Image Mosaic Research and Realization Based on LoFTR Algorithm

Aikui Tian
Shandong University of Technology
Kangtao Wang
Shandong University of Technology
liye zhang (zhangliye@sdut.edu.cn)
Shandong University of Technology https://orcid.org/0000-0002-4300-1789
Bingcai Wei
Shandong University of Technology

Research Article

Keywords: Image mosaic, LoFTR algorithm, Weighted average fusion

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1107577/v1

RESEARCH

Image mosaic research and realization based on LoFTR algorithm

Aikui Tian[1,2], Kangtao Wang[1], Liye Zhang[1,2*] and Bingcai Wei[1]

*Correspondence:
zhangliye@sdut.edu.cn
[1]School of Computer Science and Technology, Shandong University of Technology, Zibo, China
[2]Shandong Big Data Development and Innovation Laboratory, Zibo, China
Full list of author information is available at the end of the article

Abstract

Aiming at the problem of inaccurate extraction of feature points by the traditional image matching method, low robustness, and problems such as diffculty in inentifying feature points in area with poor texture. This paper proposes a new local image feature matching method, which replaces the traditional sequential image feature detection, description and matching steps. First, extract the coarse features with a resolution of 1/8 from the original image, then tile to a one-dimensional vector plus the positional encoding, feed them to the self-attention layer and cross-attention layer in the Transformer module, and finally get through the Differentiable Matching Layer and confidence matrix, after setting the threshold and the mutual closest standard, a Coarse-Level matching prediction is obtained. Secondly the fine matching is refined at the Fine-level match, after the Fine-level match is established, the image overlapped area is aligned by transforming the matrix to a unified coordinate, and finally the image is fused by the weighted fusion algorithm to realize the unification of seamless mosaic of images. This paper uses the self-attention layer and cross-attention layer in Transformers to obtain the feature descriptor of the image. Finally, experiments show that in terms of feature point extraction, LoFTR algorithm is more accurate than the traditional SIFT algorithm in both low-texture regions and regions with rich textures. At the same time, the image mosaic effect obtained by this method is more accurate than that of the traditional classic algorithms, the experimental effect is more ideal.

Keywords: Image mosaic; LoFTR algorithm; Weighted average fusion

1 Introduction

Image mosaic refers to the synthesis of two or more images with obvious overlapping areas into an image with wide viewing angle and high resolution. The fused image is similar to the multiple images before fusion, and has most of the image information. At present, image mosaic technology is widely used in many fields such as computer vision, image processing, vehicle-assisted driving, human-computer interaction and computer graphics.

Image mosaic technology includes two key links: image registration and image fusion. Image registration directly determines the quality and efficiency of image mosaic. Local feature matching between images is the cornerstone of many three-dimensional computer vision tasks. The existing matching methods include feature detection, feature description and feature matching. Due to various factors such as texture difference, viewpoint change, illumination change and motion blur, traditional algorithms may be unable to find repeatable interest points or find corresponding relationships according to descriptors. To solve the above problems,

researchers have proposed a variety of detector-free local feature matching methods, such as SIFT (Scale-Invariant Feature Transformation) algorithm. Traditional feature matching algorithms include Harris algorithm, ORB (Oriented FAST and Rotated BRIEF) algorithm, and SIFT are proposed [1] on the basis of Moravec algorithm [2]. Moravec algorithm is a corner detection operator based on gray-scale variance. The operator calculates the gray-scale variance of a pixel in the image along the horizontal, vertical, diagonal and anti-diagonal directions. The minimum value is selected as the interest value of the pixel, and then the local non-maximum suppression is used to detect whether it is a corner point. Harris algorithm is an improvement and optimization of Moravec operator. Harris has the characteristics of rotation invariance, insensitive to gray translation and scale changes, but Harris corner detection operator does not have scale invariance, resulting in poor feature matching effect. The SIFT algorithm was first proposed by David G. Lowe in 1999, and was improved and officially published in 2004 [3]. SIFT feature is the local feature of the image, which is invariant to its rotation, scale and brightness change, and also maintains a certain stability to the change of viewing angle and noise. The algorithm constructs 128-dimensional vectors for feature points, which leads to the slow operation speed. Therefore, H. Bay et al. proposed an algorithm [4] to improve the SIFT algorithm, the improved algorithm is much faster. E. Rublee et al. proposed a very fast binary descriptor based on BRIEF [5], which is a fast feature point extraction and description algorithm. In 2016, K. M. Yi et al. proposed the LIFT (Learned Invariant Feature Transform) algorithm [6], which introduces a novel network architecture that uniformly addresses three provious problems: feature detection, direction estimation and description. Many multiple visual geometry problems cannot be solved by traditional algorithms, D. DeTone et al. proposed a self-supervised framework to train key point detectors, and it is suitable for feature descriptors for multi-view geometric problems in 2018 [7].

As one of the key technologies of image mosaic, image fusion has been deeply studied by many researchers. The AKAZE-based image mosaic algorithm proposed by S. K. Sharma et al. in the reference [8] minimizes the mosaic seam and generates a perfect mosaic image. A fast sonar image mosaic method is proposed in the reference [9]. This method was composed of denoising, feature point extraction, mosaic and optimization, and the quality of mosaic image was effectively improved. The deep convolutional neural network was used in reference [10] to adaptively obtain the image features, and the results in this paper showed the robustness and effectiveness of the proposed method. Reference [11] proposed a principal component invariant feature transformation, the proposed algorithm improved the speed of image mosaic and was conducive to image fusion on the premise of ensuring the mosaic quality. Reference [12] introduced a efficient mosaic method of 6-DoF imaging model, which reduced the number of unknown variables for parameter optimization. Compared with the existing methods, this method effectively obtained more accurate mosaic results. J. Zhang et al. proposed a RANSAC algorithm based on block matching in the reference [13] to eliminate mismatch points in the process of key point matching and this algorithm had strong robustness. L. Li et al. proposed a suture detection algorithm [14], which can effectively hide the artifacts caused by dynamic objects and geometric misalignment. Y. Wang et al. proposed an image

mosaic algorithm based on empirical mode decomposition transformation [15], the computational time was reduced significantly by this method. Z. Yang et al. proposed an image serialization method on line segment acceleration robust features [16], which realized the robustness to panoramic images. H. Nejad et al. proposed a new hybrid algorithm of Gaussian weighting function [17], and this method has superiority in image mosaic and image registration. J. Kaur proposed a normalized improved SIFT algorithm in reference [18]. Compared with the traditional SIFT, the computational time was reduced, the efficiency was improved.

However, there are still some problems in real scenes, the traditional feature point matching algorithm is inaccurate in regions with poor texture, and poor robustness, resulting in poor image mosaic effects. With the continuous development of deep neural networks, a new detector free local feature matching method is proposed to generates dense descriptors or dense feature matching. By adopting this method, the matching accuracy and robustness are greatly improved. Then, the weighted average fusion algorithm is used to fuse the image. The quality of the mosaic image is significantly improved by using our method.

The block diagram of image mosaic process is shown in Fig. 1.

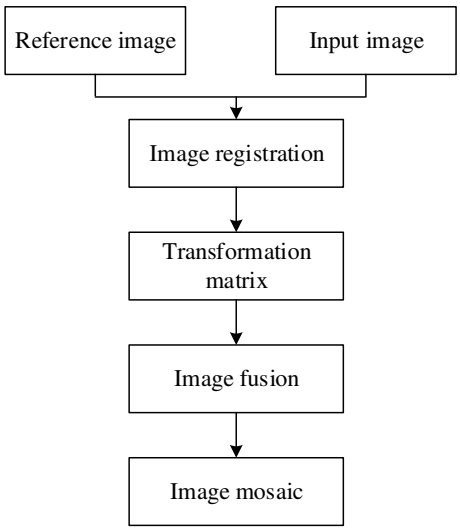


Figure 1 Block diagram of image mosaic process

2 Methods

Given the image pairs I^A and I^B, the local feature matching methods uses a feature detector to extract interest points. We propose a detector free design to solve the repeatability problem of feature detector. An overview of proposed method LoFTR is presented in Fig.2.

2.1 Local Feature Extraction function

First, a Feature Pyramid Network (FPN) [19] of local feature convolutional neural network is used to extract the multi-level features from both images. As shown in Fig.2. $\tilde{F}^A$ and $\tilde{F}^B$ are the coarse-level features represented by I^A and I^B of the

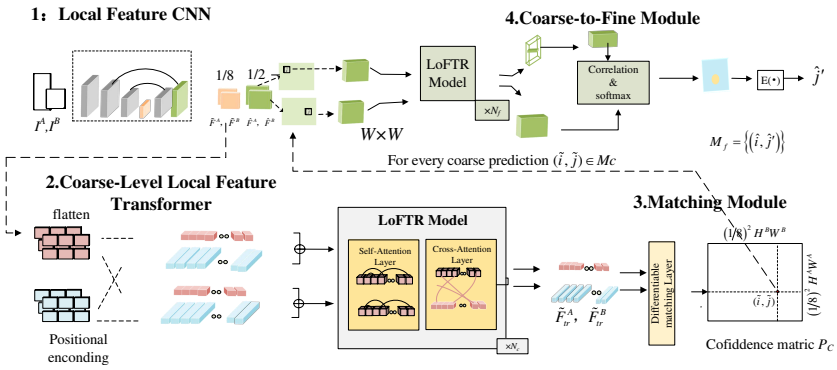


Figure 2 Principle of LoFTR algorithm

original image respectively, and their dimensions are 1/8 of the original image. $\tilde{F}^A$ and $\tilde{F}^B$ are the fine-level features represented by I^A and I^B of the original image respectively, and their dimensions are 1/2 of the original image.

2.2 Transformer module

After local feature extraction, $\tilde{F}^A$ and $\tilde{F}^B$ feature map is tiled into a one-dimensional vector, and added to the corresponding position coding, the added feature enters the LoFTR module for processing, the processed feature is expressed as $\tilde{F}^A_{tr}$ and $\tilde{F}^B_{tr}$.

Transformer consists of encoder and decoder, LoFTR Module consists of self-attention layer and cross-attention layer. The encoder of transformer consists of sequentially encoders, as shown in Fig.3 (a).The key element of the encoder layer is the attention layer. Attention layer input is generally a query vector (Query, Q), a key vector (Key, K) and a value vector (Value, V). The query vector Q retrieves information from the value vector V according to the amount of attention calculated from the dot product of Q and the key vector K corresponding to each value v. the calculation diagram of the attention layer is shown in Fig. 3 (b), and the formula of the attention layer can be denoted as:

$$\text{Attention}(Q, K, V) = \text{soft}\max(QK^T)V \tag{1}$$

The standard positional encoding proposed by DETR [20] is used in transformer. The positional encoding provides unique position information for each element in a sinusoidal format, by adding position code to $\tilde{F}^A$ and $\tilde{F}^B$, the transformed features will depend on the location information.In the linear transformer, the dot product between Q and K is represented as N, and their feature dimension as D. In the case of local matching, it is unreasonable to directly use the original version of the transformer. To solve this problem, we use an alternative kernel function with the exponential kernel used in the original attention layer, as shown in Fig.3 (c).

2.3 Establish Coarse-level Matches

The Optimal Transport (OT) layer in the Differentiable Matching Layer can be used in LoFTR. The score matrix S between the transformed features is first

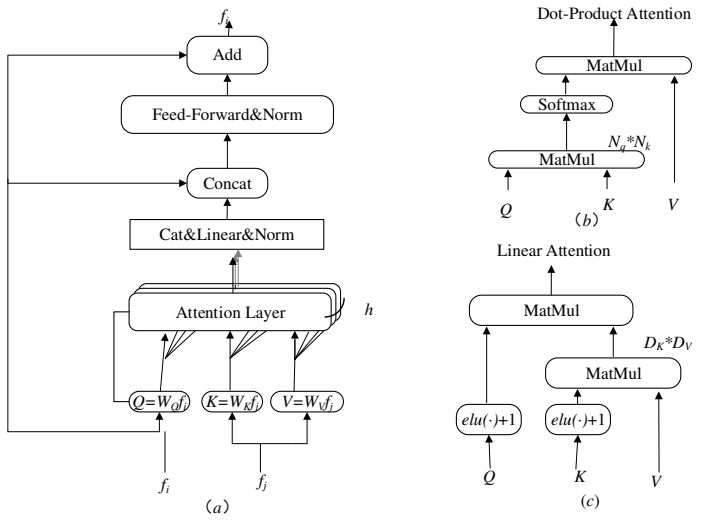


Figure 3 Encoder layer and attention layer in LoFTR

calculated by $S(i,j) = \frac{1}{\tau} \cdot \left\langle \tilde{F}_{tr}^{A}(i)\tilde{F}_{tr}^{B}(j) \right\rangle$. Where τ is the dimension of feature $\tilde{F}_{\mathrm{tr}}$.When matched with OT, -S can be used as the cost matrix for the partial allocation problem in [21]. We can also apply softmax to the two dimensions of S (hereinafter referred to as double softmax) to obtain the probability of soft mutual nearest neighbor matching. Formally, when using dual-softmax, the matching probability P_c is obtained by:

$$P_c(i,j) = \mathrm{soft\,max}\left(S(i,\cdot)\right)_j \cdot \mathrm{soft\,max}\left(S(\cdot,j)\right)_i \tag{2}$$

Matching selection is based on confidence matrix P_c. We select matches with confidence higher than the θ_c threshold and further implement the Mutual Nearest Neighbor (MNN) standard to filter the possible Coarse-level matches. We express the Coarse-level matches predictions as:

$$M_c = \left\{ \left(\tilde{i},\tilde{j}\right) | \forall \left(\tilde{i},\tilde{j}\right) \in \mathrm{MNN}\left(P_c\right), P_c\left(\tilde{i},\tilde{j}\right) \geq \theta_c \right\} \tag{3}$$

2.4 Establish Fine-level Matches

After estabishing coarse matches, these matches are re-fined to the original image resolution with Coarse-to-Fine Module. For every coarse match $(\hat{i}, \hat{j})$, we first locate its position $(\hat{i}, \hat{j})$ at the Fine-level feature maps $\hat{F}^A$ and $\hat{F}^B$,and then crop two sets of local windows of size $w \times w$. A smaller LoFTR Module transforms the cropped features within each window by N_f times,yielding two transformed local feature maps $\tilde{F}_{tr}^{A}(\hat{i})$ and $\tilde{F}_{tr}^{B}(\hat{i})$ centered on $\hat{i}$ and $\hat{j}$, respectively.Then, we associate the center vector of $\tilde{F}_{tr}^{A}(\hat{i})$ with all the vectors in $\tilde{F}_{tr}^{B}(\hat{j})$ to generate a heat map, represents the matching probability of each pixel near $\hat{j}$ with $\hat{i}$. By calculating the expectation over the probability distribution,we get the final position $\hat{j}'$ with sub-pixel accuracy on I^B.Gathering all the matches $\left\{\left(\hat{i},\hat{j}'\right)\right\}$ products the final Fine-level matches M_f.

2.5 Image fusion

Image fusion is the last step of image mosaic, two pictures after feature matching are fused into a complete picture. In this paper, the gradual in and gradual out weighted image fusion method is used image fusion, which is also called Weighted Averaging (WA), which is a relatively simple image fusion algorithm. Weighted image fusion method has the advantages of fast calculation speed, simple operation and simple implementation. In addition, the method can suppress the noise in the fused image, improve the signal to noise of the image and other advantages[22].

The main idea of the weighted image fusion method is to carry out the weighted re-fusion of the pixel values of the two images respectively, and realize the seamless mosaic of the images by setting the weight value. The formula is:

$$I(x,y) = \begin{cases} I_1(x,y) & (x,y) \in I_1 \\ w_1 I_1(x,y) + w_2 I_2(x,y) & (x,y) \in (I_1 \cap I_2) \\ I_2(x,y) & (x,y) \in I_2 \end{cases} \tag{4}$$

Where $I_1(x,y)$ and $I_2(x,y)$ are two fused images, w_1 and w_2 are the weights of image fusion,and $0 < w_1 < 1$, $0 < w_2 < 1$, $w_1 + w_2 = 1$,General take:

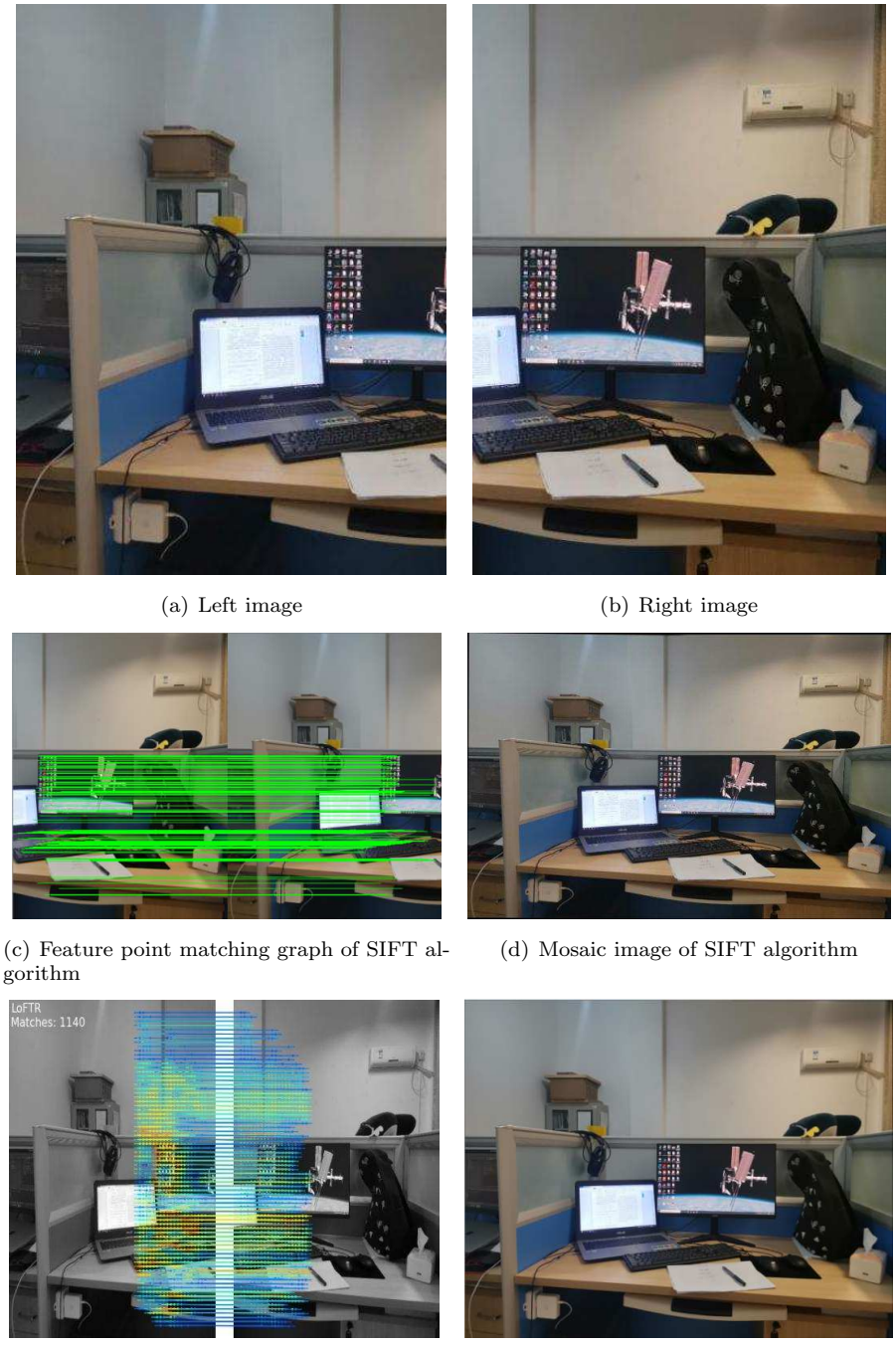
$$w_1 = d_1/width \tag{5}$$

$$w_2 = d_2/width \tag{6}$$

The d_1 and d_2 respectively represent the distance from the overlapping point (x,y) to the left and right boundaries, and $width$ represents the width of the overlap area. And $d_1 + d_2 = width$.

3 Results and discussion

In order to verify the effectiveness of the image mosaic method proposed in this paper, a comparative experiment was carried out. The experimental environment is CPU Intel Core i5-8400, the main frequency is 2.81GHz, the memory is 8GB, the operating system is Windows10, and the development environment is PyCharm Community Edition 2020.2.2. This article selects an image with a resolution of 1200×1440 for experimentation. Use the mobile phone camera to shoot in indoor scenes and outdoor scenes. Take a set of photos in each scene and use SIFT algorithm and LoFTR algorithm to match the feature points of the picture. And use the weighted average image fusion method to perform image fusion according to the result of feature point matching. The results of algorithm feature point matching and image mosaic are shown in Fig.4 and Fig.5 respectively. Take Fig.4 as an example. Figures (a) and (b) respectively represent the left and right images taken, and Figures (c) and (d) respectively represent the feature point matching picture using the SIFT algorithm and the mosaic picture using the SIFT algorithm. Figure (e) and Figure (f) respectively represent the feature point matching picture using the LoFTR algorithm and the mosaic picture using the LoFTR algorithm.



(a) Left image

(b) Right image

(c) Feature point matching graph of SIFT algorithm

(d) Mosaic image of SIFT algorithm

(e) Feature point matching graph of LoFTR algorithm

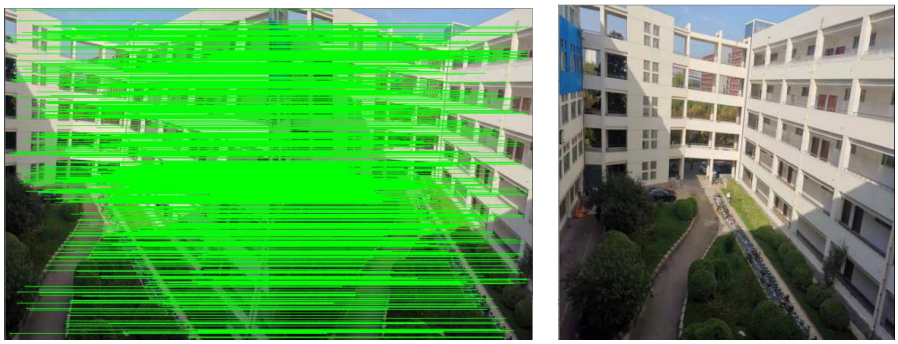
(f) Mosaic image of LoFTR algorithm

Figure 4 Comparison of the mosaic effect of the first group images (indoor)

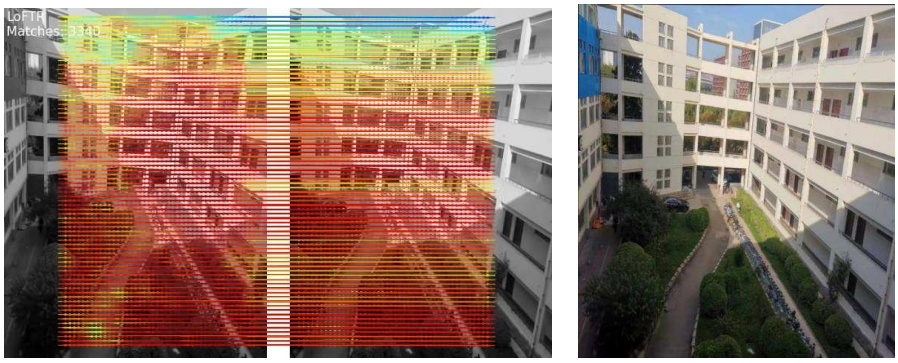
Table 1 Comparison of peak signal-to-noise ratio (PSNR) of the two algorithms

Number of groups/algorithm PSNR value(dB)	LoFTR	SIFT
First group	24.81	23.11
Second group	24.24	21.16

(a) left image (b) right image



(c) Feature point matching graph of SIFT algorithm (d) Mosaic image of SIFT algorithm



(e) Feature point matching graph of LoFTR algorithm (f) Mosaic image of LoFTR algorithm

Figure 5 Comparison of the mosaic effect of the second group images (outdoor)

It can be seen from Table 1 that the PSNR value obtained by the LoFTR algorithm is always higher than that of the SIFT whether it is indoor or outdoor. The higher the PSNR value, the better the image quality. Therefore, from the perspective of the peak signal-to-noise ratio, the images stitched by the LoFTR algorithm are better than the traditional algorithm SIFT. In addition, from the analysis of

feature matching theory, the LoFTR algorithm extracts more accurately than the traditional SIFT algorithm and recognizes dense matching points. Therefore, the weighted average fusion algorithm is used for splicing at the same time, and the LoFTR algorithm splicing is also better.

Table 2 Comparison of structural similarity (SSIM) between the two algorithms

Number of groups/algorithm SSIM value	LoFTR	SIFT
First group	0.855	0.824
Second group	0.795	0.639

It can be seen from Table 2 that the SSIM value of the LoFTR algorithm is higher than that of the traditional algorithm SIFT, whether it is indoors with low texture or outdoors with rich textures. The higher the SSIM value, the higher the image quality. Therefore, from the perspective of structural similarity, the LoFTR algorithm has better splicing quality than the traditional SIFT algorithm.

4 Conclusions

Aiming at the problem of inaccurate and low robustness of traditional algorithm image matching methods, a new local image feature matching method combined with deep neural networks is proposed in this paper. First, the pixel-intensive matching is established at the coarse level, then the good matching is refined at the fine level, and finally the image is fused by the weighted tie fusion algorithm. Finally, experiments show that the accuracy of this method for feature matching between two images is higher than that of the traditional method, especially in areas with poor texture, which can better extract feature points, and through the comparison of the peak signal-to-noise ratio and the structural similarity value, It can be clearly seen that the LoFTR algorithm is higher than the traditional SIFT algorithm, and the effect of the stitched image obtained by image fusion is also better than that of the traditional algorithm.

Acknowledgements
This work is supported by School of Computer Science and Technology, Shandong University of Technology.

Funding
This paper is supported by Shandong Provincial Natural Science Foundation, China (grant number ZR2019BF022), and National Natural Science Foundation of China (grant number 62001272).

Abbreviations
ORB: Oriented FAST and Rotated BRIEF; SIFT: scale invariant feature transformation; LoFTR: local feature transformer; PSNR: Peak Signal-to-Noise Ratio; SSIM: Structural SIMilarity; LoFTR: local feature transformer; LIFT: Learned Invariant Feature Transform; WA: Weighted Averaging

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

Consent for publication
Not applicable.

Authors' contributions
The algorithms proposed in this paper have been conceived by L. Zhang, and K. Wang. L. Zhang, K. Wang and B. Wei made the analysis and experiment and wrote the paper. A. Tian, B. Wei and K. Wang investigated, validated, and revised this paper. The authors approved the final manuscript.

Authors' information
Aikui Tian is with the School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, 255000, China. He received Ph.D. degree in Instructional Technology from East China Normal University, Shanghai, China, in 2007. He is a professor in Shandong University of Technology. His research interests include Big Data and Virtual Reality.
Kangtao Wang received his bachelor's of Engineering degree from the School of Software of Pingdingshan College in 2020. He is currently pursuing a master's degree in the School of Computer Science and Technology at Shandong University of Technology. His current research interests include machine learning and image stitching.
Liye Zhang is with the School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, 255000, China. He received the B.A.Sc. degree in Electronic Information Engineering, the M.A.Sc. and Ph.D. degree in Information and Communication Engineering from Harbin Institute of Technology, Harbin, Heilongjiang, China, in 2009, 2011 and 2018 respectively. He is a lecturer in Shandong University of Technology. His research interests include WLAN indoor localization, machine learning and Computer Vision.
Bingcai Wei received the bachelor's degree in software engineering from Qufu Normal University. He is currently pursuing the M.Sc. degree with the School of Computer Science and Technology, Shandong University of Technology. His current research interests include machine learning and image processing.

Author details
[1]School of Computer Science and Technology, Shandong University of Technology, Zibo, China. [2]Shandong Big Data Development and Innovation Laboratory, Zibo, China.

References

1. Harris, C., Stephens, M., *et al.*: A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, pp. 10–5244 (1988). Citeseer
2. Moravec, H.P.: Rover visual obstacle avoidance. In: IJCAI, vol. 81, pp. 785–790 (1981)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
4. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision & Image Understanding **110**(3), 346–359 (2008)
5. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International Conference on Computer Vision, pp. 2564–2571 (2011). Ieee
6. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European Conference on Computer Vision (2016)
7. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 224–236 (2018)
8. Sharma, S.K., Jain, K.: Image stitching using akaze features. Journal of the Indian Society of Remote Sensing **48**(10), 1389–1401 (2020)
9. Tang, Z., Ma, G., Lu, J., Wang, Z., Fu, B., Wang, Y.: Sonar image mosaic based on a new feature matching method. IET Image Processing **14**(10), 2149–2155 (2020)
10. Zeng, Y., Ning, Z., Liu, P., Luo, P., Zhang, Y., He, G.: A mosaic method for multi-temporal data registration by using convolutional neural networks for forestry remote sensing applications. Computing **102**(3), 795–811 (2020)
11. Zhu, J., Gong, C., Zhao, M., Wang, L., Luo, Y.: Image mosaic algorithm based on pca-orb feature matching. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences **42**, 83–89 (2020)
12. Tian, Y., Sun, A., Luo, N., Gao, Y.: Aerial image mosaicking based on the 6-dof imaging model. International Journal of Remote Sensing **41**(1), 1–16 (2019)
13. Zhang, J., Yin, X., Luan, J., Liu, T.: An improved vehicle panoramic image generation algorithm. Multimedia Tools and Applications **78**(19), 27663–27682 (2019)
14. Li, L., Yao, J., Li, H., Xia, M., Zhang, W.: Optimal seamline detection in dynamic scenes via graph cuts for image mosaicking. Machine Vision and Applications **28**(8), 819–837 (2017)
15. Wang, Y.-j., Wei, S.: An efficient image mosaic algorithm based on emd transform. In: 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017), pp. 605–609 (2017). Atlantis Press
16. Yang, Z., Shen, D., Yap, P.-T.: Image mosaicking using surf features of line segments. PloS One **12**(3), 0173627 (2017)
17. Hossein-Nejad, Z., Nasri, M.: Clustered redundant keypoint elimination method for image mosaicing using a new gaussian-weighted blending algorithm. The Visual Computer, 1–17 (2021)
18. Kaur, J.: A robust technique for image mosaicing using modified sift. Indian Journal of Science and Technology **9**(47) (2016)
19. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
20. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer
21. Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947 (2020)
22. De Bethune, S., Muller, F., Binard, M.: Adaptive intensity matching filters: a new tool for multi-resolution data fusion. In: AGARD CONFERENCE PROCEEDINGS AGARD CP, pp. 28–28 (1998). AGARD