Journal of Image Processing Theory and Applications (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2025.080109
ISSN 2560-6239 Vol. 8 Num. 1

Image registration with semantic information-guided focus on co-pixel points

Xuedong Liu[1,2,a], Yang Yang[1,2,b,*]

[1]*School of Information Science and Technology, Yunnan Normal University, Kunming, China*
[2]*Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming, China*
[a]*lxd20000317@163.com*, [b]*yyang_ynu@163.com*
**Corresponding author*

Keywords: Image registration, Viewpoint variations, Semantic information-guided, corresponding pixels, Irrelevant interference

Abstract: The goal of image registration is to align pixel points from images captured by different sensors, serving as a reliable foundation for subsequent information fusion and enhancing the performance of advanced vision tasks in degraded scenarios. However, current methods overlook the viewpoint variations caused by internal sensor differences, which introduce information of irrelevant interference, challenging the reliability of registration. Therefore, this paper proposes a semantic information-guided image registration algorithm that directs the network to focus on relevant pixel regions by leveraging the semantic information of corresponding pixels. Through comparative experiments, the proposed method demonstrates superior registration performance compared to existing methods.

1. Introduction

Image registration, as one of the most fundamental tasks in computer vision, refers to aligning the corresponding pixels from images captured by different sensors in the same scene. It is a crucial step that significantly impacts the quality of image fusion [1]. Additionally, it enhances the performance of subsequent visual tasks, such as tracking [2], object detection [3], and image segmentation [4], especially in degraded scenarios. However, challenges arise in achieving accurate and reliable infrared and visible light image registration due to various factors, including viewpoint changes, light intensity variations, and modality differences.

Existing image registration methods are divided into detector-based and detector-free approaches. Detector-based methods [5] first detect points in the image where gradient changes occur, which are then used as keypoints. Descriptors are constructed at these keypoints, and matching is performed between the keypoints of the two images. Finally, pixel alignment is achieved based on the matching correspondences. Clearly, such methods become unreliable in low-texture or repetitive structures. To address this issue, detector-free methods [9] have been proposed. These methods do not detect keypoints but instead perform subpixel-level dense matching. Even in low-texture situations, these methods are capable of performing a large number of dense matches. Therefore, this paper focuses

on detector-free methods. Although detector-free methods have achieved excellent results, they still face challenges due to the following factors.

Due to internal differences between sensors, images captured by different sensors often exhibit viewpoint variations. This introduces a large number of irrelevant pixels between the images, which can result in unnecessary computations when performing subpixel-level detector-free methods. Additionally, these irrelevant pixels may introduce repetitive structures related to the relevant pixels, posing a challenge for subpixel-level matching. Furthermore, the attention mechanism, which is core to implementing detector-free algorithms, may be disturbed by irrelevant pixel information, even though it excels at capturing global contextual information to describe features. This disturbance becomes problematic after viewpoint changes.

To alleviate the above issues, we propose a new detector-free image registration method. Unlike existing methods, we introduce the semantic information of relevant pixels to guide the network's focus toward the relevant pixel regions, thereby achieving more accurate registration results.

2. Related work

The standard process for feature matching typically includes four steps: feature detection, description, matching, and outlier removal. In traditional feature matching methods, SIFT [8] constructs descriptors using image gradient information, while SURF [11] simplifies several steps of SIFT to reduce computational complexity. However, these initial traditional handcrafted methods rely on relatively simple image information, such as gradients and grayscale values, leading to poor performance in feature matching. Therefore, alternative information within the image has been sought to build descriptors that improve results. MSPC [12] combines affine-invariant region extraction with image structural features, providing descriptors with affine and contrast invariance. RIFT [6] extracts repeatable feature points using FAST [13] on phase congruency (pc) maps, primarily addressing radiometric invariance, and performs well in multimodal image feature matching, although it is sensitive to scale differences. Based on RIFT, SRIFT [14] achieves scale and rotation invariance while optimizing computational complexity. Although these traditional handcrafted methods have been widely applied in many scenarios, they require complex computations and the selection of hyperparameters to achieve reliable performance.

However, with the development of deep learning, several deep learning-based methods have been proposed to address the complex computations and manual challenges associated with traditional methods. Additionally, these methods can learn more complex information directly from images through the learning process. The proposal of D2-net [15] has led to a trend where the extraction of feature points and the construction of descriptors are handled within a unified framework. SuperPoint [7] introduces a self-supervised approach that combines synthetic dataset supervision with contrastive learning to capture geometric information in images. SuperGlue [5] integrates the attention mechanism with a graph neural network (GNN) that incorporates the best matching layer, effectively establishing connections between the global contexts of images for feature matching. However, these methods still rely on the reliability of keypoint detection and feature description, making it challenging to ensure robust feature matching under large viewpoint changes, low-texture conditions, or scale variations. To overcome this challenge, a more robust detector-free feature matching strategy has been proposed. This strategy leverages rich context to establish end-to-end correspondences between images without requiring independent keypoint detection and feature description. LoFTR [10] and its variants [9], benefit from the ability of Transformers [16] to collect rich contextual information, enabling state-of-the-art performance.

3. Method

3.1. Framework

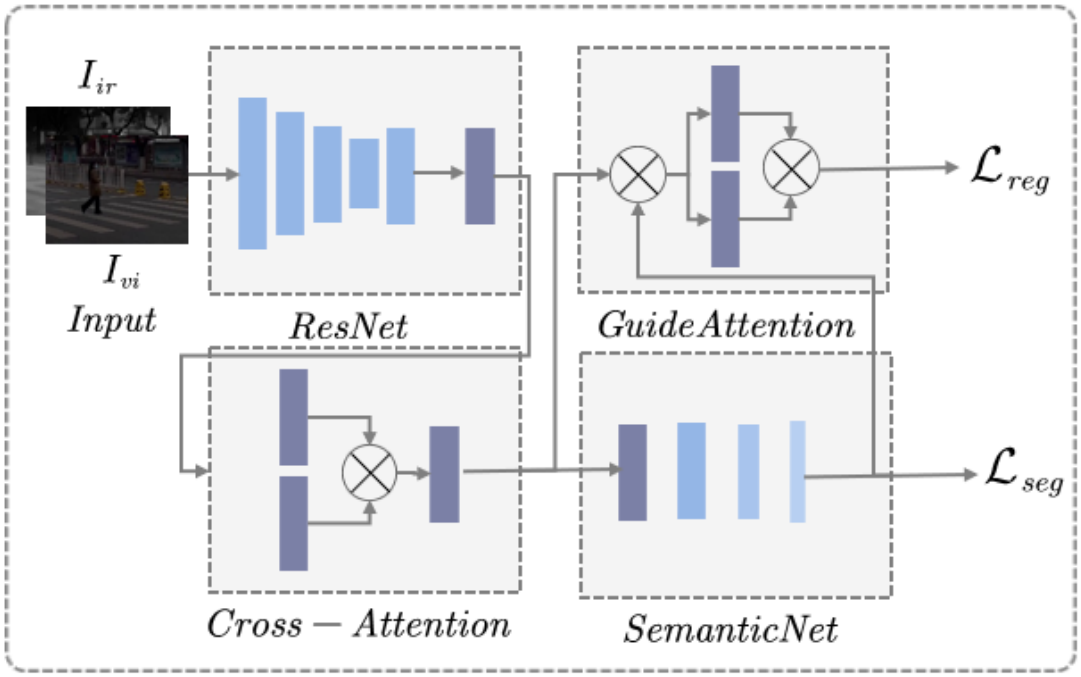

Figure 1: Framework

The network framework, as shown in Figure 1, first takes the infrared and visible light images as input. These images are then passed through ResNet to extract their features, denoted as $f_{vi}, f_{ir} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$. Next, a Cross-Attention mechanism is applied to facilitate the interaction of information between the two images, capturing the relationship between corresponding pixels, which produces the feature $f_{vi}^{at}, f_{ir}^{at} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$. The processed features are then fed into a bifurcated structure. After passing through the SemanticNet side path, the features are combined with the original input features in the GuideAttention module. SemanticNet decodes the input information to extract the semantic details of the relevant pixels, resulting in map $m_{vi}, m_{ir} \in \mathbb{R}^{1 \times \frac{H}{8} \times \frac{W}{8}}$. The GuideAttention module then uses the semantic information to guide the final output, producing the feature $f_{vi}^{reg}, f_{ir}^{reg} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$.

3.2. Cross Attention

After extracting image features using ResNet, although ResNet is capable of capturing deep features of the image, it does not consider the interaction between the images during the feature extraction process. As a result, it lacks information about the corresponding pixel relationships between the image pair. To address this issue, we introduce the Cross Attention mechanism to supplement the missing pixel-related information between the images. Specifically, we first use a linear layer to transform the input feature $f_{vi}(f_{ir})$ into $q_{vi}(q_{ir})$, $v_{vi}(v_{ir})$ and $k_{vi}(k_{ir})$. Then, the image

pair information is exchanged through the following equation (Equation 1):

$$f_{vi}^{at} = f_{vi} + \text{softmax}\left(\frac{q_{vi}{k_{ir}}^{\mathrm{T}}}{\lambda}\right)v_{ir} \tag{1}$$

The equation above represents the process of transmitting information from the infrared image to the visible light image. Similarly, we use the same approach to transmit information from the visible light image to the infrared image.

3.3. SemanticNet and GuidAttention

We treat the feature f_{vi}^{at}, f_{ir}^{at} obtained from Cross Attention as the shared feature for subsequent modules. This shared feature is first fed into SemanticNet. SemanticNet is a simple decoder composed of convolutional layers, normalization layers, and activation layers. It gradually decodes the input shared feature into a mapping that contains the semantic information of the corresponding pixels. The process can be formulated as follows:

$$x_0 = \text{ReLU}\left(\text{Conv}3\times3(f^{at})\right) \tag{2}$$

$$x_1 = \text{ReLU}\left(\text{Conv}3\times3(x_0)\right) \tag{3}$$

$$m = \text{sig}\left(\text{Conv}3\times3(x_1)\right) \tag{4}$$

where $\text{Conv}3\times3$ represents the convolution kernel of size (3,3), $\text{sig}(\cdot)$ denotes the Sigmoid activation function, and m represents the mapping $m_{vi}, m_{ir} \in \mathbb{R}^{1\times\frac{H}{8}\times\frac{W}{8}}$.

After obtaining the relevant semantic mapping m from SemanticNet, we feed it along with the shared feature into the GuideAttention module. We observe that the mapping values of irrelevant semantic pixels in m tend to approach 0 and are significantly smaller than those of the relevant region pixels. This information can be used to suppress the feature expression of irrelevant pixels, which is the key to the guiding process. Specifically, we perform matrix multiplication between m and the shared feature, then send the computed feature into the self-attention and cross-attention mechanism for information transmission. This can be formulated as follows:

$$f^{reg} = f^{at} + SCA\left(m \otimes f^{at}\right) \tag{5}$$

where $\otimes$ represents matrix represents matrix multiplication, and $SCA(\cdot)$ denotes the combination of self-attention and cross-attention.

3.4. Loss function

The loss function of this method consists of two parts: the loss for learning the semantic mapping of relevant pixels and the loss for feature point matching. Specifically, for the relevant semantic mapping, we frame the problem as a binary classification task, i.e., relevant and irrelevant. In other words, the network learns a binomial distribution, so we use a binary cross-entropy loss function. Additionally, the ground truth for training this loss is the same as the ground truth for training feature matching, thus also serving as a guiding signal during the training process to help the network focus on the relevant semantic pixels. The loss function $\mathcal{L}_{seg}$ is calculated as follows:

$$\mathcal{L}_j = -\frac{1}{N}\sum_{i=1}^{N}[\mathcal{M}_j(i)\cdot \log m_j(i) + (1-\mathcal{M}_j(i))\cdot \log m_j(i)] \tag{6}$$

$$\mathcal{L}_{seg} = 0.5\times(\mathcal{L}_{vi}+\mathcal{L}_{ir}) \tag{7}$$

where $\mathcal{M}$ represents the set of successful matches, $j \in (ir, vi)$ and $N = H \times W$.

The matching loss is calculated by the following equation, which is a loss function from [10]

$$P(A(i),B(j)) = \text{Dualsoftmax}(S(A(i),B(j))) \tag{8}$$

$$\mathcal{L}_{reg} = -\frac{1}{\mathcal{M}}\sum_{(i,j)\in\mathcal{M}} \log P(f_{vi}^{reg}(i), f_{ir}^{reg}) \tag{9}$$

where $\text{Dualsoftmax}(\cdot)$ is a dual-softmax operator, $P(A(i),B(j))$ is the set of correspondent feature points, and A and B are for f_{vi}^{reg} and f_{ir}^{reg}, respectively. The final loss $\mathcal{L}_{to}$ is obtained by summing the two losses $\mathcal{L}_{reg}$ and $\mathcal{L}_{seg}$.

$$\mathcal{L}_{to} = \mathcal{L}_{reg} + \mathcal{L}_{seg} \tag{10}$$

4. Experiments

4.1. Implementation Details

The entire network was trained on an NVIDIA GeForce GTX 3090 using PyTorch. The Adam optimizer was used with a learning rate of 3e-4, and training concluded after 30 epochs. The image size during training was 240×320, with a batch size of 12. Additionally, all comparison methods in the experiment used the default parameters of this method for their evaluations.

4.2. Datasets

Training Dataset: We use the COCO dataset as the training set.

Testing Dataset: To comprehensively evaluate the superiority of the proposed method, we will conduct experiments on three public infrared and visible light datasets: M3FD, MSRS, and RoadScene.

4.3. Evaluation Criteria

In the experiment evaluating image registration performance, we assess each network's performance from two perspectives: the number and accuracy of successfully matched features, and the registration error. For the first approach, we use the average number of corresponding feature points (NP), the average number of correctly matched feature points (NCP), and the average matching accuracy (MA) as quantitative metrics. Feature points with a reprojection error less than 8px are considered correctly matched. The formula for MA is as follows:

$$\text{MA} = \text{NCP}/\text{NP} \times 100\% \tag{11}$$

The second comparison method uses metrics such as Mean Squared Error (MSE), Normalized Cross-Correlation (NCC), and Mutual Information (MI) to evaluate the pixel reprojection error after image registration.

4.4. Performance Comparison

Table 1: Quantitative experimental results of feature matching

Dataset	MSRS			M3FD			RoadScene		
Metrics	NP	NCP	MA	NP	NCP	MA	NP	NCP	MA
RIFT	10.94	0.19	1.78	11.17	0.14	1.22	12.67	0.58	4.61
LoFTR	65.56	27.67	42.21	64.17	27.17	42.34	84.00	40.83	48.61
SP+SG	45.81	32.06	69.98	47.72	35.77	74.95	75.50	58.17	77.04
MatchFormer	110.06	69.81	63.43	83.36	56.32	67.56	159.91	117.25	73.32
Aspanformer	88.08	60.30	68.46	82.87	60.74	73.30	91.50	61.83	67.58
ReDFeat	9.03	5.75	63.68	11.39	7.59	66.70	17.50	14.0	80.00
SemLA	148.11	114.08	77.03	152.00	121.17	79.72	196.17	161.42	82.29
Our	197.97	144.69	73.08	180.50	142.75	79.08	283.83	224.17	78.97

Table 2: Quantitative experimental results of image registration

Dataset	MSRS			M3FD			RoadScene		
Metrics	MSE	NCC	MI	MSE	NCC	MI	MSE	NCC	MI
RIFT	0.114	0.084	0.262	0.098	0.037	0.234	0.139	0.050	0.248
LoFTR	0.053	0.529	0.691	0.043	0.534	0.763	0.020	0.840	1.10
SP+SG	0.045	0.610	0.803	0.038	0.617	0.952	0.013	0.874	1.223
MatchFormer	0.036	0.676	0.882	0.032	0.631	0.829	0.015	0.647	0.831
Aspanformer	0.056	0.511	0.742	0.046	0.496	0.763	0.051	0.437	0.605
ReDFeat	0.093	0.245	0.458	0.072	0.260	0.539	0.092	0.437	0.605
SemLA	0.014	0.852	1.140	0.009	0.869	1.122	0.009	0.930	1.320
Our	0.014	0.848	1.094	0.009	0.864	1.168	0.007	0.940	1.392

From the results of Table 1 and Table 2, it is obvious that our method is more effective than other methods.

5. Conclusion

In this study, we propose a novel detector-free image registration method that incorporates the semantic information of relevant pixels. By using this semantic information, the network is guided to focus on the relevant pixel regions. Experimental results demonstrate that the proposed method outperforms existing methods on all datasets, leading to the conclusion that the guidance of high-level semantic information effectively alleviates the challenges posed by viewpoint variations. This conclusion can provide further guidance for the design of future network architectures. Additionally, for future work, we plan to explore the interpretability brought by higher-level semantic information to optimize network design further.

References

[1] Tang L, Yuan J, Ma J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82: 28-42.

[2] Aharon N, Orfaig R, Bobrovsky B Z. Bot-sort: Robust associations multi-pedestrian tracking[J]. arXiv preprint arXiv:2206.14651, 2022.

[3] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[4] Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation[C]. Proceedings of the IEEE conference

on Computer Vision and Pattern Recognition. 2018: 7151-7160.
[5] Sarlin P E, DeTone D, Malisiewicz T, et al. Superglue: Learning feature matching with graph neural networks[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4938-4947.
[6] Li J, Hu Q, Ai M. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform[J]. IEEE Transactions on Image Processing, 2019, 29: 3296-3310.
[7] DeTone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description[C]. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 224-236.
[8] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60: 91-110.
[9] Wang Q, Zhang J, Yang K, et al. Matchformer: Interleaving attention in transformers for feature matching[C]. Proceedings of the Asian Conference on Computer Vision. 2022: 2746-2762.
[10] Sun J, Shen Z, Wang Y, et al. LoFTR: Detector-free local feature matching with transformers[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8922-8931.
[11] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[C]. Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. Springer Berlin Heidelberg, 2006: 404-417.
[12] Liu X, Ai Y, Zhang J, et al. A novel affine and contrast invariant descriptor for infrared and visible image registration[J]. Remote Sensing, 2018, 10(4): 658.
[13] Rosten E, Drummond T. Machine learning for high-speed corner detection[C]. Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. Springer Berlin Heidelberg, 2006: 430-443.
[14] Cui S, Xu M, Ma A, et al. Modality-free feature detector and descriptor for multimodal remote sensing image registration[J]. Remote Sensing, 2020, 12(18): 2937.
[15] Dusmanu M, Rocco I, Pajdla T, et al. D2-net: A trainable cnn for joint description and detection of local features[C]. Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2019: 8092-8101.
[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.