Pattern Recognition 66 (2017) 239–252

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Globally consistent alignment for planar mosaicking via topology analysis

Menghan Xia[a], Jian Yao[a,*], Renping Xie[a], Li Li[a], Wei Zhang[b]

[a] Computer Vision and Remote Sensing (CVRS) Lab, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, PR China
[b] School of Control Science and Engineering, Shandong University, Jinan, Shandong, PR China

ARTICLE INFO

Keywords:
Topology estimation
Reference image
Graph analysis
Global consistency
Image mosaicking

ABSTRACT

In this paper, we propose a generic framework for globally consistent alignment of images captured from approximately planar scenes via topology analysis, capable of resisting the perspective distortion meanwhile preserving the local alignment accuracy. Firstly, to estimate the topological relations of images efficiently, we search for a main chain connecting all images over a fast built similarity table of image pairs (mainly for the unordered image sequence), along which the potential overlapping pairs are incrementally detected according to the gradually recovered geometric positions and orientations. Secondly, all the sequential images are organized as a spanning tree through applying a graph algorithm on the topological graph, so as to find the optimal reference image which minimizes the total number of error propagation. Thirdly, the global alignment under topology analysis is performed in the strategy that images are initially aligned by groups via the affine model, followed by the homography refinement under the anti-perspective constraint, which manages to keep the optimal balance between aligning precision and global consistency. Finally, experimental results on two challenging aerial image sets illustrate the superiority of the proposed approach.

1. Introduction

Nowadays, satellite and aerial remote sensing are common techniques to quickly capture images for territorial monitoring in both civil and military fields. Due to the limited observing range of a single image, image mosaicking is a necessary technique to stitch multiple images into a single wide-view seamless mosaic image. The first critical step in the mosaicking process is accurately aligning images into a common coordinate system, which directly influences the mosaicking quality [1–3]. As a strict aligning model, homography is often used to describe the relationship between two images of a 3D plane or two images captured from the same camera centre [4–7]. Recently, some mosaicking methods not limited to these two geometric conditions have been proposed to extend the range of applications [8–10]. Specially, in this paper, we focus on mosaicking images from an approximately planar scene. Challenged by both pseudo-plane and accumulation error, lots of related studies have been presented in the literature of the last decade. However, the performance of both accurate alignment and global consistency still remains to be further improved.

Generally, the image alignment approaches can be divided into two categories: area-based approaches [11,12] and feature-based ones [13]. Because of the high computational cost, the area-based approaches are seldom used in the large scale mosaicking missions [14]. As for the planar mosaicking problem, such as aerial image mosaicking, the feature-based approaches are usually applied to recover the homography model between images [15–17] since the ground scene can be regarded as an approximate plane when it is observed from the aerial photographic camera. To improve the mosaicking result, many optimization algorithms have been proposed to achieve a global alignment. A typical global optimization method is “Bundle Adjustment” [18,19], which aims at finding an optimal solution minimizing the total reprojection error [20]. To provide a good initial solution for global optimization, Xing et al. [21] proposed to first apply the Extended Kalman Filter [22] onto the local area, and then refine all the parameters globally. To avoid the non-linear optimization, Kekec et al. [23] employed the affine model to optimize the initial alignment recovered by the homography model in the global optimization. To prevent the image suffering down-scaling effect, Elibol et al. [24] proposed to optimize point positions in the mosaicked frame and the alignment model in an alternate iteration scheme.

All those methods merely seek for an alignment with the least registration error, which can usually composite a satisfactory mosaic image from dozens of images. However, when sequential images are taken from a wide-range area, the global consistency of mosaicking images will be inaccessible for them, because in the case of pseudo-plane violating the strict geometric model, the least-registration-error principle is prone to inducing severe accumulation of perspective

* Corresponding author.
E-mail addresses: jian.yao@whu.edu.cn (J. Yao).

http://dx.doi.org/10.1016/j.patcog.2017.01.020
Received 17 July 2016; Received in revised form 1 November 2016; Accepted 11 January 2017
Available online 13 January 2017

distortions. To avoid this problem, Caballero et al. [22] proposed to use the hierarchical models according to the alignment quality of images, where the model with the less degree of freedom (DoF) was used for images with the bigger parallax. The essence of this method is to make a trade-off between improving the aligning precision and resisting the perspective distortion. In fact, a more reasonable solution is to allow continuous transition between aligning models according the predefined constraint, which is detailed in our previous work [25].

As to the large-scale mosaicking problem, utilizing the topology among images is another effective way to improve the mosaicking result [26,27]. To estimate the topology efficiently, Elibol et al. [28] used the low-cost tentative matching combined with the Minimum Spanning Tree (MST) solution to detect overlapping relations in an iterative scheme and decide when to update the topological estimation via the information-theory principles. As for the reference image selection, Richard et al. [29] stated that a reasonable choice is the most central image geometrically. This idea is obviously reasonable due to the fact that the central image usually has the shortest distance to all other images on average. To implement this idea, Choe et al. [30] applied a graph algorithm to select the reference image with the lowest cumulative registration error, but the registration error between each image pair have to be calculated in advance.

In this paper, we propose to obtain a visually satisfactory mosaic image with both accurate alignment and global consistency through two technical means: (1) utilizing the topology analysis to strengthen the registration constraints and reduce the error propagation; (2) adopting the alignment strategy of allowing continuous transition between different aligning models, to adaptively keep the optimal balance between alignment accuracy and global consistency. Firstly, we initialize an approximate similarity matrix for image pairs efficiently, which is combined with the Minimum Spanning Tree (MST) to find the main chain for an unordered image sequence. Then, other potential overlapping relationships are detected incrementally with the gradually recovered geometric positions along the main chain. Because of the synchronism of overlap detection and image location, our topology estimation strategy is more efficient than the Elibol et al.'s method [28]. Secondly, all the sequential images are organized as a spanning tree through the classical Floyd-Warshall algorithm, so as to find the optimal reference image with the least cascading times. Finally, a globally consistent alignment strategy is applied, which combines the affine model with the homography model effectively. The initial alignment is recovered by the robust affine model by groups and the globally homography refinement is followed under the anti-perspective constraint. The proposed approach was tested by several experiments on two challenging aerial image datasets and the performances were comprehensively evaluated by comparing with the state-of-the-art algorithm and a famous commercial software.

The remainder of this paper is organized as follows. The proposed framework is detailed in Section 2, which is comprised of the topology estimation, the selection of reference image, and the global alignment. Experimental results are provided in Section 3, followed by the conclusion and future work presented in Section 4.

2. Our approach

We propose a generic framework for globally consistent alignment of images captured from an approximately planar scene as shown in Fig. 1, which includes three modules: topology estimation, reference image selection, and global alignment. First, the sequential images are inputted for topology estimation, through which the obtained topological graph and matching results are utilized to search the optimal reference image and to provide feature correspondences for the global alignment respectively. Finally, according to the reorganized hierarchy, all the images are aligned through a specially designed double-model optimization strategy. Due to the versatile topology estimation, the proposed framework is suitable for both time-consecutive image sequence and unordered one.

For the description convenience in the following, the frequently used notations in this paper are summarized below:

- $\mathbf{I}_i$ - the i-th image in the sequential images.
- $\mathbf{A}_i$ - the 3×3 affine transformation matrix relating $\mathbf{I}_i$ to the reference frame.
- $\mathbf{H}_i$ - the 3×3 homography transformation matrix relating $\mathbf{I}_i$ to the reference frame.
- $\mathbf{x} = [\mathbf{x}, \mathbf{y}, \mathbf{w}]^T$ - the homogeneous coordinate of a 2D image feature point.
- $\mathbf{x}_{i,j}^k$ - the 2D coordinate of the k-th matched feature in $\mathbf{I}_i$ corresponding to the k-th matched feature $\mathbf{x}_{j,i}^k$ in $\mathbf{I}_j$.
- $M_{i,j}$ - the total number of feature matches between $\mathbf{I}_i$ and $\mathbf{I}_j$.
- $\varpi(\mathbf{x}) = [\mathbf{x}/\mathbf{w}, \mathbf{y}/\mathbf{w}]^T$ - the function transforming the homogeneous coordinate of a 2D point into the non-homogeneous coordinate.

2.1. Fast topology estimation

The image topology of the surveyed area is usually represented by a graph where each image is depicted as a node and the overlapping relationship between image pair is denoted by an edge or a link. Topology estimation means to find the existing overlapping relationships among all images. In this section, we try to find all the potential overlapping image pairs by utilizing the gradually recovered geometric positions of images in the time-consecutive order on the mosaicking plane, instead of blindly doing matching attempts. As for an unordered image sequence, finding a main chain connecting all images can make the problem the same as that of the time-consecutive image sequence. Therefore, an efficient strategy can be proposed to find the complete topology with the minimum image matching attempts.

2.1.1. Finding main chain with most reliability

For a sequence of n images, the main chain consisting of $(n-1)$ edges connects all the nodes/images in the graph. More strictly, it is defined as a spanning tree of an undirected graph in graphic theory [31,32]. Usually, there is no need to find a main chain for a time-consecutive image sequence since their time-consecutive links have implied a main chain already. Thus, this step is mainly set for an unordered image set in topology estimation.

Given an unordered image set, we have to measure the similarities between image pairs in advance for finding a main chain. Here, the initial similarity information is intended to be computed in an approximate and efficient way. For each image, we first extract SURF [33] features and only select a particular feature subset to represent this image. Then, the similarity between image pair is defined as the number of candidate point matches whose descriptor vector distances are less than some given distance. Specially, to increase the corresponding probability, the feature subset of every image consists of the features extracted on the same scale layer defined in the SURF detector, instead of sampling randomly. In our approach, the features from the second scale layer of the total four octaves are selected as the subset representing each image, because features from this layer hold a stable ratio of $22 \pm 3\%$ (an appropriate sampling ratio) almost for all kinds of images. The computational cost of obtaining the initial similarity information is comparatively low, since it mainly involves computing the distances between descriptor vectors of feature subsets. Over the exhaustive comparison, all the similarity values between image pairs are organized in the form of a matrix $\mathbf{S}$, where $\mathbf{S}(\mathbf{i}, \mathbf{j})$ represents the similarity between images $\mathbf{I}_i$ and $\mathbf{I}_j$. The value of $\mathbf{S}(\mathbf{i}, \mathbf{j})$ from small to large means an increasing similarity between images $\mathbf{I}_i$ and $\mathbf{I}_j$, which can be regarded as the probability of images $\mathbf{I}_i$ and $\mathbf{I}_j$ sharing an overlap.

Based on the similarity matrix, the reciprocals of those non-zero similarity values are set as the weights for the edges of the graph, i.e.,

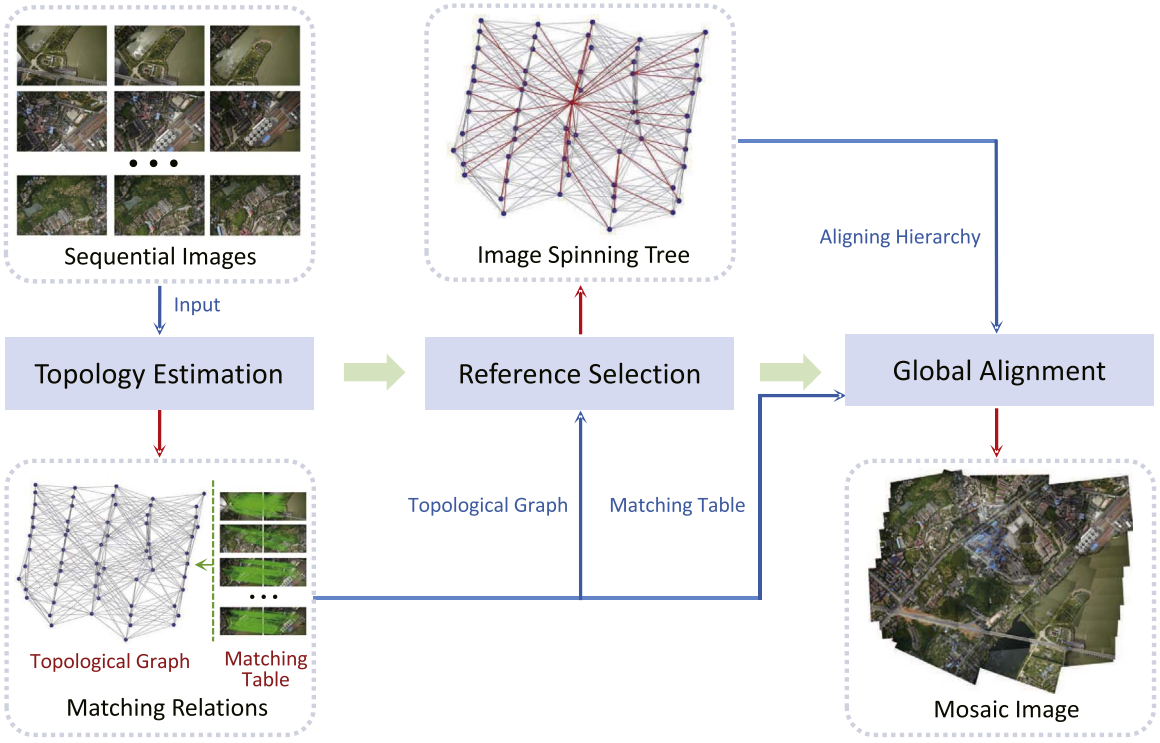


Fig. 1. The flowchart of our proposed framework for globally consistent alignment of images. The blue and red thin arrows denote the input and output of each processing module, respectively, and the wide green arrows indicate the execution sequence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$W(i,j) = \frac{1}{S(i,j)}$. Considering the similarity is not reliable, we try to find a valid main chain through an iterative scheme between selecting the main chain candidate and verifying its connectivity:

Maximum reliability: Given such a weighted graph, we try to select a linkage path that connects all the nodes with the highest total reliability, i.e., the lowest sum of weights. This is realized by finding the Minimum Spanning Tree (MST) of the current weighted graph. The MST is a spanning tree whose edges have the minimum total weight in all the spanning trees of the graph. So, the edges of MST represent the most possible overlapping relationships between image pairs.

Connectivity verification: To verify the real connectivity depicted by edges of the MST, a more reliable feature matching algorithm is employed to check the overlapping relationships between these image pairs. In such matching, all the SURF features are used and both epipolar constraint and appropriately homography constraint are applied to remove outliers. If all the matching attempts succeed, the MST is a valid main chain and the iteration terminates. Reversely, when there exists any failed matching attempt, we modify the weights of the graph where the weights of successfully matched pairs are set as zero while the weights of matching-failed image pairs are set as an infinite value, then it turns to the next iteration.

Specifically, when it fails to find the MST or the iteration reaches a given threshold, it means no connected graph exists, and the procedure quits. To illustrate the property of main chain, a subset of the first dataset described in Section 3 was selected to demonstrate the results of topology estimation, which was tested in the time-consecutive mode and in the unordered mode, respectively, as shown in Fig. 2. Apart from the difference of the main chains, the topologies estimated in the two different modes are almost the same, which are compared quantitatively in Table 1.

Algorithm 1. Detecting potential overlapping pairs.

Input: The image set $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{n}$ arranged in the order of breadth-first searching the main chain spanning tree (with the reference image as the root node).

Output: The set of overlapping image pairs $\mathcal{P} = \{\mathbf{P}_{ij}\}_{i \neq j}$.

1: Initialize the located image set $\widehat{\mathcal{I}} = \{\mathbf{I}_1\}$.
2: **for** each image $\mathbf{I}_i \in \mathcal{I} \backslash \{\mathbf{I}_1\}$ **do**
3: Align $\mathbf{I}_i$ with its direct reference image $\mathbf{I}_{\rho(i)}$.
4: Initialize the overlapping pairs set $\mathcal{P}_i = \{\mathbf{P}_{i\rho(i)}\}$.
5: **for** each image $\mathbf{I}_j \in \widehat{\mathcal{I}} \backslash \{\mathbf{I}_{\rho(i)}\}$ **do**
6: yes/no ← Detect the overlap between $\mathbf{I}_i$ and $\mathbf{I}_j$.
7: **if** yes **then**
8: $\mathcal{P}_i = \mathcal{P}_i \cup \{\mathbf{P}_{ij}\}$.
9:**end if**
10: **end for**
11: Realign $\mathbf{I}_i$ with its neighborhood image set $\mathcal{P}_i$.
12: $\mathcal{P} = \mathcal{P} \cup \mathcal{P}_i$
13: $\widehat{\mathcal{I}} = \widehat{\mathcal{I}} \cup \{\mathbf{I}_i\}$
14: **end for**
15: **return** $\mathcal{P}$

2.1.2. Detecting potential overlapping pairs

According to overlapping relationships indicated by the main chain, we can recover the comparative geometric positions of sequential images by projecting them into a common coordinate system. Then, based on the geometric information, the potential overlapping pairs can be detected easily. In our approach, the two operations are conducted in a synchronously collaborative way, instead of an independently serial way.

Firstly, we temporarily select a reference image as the mosaicking plane through applying the algorithm detailed in Section 2.2 on the main chain. To recover the comparatively geometric positions, we employ the affine model to align images, which is robust in locating the

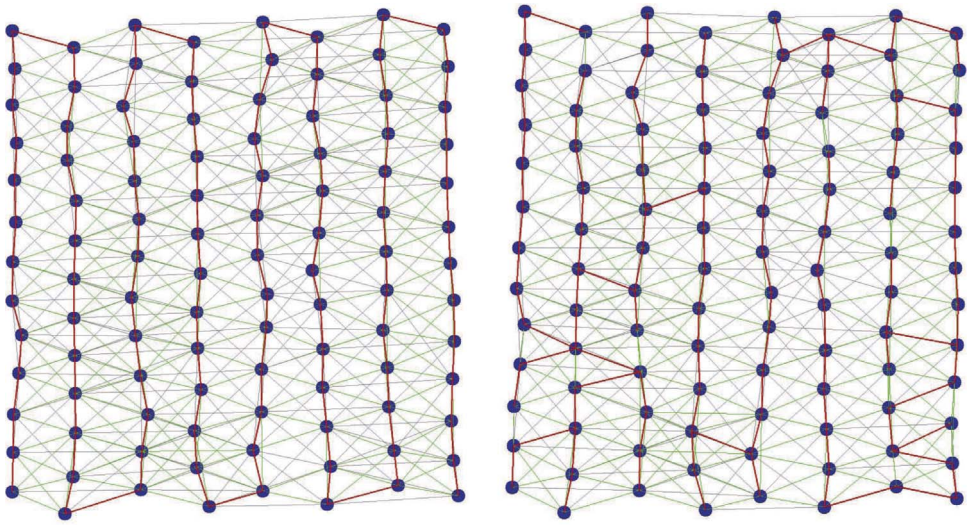


Fig. 2. The estimated topologies of an image sequence (104 images) in the time-consecutive mode and the unordered one, respectively: (a) the topology estimated in the time-consecutive mode, where the red edges represent the prior main chain in the time-consecutive order; (b) the topology estimated in the unordered mode, where the red edges represent the main chain linked by the proposed iterative scheme. Besides, the green edge indicate more than 100 matches between an image pair, while the gray edge indicate the number of the matches are less than 100. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Comparisons of our topology estimation running in both the time-consecutive mode (a) and the unordered mode (b) (with All-against-all as the ground truth).

Strategy	Successful Attempts	Total Attempts	% of Recall	% of Computation on Feature Matching
Proposed Approach (a)	606	896	94.71	99.42
Proposed Approach (b)	595	905	92.10	84.14
All-against-all	646	5356	100.00	100.00

centroids of images. Compared with the affine model, the homography model is prone to suffering from the perspective distortion and the 2D rigid model tends to cause a bending trajectory because of the error accumulation, which was testified in Section 3.2. To improve the reliability of the image locations, the images on the main chain will be aligned starting from the reference image one by one. As the images are located gradually, the potential overlapping relationships around the newly located image are detected, which then is used for optimizing the position of this newly aligned image. This strategy benefits in improving the accuracy of the recovered geometric positions, because the simultaneously detected overlapping pairs can provide extra constraints for alignment. Given a newly aligned image $\mathbf{I}_i$, we check whether it shares an overlap with all the previously aligned images $\widehat{\mathcal{I}} = \{\mathbf{I}_j\}_{j=1}^{m}$. Specially, the overlap detection between $\mathbf{I}_i$ and $\mathbf{I}_j$ is performed by calculating the distance between their centroids as follows:

$$\delta_{ij} = \frac{\max(0, |c_i - c_j| - |d_i - d_j|/2)}{\min(d_i, d_j)}, \tag{1}$$

where c_i, c_j, d_i and d_j are the image centroids and the diameters of the minimum boundary circles of the projection onto the mosaicking plane of $\mathbf{I}_i$ and $\mathbf{I}_j$, respectively. If $\delta_{ij} > 1$, there is no overlap. Otherwise, there may exist an overlap between $\mathbf{I}_i$ and $\mathbf{I}_j$, and we attempt to match them for verification. Of course, if the matching between $\mathbf{I}_i$ and $\mathbf{I}_j$ has been done in the stage of finding the main chain, it is no need to do the matching attempt again. The procedure of our topology estimation approach is described in Algorithm 1. When all the overlapping pairs are obtained, we redefine the similarity matrix as the final topological representation. The original similarity matrix is reset as a zero matrix firstly, and the value of $\mathbf{S}(i, j)$ is replaced with the number of matched points only if $\mathbf{I}_i$ and $\mathbf{I}_j$ have been matched successfully.

It should be noted that the major computation cost of the topology estimation is on feature matching between images, as listed in the fourth column of Table 1. The image alignment and potential overlapping detection have a relatively low computation cost, since there is no global optimization or iterative detection involved. Besides, as for an unordered image set, the initialization of the similarity matrix occupies the majority of the rest computation.

2.2. Optimal reference image selection

As we known, the images alignment is realized through warping each image onto the mosaicking plane which is usually one of the input images (named as the reference image). Projecting an image without direct overlap with the reference image to the mosaicking plane involves cascading a series of relative transformation models of other intermediate images. Obviously, less intermediate images used for cascading makes less error accumulation. In fact, there might exist more than one path of the same cascading number from an image to another. Considering each cascading induces a different error, we manage to select the path with the least accumulation error. In terms of this, the optimal reference image should give the lowest sum of accumulation errors from all the other images to the reference image plane. To address this problem, we construct an undirected weighted graph based on the estimated topology in Section 2.1. According to the final similarity matrix, image pairs with non-zero values of similarities are linked with edges. As to the weight (or cost) of an edge, there are two kinds of settings in the existing literature: the reciprocal of the number of matched features [28] and the registration error between the image pair [30]. The former is intuitive and efficient while the latter depicts the error directly at the cost of calculating the registration error between all available image pairs in advance. Considering the association between the number of matched features and the registration error, we creatively set the edge weight as follows:

$$w_{ij} = \begin{cases} \inf, & \text{if } M_{i,j} = 0, \\ \dfrac{1}{\log(M_{i,j} + \varepsilon)}, & \text{if } M_{i,j} > 0, \end{cases} \tag{2}$$

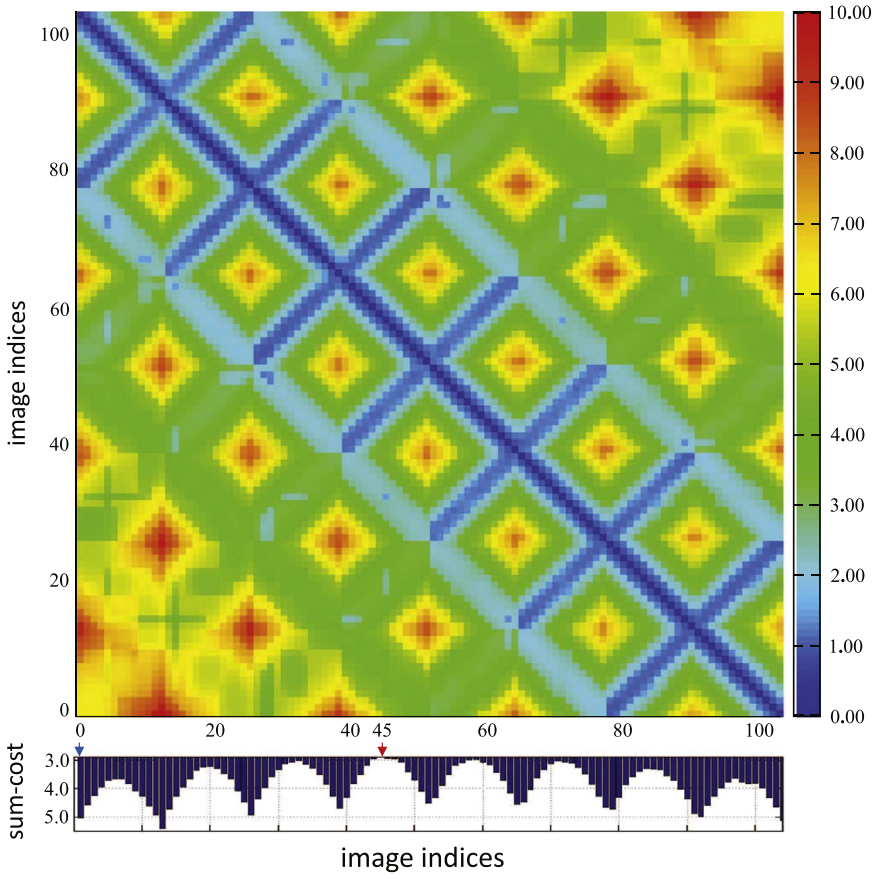


Fig. 3. The cost matrix of all-pairs shortest path calculated from a sequence of 104 images. Below is attached by the bar chart depicting the mean cumulative cost of each column of the cost matrix. A red arrow in the bottom indices labels the 45-th column with the minimum mean cumulative cost of 3.04. For comparison, the first column labeled with a blue arrow has a much higher mean cumulative cost of 5.23. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $M_{i,j}$ denotes the total number of matches between $\mathbf{I}_i$ and $\mathbf{I}_j$, and ε is a constant for regularization ($\varepsilon = 50$ by default). This weight setting equation, which describes the contribution of matched features to the registration accuracy, compromises between efficiency and effectiveness.

Based on the weighted graph, the optimal reference image selection problem is formulated as finding a node with the least total weight of the shortest paths to all the other nodes, which can be solved by the Floyd Warshalls all-pairs shortest path algorithm [34,35]. This algorithm is more efficient than running n times of single source shortest path algorithm, because the dynamic programming strategy is applied in it with the computation complexity of $O(n^3)$. With this algorithm, the shortest paths between any two nodes can be obtained. For a sequence of n images, we build a $n \times n$ size symmetric cost matrix $\mathbf{W}$ where $\mathbf{W}(i, j)$ records the cost of the shortest path between $\mathbf{I}_i$ to $\mathbf{I}_j$. Thus, the i-th row or column of matrix $\mathbf{W}$ indicates the cost of every shortest path from other images to $\mathbf{I}_i$, and the column with the minimum cumulative cost is selected as the reference image. To demonstrate the procedure, the cost matrix $\mathbf{W}$ of a sequence of 104 images is visualized in Fig. 3. The 45-th column with the minimum total cost is marked with a red arrow in the bottom indices. As a comparison, the conventional strategy of naively selecting the first image as the reference image is marked with a blue arrow. Considering the amount of images, the gap between the mean cumulative costs of the two strategies can make a big difference to the mosaicking result.

Actually, each row in $\mathbf{W}$, e.g. the i-th row, corresponds to a spanning tree with the node i as the root one, which describes the hierarchical relationships of all nodes. With the selected reference image, the spanning tree of the image sequence used in Fig. 3, is displayed in Fig. 4. The spanning tree indicates the direct reference

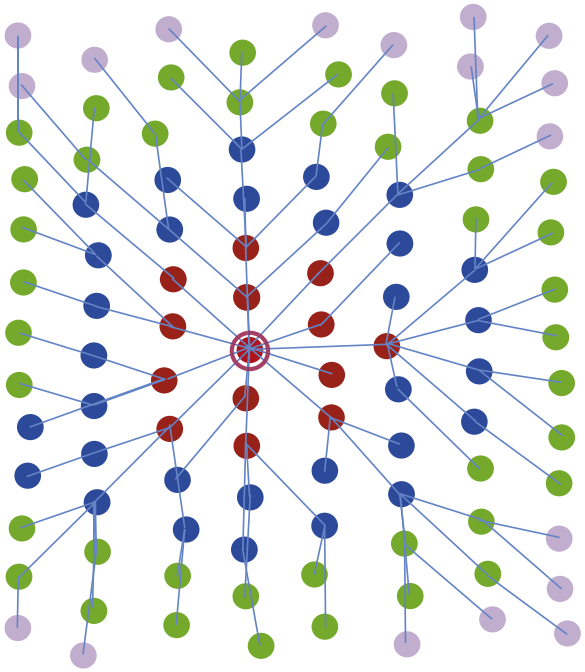


Fig. 4. The spanning tree of the graph with the optimal reference image as the root node (marked with red ring). Nodes in different levels of the tree are marked with different colors, and the blue lines imply the shortest path from every image to the reference image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

image (i.e., the parent node) of each image, which determines the aligning order of images in the following global alignment.

2.3. Globally consistent alignment

In general, the local aligning accuracy and the global consistency are two basic factors determining the quality of mosaicking result. Under a strict transformation model, these two factors can be guaranteed in a coherent way, where the higher aligning precision contributes on the better global consistency. However, in most applications, the observing scenes of pesudo-planes make the frequently-used homographic transformation just an approximate aligning model. In this case, the model of higher degrees of freedom (DoFs) usually makes more accurate alignment but suffers more severe perspective distortion meanwhile, and vice versa. Therefore, we have to deal with these two factors in a trade-off way. To keep the optimal balance between them, the model with a relatively low DoFs is employed for initial alignment which is robust to the perspective distortion, it then is refined with a higher DoFs to improve the aligning precision under the anti-perspective constraint.

2.3.1. Robust alignment by affine model

The affine model, compromising between the 2D rigid transformation and the homographic transformation, is used to make a robustly initial alignment. On the one hand, the approximately coplanar constraint of images is partly implied in the six-parameter affine model which can suppress the perspective distortion in some extent, on the other hand, the affine transformation is able to provide a qualified initial solution for the following homography refinement.

According to the spanning tree mentioned in Section 2.2, the sequential images are aligned group by group in the order of breadth-first search, which reduces the accumulation error of alignment compared to the way one by one. When aligning a new group of images to the reference frame, the overlapping relations between all the previously aligned images and the newly added ones, and the overlapping relations inside this group will be jointly used in the cost function. Let $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{s}$ be the set of previously aligned images. The affine model parameters $\mathcal{A} = \{\mathbf{A}_i\}_{i=s+1}^{s+m}$ of the newly added image group $\mathcal{G} = \{\mathbf{I}_i\}_{i=s+1}^{s+m}$ will be solved by minimizing the cost function as below:

$$E(\mathcal{A}) = E_1(\mathcal{A}|\mathcal{I}, \mathcal{G}) + E_2(\mathcal{A}|\mathcal{G}), \tag{3}$$

where the first energy term $E_1(\mathcal{A}|\mathcal{I}, \mathcal{G})$ corresponds to the overlapping relations between $\mathcal{I}$ and $\mathcal{G}$ as:

$$E_1(\mathcal{A}|\mathcal{I}, \mathcal{G}) = \sum_{\mathbf{I}_i\in\mathcal{I},\mathbf{I}_j\in\mathcal{G}} \sum_{k=1}^{M_{i,j}} \| \varpi(\mathbf{A}_i\mathbf{x}_{i,j}^k) - \varpi(\mathbf{A}_j)\mathbf{x}_{j,i}^k \|^2, \tag{4}$$

and the second energy term $E_2(\mathcal{A}|\mathcal{G})$ corresponds to the intra-group overlapping relations in $\mathcal{G}$ as:

$$E_2(\mathcal{A}|\mathcal{G}) = \sum_{\mathbf{I}_i,\mathbf{I}_j\in\mathcal{G}} \sum_{k=1}^{M_{i,j}} \| \varpi(\mathbf{A}_i\mathbf{x}_{i,j}^k) - \varpi(\mathbf{A}_j\mathbf{x}_{j,i}^k) \|^2, \tag{5}$$

where the meanings of the notations $\varpi(\cdot)$, $\mathbf{A}_i$, $M_{i,j}$ and $\mathbf{x}_{i,j}^k$ are given in the beginning of Section 2.

As the linear equations, Eq. (3) can be solved easily by the Singular Value Decomposition (SVD) method. To increase the numerical solution stability, we normalize the coordinates of matched points [36] to build the coefficient matrix. In addition, the robust estimator MLESAC [37] is used to exclude outliers for affine estimation because it is beneficial for the image mosaicking of quasi-planar scenes.

2.3.2. Model refinement under anti-perspective constraint

The affine models recovered by groups are mainly used to make the robust initial alignment, which well preserves the mosaicking result from the perspective distortion. However, due to the fact that the DoF of the model is limited and no global optimization is performed, it is necessary and probable to improve the aligning precision further. To increase the aligning accuracy to the extent not inducing the perspective distortion, the energy function should allow the transition from the affine model to the homography model under some reasonable constraint. In fact, such constraint has been implied in the affine model which has the anti-perspective property. So, the deviation between the optimal homography transformation and the initially estimated affine transformation is set as a regularization term in the optimization function.

As the images are aligned by groups, the affine models of all the images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{n}$ can be obtained, denoted as $\mathcal{A} = \{\mathbf{A}_i\}_{i=1}^{n}$, which are used as the initial parameters for the homography models in the global optimization. The homography models $\mathcal{H} = \{\mathbf{H}_i\}_{i=1}^{n}$ with respective to the reference plane will be optimized in the energy function composed of two mutually contrary terms. The data term targeting to minimize the sum of squares of the feature registration errors between images is defined as:

$$E_d(\mathcal{H}) = \sum_{\mathbf{I}_p,\mathbf{I}_q\in\mathcal{I}} \sum_{k=1}^{M_{p,q}} \| \varpi(\mathbf{H}_p\mathbf{x}_{p,q}^k) - \varpi(\mathbf{H}_q\mathbf{x}_{q,p}^k) \|^2, \tag{6}$$

where all the aligning models have more free parameters to adjust the positions of points on the mosaicking plane, which is bound to increase the whole alignment precision. Besides, the residual error is prone to distributing evenly under an uniform energy framework.

Another optimization objective is to keep the global consistency by suppressing the accumulation of the perspective distortions which may emerge in the transition from the affine model to the homography model. The regularization term, derived from the idea that the optimal homography transformation should be close to the initially estimated affine transformation, is expressed as the displacements of the warped features from their initial positions as:

$$E_r(\mathcal{H}) = \sum_{\mathbf{I}_p,\mathbf{I}_q\in\mathcal{I}} \sum_{k=1}^{M_{p,q}} (\| \varpi(\mathbf{H}_p\mathbf{x}_{p,q}^k) - \mathbf{A}_p\mathbf{x}_{p,q}^k \|^2 + \| \varpi(\mathbf{H}_q\mathbf{x}_{q,p}^k) - \mathbf{A}_q\mathbf{x}_{q,p}^k \|^2). \tag{7}$$

As depicted in Eq. (7), the regularization term is also denoted by the distances of image feature points as the data term does, which saves the troublesome normalized problem between different kinds of energy terms. So far, the energy terms defined in Eqs. (6) and (7) can be linearly combined to define the final energy function as:

$$E(\mathcal{H}) = E_d(\mathcal{H}) + \lambda E_r(\mathcal{H}), \tag{8}$$

where λ is the weight coefficient used for balancing the two terms E_d and E_r, which should be set to an appropriate small value since the constraint is not a strict one. Theoretically, a bigger value of λ strengthens the global consistency while decreases the accuracy of the local alignment. For instance, for images of large-depth-difference ground, a slightly bigger λ is needed to resist perspective distortion. In our experiments, we set its value from 0.01 to 0.05 according to data traits. As a typical non-linear least squares problem, Eq. (8) can be solved by the Levenberg-Marquardt (LM) algorithm. However, considering the specialty of this problem, we employ the sparse LM algorithm [38] to save memory and to speed up the computation, which is stated detailed in Appendix A.

3. Experimental results

In this section, two sets of representative aerial images acquired by different flight platforms and over different landforms, respectively, were used as the experimental dataset. The first dataset, consisting of 744 images from 24 sequentially ordered strips, was captured at a flight height of about 780 m over an urban area. The original images with a forward overlapping rate of about 60%, are down-sampled to the size of 1000×642 in our experiments. The second dataset, consisting of 130

Table 2
Comparisons of the topology estimation obtained by different approaches on the first dataset (with All-against-all as the ground truth).

Strategy	Successful Attempts	Total Attempts	% of Recall	% of Attempts As to All-against-all
Our Approach	5197	7771	97.83	2.81
Fast-Topology [28]	5229	9601	98.43	3.47
All-against-all	5312	276396	100.00	100.00

Table 3
Comparisons of the topology estimation obtained by different approaches on the second dataset (with All-against-all as the ground truth).

Strategy	Successful Attempts	Total Attempts	% of Recall	% of Attempts As to All-against-all
Our Approach	781	934	95.36	11.14
Fast-Topology [28]	793	1336	96.83	15.93
All-against-all	819	8385	100.00	100.00

images with the down-sampling size of 800×533, was captured by an unmanned aerial vehicle (UAV) with a forward overlapping rate of about 70%, which observes a suburb area containing mountains.

Due to the limit of pages, more experimental results and analysis are presented at http://cvrs.whu.edu.cn/projects/PlanarMosaicking/, where the dataset and the source code are publicly available for download.

3.1. Evaluation on topology estimation

Our topology estimation approach was compared with the classic all-against-all strategy and the state-of-the-art algorithm implemented according to [28] (we name it as Fast-Topology hereafter). The comparisons were made on the estimated topologies of the aforementioned datasets. To test our approach diversely, the aerial image sequence and the UAV image sequence were respectively processed in two different modes for topology estimation: the time-consecutive mode and the unordered mode. As a robust but exhaustive strategy, matching all-against-all can give the topology estimation result which can be regarded as the ground truth. Moreover, the successfully matched image pairs and the total matching attempts were combined to evaluate the topology estimation algorithm as quantitative metrics in accuracy and efficiency.

The topology estimation results of the two datasets are summarized in Tables 2, 3, respectively. As the tables show, both our approach and Fast-Topology [28] almost recovered the complete topology as the all-against-all strategy does, but with much less matching attempts. Although there are some omissions with respect to all-against-all, the major overlapping relations have been detected successfully in our approach, which can be observed in the topological graph depicted in Fig. 5(a) and Fig. 6(a). This implies that most of the undetected overlapping pairs probably share very small overlapping areas, and usually make little contribution to the mosaicking results. Compared to Fast-Topology [28], our approach has roughly the same recall rates but less total matching attempts, which benefits from two key strategies used in the potential overlapping pairs detection. The one is selecting a suitable temporary reference image by applying the strategy detailed in Section 2.2, instead of selecting the first image simply as in Fast-Topology. The other is that the position of the newly added image is simultaneously adapted along with the potential overlapping relations being detected, which improves the alignment accuracy and so does the efficiency. However, in Fast-Topology, detecting the potential overlapping pairs and adapting alignment of images with the detecting results are divided into two independent steps. Thus, it inevitably introduces many unnecessary matching attempts because of the inaccurate alignment in the first few iterations, though it can find most of the existing overlapping relations after several iterations.

As mentioned in Section 2.2, the estimated topology is used to search for the optimal reference image, by the way of which the images are organized as a spanning tree implying the aligning order for the global alignment. Here, the spanning trees with the reference image as the root node, are expressed by a group of red edges of the topological graph in Fig. 5(b) and Fig. 6(b), corresponding to the first and second datasets, respectively. It's easy to find that the selected reference images can always locate in the central part geometrically, no matter of the square-shaped aerial data or the strip-shaped UAV data.

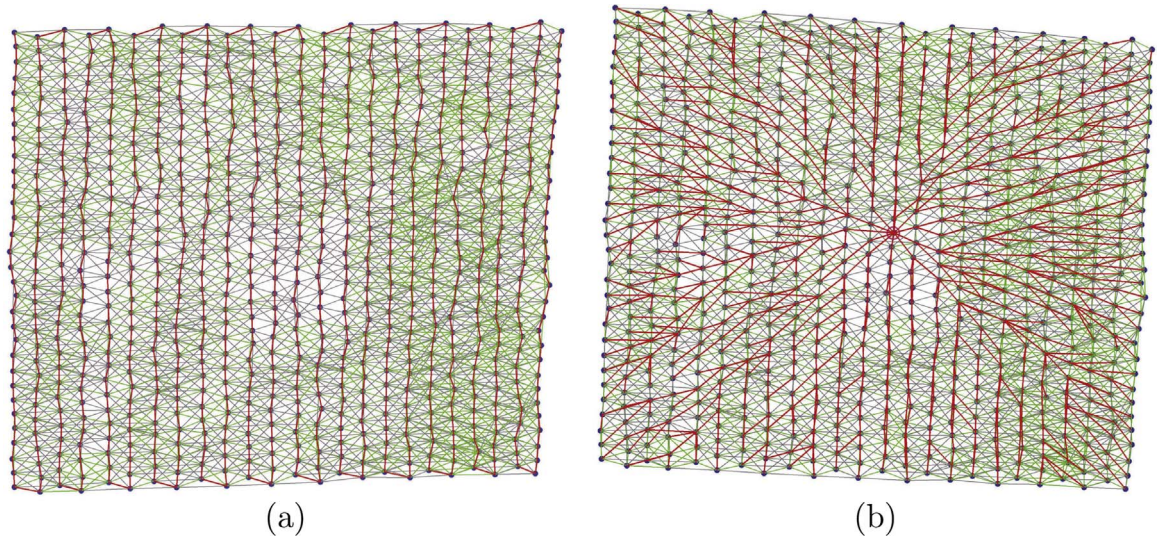


Fig. 5. The estimated topology of the first dataset (744 images) highlighted for different aims: (a) the estimated topology with the prior main chain marked with red edges; (b) the spanning tree generated by searching for the optimal reference image, marked with red edges on the estimated topological graph. Different from (a), the geometric positions in (b) are recovered by the final global alignment. The edge in green and gray indicate the matches between the image pair more and less than 100 respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

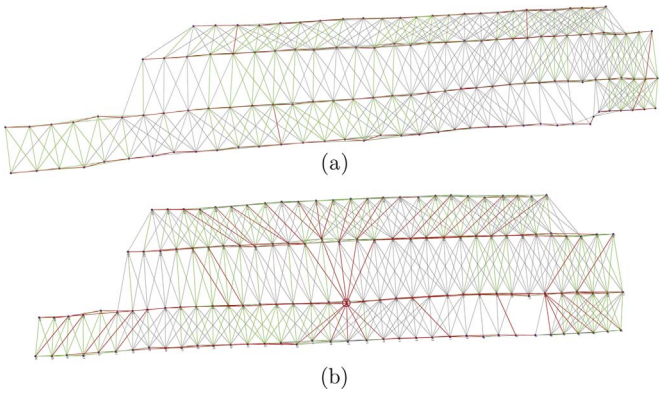


Fig. 6. The estimated topology of the second dataset (130 images) highlighted for different aims: (a) the estimated topology with the searched main chain marked with red edges; (b) the spanning tree generated by searching for the optimal reference image, marked with red edges on the estimated topological graph. Different from (a), the geometric positions in (b) are recovered by the final global alignment. The edge in green and gray indicate the matches between the image pair more and less than 100 respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The Root-Mean-Square (RMS) errors through selecting different transformation models for initial alignment in the proposed approach (GR: Global Refinement; Unit: pixel).

Models	Strip Aerial Images			Block UAV Images		
	#Matches	RMS	RMS (GR)	#Matches	RMS	RMS (GR)
Rigid	131279	3.142	1.247	48783	5.112	1.985
Affine	131279	2.825	1.117	48783	4.421	1.743
Homography	131279	2.459	0.808	48783	3.605	1.485

3.2. Evaluation on initial model selection

In the period of recovering initial alignment described in Section 2.3.1, the selection of the transformation model among *rigid*, *affine* and *homography* models can make differences to the final mosaicking result. To amplify the influence of error factors, we specially selected a strip-shaped aerial image subset and a block UAV image subset from the first dataset and the second one, respectively, and the image on the end was set as the reference image. The comparative analyses were made on both alignment precision and global consistency, where the numerical results are shown in Table 4 while the global consistency can be judged via the visual results shown in Fig. 7.

As for the strip aerial images, the homography model as the initial model has the best alignment precision, but suffers severe an accumulation of the perspective distortions because it has the highest DoF for alignment. However, the mosaicking result of rigid transformation shows a bending tendency with the lowest accuracy although it doesn't induce any perspective distortion. This is because the rigid model of 3 free parameters just allows the image translation and rotation, which is not enough to describe the truly geometric relations and easy to cause accumulation error in rotation or translation. Compromising between them, the affine model with a moderate DoF, has made a good balance between the aligning accuracy and the global consistency, which gives the most visually satisfactory mosaicking result. Additionally, because of the low flight altitude, the comparatively large-depth-difference ground greatly decreases the aligning precision of the UAV image sequence. In such case, the affine model still shows the similar ability as it does in the aerial image sequence. Conclusively, the affine model has the best comprehensive property to provide a robust initial alignment.

3.3. Comprehensive evaluation on mosaicking results

The final mosaicking results of our approach were evaluated in both qualitative and quantitative forms. Firstly, we compare the mosaic

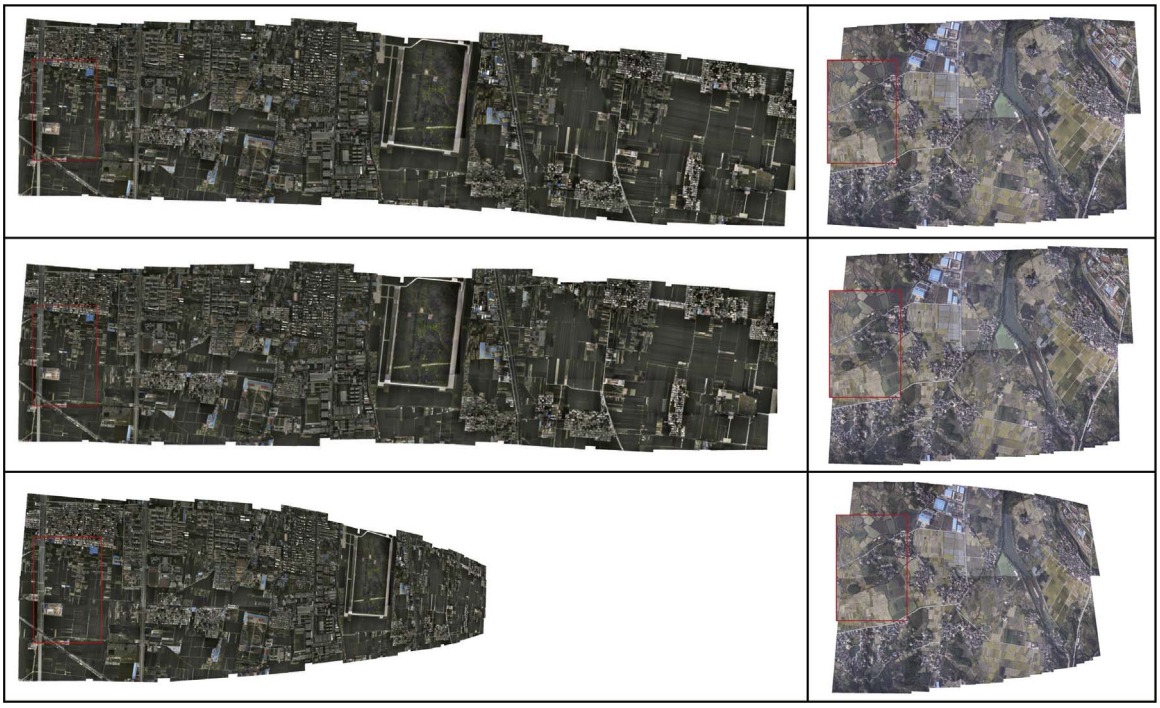


Fig. 7. The thumbnails of the mosaicking results on the aerial images (Left) and the UAV images (Right) where the rigid model in the first row, the affine model in the second row, and the homography model in the last row are chosen for initial alignment, respectively. Notice that the reference image of each mosaic is marked with a red rectangular box.

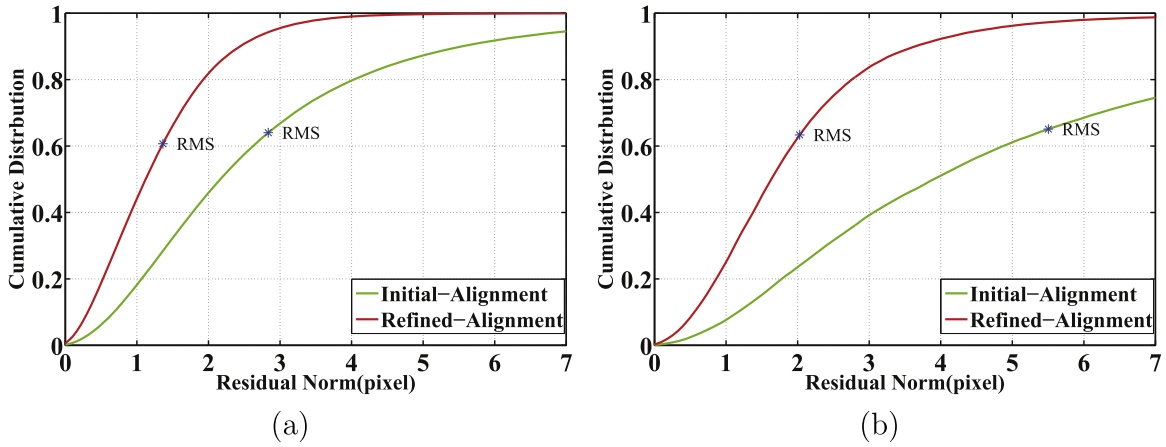


Fig. 8. Cumulative probability distributions of the residual error norms with and without the global refinement performed in our approach: (a) the error analysis for the first dataset; (b) the error analysis for the second dataset. The green curve depicts the precision of initial alignment, while the red curve depicts the precision of global alignment. The blue marks on curves indicate the RMS errors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

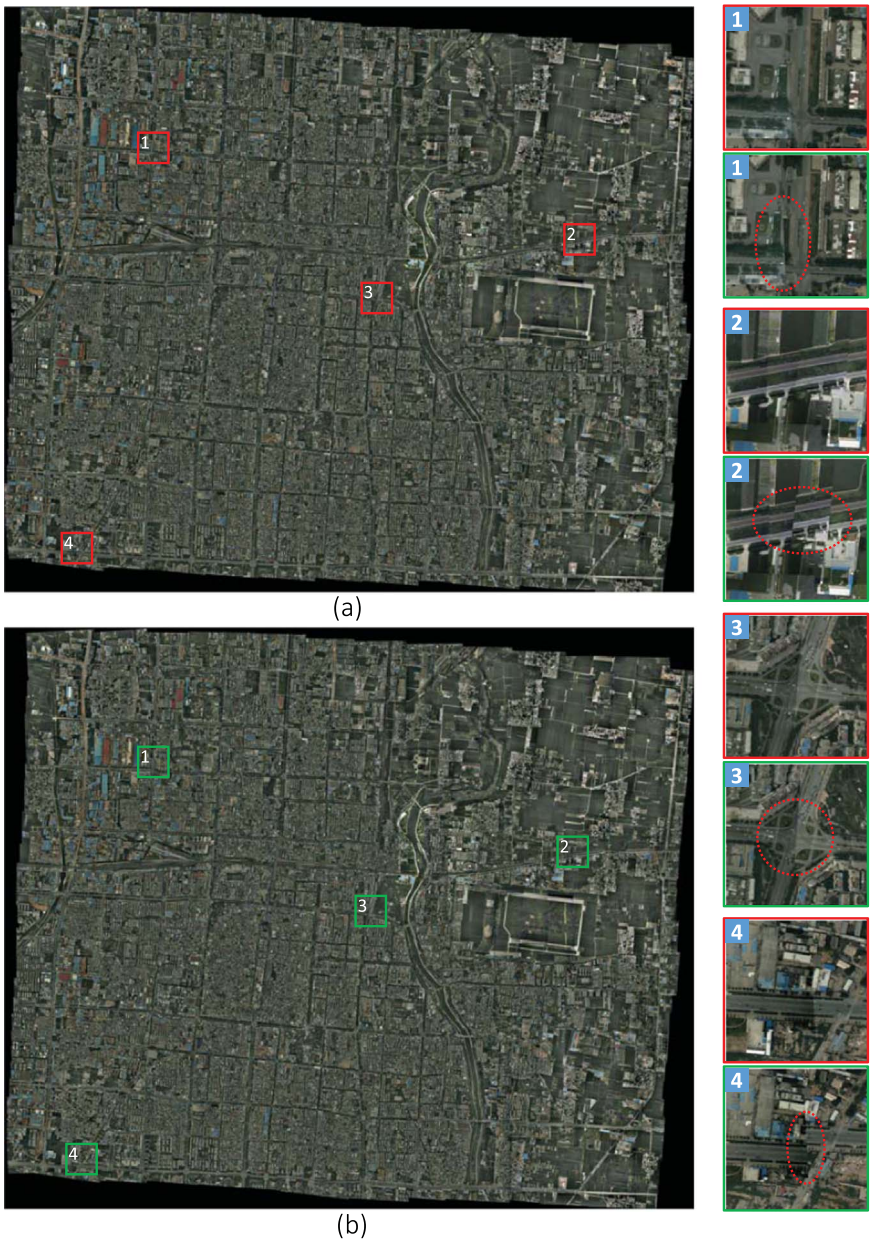


Fig. 9. The mosaics composited from the first dataset (744 images) by: (a) our approach and (b) PTGui, respectively. Several typical regions grabbed from the mosaics are enlarged in pairs in the right column.

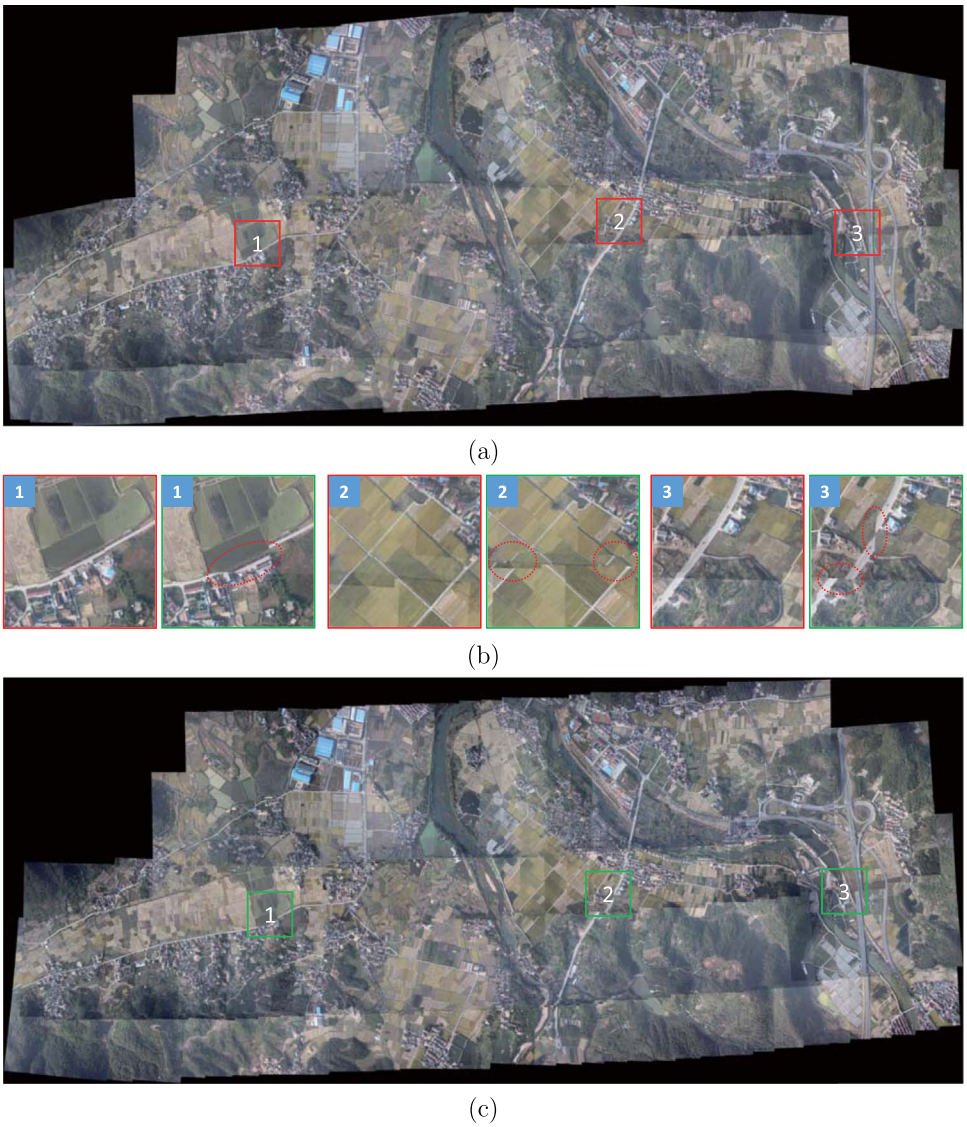


Fig. 10. The mosaics composited from the second dataset (130 images) by: (a) our approach and (c) PTGui, respectively. Several typical regions grabbed from the mosaics are enlarged in pairs in (b).

images generated by our approach with those composited by a commercial software named PTGui[1] on visual effects. Aiming at comparing the alignment results only, the following seamline detection and tonal correction were skipped in PTGui and our image stacking order was also made consistent with that of PTGui. The comparative results of the first and the second dataset are illustrated in Figs. 9 and 10, respectively.

From the mosaics shown in Fig. 9, the two mosaics have similar visual effects as a whole, both of which take on a pretty good global consistency. However, when it comes to the local aligning accuracy, our approach has an obvious superiority over PTGui, which can be observed from some enlarged regions listed in the right column of Fig. 9. As for the UAV data, the large-depth-difference ground makes the assumption of planarity of the scene weaker, which increases the difficulty to keep the global consistency. A slightly down-scale tendency in the left part can be found in the mosaicking result of our approach in Fig. 10(a). Since some strong constraints were employed for keeping the scale of each image consistent, the mosaicking result of PTGui nearly suffered no perspective distortions, but meanwhile, its alignment precision is destroyed greatly. For a detail comparison, a serial of enlarged typical regions are listed in the middle line of Fig. 10, which illustrate the better performance of our approach in the aspect of aligning accuracy.

Without precision analysis in PTGui, the quantitative evaluation of our approach was performed in two aspects. As an alignment precision, the registration error of the initial alignment and the finally global alignment in our approach, were compared in the form of cumulative probability distribution, as displayed in Fig. 8. From the comparisons, it is easy to find that the aligning precision increases a lot with the help of the homography refinement, while the global consistency is not affected during the transition from the affine model to the homography model, as can be observed in Fig. 9(a) and Fig. 10(a). This is what we aim at, namely to keep an optimal balance between the alignment accuracy and the global consistency.

In addition, the available poses of the first dataset, recovered by the rigid block adjustment of photogrammetry field, were used to calculate the homography models according to the formula in [4] under the assumption of the ground being a plane. Considering the pesudo-planarity of the ground scene, they are not accurate enough to be used

[1] http://www.ptgui.com/

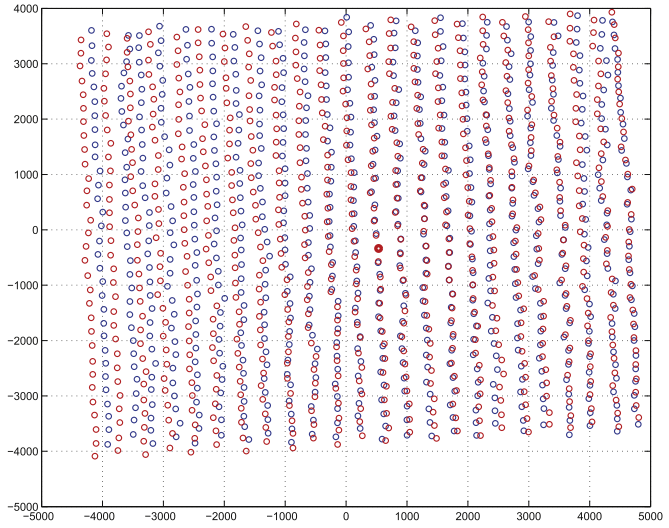


Fig. 11. Distributions of image centroids on the mosaic computed by two different approaches. The red circles are the centroids recovered by the proposed approach, and the blue ones represent the result of the pose-based approach. The solid red circle stands for the centroid of the reference image, via which the two groups are strictly superimposed as a base point. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

as the ground truth, but they are qualified to be a reference of global consistency, since the pose parameters can be regarded as no accumulation error. The recovered image centroids of the first dataset obtained by our approach and the reference models, are illustrated in Fig. 11. It shows that the two groups of centroids have a similar distribution form, except for some displacements which average at 5.16 pixels. In fact, as an image mosaicking algorithm based on feature registration, the recovered geometric positions are accurate enough to keep the global consistency of a mosaic, which emphasizes more on the visual effects than the geometric accuracy. Besides, because of no image registration based optimization performed, the pose-based approach has a terrible image aligning accuracy with the RMS error of 103.9 pixels, which is much inferior to 1.36 pixels obtained by our approach. Therefore, our approach holds a good property of alignment accuracy and global consistency in image mosaicking.

4. Conclusion and future works

In this paper, a topology analysis based generic framework is presented for mosaicking sequential images of an approximately planar scene, which contains three steps: topology estimation, reference image selection, and global alignment. Specifically, it is adapted to both ordered and unordered image sequences. To estimate the topology robustly, we perform the image location and the potential overlapping pairs detection in a collaborative way, which makes our approach significantly outperform the state-of-the-art method in efficiency. Based on the topological graph, the optimal reference image is found by graph analysis and all the images are organized as a spanning tree which gives the reference relationships for each image. With the result of topological analysis, we propose a global alignment strategy of allowing the continuous transition between the affine model and the homography one according to the energy definition, which can keep the optimal balance between the global consistency and the aligning accuracy adaptively. The proposed framework is tested with several datasets and the experimental results illustrate the superiority of our approach. However, the strategy of selecting the reference image, not taking the visual angle of image into account, may make a mosaic image of squint angle. This problem will be studied in our future work.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (Project No. 41571436), the Hubei Province Science and Technology Support Program, China (Project No. 2015BAA027), the National Natural Science Foundation of China (Project No. 41271431), and the Jiangsu Province Science and Technology Support Program, China (Project No. BE2014866).

Appendix A. Optimization derivation for model refinement under anti-perspective constraint

All the terms in the energy definition in Eq. (8) for model refinement under anti-perspective constraint are quadratic, which need to be linearized by the Taylor expansion for the iterative optimization. Generally, the first-order Taylor series expansion is accurate enough for the optimization problem of quadratic functions.

Here, we define the parameter vector of the homography matrix $\mathbf{H}_i$ as $\theta_i = [h_1^i, h_2^i, h_3^i, h_4^i, h_5^i, h_6^i, h_7^i, h_8^i]^\top$, $i \in [1, n]$, and the initial value of θ_i is defined as $\overline{\theta}_i = [\overline{h}_1^i, \overline{h}_2^i, \overline{h}_3^i, \overline{h}_4^i, \overline{h}_5^i, \overline{h}_6^i, \overline{h}_7^i, \overline{h}_8^i]^\top$. Taking a pair of matching points $\{\varpi(\mathbf{x}_{ij}^k) = (\mathbf{x}, \mathbf{y}), \varpi(\mathbf{x}_{ji}^k) = (\mathbf{x}', \mathbf{y}')\}$ from $\mathbf{I}_i$ and $\mathbf{I}_j$ for example, Eq. (8) can be written as:

$$f_k = \left(\frac{h_1^i x + h_2^i y + h_3^i}{h_7^i x + h_8^i y + 1} - \frac{h_1^j x' + h_2^j y' + h_3^j}{h_7^j x' + h_8^j y' + 1}\right)^2 + \left(\frac{h_4^i x + h_5^i y + h_6^i}{h_7^i x + h_8^i y + 1} - \frac{h_4^j x' + h_5^j y' + h_6^j}{h_7^j x' + h_8^j y' + 1}\right)^2 + \lambda\left[\left(\frac{h_1^i x + h_2^i y + h_3^i}{h_7^i x + h_8^i y + 1} - x_0\right)^2 + \left(\frac{h_1^j x' + h_2^j y' + h_3^j}{h_7^j x' + h_8^j y' + 1} - x'_0\right)^2 + \left(\frac{h_4^i x + h_5^i y + h_6^i}{h_7^i x + h_8^i y + 1} - y_0\right)^2 + \left(\frac{h_4^j x' + h_5^j y' + h_6^j}{h_7^j x' + h_8^j y' + 1} - y'_0\right)^2\right], \tag{A.1}$$

where $[x_0, y_0]^\top = \varpi(\mathbf{A_i}\mathbf{x}_{ij}^k)$ and $[x'_0, y'_0]^\top = \varpi(\mathbf{A_j}\mathbf{x}_{ji}^k)$, are the constant terms which can be calculated in advance. Eq. (A.1) is expanded in the form of the first-order Taylor series as:

$$f_k \approx \overline{f_k} + \frac{\partial f_k}{\partial h_1^k}dh_1^i + \frac{\partial f_k}{\partial h_2^i}dh_2^i + \frac{\partial f_k}{\partial h_3^i}dh_3^i + \frac{\partial f_k}{\partial h_4^i}dh_4^i + \frac{\partial f_k}{\partial h_5^i}dh_5^i + \frac{\partial f_k}{\partial h_6^i}dh_6^i + \frac{\partial f_k}{\partial h_7^i}dh_7^i + \frac{\partial f_k}{\partial h_8^i}dh_8^i + \frac{\partial f_k}{\partial h_1^j}dh_1^j + \frac{\partial f_k}{\partial h_2^j}dh_2^j + \frac{\partial f_k}{\partial h_3^j}dh_3^j + \frac{\partial f_k}{\partial h_4^j}dh_4^j + \frac{\partial f_k}{\partial h_5^j}dh_5^j + \frac{\partial f_k}{\partial h_6^j}dh_6^j + \frac{\partial f_k}{\partial h_7^j}dh_7^j + \frac{\partial f_k}{\partial h_8^j}dh_8^j, \tag{A.2}$$

where $\overline{f_k}$ is the values of f_k when substituting $\overline{\theta}_i$ and $\overline{\theta}_j$ into Eq. (A.1). $d\theta_i = [dh_1^i, dh_2^i, dh_3^i, dh_4^i, dh_5^i, dh_6^i, dh_7^i, dh_8^i]^\top$ represents the delta value of θ_i, $i \in [1, n]$. The partial derivatives of functions f_k with respect to θ_i and θ_j are listed as below:

$$
\begin{cases}
\frac{\partial f_k}{\partial h_1^i} = \frac{K_1 x}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \quad \frac{\partial f_k}{\partial h_2^i} = \frac{K_1 y}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \quad \frac{\partial f_k}{\partial h_3^i} = \frac{K_1}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \\
\frac{\partial f_k}{\partial h_4^i} = \frac{K_2 x}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \quad \frac{\partial f_k}{\partial h_5^i} = \frac{K_2 y}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \quad \frac{\partial f_k}{\partial h_6^i} = \frac{K_2}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \\
\frac{\partial f_k}{\partial h_7^i} = \frac{-K_1(\bar{h}_1^i x + \bar{h}_2^i y + \bar{h}_3^i)x}{(\bar{h}_7^i x + \bar{h}_8^i y + 1)^2} + \frac{-K_2(\bar{h}_3^i x + \bar{h}_4^i y + \bar{h}_5^i)x}{(\bar{h}_7^i x + \bar{h}_8^i y + 1)^2}, \\
\frac{\partial f_k}{\partial h_8^i} = \frac{-K_1(\bar{h}_1^i x + \bar{h}_2^i y + \bar{h}_3^i)y}{(\bar{h}_7^i x + \bar{h}_8^i y + 1)^2} + \frac{-K_2(\bar{h}_3^i x + \bar{h}_4^i y + \bar{h}_5^i)y}{(\bar{h}_7^i x + \bar{h}_8^i y + 1)^2}, \\
\frac{\partial f_k}{\partial h_1^j} = \frac{K_3 x}{\bar{h}_7^j x + \bar{h}_8^j y + 1}, \quad \frac{\partial f_k}{\partial h_2^j} = \frac{K_3 y}{\bar{h}_7^j x + \bar{h}_8^j y + 1}, \quad \frac{\partial f_k}{\partial h_3^j} = \frac{K_3}{\bar{h}_7^j x + \bar{h}_8^j y + 1}, \\
\frac{\partial f_k}{\partial h_4^j} = \frac{K_4 x}{\bar{h}_7^j x + \bar{h}_8^j y + 1}, \quad \frac{\partial f_k}{\partial h_5^j} = \frac{K_4 y}{\bar{h}_7^j x + \bar{h}_8^j y + 1}, \quad \frac{\partial f_k}{\partial h_6^j} = \frac{K_4}{\bar{h}_7^i x + \bar{h}_8^i y + 1}, \\
\frac{\partial f_k}{\partial h_7^j} = \frac{-K_3(\bar{h}_1^j x + \bar{h}_2^j y + \bar{h}_3^j)x}{(\bar{h}_7^j x + \bar{h}_8^j y + 1)^2} + \frac{-K_4(\bar{h}_3^j x + \bar{h}_4^j y + \bar{h}_5^j)x}{(\bar{h}_7^j x + \bar{h}_8^j y + 1)^2}, \\
\frac{\partial f_k}{\partial h_8^j} = \frac{-K_3(\bar{h}_1^j x + \bar{h}_2^j y + \bar{h}_3^j)y}{(\bar{h}_7^j x + \bar{h}_8^j y + 1)^2} + \frac{-K_4(\bar{h}_3^j x + \bar{h}_4^j y + \bar{h}_5^j)y}{(\bar{h}_7^j x + \bar{h}_8^j y + 1)^2},
\end{cases}
$$

where K_1, K_2, K_3, and K_4 are computed as:

$$
\begin{cases}
K_1 = \frac{2(\bar{h}_1^i x + \bar{h}_2^i y + h_3^i)}{\bar{h}_7^i x + h_8^i y + 1} - \frac{2(\bar{h}_1^j x' + \bar{h}_2^j y' + \bar{h}_3^j)}{\bar{h}_7^j x' + \bar{h}_8^j y' + 1} + 2\lambda\left(\frac{\bar{h}_1^i x + \bar{h}_2^i y + \bar{h}_3^i}{\bar{h}_7^i x + \bar{h}_8^i y + 1} - x_0\right), \\
K_2 = \frac{2(\bar{h}_4^i x + \bar{h}_5^i y + \bar{h}_6^i)}{\bar{h}_7^i x + \bar{h}_8^i y + 1} - \frac{2(\bar{h}_4^j x' + \bar{h}_5^j y' + \bar{h}_6^j)}{\bar{h}_7^j x' + \bar{h}_8^j y' + 1} + 2\lambda\left(\frac{\bar{h}_4^i x + \bar{h}_5^i y + \bar{h}_6^i}{\bar{h}_7^i x + \bar{h}_8^i y + 1} - y_0\right), \\
K_3 = -\frac{2(\bar{h}_1^i x + \bar{h}_2^i y + \bar{h}_3^i)}{\bar{h}_7^i x + \bar{h}_8^i y + 1} + \frac{2(\bar{h}_1^j x' + \bar{h}_2^j y' + \bar{h}_3^j)}{\bar{h}_7^j x' + \bar{h}_8^j y' + 1} + 2\lambda\left(\frac{\bar{h}_1^j x' + \bar{h}_2^j y + \bar{h}_3^j}{\bar{h}_7^j x' + \bar{h}_8^j y' + 1} - x'_0\right), \\
K_4 = -\frac{2(\bar{h}_4^i x + \bar{h}_5^i y + \bar{h}_6^i)}{\bar{h}_7^i x + \bar{h}_8^i y + 1} + \frac{2(\bar{h}_4^j x' + \bar{h}_5^j y' + \bar{h}_6^j)}{\bar{h}_7^j x' + \bar{h}_8^j y' + 1} + 2\lambda\left(\frac{\bar{h}_4^j x' + \bar{h}_5^j y' + \bar{h}_6^i}{\bar{h}_7^j x' + \bar{h}_8^j y' + 1} - y'_0\right).
\end{cases}
$$

For the convenience of descriptions in the following, the matrix form of Eq. (A.2) are written as the standard equation of the Least Square optimization:

$$
[v_k] = \begin{bmatrix} \cdots & \frac{\partial f_k}{\partial h_1^i} & \frac{\partial f_k}{\partial h_2^i} & \frac{\partial f_k}{\partial h_3^i} & \frac{\partial f_k}{\partial h_4^i} & \frac{\partial f_k}{\partial h_5^i} & \frac{\partial f_k}{\partial h_6^i} & \frac{\partial f_k}{\partial h_7^i} & \frac{\partial f_k}{\partial h_8^i} & \cdots \\ \cdots & \frac{\partial f_i}{\partial h_1^j} & \frac{\partial f_i}{\partial h_2^j} & \frac{\partial f_k}{\partial h_3^j} & \frac{\partial f_k}{\partial h_4^j} & \frac{\partial f_k}{\partial h_5^j} & \frac{\partial f_k}{\partial h_6^j} & \frac{\partial f_k}{\partial h_7^j} & \frac{\partial f_k}{\partial h_8^j} & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ d\theta_i \\ \vdots \\ d\theta_j \\ \vdots \end{bmatrix} - [-\overline{f_k}].
$$

The above equation is expressed with the corresponding matrix labels as:

$$
\mathbf{V}^k = \mathbf{J}^k \mathbf{X} - \mathbf{L}^k, \tag{A.3}
$$

where the dots in the Jacobi matrix $\mathbf{J}^k$ represent a series of zeros, and the dots in $\mathbf{X}$ indicate the other unknown parameters in $\{d\theta_i\}_{i=1}^n$. $\mathbf{V}^k$ is the residual error of a pair of matching points. Hereafter, we name $\mathbf{J}^k$ and $\mathbf{L}^k$ as the coefficient matrix and the constant matrix, respectively.

As can be seen, a pair of matching points from two images provides an equation with 16 unknown parameters. Supposing that n images have m pairs of overlapping relations and there are s matching points of each image pair in average, then we obtain a Jacobi matrix with the size of $m \times s$ rows and $8 \times m$ columns and a constant matrix with the size of $m \times s$ rows and 1 column. In each iteration, $\mathbf{J}_{ms\times 8n}$ and $\mathbf{L}_{ms\times 1}$ have to be recalculated and the corresponding solution vector $\mathbf{X}_{8n\times 1} = [d\theta_1^\top, \ldots, d\theta_n^\top]^\top$ can be solved with the following equation:

$$
\mathbf{X}_{8n\times 1} = (\mathbf{J}_{ms\times 8n}^\top \mathbf{J}_{ms\times 8n})^{-1}(\mathbf{J}_{ms\times 8n}^\top \mathbf{L}_{ms\times 1}). \tag{A.4}
$$

The initial solution of $\{\theta_i\}_{i=1}^n$ for next iteration is updated by adding up $\mathbf{X}_{8n\times 1}$ and the initial solution used in this iteration. As the iteration goes, the updated solution will converge to the optimal solution gradually unless the initial solution provided at the very beginning is not accurate enough. However, when the amount of images is large, the size of the Jacobi matrix will be very huge and makes a challenge to the memory of computer.

In fact, we can calculate $\{\theta_i\}_{i=1}^n$ directly if $\mathbf{TJ}_{8n\times 8n} = \mathbf{J}_{ms\times 8n}^\top \mathbf{J}_{ms\times 8n}$ and $\mathbf{TL}_{8n\times 1} = \mathbf{J}_{ms\times 8n}^\top \mathbf{L}_{ms\times 1}$ have been obtained. So, to reduce the required memory space and the computation time, we manage to compute $\mathbf{TJ}_{8n\times 8n}$ and $\mathbf{TL}_{8n\times 1}$ by adding up the matrix $\mathbf{J}^{i\top}\mathbf{J}^i$ and the matrix $\mathbf{J}^{i\top}\mathbf{L}^i$ calculated from each pair of matching points, instead of building the large $\mathbf{J}$ and $\mathbf{L}$ beforehand. The improved computation formula is defined as:

$$
\begin{cases}
\mathbf{TJ}_{8n\times 8n} = \sum_{i=1}^{ms} \mathbf{J}^{i\top}_{1\times 8n} \mathbf{J}^i_{1\times 8n}, \\
\mathbf{TL}_{8n\times 1} = \sum_{i=1}^{ms} \mathbf{J}^{i\top}_{1\times 8n} \mathbf{L}^i_{2\times 1}.
\end{cases} \tag{A.5}
$$

Then, the solution can be obtained in this way as:

$$\mathbf{X}_{8n\times 1} = \mathbf{T}\mathbf{J}^{-1}_{8n\times 8n}\mathbf{T}\mathbf{L}_{8n\times 1}. \tag{A.6}$$

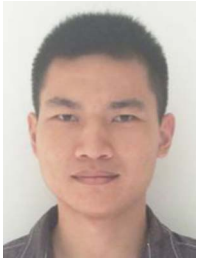
Additionally, considering the sparsity of $\mathbf{J}^i$, the computation of the matrix multiplication in Eq. (A.5) can be improved further in the complexity of both time and space.

Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.patcog.2017.01.020.

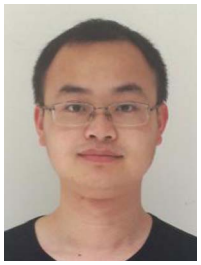
References

[1] E. Zagrouba, W. Barhoumi, S. Amri, An efficient image-mosaicing method based on multifeature matching, Mach. Vis. Appl. 20 (3) (2009) 139–162.
[2] J. Chen, H. Feng, K. Pan, Z. Xu, Q. Li, An optimization method for registration and mosaicking of remote sensing images, Opt. - Int. J. Light Electron Opt. 125 (2) (2014) 697–703.
[3] B. Zitova, J. Flusser, Image registration methods: a survey, Image Vis. Comput. 21 (11) (2003) 977–1000.
[4] L. Kang, L. Wu, Y. Wei, B. Yang, H. Song, A highly accurate dense approach for homography estimation using modified differential evolution, Eng. Appl. Artif. Intell. 31 (4) (2014) 68–77.
[5] Z. Wang, Y. Chen, Z. Zhu, W. Zhao, An automatic panoramic image mosaic method based on graph model, Multimed. Tools Appl. 75 (5) (2015) 2725–2740.
[6] N.R. Gracias, S. Van Der Zwaan, A. Bernardino, S.-.V. Jos, Mosaic-based navigation for autonomous underwater vehicles, IEEE J. Ocean. Eng. 28 (4) (2003) 609–624.
[7] Y. Xu, J. Ou, H. He, X. Zhang, J. Mills, Mosaicking of unmanned aerial vehicle imagery in the absence of camera poses, 2016, 8, 3, (2016), 204.
[8] Y. He, R. Chung, Image mosaicking for polyhedral scene and in particular singly visible surfaces, Pattern Recognit. 41 (3) (2008) 1200–1213.
[9] J. Zaragoza, T.-J. Chin, M. Brown, D. Suter, As-projective-as-possible image stitching with moving DLT, IEEE Transact. Pattern Anal. Mach. Intell. 36 (7) (2014) 1285–1298.
[10] C.-H. Chang, Y. Sato, Y.-Y. Chuang, Shape-preserving half-projective warps for image stitching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3254–3261.
[11] D. Patidar, A. Jain, Automatic image mosaicing: an approach based on FFT, Int. J. Sci. Eng. Technol. 1 (1) (2011) 01–04.
[12] S. Ghannam, A.L. Abbott, Cross correlation versus mutual information for image mosaicing, Int. J. Adv. Comput. Sci. Appl. 4 (11) (2013) 94–102.
[13] A. Elibol, R. Garcia, O. Delaunoy, N. Gracias, Efficient Topology Estimation for Large Scale Optical Mapping, Springer, 2013, Ch. A New Global Alignment Method for Feature Based Image Mosaicing, pp. 25–39.
[14] S. Ali, C. Daul, E. Galbrun, Guillemin, Anisotropic motion estimation on edge preserving Riesz wavelets for robust video mosaicing, Pattern Recognit. 51 (2016) 425–442.
[15] Y. Chen, J. Sun, G. Wang, Minimizing geometric distance by iterative linear optimization, in: Proceedings of the IEEE International Conference on Pattern Recognition, Vol. 29, 2010, pp. 1–4.
[16] W. Mou, H. Wang, G. Seet, L. Zhou, Robust homography estimation based on nonlinear least squares optimization, Math. Probl. Eng. 2014 (1) (2013) 372–377.
[17] L. Zhang, Y. Li, J. Zhang, Y. Hu, Homography estimation in omnidirectional vision under the L_∞-norm, in: Proceedings of the IEEE International Conference on Robotics and Biomimetics, 2010, pp. 1468–1473.
[18] B. Triggs, P.F. McLauchlan, R.I. Hartley, A.W. Fitzgibbon, Vision Algorithms: Theory and Practice, Lecture Notes in Computer Science, Springer, 2000, Ch. Bundle adjustmenta modern synthesis, pp. 298–372.
[19] K. Konolige, Sparse sparse bundle adjustment, in: British Machine Vision Conference, 2010, pp. 1–10.
[20] M. Li, D. Li, D. Fan, A study on automatic UAV image mosaic method for paroxysmal disaster, in: Proceedings of the International Society of Photogrammetry and Remote Sensing Congress, 2012.
[21] C. Xing, J. Wang, Y. Xu, A robust method for mosaicking sequence images obtained from UAV, in: Proceedings of the International Conference on Information Engineering and Computer Science (ICIECS), 2010.
[22] F. Caballero, L. Merino, J. Ferruz, A. Ollero, Homography based Kalman filter for mosaic building. applications to UAV position estimation, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2007, pp. 2004–2009.
[23] A.Y. Taygun Kekec, M. Unel, A new approach to real-time mosaicing of aerial images, Robot. Auton. Syst. 62 (12) (2014) 1755–1767.
[24] A. Elibol, R. Garcia, R. Elibol, A. Gracias Na, O. Delaunoy, N. Gracias, A new global alignment method for feature based image mosaicing, Adv. Vis. Comput., Lect. Notes Comput. Sci. 5359 (7) (2008) 257–266.
[25] M. Xia, J. Yao, L. Li, X. Lu, Globally consistent alignment for mosaicking aerial images, in: Proceedings of the IEEE International Conference on Image Processing, 2015.
[26] E.-Y. Kang, I. Cohen, G. Medioni, A graph-based global registration for 2D mosaics, in: Proceedings of the International Conference on Pattern Recognition, Vol. 1, 2000, pp. 257–260.
[27] R. Marzotto, A. Fusiello, V. Murino, High resolution video mosaicing with global alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2004, pp. 692–698.
[28] A. Elibol, N. Gracias, R. Garcia, Fast topology estimation for image mosaicing using adaptive information thresholding, Robot. Auton. Syst. 61 (2) (2013) 125–136.
[29] R. Szeliski, Image alignment and stitching: a tutorial, Found. Trends Comput. Graph. Vis. 2 (1) (2006) 1–104.
[30] T.E. Choe, I. Cohen, M. Lee, G. Medioni, Optimal global mosaic generation from retinal images, in: Proceedings of the IEEE International Conference on Pattern Recognition, Vol. 3, 2006, pp. 681–684.
[31] B. Bollobs, Modern Graph Theory, Springer Science & Business Media, 2013.
[32] R.L. Graham, P. Hell, On the history of the minimum spanning tree problem, Ann. Hist. Comput. 7 (1) (1985) 43–57.
[33] T.T. Bay, Herbert, L.V. Gool, Surf: Speeded up robust features, in: Proceedings of the IEEE European Conference on Computer Vision, vol. 3951, 2006, pp. 404–417.
[34] R.W. Floyd, Algorithm 97: shortest path, Commun. ACM 5 (6) (1962) 345.
[35] T.H. Cormen, Introduction to Algorithms, MIT Press, 2009.
[36] R.I. Hartley, In defense of the eight-point algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 19 (6) (1997) 580–593.
[37] P. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156.
[38] M.I.A. Lourakis, Sparse non-linear least squares optimization for geometric vision, in: Proceedings of the IEEE European Conference on Computer Vision, vol. 6312, 2010, pp. 43–56.



Menghan Xia received the B.E. degree from Wuhan University, China, in 2014. He is pursuing a M.E. degree at School of Remote Sensing and Information Engineering, Wuhan University, China. He has published several international conference papers and journal ones. Currently he mainly works on Image Processing, Computer Vision, 3D Reconstruction, etc.

Jian Yao received the B.Sc. degree in Automation in 1997 from Xiamen University, China, the M.Sc. degree in Computer Science from Wuhan University, China, and the Ph.D. degree in Electronic Engineering in 2006 from The Chinese University of Hong Kong. From 2001–2002, he has ever worked as a Research Assistant at Shenzhen R & D Centre of City University of Hong Kong. From 2006–2008, he worked as a Postdoctoral Fellow in Computer Vision Group of IDIAP Research Institute, Martigny, Switzerland. From 2009–2011, he worked as a Research Grantholder in the Institute for the Protection and Security of the Citizen, European CommissionJoint Research Centre (JRC), Ispra, Italy. From 2011–2012, he worked as a Professor in Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, China. Since April 2012, he has been a Hubei "Chutian Scholar" Distinguished Professor with School of Remote Sensing and Information Engineering, Wuhan University, China, and the director of Computer Vision and Remote Sensing (CVRS) Lab (CVRS Website: http://cvrs.whu.edu.cn/), Wuhan University, China. He is the member of IEEE, has published over 90 papers in international journals and proceedings of major conferences and is the inventor of over 20 patents. His current research interests mainly include Computer Vision, Image Processing, Machine Learning, and LiDAR Data Processing, Robotics, etc.



Renping Xie received the B.E. degree from Shanxi University of Finance & Economics in June 2012. He is a successive postgraduate and doctoral program graduate student majoring in Photogrammetry and Remote Sensing in Wuhan University. He has published several international conference papers and journal ones, and is the inventor of several patents. His current research interests include SLAM (simultaneous localization and mapping), Robotics, Image Processing, etc.

Li Li received the B.E. degree from Wuhan University, China, in 2013. He is pursuing a Ph.D. degree at School of Remote Sensing and Information Engineering, Wuhan University, China. He has published several international conference papers and journal ones, and is the inventor of several patents. Currently he mainly works on LiDAR Data Processing, Robotics, Machine Learning, etc.

Wei Zhang is with School of Control Science and Engineering of Shandong University as an associate professor. He received his Ph.D. at The Chinese University of Hong Kong (CUHK). He previously worked as a Postdoc scholar at University of California, Berkeley (UC Berkeley). He is a member of IEEE and has published over 40 papers in computer vision, image processing and machine learning. He received several international awards from IEEE, and served for many famous international conferences such as CVPR, ICCV, ECCV, ICIP, ROBIO, ICIA, ICAL, etc.