Journal of Systems Engineering and Electronics
Vol. 33, No. 2, April 2022, pp.259 – 268

Vision-based aerial image mosaicking algorithm with object detection

HAN Jun, LI Weixing*, FENG Kai, and PAN Feng

School of Automation, Beijing Institute of Technology, Beijing 100081, China

Abstract: Aerial image sequence mosaicking is one of the challenging research fields in computer vision. To obtain large-scale orthophoto maps with object detection information, we propose a vision-based image mosaicking algorithm without any extra location data. According to object detection results, we define a complexity factor to describe the importance of each input image and dynamically optimize the feature extraction process. The feature points extraction and matching processes are mainly guided by the speeded-up robust features (SURF) and the grid motion statistic (GMS) algorithm respectively. A robust reference frame selection method is proposed to eliminate the transformation distortion by searching for the center area based on overlaps. Besides, the sparse Levenberg-Marquardt (LM) algorithm and the heavy occluded frames removal method are applied to reduce accumulated errors and further improve the mosaicking performance. The proposed algorithm is performed by using multithreading and graphics processing unit (GPU) acceleration on several aerial image datasets. Extensive experiment results demonstrate that our algorithm outperforms most of the existing aerial image mosaicking methods in visual quality while guaranteeing a high calculation speed.

Keywords: image mosaicking, object detection, grid motion statistic (GMS), mapping.

DOI: 10.23919/JSEE.2022.000026

1. Introduction

Aerial imagery captured by unmanned aerial vehicle (UAV) has already been adopted in many tasks such as image mosaicking, object detection and semantic segmentation. The low-altitude UAV remote sensing images have the advantages of high resolution, real time and convenience, which meet the application demands of terrain mapping and disaster monitoring by UAV platform. To take the global information by an image sequence, image mosaicking methods are designed to stitch a sequence of images and generate a large global image. Most existing aerial image mosaicking algorithms [1−3] require auxiliary information, such as ground control points, the position and orientation information from global navigation satellite system (GNSS) and inertial measurement unit (IMU). Undoubtedly, the extra information limits their practical applications and introduces computing complexities. While for the purely vision-based method, it seems difficult to generate satisfying ortho-mosaics due to the lack of sufficient cues. To address this issue, we present a vision-based aerial image mosaicking method with object detection cues.

The applications of deep learning techniques in aerial images have shown great performance superiority in many fields [4,5], like forest fire early warning and geological environment monitoring. However, most existing methods only take one image into consideration, which fail to collect global information. With the combination of image mosaicking and object detection, the proposed method can make full use of the image sequence to integrate global information. Besides, the performance of purely vision-based image mosaicking methods is largely dependent upon feature points extraction and matching. Object detection results play an important role in offering semantic cues for image mosaicking [6]. We dynamically adjust the feature extraction and matching processes based on detection information and further optimize the image mosaicking workflow.

To summarize, our main contributions are as follows:

(i) We combine object detection and aerial image mosaicking into a unified task. It is convenient to produce a global orthophoto image with object detection information of the large-scale area.

(ii) We dynamically adjust the feature extraction and the matching processes based on object detection results. By scaling the complexity of different scenes with detection information, we dynamically optimize the feature extraction process.

(iii) We present an adaptive reference frame selection method. The reference frame can be dynamically picked

Manuscript received September 25, 2020.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61603040; 61973036).

through calculating overlaps of aerial image sequence.

2. Related work

The aerial image mosaicking methods can be roughly divided into two categories. One is used to generate a large scale orthophoto map under the camera pure translation assumption; the other is the panoramic image stitching method, which takes camera rotation into consideration. Szeliski [7] and Zitova et al. [8] discussed the related work of image stitching in details and proposed standard steps for image stitching. The main contents usually include feature points extraction and matching, projection transformation and image blending. Feature points extraction and matching are used to establish corresponding relationship between two images with extracted feature points. The main process of projection transformation is to transform aerial images to the coordinate system of the reference frame according to feature matching results. Image blending primarily eliminates stitching gaps and obtains more visually satisfying results.

The most commonly used feature points extraction and description methods mainly include scale-invariant feature transform (SIFT) [9], speed up robust features (SURF) [10], oriented fast and rotated brief (ORB) [11], features from acceletated segment test (FAST) [12], etc. As for feature matching, brute force (BF), fast library for approximate nearest neighbors and grid motion statistic (GMS) [13] algorithms are the most popular choices. With the optimization of the SURF-Harris algorithm, An et al. [14] proposed a high speed robust image registration and localization method for feature points generating and matching. Amiri et al. [15] used the SURF algorithm to extract feature points and described them with binary robust invariant scalable keypoints (BRISK) descriptors. Coupled with random sample consensus (RANSAC), a real-time UAV video mosaicking system was made into reality. Li et al. [16] utilized GMS to perform high-quality feature matching on aerial images, achieving fast and accurate image sequence stitching. However, these methods cannot dynamically adjust the stitching process for different scenes.

Due to coordinate transformation biases, it is inevitable to introduce stitching errors. Therefore, more and more optimization algorithms are designed for eliminating transformation errors. Mclauchlan et al. [17] carried out a refinement task using bundle adjustment method and further reduced the projection transformation errors. Xu et al. [18] proposed a stitching method based on multiregion guided local projection deformation to reduce ghosting. Wu et al. [19] analyzed the projective and similarity transformation and then proposed a mosaicking method combination of as-projective-as-possible warps and similarity transformation. To improve the visual quality of mosaicking results, many researchers have paid much attention to fusion works. Li et al. [20] took full advantage of the graph cut method to process stitching gaps and acquired the seamless global orthophoto map. Yuan et al. [21] utilized a superpixel-based color blending method to obtain seamless image stitching results. Despite these works' efforts to eliminate stitching errors, their time cost is too expensive to be practical. We propose a novel reference frame selection method to avoid obvious stitching errors, which is simple and effective.

Ge et al. [22] realized rapid stitching of aerial images assisted by ground control points. Integrated with camera poses and Global Positioning System (GPS) coordinates, Bu et al. [3] designed a monocular simultaneous localization and mapping (SLAM) framework, which can generate 3D point cloud and mosaicking results. Different from these methods, our proposed algorithm only relies on vision cues and is more convenient for practical applications.

3. Methodology

In this paper, we propose a novel vision-based aerial image mosaicking method, which integrates image mosaicking and object detection into a unified framework. Based on object detection results, we dynamically optimize the mosaicking process, contributing to improving the visual quality. An adaptive reference frame selection method is designed to eliminate accumulated transformation errors. The overall structure of our proposed algorithm is shown in Fig. 1.

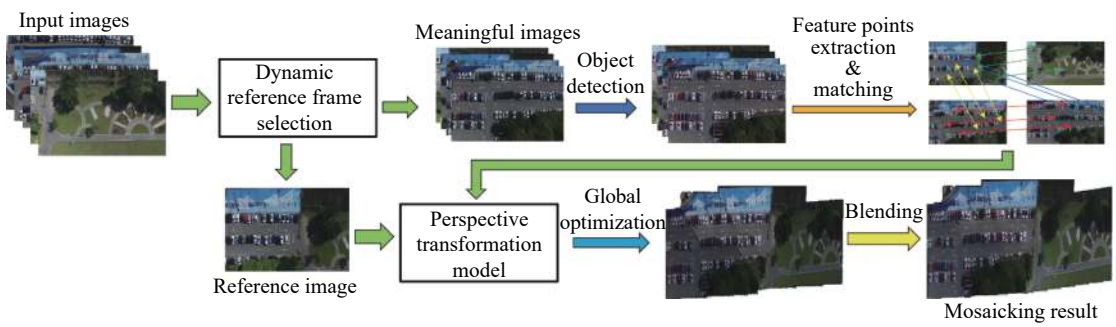


Fig. 1 Overall workflow of our proposed algorithm

3.1 Object detection in mapping

YOLOv3 [23] is one of the most popular real-time object detectors in computer vision. It adopts a much more powerful and efficient backbone network, i.e., Darknet-53. Inspired by the idea of feature pyramid network, it predicts bounding boxes at three feature maps of different scales. Benefiting from these detection headers, YOLOv3 is conductive to small objects detection and counting in aerial images. Supported by YOLOv3, our proposed method can simultaneously detect objects of specific classes and generate the orthophoto map with detection information.

In most cases, there are some meaningful areas in aerial images, such as buildings in city and vehicles in parking lot. We call them "regions of interest" (ROIs), which have a considerable influence on mosaicking results. To focus on ROIs, we give the following definition "complexity" of aerial images: an aerial image is complex if it covers ROIs and its complexity is determined by the number of detected objects in ROIs. The more objects of interest are detected, the more complex this image should be.

It is a favorite choice to briefly describe the complexity of aerial image scenes with the help of detection information. A human experience-determined threshold value σ is set to distinguish whether the scene is complicated or not. We adjust the mosaicking process dynamically according to the complexity.

3.2 Mosaicking workflow

3.2.1 Feature points extraction and matching

Speeded-up robust features (SURF) is one of the most popular feature points extraction and description algorithms for aerial image mosaicking. Based on the core idea of the SIFT algorithm, there is a great improvement on feature extraction speed. Due to the application of the Hessian matrix, the feature detectors can be more stable and repeatable. However, the time cost and the quantity of extracted feature points are thus heavily influenced by the Hessian matrix.

Most existing aerial image mosaicking algorithms still use the fixed Hessian threshold for all images to extract features. To further improve the feature quality of key frames, we evaluate the complexity of image scenes based on detection results and then adjust the feature points extraction process dynamically. The Hessian threshold adjustment rule is shown as

$$\min \text{Hessian} = \begin{cases} 800 - \dfrac{s}{\sigma} \times 600, & s < \sigma \\ 200, & s \geqslant \sigma \end{cases} \tag{1}$$

where s means the detection information of ROIs and σ represents a constant complexity threshold. In this paper, σ is set as 15 for all test aerial image sequences.

According to experience, the trade-off between speed and accuracy can be achieved by limiting the Hessian threshold to 200−800. Owing to the defined complexity, we can increase the quantity of feature points for more complicated scenes by reducing the threshold. While for those simpler images, we do quick feature extraction with a higher threshold.

For feature matching, we prefer the robust GMS algorithm, which filters mismatches and picks correct matches based on the results of the BF algorithm. It is considered in GMS that one truly matched point should have more true matching points in its neighborhood region. And the GMS algorithm assumes that the score of the ith region satisfies the Bernoulli distribution

$$S_i \sim \begin{cases} \mathrm{B}(Kn, p_t), & \text{if } x_i \text{ is true} \\ \mathrm{B}(Kn, p_f), & \text{if } x_i \text{ is false} \end{cases} \tag{2}$$

where x_i represents the feature point, K is the number of disjoint regions, n is the number of features in the ith region, and p_t and p_f represent true and false probability respectively. The mean and standard deviations of (2) are

$$\begin{cases} m_t = Knp_t, \ s_t = \sqrt{Knp_t(1-p_t)} \\ m_f = Knp_f, \ s_f = \sqrt{Knp_f(1-p_f)} \end{cases}. \tag{3}$$

The main idea of GMS is to maximize the following function:

$$P = \frac{m_t - m_f}{s_t + s_f} = \frac{Knp_t - Knp_f}{\sqrt{Knp_t(1-p_t)} + \sqrt{Knp_f(1-p_f)}}. \tag{4}$$

In practice, image pairs are divided into several grids and the neighborhood score is calculated based on (5). As seen in (6), true and false matches are potentially separable with score-based thresholding.

$$S_{ij} = \sum_{k=1}^{K=9} \left| \mathcal{X}_{i^k j^k} \right|, \tag{5}$$

$$\{i, j\} \in \begin{cases} \mathcal{T}, & S_{ij} > \tau_i = \alpha \sqrt{n_i} \\ \mathcal{F}, & \text{otherwise} \end{cases}, \tag{6}$$

where $\left| \mathcal{X}_{i^k j^k} \right|$ is the number of matches between cells $\{i^k, j^k\}$, n_i is the average number of features present in a single grid-cell, and α is a constant threshold.

The feature extraction and matching results are shown in Fig. 2. It can be obviously seen that there seems to be few mismatching points between two continuous frames. Coupled with our proposed dynamic Hessian threshold adjustment strategy, we can focus more on the informative regions, like the buildings in Fig. 2.

Fig. 2 Results of our feature extraction and matching method

3.2.2 Transformation model and global optimization

After feature extraction and matching, the next step is to transform the aerial image sequence into the same coordinate system by the projection transformation model

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \boldsymbol{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{7}$$

where (x', y') and (x, y) represent the point coordinate in the image after and before the transformation respectively. And $\boldsymbol{H}$ is the homography matrix that has eight independent variables, which can be defined as

$$\boldsymbol{H} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix}. \tag{8}$$

Furthermore, we apply RANSAC to eliminate possible mismatch points and estimate the homography transformation matrix. To reduce the cumulative error of long-sequences, the transformation parameters are optimized as follows.

Assume that there are M images in the aerial image sequence $\boldsymbol{I}$, and $\boldsymbol{N}$ pairs of features are extracted. And the transformation parameters of the ith image are expressed as $\boldsymbol{X}_i = (a_{11}, a_{12}, \cdots, a_{32})^{\mathrm{T}}$. Following the strategy in [24], we construct a typical nonlinear least squares problem as follows:

$$E(\boldsymbol{X}) = \mathrm{Ecor}(\boldsymbol{X}) + \omega \mathrm{Ereg}(\boldsymbol{X}), \tag{9}$$

$$\boldsymbol{X} = \left[\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}}, \cdots, \boldsymbol{X}_M^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{10}$$

where $\mathrm{Ecor}(\boldsymbol{X})$ means the correlation of features, which contributes to aligning the images geometrically and minimizing the sum of squared distances between features. Much more details are given by (11), (12), and (13). And $\mathrm{Ereg}(\boldsymbol{X})$ represents the regularization part to prevent distortion of the mosaicking result. It is introduced minutely by (14) and (15). ω is a constant weight value to balance the accumulation error and mosaicking distortion.

$$\mathrm{Ecor}(\boldsymbol{X}) = \sum_{i=1}^{N} \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{e}_i + \sum_{i=1}^{\tilde{N}} \tilde{\boldsymbol{e}}_i^{\mathrm{T}} \tilde{\boldsymbol{e}}_i, \tag{11}$$

$$\boldsymbol{e}_i = T_m(\boldsymbol{p}_{i,m}) - T_n(\boldsymbol{p}_{i,n}), 1 \leqslant m, n \leqslant M, \tag{12}$$

$$\tilde{\boldsymbol{e}}_i = T_{n\_\mathrm{ref}}(\boldsymbol{p}_{i,n\_\mathrm{ref}}) - \boldsymbol{p}_{i,n\_\mathrm{ref}}, \tag{13}$$

where $\tilde{N}$ is the counts of feature correspondences in the reference frame, T_m denotes the transformations of $\boldsymbol{X}_m$, and $(\boldsymbol{p}_{i,m}, \boldsymbol{p}_{i,n})$ represents the ith feature pairs.

$$\mathrm{Ereg}(\boldsymbol{X}) = \sum_{i=1}^{N} p_i E_{\mathrm{reg}}(\boldsymbol{X}_i), \tag{14}$$

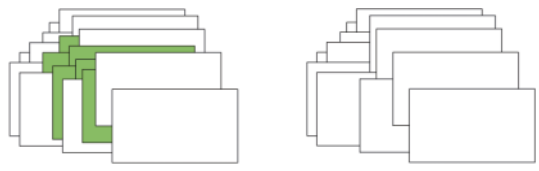
$$E_{\mathrm{reg}}(\boldsymbol{X}_i) = \left(a_{11}^i a_{12}^i + a_{21}^i a_{22}^i\right)^2 + \left(a_{11}^{i\ 2} + a_{21}^{i\ 2} - 1\right)^2 + \left(a_{12}^{i\ 2} + a_{22}^{i\ 2} - 1\right)^2 + \left(a_{31}^{i\ 2} + a_{32}^{i\ 2}\right)^2, \tag{15}$$

where p_i is the weight coefficient determined by the number of feature matching results. For long-sequence images, the value of $\mathrm{Ecor}(\boldsymbol{X})$ will increase, and to make $\mathrm{Ecor}(\boldsymbol{X})$ and $\mathrm{Ereg}(\boldsymbol{X})$ equivalent, the value of p_i should be set larger.

To solve the constructed nonlinear least squares problem efficiently, we apply the sparse Levenberg-Marquardt (LM) algorithm to optimize the transformation parameters.

3.2.3 Self-adaptive reference frame selection based on overlap

Since the aerial image sequences are consecutive and have a high degree of overlap, it is unwise to deal with all images. As shown in Fig. 3, excessively high overlaps have no benefit for improving the quality of mosaicking results. On the contrary, it increases calculation time cost and accumulated errors.



(a) Excessive (b) Appropriate

Fig. 3 Excessive overlap v.s. appropriate overlap

From Fig. 3, we can find that Fig. 3(a) describes the same size area with Fig. 3(b), while Fig. 3(b) uses fewer frames. Fig. 3(a) introduces redundant calculations because of excessive overlap. Our goal is to remove unne-

cessary images and reduce the computational complexity without damaging mosaicking results. Based on coordinate transformation results, we calculate the non-overlapped area between two images by (16). Then unnecessary images with excessive overlap are removed to improve the operation efficiency.

$$d_{\alpha\beta} = s_\alpha - (s_\alpha \cap s_\beta) \tag{16}$$

where $d_{\alpha\beta}$ represents the left non-overlapped area of α between frame α and β. s_α and s_β describe the areas of frames α and β, respectively.

The reference frame selection method has a significant effect on the visual quality of mosaicking results. Most current aerial image mosaicking algorithms simply select the first frame or the intermediate frame [25] as the reference image. This simple selection method results in large transformation error and poor robustness [26]. As shown in Fig. 4, if the first or intermediate image is selected, the edge image will be damaged, causing large transformation distortion and poor visual quality. Inspired by [27], we propose a self-adaptive reference frame selection algorithm according to the overlap and greatly improve the visual quality of mosaicking results.

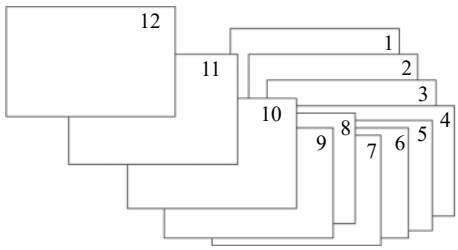


Fig. 4 Adaptively selecting Frame 10 as the reference

According to the calculated overlap, an intuitive idea is that the closer one image is to the orthophoto map's center area, the more images should overlap with it. Besides, selecting the most overlapped image can reduce the accumulated error when transformed to the reference coordinate system. The selection method is depicted as follows:

$$\delta_{ij} = \begin{cases} 1, & \text{if } (i,j) \text{ is matched pair} \\ 0, & \text{else} \end{cases}, \tag{17}$$

$$i_{\text{ref}} = \arg\max_{i \in I} \sum_{j=1}^{M} \delta_{ij}, \quad i = 1,2,\cdots,M, \tag{18}$$

where δ_{ij} is the indicator function that is used to discriminate between matched pairs and unmatched ones. Coupled with (17) and (18), we choose the i_{ref}th frame as the reference frame, which has the most intersection with the other frames. Fig. 5 illustrates the significant effect of different reference frame selection algorithms on the final mosaicking result. There is no doubt that our proposed method achieves better visual quality and performance.

The detail of our proposed algorithm is illustrated in Algorithm 1.

Algorithm1 The proposed mosaicking method

Input: A sequence of aerial images $I = \{I_1, I_2, I_3, \cdots, I_M\}$
Output: Stitched image with detection results
for $i = 1$ to M **do**
  Compute complexity score s_i based on detection results
  if $s_i < \sigma$ **then**
    increase Hessian threshold dynamically
  else
    decrease Hessian threshold dynamically
  end if
  Compute matches between frame I_i and $I_j (j \neq i, j = 1,2,\cdots,M)$
  Compute non-overlap area d_i of frame I_i
  if $d_i \leqslant 0$ **then**
    Delete frame I_i
  end if
end for
Select reference frame adaptively based on the overlap
for $i = 1$ to M **do**
  Compute homography metric $E(x)$
end for
Global optimization by sparse LM algorithm
Multiband blending

(a) Dynamic selection (b) First as reference (c) Half as reference

Fig. 5 Mosaicking results of different reference frame selections

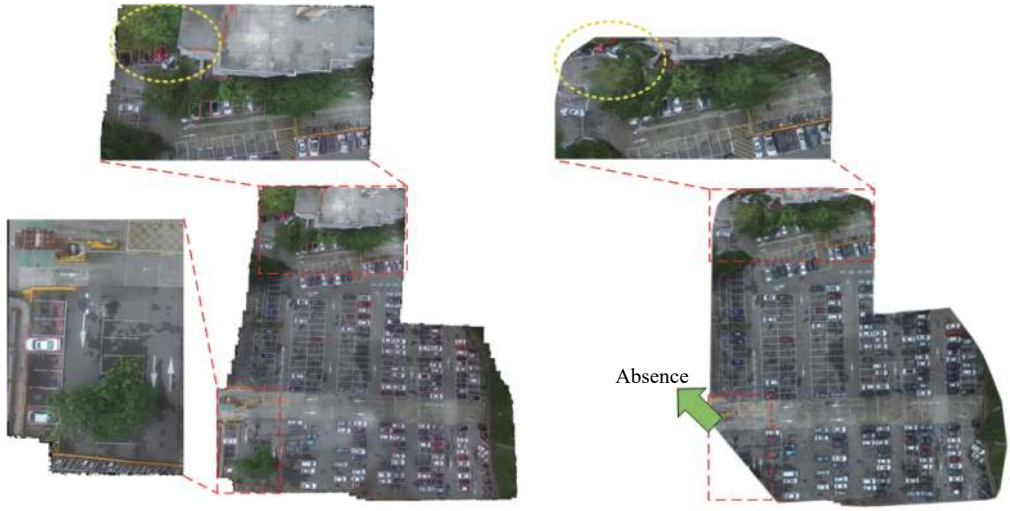
4. Experiments

4.1 Object detection training details

To detect vehicles in aerial images, off-the-shelf YOLOv3 detector is trained on the Car Parking Lot (CARPK) dataset [28]. And we divide the CARPK dataset into the training set and the test set. The training set is used for vehicle detection training, while the test set is utilized to generate global orthophoto map. We choose stochastic gradient descent (SGD) as the optimizer and train the network for 20k iterations with a batch size of 64. The learning rate is set as 1e−3 and decays by a factor of 0.000 5 at iteration 13k and 17k respectively. The value of the momentum is set to 0.9. As for evaluation on the test set, we got 90.83 mean average precision (mAP) as the final detection results.

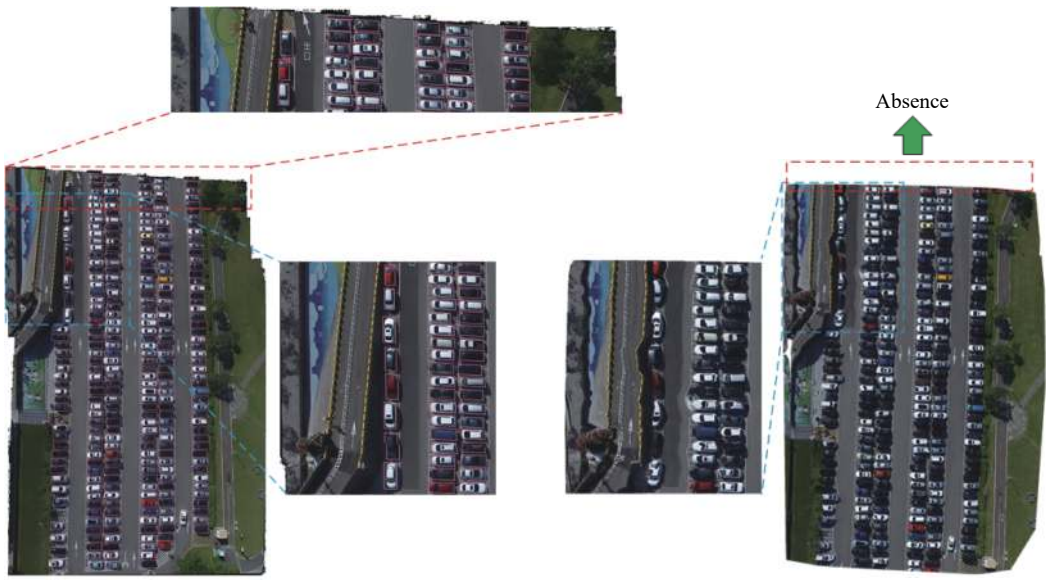
4.2 Mosaicking results and analysis

Our algorithm is developed on Ubuntu18.04 with Intel Core i7-9750H 2.60 GHz CPU, NVIDIA GTX1650 GPU. OpenCV4.2.0 and CUDA10.0 libraries are essential. To verify the effectiveness and robustness of our proposed mosaicking algorithm, we perform several controlled experiments with professional mapping software Pix4Dmapper (high-accuracy mode is selected). The mosaicking results of our proposed method for the CARPK dataset are illustrated in Fig. 6 to Fig. 9.



(a) Mosaicking result of the proposed algorithm (b) Mosaicking result via Pix4Dmapper

Fig. 6 Mosaicking results of Dataset 1



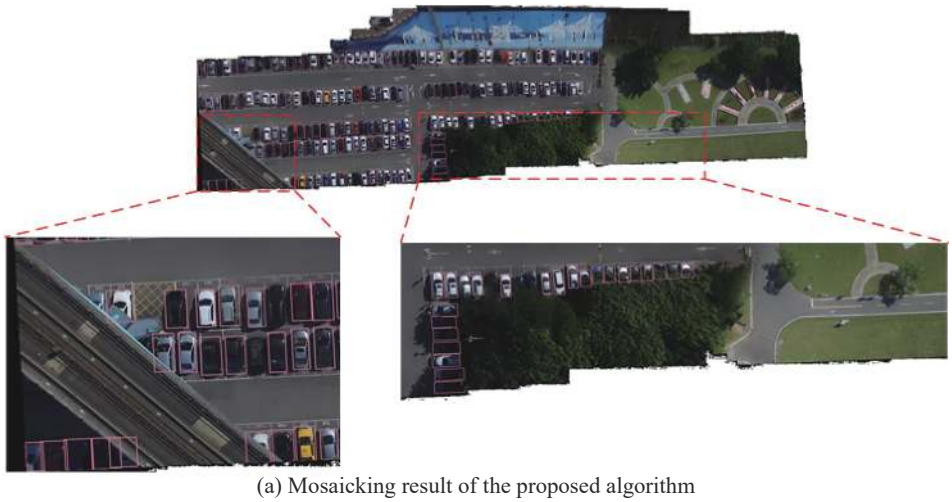
(a) Mosaicking result of the proposed algorithm (b) Mosaicking result via Pix4Dmapper

Fig. 7 Mosaicking results of Dataset 2

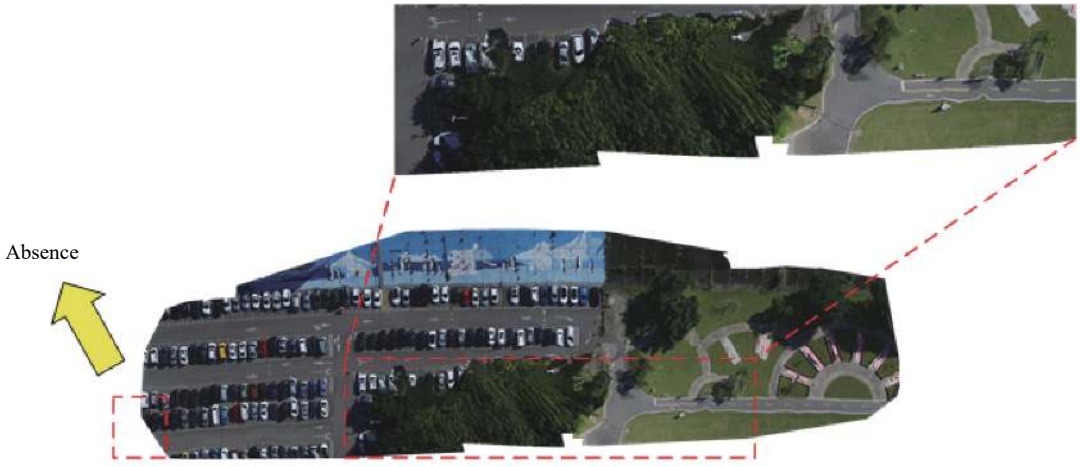
(a) Mosaicking result of the proposed algorithm

(b) Mosaicking result via Pix4Dmapper

Fig. 8 Mosaicking results of Dataset 3



(a) Mosaicking result of the proposed algorithm



(b) Mosaicking result via Pix4Dmapper

Fig. 9 Mosaicking results of Dataset 4

Based on the mosaicking results shown in Fig. 6 to Fig. 9, we can conclude that our proposed algorithm achieves better visual quality and much more satisfying performance than the professional software Pix4Dmapper. The mosaicking results via the proposed method have almost no distortion and can keep the image edge details. Moreover, our algorithm makes it accessible to collect the car detection information of the large-scale area. Besides, we evaluate the robustness of the proposed mosaicking algorithm on our own data, as shown in Fig. 10.

Fig. 10 Mosaicking result of data in the paper

4.3 Speed analysis

We conduct controlled experiments on Dataset 1 to Dataset 4. Table 1 shows the time cost compared between our proposed method and the Pix4Dmapper. According to the time cost results, we can easily observe that the designed method takes less time to obtain the more satisfying mosaicking results than Pix4Dmapper. Besides, our proposed algorithm can acquire the detection results and can be easily extended to other fields, such as car counts and fire monitoring.

Benefiting from the lightweight real-time object detection network, our proposed image mosaicking method can be easily deployed to embedded platforms, like NVIDIA Jetson TX2 and Xavier. Thus, it is convenient to build an intelligent UAV system, equipped with real-time object detection and fast image mosaicking abilities.

Table 1 Time cost comparison

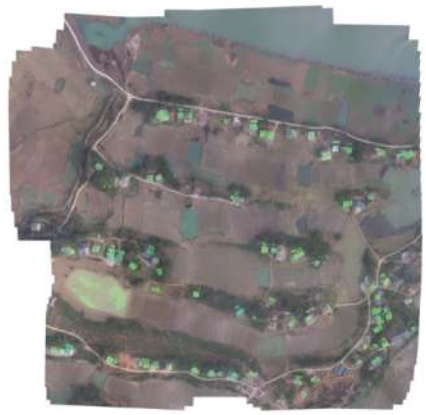
Image data	Sequence length	Mosaicking time/s	
		Pix4Dmapper	Ours
Dataset 1	71	169	14
Dataset 2	37	92	8
Dataset 3	40	113	7
Dataset 4	24	58	4

4.4 Application analysis

To further demonstrate the actual application value of our proposed method, we extend our algorithm to more application scenarios. Purely relying on visual information, our proposed method can be utilized in several computer vision fields without the constraint of location priors. With a simple modification, our proposed method can be extended to panoramic image stitching and building segmentation, as shown in Fig. 11.

(a) Panoramic image stitching

(b) Orthophoto map mosaicking



(c) Aerial image mosaicking & building segmentation

Fig. 11 Extended applications of our proposed method

5. Conclusions

We propose a vision-based algorithm that combines object detection task for aerial image mosaicking. By defining the complexity of aerial images according to object detection results, we dynamically optimize the feature extraction process. For the sake of distortion reduction, we introduce a self-adaptive reference frame selection algorithm based on overlaps. The controlled experiments have demonstrated the feasibility, effectiveness and robustness of our proposed algorithm. Our proposed method can generate the global orthophoto map and collect the detection information simultaneously.

What we need to be aware of is that there still lacks enough quantitative analysis of mosaicking results. Therefore, future research will concentrate on seeking for several mathematical quantitative metrics to evaluate the quality of the final mosaicking results.

References

[1] CUI J G, LIU M, ZHANG Z T, et al. Robust UAV thermal infrared remote sensing images stitching via overlap-prior-based global similarity prior model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 270–282.

[2] LIU C, ZHANG S H, AKBAR A. Ground feature oriented path planning for unmanned aerial vehicle mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(4): 1175–1187.

[3] BU S H, ZHAO Y, WAN G, et al. Map2DFusion: real-time incremental UAV image mosaicing based on monocular SLAM. Proc. of the IEEE International Conference on Intelligent Robots and Systems, 2016: 4564–4571.

[4] ZHANG J M, ZHU H Q, WANG P Y, et al. ATT squeeze U-Net: a lightweight network for forest fire detection and recognition. IEEE Access, 2021, 9: 10858–10870.

[5] CAO Y C, YANG F, TANG Q F, et al. An attention enhanced bidirectional LSTM for early forest fire smoke recognition. IEEE Access, 2019, 7: 154732–154742.

[6] AVOLA D, CINQUE L, FORESTI G L, et al. A UAV video dataset for mosaicking and change detection from low-altitude flights. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2020, 50(6): 2139–2149.

[7] SZELISKI R. Image alignment and stitching: a tutorial. Boston: Now Foundations and Trends, 2006.

[8] ZITOVA B, FLUSSER J. Image registration methods: a survey. Image and Vision Computing, 2003, 21(11): 977–1000.

[9] LOWE D G. Distinctive image features from scale-invariant

keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110.
[10] BAY H, ESS A, TUYTELAARS T, et al. Speeded-up robust features (SURF). Computer Vision and Image Understanding, 2008, 110(3): 346–359.
[11] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF. Proc. of the IEEE International Conference on Computer Vision, 2011: 2564–2571.
[12] EDWARD R, REID B P, TOM D. Faster and better: a machine learning approach to corner detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(1): 105–119.
[13] BIAN J W, LIN W Y, LIU Y, et al. GMS: grid-based motion statistics for fast, ultra-robust feature correspondence. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2828–2837.
[14] AN M, JIANG Z G, ZHAO D P. High speed robust image registration and localization using optimized algorithm and its performances evaluation. Journal of Systems Engineering and Electronics, 2010, 21(3): 520–526.
[15] AMIRI A J, MORADI H. Real-time video stabilization and mosaicking for monitoring and surveillance. Proc. of the 4th International Conference on Robotics and Mechatronics, 2016: 613–618.
[16] LI C, GUO B L, GUO X X, et al. Real-time UAV imagery stitching based on grid-based motion statistics. Journal of Physics: Conference Series, 2018(1069): 012163.
[17] MCLAUCHLAN P F, JAENICKE A. Image mosaicing using sequential bundle adjustment. Image and Vision Computing, 2002, 20(9/10): 751–759.
[18] XU Q, CHEN J, LUO L B, et al. UAV image mosaicking based on multiregion guided local projection deformation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3844–3855.
[19] WU F L, GUAN H N, YAN D J, et al. Precise geometric correction and robust mosaicking for airborne lightweight optical butting infrared imaging system. IEEE Access, 2019, 7: 93569–93579.
[20] LI M, LI D R, GUO B X, et al. Automatic seam-line detection in UAV remote sensing image mosaicking by use of graph cuts. International Journal of Geo-Information, 2018, 7(9): 361.
[21] YUAN Y T, FANG F M, ZHANG G X. Superpixel-based seamless image stitching for UAV images. IEEE Trans. on Geoscience and Remote Sensing, 2021, 59(2): 1565–1576.
[22] GE Y, WEN G J, YANG X L. A fast mosaicking method for small UAV image sequence using a small number of ground control points. Proc. of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2016: 90–94.
[23] REDMON J, FARHADI A. YOLOv3: an incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.
[24] XU Y H, OU J L, HE H, et al. Mosaicking of unmanned aerial vehicle imagery in the absence of camera poses. Remote Sensing, 2016, 8(3): 204.
[25] HU J W, ZHOU Y H, ZHAO C H, et al. An application of panoramic mosaic in UAV aerial image. Proc. of the IEEE 13th International Conference on Control and Automation, 2017: 1049–1053.
[26] QU Z, LI J, BAO K H. An unordered image stitching method based on binary tree and estimated overlapping area. IEEE Trans. on Image Processing, 2020, 29: 6734–6744.
[27] JI Y F, LI W X, FENG K, et al. Automatic video mosaicking algorithm via dynamic key-frame. Journal of Systems Engineering and Electronics, 2020, 31(2): 272–278.
[28] HSIEH M R, LIN Y L, HSU W H. Drone-based object counting by spatially regularized regional proposal network. Proc. of the IEEE International Conference on Computer Vision, 2017: 4165–4173.

Biographies

HAN Jun was born in 1997. He received his B.S. degree from Beijing Institute of Technology (BIT) in 2019. He is a postgraduate student majoring in control science and engineering in BIT. His research interests include computer vision and multi-object tracking.
E-mail: 3120190878@bit.edu.cn

LI Weixing was born in 1976. He is a senior experimentalist at the School of Automation in Beijing Institute of Technology (BIT). He received his M.S. degree in pattern recognition and intelligent control from BIT. His research interests include pattern recognition, video analysis, and machine learning.
E-mail: liweixing@bit.edu.cn

FENG Kai was born in 1996. He is a postgraduate student majoring in control science and engineering in Beijing Institute of Technology (BIT). He received his B.S. degree from BIT in 2018. His research interests include computer vision and object detection.
E-mail: fengkai_bit@outlook.com

PAN Feng was born in 1978. He is an associate professor at the School of Automation in Beijing Institute of Technology (BIT). He received his M.S. and Ph.D. degrees in pattern recognition and intelligent control from BIT. His research interests include intelligent computing and robotics and control.
E-mail: andropanfeng@126.com