, Z. (2024). Multimodal registration network with multi-scale feature-crossing. International Journal of Computer Assisted Radiology and Surgery, 19(11), 2269-2278.

[26] N. B. Alkzir,M. S. Yarykina,D. P. Nikolaev & I. P. Nikolaev. (2024). Development of Image Preprocessing Methods for Software Compensation of Refraction Anomalies of an Observer's Eyes. Neuroscience and Behavioral Physiology(prepublish),1-14.

[27] Aarthi D,Panimalar A,Santhosh Kumar S & Anitha K. (2024). Development of Adaptive Gaussian Filter Based Denoising as an Image Enhancement Technique. Mathematical Modelling of Engineering Problems(10).