Mathematical Biosciences and Engineering

MBE, 18(3): 2274–2287.
DOI: 10.3934/mbe.2021115
Received: 22 December 2020
Accepted: 03 March 2021
Published: 05 March 2021

http://www.aimspress.com/journal/MBE

Research article

A multi-scale UAV image matching method applied to large-scale landslide reconstruction

Chaofeng Ren[1,*], Xiaodong Zhi[2], Yuchi Pu[1] and Fuqiang Zhang[1]

[1] College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China
[2] eFly of China Electronic Science Technology Co., Ltd., Chongqing 401332, China

*** Correspondence:** Email: ren_cf@163.com; Tel: +8618729981156.

Abstract: Three-dimensional (3D) sparse reconstruction of landslide topography based on unmanned aerial vehicle (UAV) images has been widely used for landslide monitoring and geomorphological analysis. In order to solve the isolated island phenomenon caused by multi-scale image matching, which means that there is no connection between the images of different scales, we herein propose a method that selects UAV image pairs based on image retrieval. In this method, sparse reconstruction was obtained via the sequential structure-from-motion (SfM) pipeline. First, principal component analysis (PCA) was used to reduce high-dimensional features to low-dimensional features to improve the efficiency of retrieval vocabulary construction. Second, by calculating the query depth threshold and discarding the invalid image pairs, we improved the efficiency of image matching. Third, the connected network of the dataset was constructed based on the initial matching of image pairs. The lost multi-scale image pairs were identified and matched through the image query between the connection components, which further improved the integrity of image matching. Our experimental results show that, compared with the traditional image retrieval method, the efficiency of the proposed method is improved by 25.9%.

Keywords: unmanned aerial vehicle (UAV); image retrieval; image matching, image pairs; 3D reconstruction; Structure-from-Motion (SfM) reconstruction

1. Introduction

Unmanned aerial vehicles (UAV) are a convenient and cost-efficient platform for mapping

inaccessible or dangerous areas [1], and they have been widely used in landslide mapping [2], heritage site reconstruction [3], emergency response measures[4] , and three-dimensional (3D) urban modelling [5]. The payload of a typical UAV mapping system is a camera, which can take images according to a pre-planned route. Based on the structure-from-motion (SfM) pipeline method [6,7], the poses of the acquired two-dimensional (2D) sequence images can be accurately recovered. The reconstruction of large-scale landslide surfaces based on UAV is important for landslide monitoring and geomorphological analysis [8]. In order to yield more accurate and detailed reconstruction results in key areas, it is necessary to control the UAV to fly lower and obtain multi-view images. However, the completeness and efficiency of 3D sparse reconstruction is a great challenge for multi-scale UAV images.

The 3D sparse reconstruction-based sequence images can be divided into two tasks: correspondence search and reconstruction. The first task is to select image pairs with overlap regions and then obtain the corresponding points by image matching. Although the exhaustive method (EM) can find all potential image pairs with overlap regions [9,10], it is time-consuming and proportional to the square number of images. Further, it is difficult to meet the application needs of a large number of image data. The second method calculates the approximate footprint of the image via measurements found using a position and orientation system (POS) equipped with UAV [11,12]. If two image footprints intersect, image matching is performed for the image pair. However, the method is difficult to apply when the optical axis of the camera swings around the vertical direction. The third method selects image pairs based on image retrieval [13], which is when relevant images from the database for a query image are found. The image pair selection method based on image retrieval constructs the bag-of-words (BoW) model [14], which represents the image as a set of weighted visual words by quantizing the local feature space. All visual words are composed of a hierarchical tree that generates a vocabulary tree [15]. Then, one can determine whether the two images share overlap regions by evaluating the similarity of the image visual words vector. It is common to apply a weighting to visual words, such as the frequency-inverse document frequency (TF-IDF), rather than applying frequency directly. For large-scale image retrieval, the Hamming embedding (HE) [16] technique represents local descriptors as binary codes, which improves efficiency and saves memory. These methods are directly applied to hybrid and multi-scale UAV images that result in insufficient image pairs and in sufficient image pairs, all of which may lead to disconnected structures or incomplete components in the SfM pipeline [17]. The second task is to reconstruct a scene model that computes both the 3D sparse points and camera poses. Methods for this task can be divided into two classes: sequential and global [18]. Sequential SfM reconstruction starts from a minimal reconstruction based on two or three views, then incrementally adds new views into a merged representation (e.g., Bundler [19], OpenMVG [20], etc.). The sequential SfM method needs to perform multiple bundle adjustments (BA) and is a rather slow procedure when too many images (more than 10,000) are involved in the reconstruction. Global SfM reconstruction generally adopts the method of divide and conquer to “bundle” data from a large area [21]. Large-scale SfM reconstruction data is divided into small pieces and processed separately. Then, the results of small pieces are fused into a whole result. Although the global reconstruction method is more efficient, the sequential SfM reconstruction method algorithm is simpler, and the results are stable and reliable [22].

In this paper we propose an image pair selection method for multi-scale UAV images based on Hamming Embedding (HE) and TF-IDF weighting. This method greatly reduces the invalid calculation in the image matching and achieves complete sparse reconstruction results. Based on the

sequential SfM reconstruction method, the 3D sparse points of landslide topography are completely reconstructed. The remainder of this paper is structured as follows. In section 2, we present the main method, and describe in detail the two key components of the method: threshold segmentation and missing image pair search. Experiments and results are shown in section 3. We summarize the conclusions and contributions of the paper in section 4.

2. Methodology

Figure 1 shows the image pair selection method used for multi-scale UAV image matching. The method contains six steps: feature extraction, vocabulary tree construction, visual words indexing, image similarity query, two view matching, and missing image pair searching.

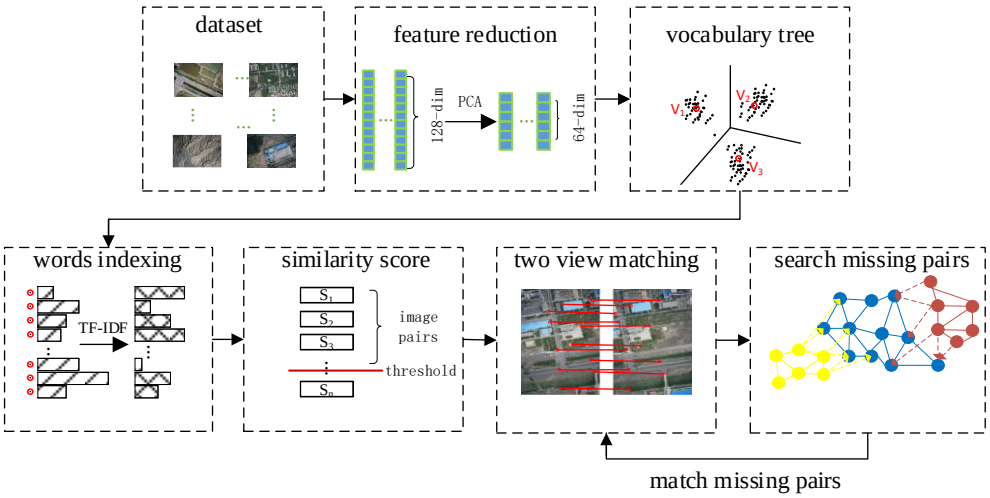


Figure 1. Overall workflow for multi-scale UAV image matching.

2.1. Feature extraction

The main idea of selecting image pairs based on image retrieval is to represent the image with visual feature information and then determine whether the image overlaps according to the similarity of the feature information. In most image retrieval works, the features vectors extracted by SIFT [23] or SURF [24] are used. These local features are widely used because they are invariant to lighting, scale, and rotation changes, and they can overcome affine image changes. ASIFT [25] is based on SIFT and is more suitable for affine transformation when its efficiency is too low to be practical. Although SURF is more efficient, SIFT proves to be the most robust to changes in perspective [26].

Since the original SIFT feature contains 128-dimensional feature vectors, a large number of high-dimensional feature operations can also reduce the efficiency of image retrieval [27]. Therefore, this paper uses the principal component analysis (PCA) [28] method to reduce the high-dimensional features to low-dimensional features and to improve efficiency while maintaining most of the feature information. Since the PCA-SIFT [29] method can also achieve a similar effect, SIFT features obtained better results than the PCA-SIFT with regard to image matching [30]. Thus, the features after dimension reduction are only used in the image pair selection stage.

Suppose that a total of n features are obtained and represented by a matrix $\mathbf{F}$. The singular

value decomposition (SVD) of the matrix $\mathbf{F}$ is performed according to the PCA algorithm and the result is shown in Eq (1).

$$\mathbf{F}_{n\times 128} = \mathbf{U}_{n\times n} \sum_{n\times 128} \mathbf{V}^{\mathrm{T}}_{128\times 128} \tag{1}$$

n is the number of SIFT features; $\mathbf{U}_{n\times n}$ and $\mathbf{V}_{128\times 128}$ are the n and 128-order orthonormal matrices, respectively; $\{\sigma_1 \mathrm{L}\ \sigma_r\}$ represents the r eigenvalues of SVD decomposition of matrix $\mathbf{F}_{n\times 128}$; and $\sum_{n\times 128}$ is a diagonal matrix formed by r eigenvalues arranged in descending order.

If the SIFT feature is reduced to the d-dimensional feature, the first d row vectors of the matrix $\mathbf{V}^{\mathrm{T}}_{128\times 128}$ are converted into a matrix $\mathbf{V}^{\mathrm{T}}_{d\times 128}$, and the reduced-dimensional SIFT feature $\mathbf{F}'_{n\times d}$ is as follows:

$$\mathbf{F}'_{n\times d} = \mathbf{F}_{n\times 128} \mathbf{V}^{\mathrm{T}}_{d\times 128} \tag{2}$$

2.2. Words indexing

The vocabulary tree is composed of visual words generally obtained by clustering SIFT features with *k*-means [31]. In order to evaluate the importance of visual words in different images, the weights are calculated by TF-IDF method, as shown in Eq (3).

$$t_{id} = \frac{n_{id}}{n_d} \ln \frac{\mathrm{N}}{n_i} \tag{3}$$

n_{id} is the number of the occurrences of visual word (i) in the image (d); n_d is the number of words in the image (d); n_i is the number of occurrences of visual word i in the image; N is the number of images; and t_{id} is the weight of the visual word (i) in the image (d).

After the vocabulary tree is created, the occurrences number of all visual words can be indexed in the entire image dataset. After, Eq (3) calculates the weights of each visual word in different images. Supposing that there are k visual words in the vocabulary tree, the *i*th image can be represented as a vector of k weights, as shown in Eq (4):

$$\mathbf{V_i} = \{t_{1i}, \mathrm{L}\ t_{ki}\} \tag{4}$$

where $\mathbf{V_i}$ are the k weights of image i.

2.3. Image query

The purpose of image query is to find the most similar image sequence in the dataset. If there are overlap regions between the two images, the extracted SIFT features should be similar, and the visual word weight vectors of the two images should also be similar. The similarity between them can be obtained by calculating the angle between the weight vectors, as shown in Eq (5):

$$S_{(i,j)} = \cos\theta = \frac{\mathbf{V_i} \cdot \mathbf{V}_j}{|\mathbf{V}_i| \cdot |\mathbf{V_j}|} \tag{5}$$

where $S_{(i,j)}$ is the similarity between the *i*th image and the *j*th image, θ is the angle between

the two weight vectors.

The query depth means that the top Q images with the highest similarity in the dataset for the query image, as shown in Figure 2.

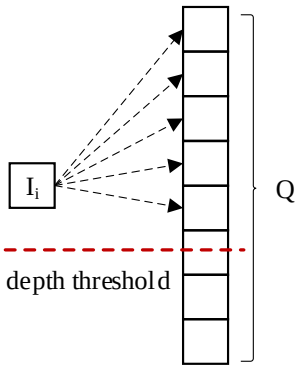


Figure 2. The diagram of query depth threshold.

In Figure 2, the conventional method forms image pairs by the query image (I_i) and all images in query depth Q. However, if the value of Q is too small, the image pairs with overlap regions will be lost. In contrast, a large number of image pairs without overlap regions can be added, which greatly reduces the efficiency of image matching. Therefore, we proposed a method of query depth threshold based on that proposed by Otsu [32], using images above the threshold and query images that form image pairs. The threshold calculation method is shown in Eq (6):

$$g(t) = \frac{N_1}{\mathrm{Q}}\left(1-\frac{N_1}{\mathrm{Q}}\right)\left(\sum_{j=0}^{t} jS_{(i,j)}^{n} - \sum_{j=t+1}^{\mathrm{Q}} jS_{(i,j)}^{n}\right)^2 \tag{6}$$

where $S_{(i,j)}^{n}$ is the normalization of $S_{(i,j)}$, which is defined as $S_{(i,j)}^{n} = \frac{S_{(i,j)}}{\sum_{j=0}^{\mathrm{Q}} S_{(i,j)}}$, and N_1 is the number of images above the threshold, $g(t)$ is the corresponding variance value at the threshold t.

Equation (6) divides the image sequence within the query depth Q into two parts: above the threshold t and below the threshold t based on the maximum variance between two parts. Within the query depth range $t \in [0, \mathrm{Q})$, the variance values are calculated according to Eq (6), and the position corresponding to the largest variance is the selected threshold t^*, as shown in Eq (7).

$$t^* = \arg\max_{0 \le t < \mathrm{Q}} \{g(t)\} \tag{7}$$

2.4. Search missing pairs

After querying every image in the dataset, we performed two-view image matching on all candidate image pairs based on the methods of SIFTGPU [33]. We removed outliers using the fundamental matrix estimation within the RANSAC [34] algorithm. Unlike conventional

similar-scale UAV images, the retrieval results of multi-scale UAV images showed aggregation. The reason is that image retrieval is based on image similarity and the similarity of similar-scale images are much higher than cross-scale images. Under a fixed query depth, the query images preferentially select images of a similar scale and the obtained matching relationship may show an isolated island phenomenon, which means that there is no connection between the images of different scales. To this end, we proposed a fusion method of local connected components, as shown in Figure 3.

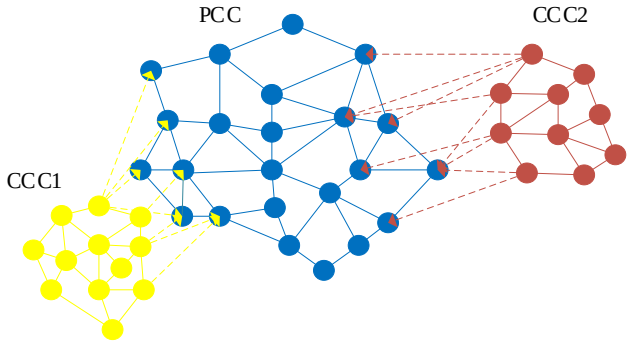


Figure 3. Adjacent connected components. PCC denotes the parent connected component, whereas CCC denotes the child connected component. The blue dot represents an image in the parent connected components; red dots and yellow dots represent the images of two different child connected components; the solid line indicates that there is an overlap region between the two images after the geometrical verification in child connected component; the dashed line indicates that there is an overlap region between two images of adjacent connected components after the geometrical verification.

We considered the image pairs after two-view geometrical verification with more than 16 correspondence points as one edge. The two images were two vertices of the same edge. An undirected graph was constructed based on the initial matching graph by the open source library lemon [35]. The local connected components were searched in the undirected graph and the connected component with the largest number of images were considered as the parent connected component (PCC), such as in Figure 3, while the other connected components were considered as child connected components (CCC), such as CCC1 and CCC2 in Figure 3. Next, we selected each image in the CCC and followed the method presented in section 2.3 to query only in the vocabulary tree of the PCC. Then, the PCC and CCC were connected by the image pairs between the PCC and CCC, as shown by the dashed line in Figure 3.

Using the same method as the image pairs obtained using image retrieval, the missing image pairs obtained by searching missing pairs were performed using two-view geometrical verification, and the complete matching relationship of the obtained dataset.

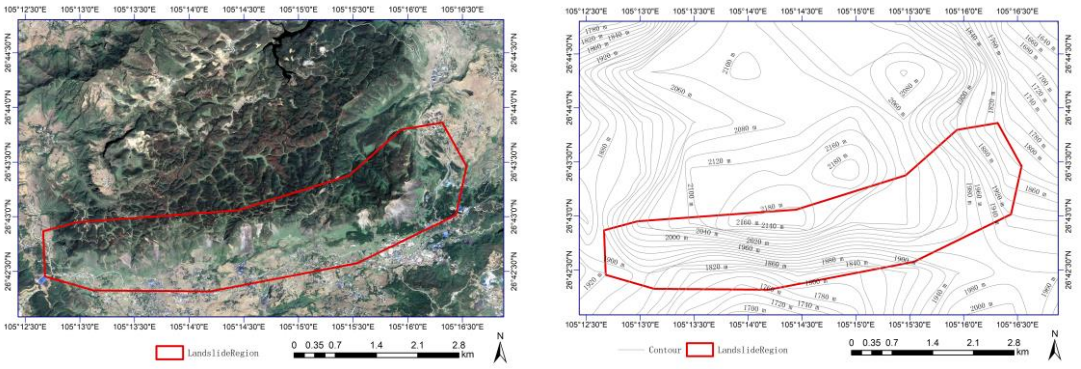
3. Experiments

In our experiments, we used a dataset to evaluate the proposed method. Image match pairs were firstly selected using the proposed method, whose performance would be compared with two state-of-the-art methods, i.e., the vocabulary tree (VT) [36] and adaptive vocabulary tree (AVT) [37] methods. Then, two-view geometrical verification were performed based on SIFTGPU and

RANSAC. Next, in order to access efficiency, completeness, and accuracy of the SfM pipeline, the sequential SfM pipeline was used. The proposed method was developed using Visual Studio 2015 C++. All experiments were conducted on the same machine with a 12-core Intel i7-8700K, 3.7GHz CPU, 64GB of RAM, 512GB of SSD, and a 12 GB NVIDIA Titan XP graphics card.

3.1. Study area

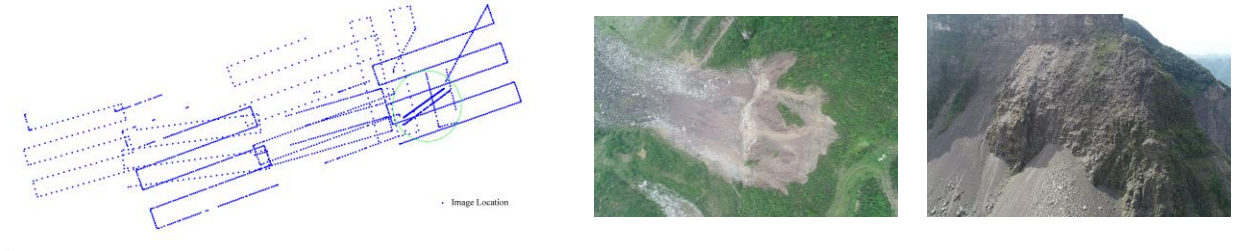
In order to demonstrate the application of the proposed method, we used UAV image data obtained from a large-scale landslide area for experimental and analysis. The study area in Figure 4 is the Zongling landslide, which is located in the coal mining district of Liupanshui, in the Guizhou province in southwest China. The total area of the study area is about 8.72 km^2 and the altitude difference within the area is 603.16 m.



(a) (b)

Figure 4. The location of the study area: (a) the location of the landslide; (b) the contour of the landslide.

To investigate and analyze the landslide body, we used the DJI phantom 4 Pro UAV to obtain nadir images of the entire landslide district and the oblique images of the key area by manually controlling the UAV. A total of 2137 high-resolution UAV images were obtained in the Zongling landslide. Their distributions are shown in Figure 5.



(a) (b) (c)

Figure 5. The distribution of obtained UAV images: (a) the nadir images distribution; (b) a nadir image; (c) an oblique image.

Figure 5a shows the distribution of UAV images in the dataset and where the oblique images were distributed in the green circle area. The nadir images completed most of the Zongling landslide, while the oblique images obtained details of the key area of the landslide.

3.2. Performance evaluation of feature reduction

To evaluate the performance of the feature reduction for image retrieval, we used PCA to convert the original 128-dimensional SIFT features into 112, 96, 80, 64, 48, 32, and 16 and generated the vocabulary tree using the VT method. The features of the vocabulary tree were derived from randomly selected images in the dataset (10%). The initial clustering center number k of the k-means was uniformly set to 42,600, which is about 200 times the number of vocabulary tree images. In order to evaluate the image query accuracy at different query depth, the query depth Q was set to 20, 40, 60, 80, and 100. At the beginning, the EM method was used for image matching and the results were used as the ground truth to evaluate the retrieval accuracy of other methods. Figure 6 shows the precision and recall curve for different feature dimensions and query depth using the VT method.

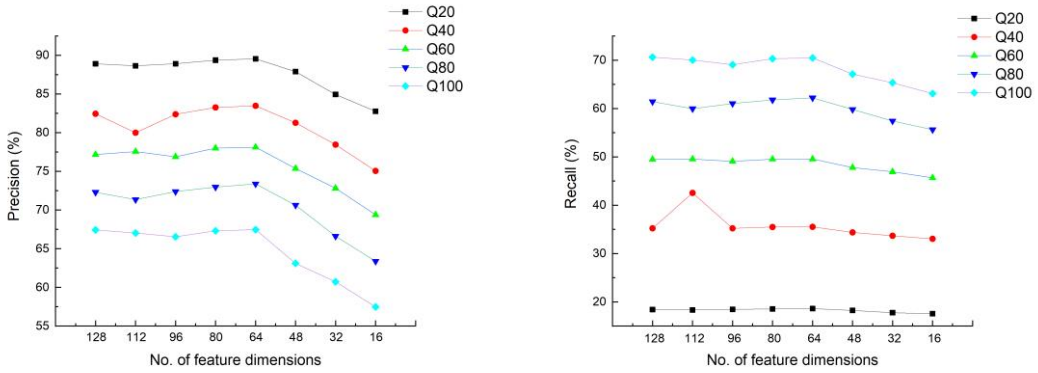


Figure 6. Comparison of precision and recall curve for the query depth: (a) indicates the precision curve of different query depth and feature dimensions; (b) indicates the recall curve of different query depth and feature dimensions; Q20 indicates the query depth as 20.

Figure 6a illustrates the precision curve of image retrieval versus the VT method's different feature dimensional and different query depth. It is shown that as precision decreased, query depth increased. Under the same query depth condition, its precision increased first and then decreased as the feature dimension decreased, reaching its peak when the feature dimension had a dimension of 64. Figure 6b illustrates the recall curve of image retrieval versus the different feature dimensional and different query depth by the VT method. In contrast to Figure 6a, the recall decreased when query depth increased. Under the same query depth condition, the recall was stable with a dimension of 64. The results in Figure 6a,b show that the precision and the recall were mutually restricted, and that the feature dimension was reduced to 64 dimensions to obtain better retrieval accuracy.

In order to express the overall performance of a method, we compare the recognition performance in terms of *F1-measure* [38], which represents the combination of precision and recall. The detailed results are shown in Figure 7.

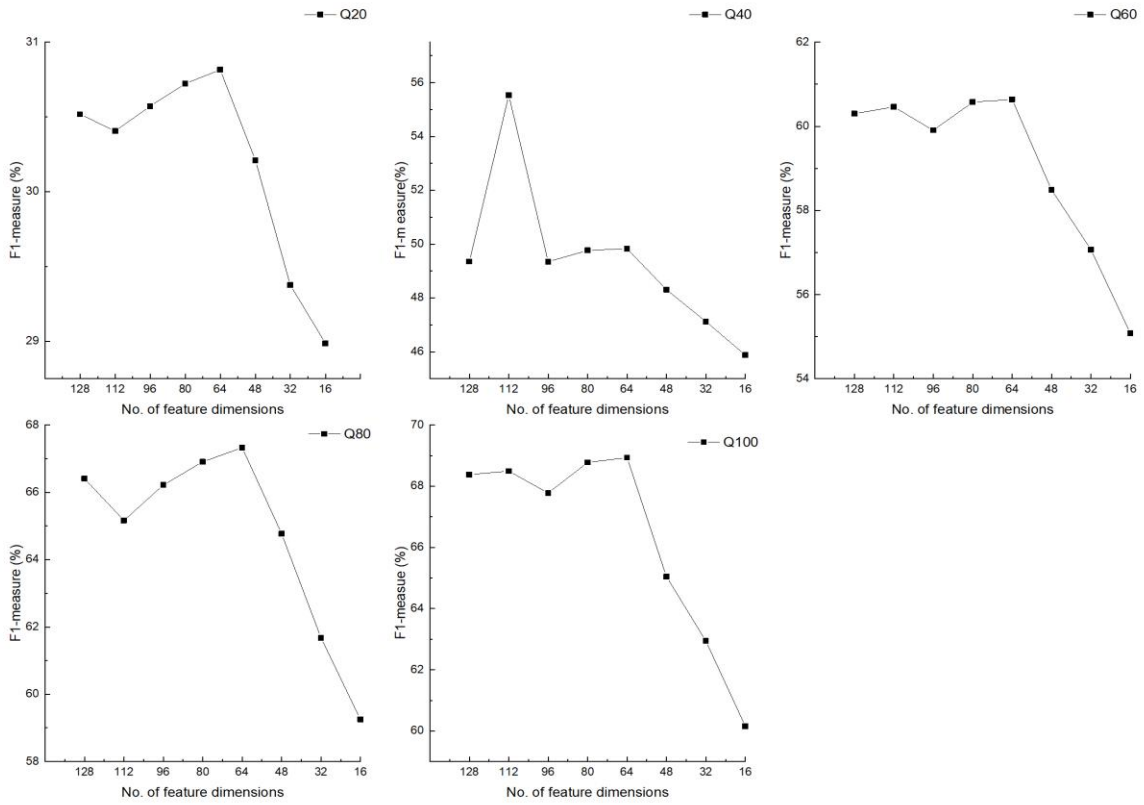


Figure 7. Comparison the terms of *F1-measure* with different query depth.

The higher the *F1-meansure* comprehensive term, the better the performance of the method. As shown in Figure 7, the curve of *F1-meansure* shows the similar tendency with five different query depth, and the maximum value is obtained when the feature dimension is reduced to 64. Only when the query depth is set to 40 and the feature dimension is reduced to 112, the *F1-measure* value is a special case. As the query depth increases, the overall performance of the method tends to be stable, and the optimal *F1-measure* value is about 67%. Experimental results show that the proposed method can achieve the best performance when the feature dimension is 64, but the continuous increase of query depth cannot further improve the method performance.

3.3. Performance evaluation of the integrity of image matching

The integrity of image matching was evaluated using the integrity of the undirected graph, which was constructed by initial matching pairs. If the image pairs were selected sufficiently, the undirected graph constructed based on the matched pairs should be a connected component. Otherwise, the 3D reconstruction would result in a disconnected structure or an incomplete component. Thus, we evaluate the integrity of the image pairs selected by the four methods of EM, VT, AVT, and the method proposed here. Among these, the EM, VT, and AVT used 128-dimensional features, while ours used 64-dimensional features. The query depth Q was uniformly set to 50. Since both the AVT and our method performed threshold segmentation on the query curve, the impact on the image query accuracy was evaluated. The statistical results of the thresholds of all query images in the dataset are shown in Figure 8.

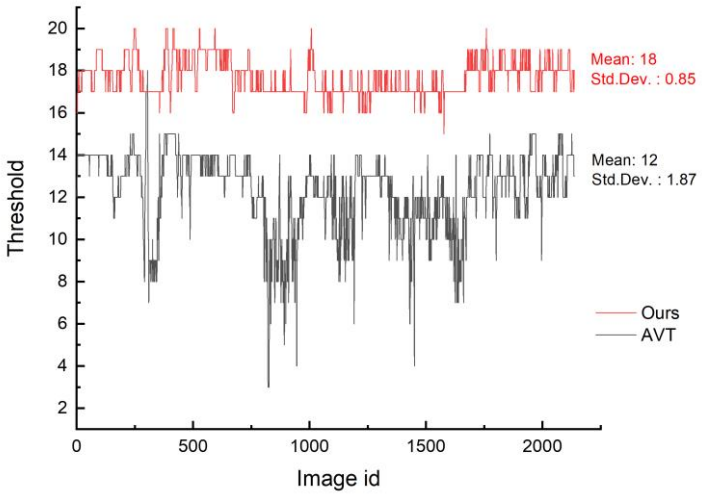


Figure 8. Statistical results of the threshold of image query using the adaptive vocabulary tree (AVT) and the method proposed here.

In Figure 8, the horizontal axis represents the query image ID and the vertical axis represents the threshold of the query image. As shown in Figure 8, the standard deviation of the threshold curve obtained by our method is 0.85, while by AVT method is 1.87. The experimental results show that the threshold curve obtained by our method was stable and the standard deviation was smaller. In order to analyze the threshold curves of different methods, the curves with ID 944 in the dataset are shown in Figure 9.

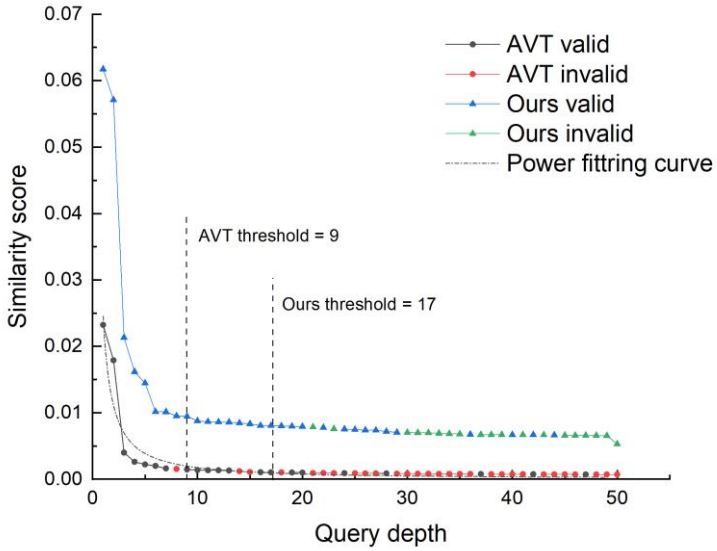


Figure 9. Comparison of threshold segmentation between the AVT and our method for image 944.

In Figure 9, the valid means that the image pair had overlap regions and the invalid means that the image pair did not have overlap regions. The AVT method fit the similarity curve as a power function and the threshold calculated based on the mean and standard deviation of the curve scores is 9. By analyzing the similarity curve obtained by the AVT method, it is found that the similarity curve of multi-scale UAV images does not conform to the power function distribution perfectly, which results in a large number of valid image pairs being lost. After feature reduction, the threshold calculated by Ours method based on the maximum variance is 17, the valid image pairs are more concentrated at the front of the curve, and more valid image pairs are obtained.

According to the image pairs selected by the four methods of EM, VT, AVT and the proposed method, the two view image matching completed and the connection relationship network was generated.

Table 1. Statistics of image selection accuracy of different methods.

Method	No. of Valid Image Pairs	No. of Invalid Image Pairs	Precision (%)	Recall (%)	No. of CC
EM	113189	2169127	—	100	2
VT	48251	56462	79.69	42.63	4
AVT	12901	11216	95.29	11.40	4
Ours	18893	28274	91.21	16.69	2

Table 1 shows the image selection accuracy of four different methods. Compared with our method, the VT method obtained the largest number of valid image pairs. The AVT method obtained the highest precision, but the connection components of both methods were 4. If the number of CC was more than 1, there were isolated CCC in the image connection network and unconnected structures in the SfM pipeline. Our method obtained the same number of CC as the EM method and achieved a balance between precision and recall.

3.4. Efficiency and accuracy of the reconstructions

In order to evaluate the efficiency and accuracy of different 3D reconstruction methods, the image pairs selected using EM, VT, and AVT methods, as well as the method proposed here, used the sequential SfM pipeline. The statistical results are shown in Table 2.

Table 2. Statistical results of SfM.

Method	Query Depth	Matching Time (min)	SfM Time (min)	No. of 3D Points	No. of Images	RMS of Reprojection Error (pixels)
EM	—	121.38	65.10	309,236	2100	0.72
VT	50	26.73	45.80	282,444	1634	0.71
	100	37.41	61.22	317,218	2100	0.72
AVT	50	20.78	40.71	323,436	1612	0.71
	100	25.72	48.74	299,552	1634	0.72
Ours	50	23.45	49.59	366,199	2100	0.73
	100	25.43	59.37	341,780	2100	0.73

As shown in Table 2, SfM time in the EM method was the longest yet had the lowest efficiency. Nonetheless, the number of connected images was the largest. When the query depth was set to 50, both the VT and the AVT methods had a large number of disconnected images. Due to the incomplete reconstruction results, its efficiency and accuracy were not evaluated. When the query depth increased to 100, the VT method obtained complete results, and the number of connected images in the AVT method also increased by 22. The experimental results show that more complete reconstruction results could be obtained by increasing the query depth, because a larger query depth can establish the connection relationship between UAV images at different scales. The disadvantage was that the efficiency decreased when query depth increased. It is difficult to determine a suitable query depth for all UAV image matching. Compared with other methods, our method obtained complete reconstruction results when query depth was 50 and 100. Moreover, it obtained the highest efficiency. Compared with the EM method, the efficiency improved by 60.8%. Compared with the VT method, which had a query depth of 100, its efficiency improved by 25.9%. Based on the comprehensive analysis presented in Tables 1 and 2, the EM method can be observed to have the highest recall, but the least number of 3D sparse points, thus reducing the SfM efficiency. Therefore, it was unnecessary to obtain the most complete recall for UAV image reconstruction. After the bundle adjustment, the root mean square reprojection of the four methods was about 0.72 pixels, which shows that the image pairs selected by our method were reasonable and that the reconstruction results were reliable.

4. Conclusions

In this paper, an image pair selection method based on the BoW retrieval method was designed for multi-scale UAV image matching. The sparse reconstruction of a landslide was obtained using a sequential SfM pipeline. First, the PCA method was used to reduce high-dimensional features to low-dimensional features, which improved the efficiency of vocabulary tree construction. Second, by calculating the query depth threshold, a large number of invalid matching image pairs within the retrieval depth were discarded, and the accuracy of image retrieval was improved. Third, through searching for missing image pairs, some image pairs that connected the parent were connected to the component, and the child connected component was automatically identified and matched. Thus, this further improved integrity of multi-scale UAV image matching. The experimental results demonstrate that the proposed method can be efficient and reliable for multi-scale UAV image matching and able to obtain complete reconstruction results.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC) (41801383); the National Key Research and Development Project of China (2018YFC1504805); the Fundamental Research Funds for the Central Universities, CHD, (300102269206, 300102269304). We are grateful to the referees for their valuable comments and suggestions which have led to an improved paper.

Conflict of interest

The authors declare no conflicts of interest.

References

1. J. L. Schönberger, F. Fraundorfer, J. M. Frahm, Structure-from-motion for MAV image sequence analysis with photogrammetric applications, *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, **40** (2014), 305.
2. G. Rossi, L. Tanteri, V. Tofani, P. Vannocci, S. Moretti, N. Casagli, Multitemporal UAV surveys for landslide mapping and characterization, *Landslides*, **15** (2018), 1045–1052.
3. D. Meyer, E. Fraijo, E. Lo, D. Rissolo, F. Kuester, *Optimizing UAV systems for rapid survey and reconstruction of large scale cultural heritage sites*, 2015 Digital Heritage, 2015.
4. P. Boccardo, F. Chiabrando, F. Dutto, F. G. Tonolo, A. Lingua, UAV deployment exercise for mapping purposes: evaluation of emergency response applications, *Sensors*, **15** (2015), 15717–15737.
5. J. A. Tenedório, R. Estanqueiro, A. Matos Lima, J. Marques, *Remote sensing from unmanned aerial vehicles for 3D urban modelling: case study of Loulé, Portugal*, International Conference Virtual City and Territory-11th Congress Virtual City and Territory, 2016.
6. C. Wu, *Towards linear-time incremental structure from motion*, 2013 International Conference on 3D Vision-3DV 2013, 2013.
7. J. L. Schonberger, J. M. Frahm, *Structure-from-motion revisited*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
8. S. Rothmund, U. Niethammer, J. P. Malet, M. Joswig, Landslide surface monitoring based on UAV- and ground-based images and terrestrial laser acanning: accuracy analysis and morphological interpretation, *First Break*, **31** (2013).
9. N. Snavely, S. M. Seitz, R. Szeliski, Modeling the world from internet photo collections, *Int. J. Comput. Vision*, **80** (2008), 189–210.
10. S. Fuhrmann, F. Langguth, N. Moehrle, M. Waechter, M. Goesele, MVE—An image-based reconstruction environment, *Comput. Graphics*, **53** (2015), 44–53.
11. J. P. Li, T. Jiang, D. Xiao, J. C. Wang, On diagram-based three-dimensional reconstruction of UAV image, *Opt. Precis. Eng.*, **24** (2016), 1501–1509.
12. S. Jiang, W. Jiang, Efficient structure from motion for oblique UAV images based on maximal spanning tree expansion, *ISPRS J. Photogramm. Remote Sens.*, **132** (2017), 140–161.
13. D. Gálvez-López, J. D. Tardos, Bags of binary words for fast place recognition in Image sequences, *IEEE Trans. Rob.*, **28** (2012), 1188–1197.
14. J. Sivic, A. Zisserman, *Video google: a text retrieval approach to object matching in videos*, Computer Vision IEEE International Conference on IEEE Computer Society, 2003.
15. D. Nister, H. Stewenius, *Scalable recognition with a vocabulary tree*, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006.
16. H. Jegou, M. Douze, C. Schmid, *Hamming embedding and weak geometric consistency for large scale image search*, European conference on computer vision, 2008.
17. T. Shen, S. Zhu, T. Fang, R. Zhang, L. Quan, *Graph-based consistent matching for structure-from-motion*, European conference on computer vision, 2016.
18. P. Moulon, P. Monasse, R. Marlet, *Global Fusion of Relative Motions for Robust*, Proceedings of the IEEE International Conference on Computer Vision, 2013.
19. N. Snavely, S. M. Seitz, R. Szeliski, Photo Tourism: Exploring Photo Collections In 3D, in *ACM siggraph 2006 papers*, 2006.

20. P. Moulon, P. Monasse, R. Perrot, R. Marlet, OpenMVG: Open Multiple View Geometry, International Workshop on Reproducible Research in Pattern Recognition, 2016.
21. K. Ni, F. Dellaert, *HyperSfM*, Georgia Institute of Technology, 2012.
22. S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, et al., Building Rome in a day, *Commun. ACM*, **54** (2011), 105–112.
23. D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision*, **60** (2014), 91–110.
24. H. Bay, T. Tuytelaars, L. Van Gool, *Surf: Speeded up robust features*, European Conference on Computer Vision, 2006.
25. G. Yu, J. M. Morel, ASIFT: An Algorithm for Fully Affine Invariant Comparison, *Image. Process. Line*, **1** (2011), 11–38.
26. J. Heinly, E. Dunn, J. M. Frahm, *Comparative evaluation of binary features*, European Conference on Computer Vision, 2012.
27. H. Yimin, S. Wenxiu, S. Xiaoxue, SIFT feature dimension reduction method and its application in image retrieval, *Chin. J. Lasers*, **42** (2015), 216–221.
28. I. T. Joliffe, *Principal component analysis*, Springer-Verlag, (1986), 1061–1065.
29. Y. Ke, R. Sukthankar, *PCA-SIFT: a more distinctive representation for local image descriptors*, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.
30. K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, *IEEE Trans. Pattern Anal. Mach. Intell.*, **27** (2005), 1615–1630.
31. H. Wang, L. Zhang, A. I. Haibin, A. N. Hong, Large scale aerial image retrieval method in 3D reconstruction, *Sci. Surv. Mapp.*, **44** (2019), 136–144.
32. N Otsu, A Threshold Selection Method from Gray-Level Histograms, *IEEE Trans. Syst. Man Cybern.*, **9** (1979), 62–66.
33. C. Wu, *SiftGPU: A GPU Implementation of scale invariant feature transform*, 2013. Available from: http://ccwu.me/vsfm/.
34. M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Commun. ACM*, **24** (1981), 381–395.
35. B. Dezső, A. Jüttner, P. Kovács, LEMON-an Open Source C++ Graph Template Library, *Elect. Notes Theor. Comput. Sci.*, **264** (2011), 23–45.
36. J. L. Schönberger, T. Price, T. Sattler, J. M. Frahm, M. Pollefeys, *A vote-and-verify strategy for fast spatial verification in image retrieval*, Asian Conference on Computer Vision, 2016.
37. S. Jiang, W. Jiang, Efficient match pair selection for oblique UAV images based on adaptive vocabulary tree, *ISPRS J. Photogramm. Remote Sens.*, **161** (2020), 61–75.
38. G. Ioannakis, A. Koutsoudis, I. Pratikakis, C. Chamzas, RETRIEVAL—An Online Performance Evaluation Tool for Information Retrieval Methods, *IEEE Trans. Multimedia*, **20** (2017), 119–127.