Remote Sens. **2015**, *7*, 6932-6949; doi:10.3390/rs70606932

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Diverse Scene Stitching from a Large-Scale Aerial Video Dataset

Tao Yang [1,*], Jing Li [2,*], Jingyi Yu [3], Sibing Wang [1] and Yanning Zhang [1]

[1] Shaanxi Provincial Key Lab of Speech and Image Information Processing,School of Computer Science, Northwestern Polytechnical University, Xi'an, China;
E-Mails: wangsiminnwpu@gmail.com (S.W.); ynzhangnwpu@gmail.com (Y.Z.)

[2] School of Telecommunications Engineering, Xidian University, Xi'an, China

[3] Department of Computer and Information Science, University of Delaware, Newark, DE 19711, USA;
E-Mail: yu@eecis.udel.edu

* Authors to whom correspondence should be addressed; E-Mail:tyang@nwpu.edu.cn (T.Y.); jinglixd@mail.xidian.edu.cn (J.L.); Tel.: +86-15002919079 (T.Y.); +86-13991320168 (J.L.).

Academic Editors: Devrim Akca and Prasad S. Thenkabail

Received: 19 March 2015 / Accepted: 22 May 2015 / Published: 28 May 2015

Abstract: Diverse scene stitching is a challenging task in aerial video surveillance. This paper presents a hybrid stitching method based on the observation that aerial videos captured in real surveillance settings are neither totally ordered nor completely unordered. Often, human operators apply continuous monitoring of the drone to revisit the same area of interest. This monitoring mechanism yields to multiple short, successive video clips that overlap in either time or space. We exploit this property and treat the aerial image stitching problem as temporal sequential grouping and spatial cross-group retrieval. We develop an effective graph-based framework that can robustly conduct the grouping, retrieval and stitching tasks. To evaluate the proposed approach, we experiment on the large-scale VIRATaerial surveillance dataset, which is challenging for its heterogeneity in image quality and diversity of the scene. Quantitative and qualitative comparisons with state-of-the-art algorithms show the efficiency and robustness of our technique.

Keywords: diverse scene stitching; cross-group retrieval; aerial image stitching; aerial video surveillance

---

1. Introduction

Diverse scene stitching from a large-scale aerial video dataset is highly desired in the field of remote sensing. Different from single scene stitching, diverse scene stitching simultaneously processes data from multiple scenes. For each scene, the data may come from revisited observations of the same aerial platform at different times or from co-observations of multiple aerial platforms. The purpose of diverse scene stitching is to combine all observations with various times, illuminations, viewpoints and even platforms together to generate scene oriented panoramas. The biggest advantage is the ability to automatically explore multiple scenes and to generate panoramas from large-scale aerial surveillance data. Further, the results also provide useful inputs for high-level scene understanding.

Image stitching has been well studied over the past decade [1–10]. We refer the readers to the comprehensive survey [11] for a background. In the field of remote sensing, image stitching or image mosaicking is one of the main tasks, and there are many relevant works, such as environmental monitoring, land, water or marine resources survey, aerial video surveillance, *etc*. Here, we mainly focus on stitching methods related to aerial video surveillance, and the related stitching methods [1,5,6,9,12] can be classified into two categories, including sequential stitching and retrieval stitching.

Sequential stitching algorithms [1,5,12] are based on the assumption that the input data are ordered. Small baseline algorithm, such as optical flow [13,14], can then be used to efficiently estimate the image transformation between consecutive frames. Assuming the time consistency cannot be broken, these sequential methods require that the input data should be taken on a stable aerial platform without sudden viewpoint changes or large motion blurs. Otherwise, no time consistency can be used to align these sequential images, resulting in discontinuous panoramas. While advanced methods based on wide baseline image matching techniques, such as the well-known scale-invariant feature transform [15] and speeded-Up robust features (SURF) [16], can reduce these artifacts, they cannot handle large changes of aerial data. For example, Figure 1 shows samples from the benchmark VIRATdataset [17] of a diverse scene (as shown in Figure 1a). Motion blurs, low contrast and significant viewpoint changes (as shown in Figure 1b,c) are common in the acquired images. The continuity assumption hence is not valid in this case.

The stitching problem can also resort to image matching techniques. Techniques, such as [2,4,6,18–20], assume that the input data are completely unordered. Brown *et al.* [4,19] formulate stitching as a multi-image matching problem and use brute-force searching to find the overlapping relationships among all images. These methods can recognize multiple panoramas from a small-scale image dataset.

However, given a dataset with n images, this results in $n(n-1)/2$ possible image pairs and, hence, leads to $O(n^2)$ complexity. Consequently, the computational cost of these approaches is generally very high and becomes a bottleneck in real-time surveillance.

There is also an emerging trend for using retrieval techniques [6,18,20] to find one-to-many image matches. These solutions then apply two-view matching on all selected image pairs. The retrieval stitching method is based on feature indexing. Although using indexing is much less expensive than matching, it requires sufficient distinctive image features. If the observed scene lacks enough unique features, the variance of different images is not enough for the method to reliably find correct matches.

As a result, the retrieval stitching method cannot find all overlapping aerial images from the same scene and yields low performance: typically about 80% correct matches can be found [18]. For instance, a large number of indistinctive or self-similar scenes in the VIRAT dataset (as shown in Figure 1b) will lead to retrieval failure.

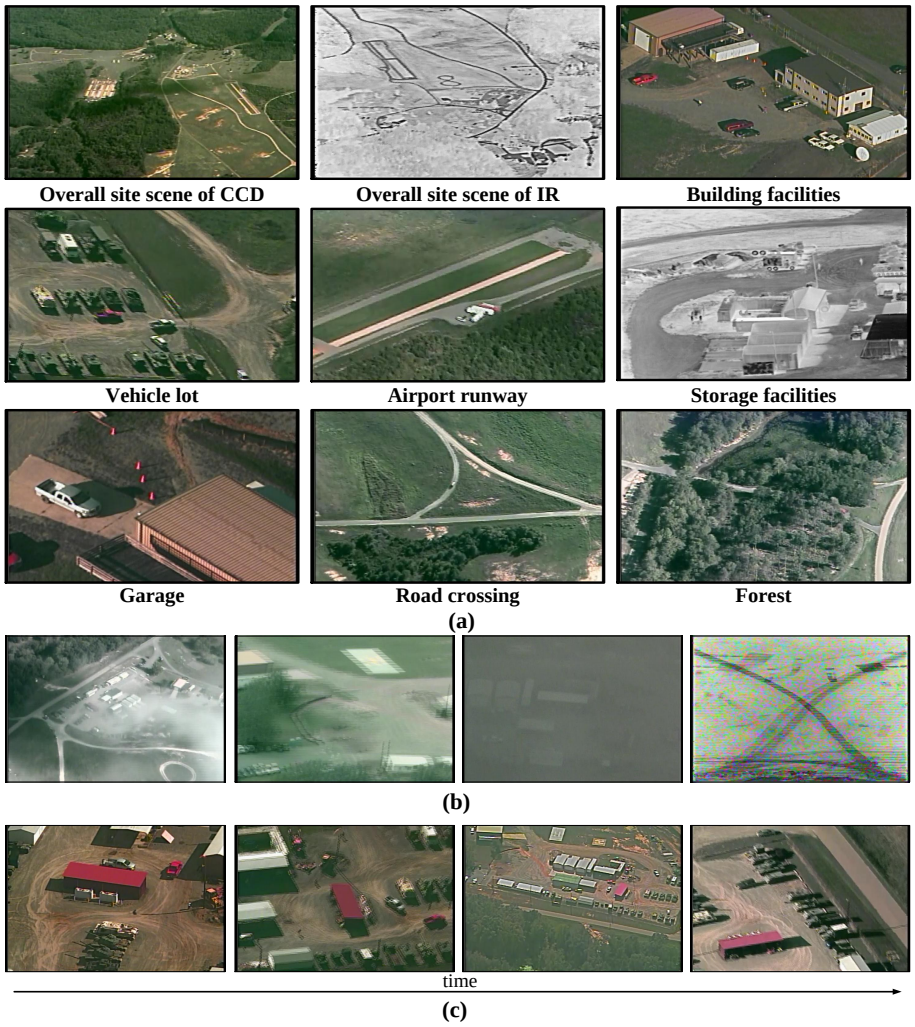

Figure 1. Examples of the large-scale VIRATaerial dataset [17], which includes images selected from 25-h realistic aerial videos. The dataset is published in [17] and available from www.viratdata.org. (**a**) Diverse scenes of VIRAT aerial dataset. (**b**) Sample image shots with clouds, motion blur, low contrast and camera noise. (**c**) Sample image shots with varying scales and viewpoints over time

The main contribution of this paper is a hybrid stitching method that unifies the temporal continuity and spatial repeatability of aerial videos in a graph-based framework. The heart of our method is built on the observation that aerial videos captured in diverse realistic scene are neither totally ordered nor completely unordered. Even though the aerial videos are corrupted by large motion blurs, sudden scene changes, low contrast and high camera noise, we observe that for each scene, there always exists short continuous videos from different times. We therefore adopt sequential grouping to first roughly partition the entire video into small continuous groups and then present a cross-group retrieval method to efficiently find spatially overlapping images among different groups. Finally, a graph-based method

is applied to find global optimal paths for stitching the images into panoramas. Experiments show that our method can robustly stitch VIRAT aerial surveillance video and achieve a few orders of magnitude accelerations over the state-of-the-art stitching systems, such as PTGui [21], AutoStitch [19] and the most recent scheme by Autopano [22].

2. Hybrid Stitching Model

2.1. Problem Formulation

In this section, we tackle the diverse scene stitching problem by proposing a reliable hybrid stitching model to automatically find overlapping images from a large-scale aerial dataset. The proposed model is built upon two characteristics of aerial video. (1) Temporal continuity: the human operator usually applies continuous monitoring on the area of interest, which results in short, but temporal continuous aerial video clips. (2) Spatial repeatability: in many surveillance task, especially long time persistent surveillance, same spatial area may be repeatedly visited at various time from one or multiple aerial platforms. We term this method hybrid stitching and integrate both sequential grouping and cross-group retrieval in a graph-based framework to solve it (as shown in Figure 2).

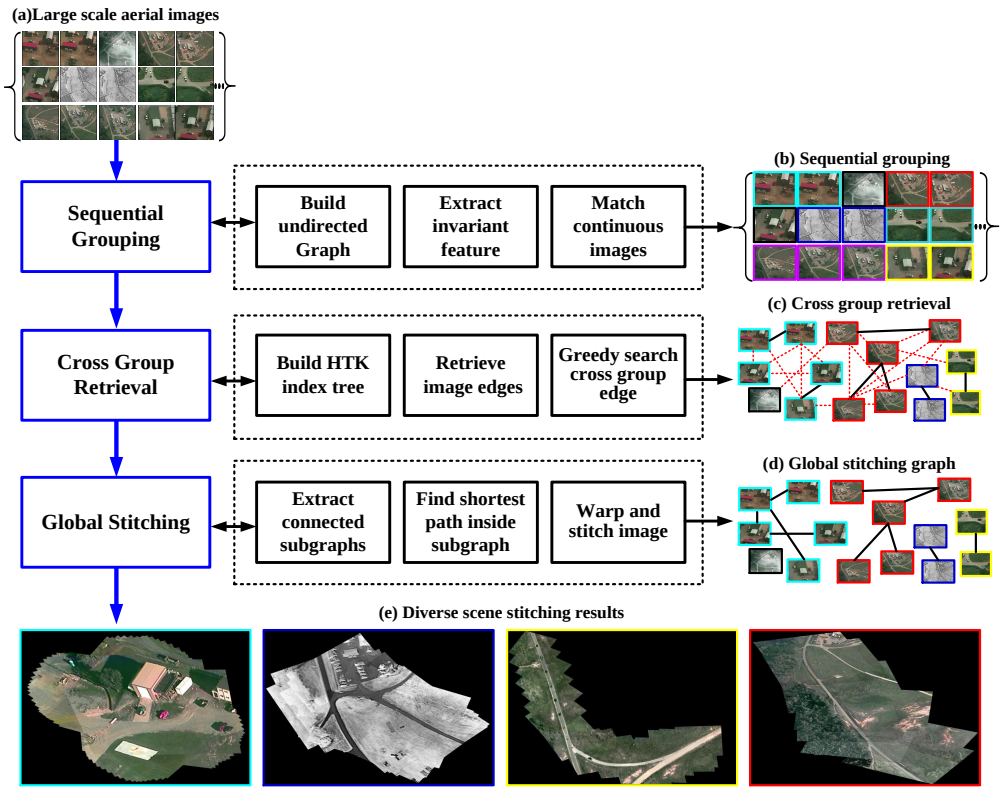

Figure 2. Framework of our method, which mainly contains three parts: sequential grouping, cross-group retrieval and global stitching. (**a**) Input large-scale aerial images; (**b**) example results of sequential grouping; the same color represents the same group; (**c**) example results of cross-group retrieval; black lines denote the sequential edges, and red dotted lines represent the candidate edges generated by retrieval; (**d**) example results of global stitching graph; black lines denote the final edges after optimization; (**e**) example results of diverse scene stitching.

We use a undirected graph $G = (V, E)$ to represent the pairwise image relations of the unordered image set. The graph vertices Vdenote the aerial image, and graph edges E represent the relations of two aerial images. If an image pair $< I_i, I_j >$ is verified as overlapping images, there exists an edge between vertices V_i and V_j. For example, a graph is visualized in Figure 2c, in which graph vertices V represent aerial images, and graph edges E are shown as a black edge between overlapping images.

A similar graph-based framework [23] can be found in the field of remote sensing, but with quite different applications. In this paper, the key objective of our hybrid stitching is to explore edges in the graph completely and efficiently for stitching.

2.2. Sequential Grouping

In this section, we will introduce the method to generate initial small groups or sequential subgraphs. As mentioned above, we do not expect to use sequence information to handle all challenges of the VIRAT aerial video, because many kinds of disturbances, such as sudden viewpoint changes and large motion blur, can cause failure of sequential stitching. Moreover, sequential methods cannot deal with repeated observation of the same scene at different times.

However, we observe that sequential information often exists in aerial video sequence. Even if not providing a complete global picture of the entire surveillance scene, small sequential fragments are very valuable for further cross-group retrieval.

Standard optical flow tracker [13,14] or wide baseline matching methods [15,16] can be used to get frame-to-frame alignment between continuous image pairs. In this work, we select scale-invariant feature transform (SIFT) [15] features for the following two reasons: First, compared against other local invariant features, SIFT is proven to have the most robust features under geometric and illumination changes. Second, the extracted SIFT features are also used for further feature indexing and retrieval.

Given a set of VIRAT images $I = \{I_1, I_2, ..., I_n\}$ (as shown in Figure 2a), we firstly extract the SIFT features $F = \{f_1, f_2, ..., f_m\}$ from the entire image set. Then, for each continuous image pair, we apply SIFT matching and outlier removal [24] to check if they are overlapped. After that, new edges are added into $G = (V, E)$ between overlapped image vertices. Finally, the original isolate notes are classified as many small sequential subgraphs $\{G_1^s, G_2^s, ..., G_w^s\}$ (As shown in Figure 2(b) with the same color).

With the sequential stitching, we can quickly set up the relationships between successive overlapping images. More importantly, we can transfer the problem from complicated image-to-image matching into group-to-group retrieval, which means that even if only one connected edge between two sequential groups is found, the big panorama will be generated completely without breaks. Thus, we can significantly improve the probability of finding all overlapping images, meanwhile reducing the most time-consuming part of image stitching, which is feature matching between candidate image pairs.

2.3. Cross-Group Retrieval

In this section, we propose a novel cross-group retrieval method to find an optimal edge between candidate groups quickly. Our method contains two steps: (1) feature indexing; and (2) greedy searching-based optimal edge selection.

• Feature Indexing

In the first step, the hierarchical K-means tree (HKT) is adopted to find the overlapping image pairs from the unordered image set. HKT has proven to be powerful in image recognition, image classification and retrieval [25], demonstrating that the searching hierarchical tree can speed the matching of high-dimensional vectors by up to several orders of magnitude compared to a linear search.

Given a set of SIFT features $F = \{f_1, f_2, ..., f_m\}$ from the entire image set, HKT is constructed by splitting the SIFT feature points at each level into k distinct clusters using a k-means clustering. We apply the same method recursively to the points in each cluster. The recursion will be stopped when the number of feature points in a cluster is smaller than k. Finally, we save the index correspondences between the feature points and images in a look-up table for efficient retrieval. In all experiments, the k used for the hierarchical k-means tree is 32; the maximum number of iterations to use in the k-means clustering stage when building the k-means tree is 100; and we pick the initial cluster centers randomly when performing k-means clustering.

• Greedy Searching-Based Optimal Edge Selection

To find the relationships between two groups, we define a $n \times n$ accumulator matrix A, which is initialized by zeros. For each query feature point f in image I_i, we search for the HKT to find k closest points in the high-dimensional feature space, then we check the look-up table to find the index set $\ell = \{\ell_1, \ell_2, ..., \ell_t\}$ of corresponding images and increase the accumulator matrix $A_{i,c}$ by one, where $c \in \ell$. After querying all feature points, image pairs $< I_i, I_j >$ with sufficient high matching times as $A_{i,j} \geq \frac{1}{n}\sum_{c=1}^{n} A_{i,c}$ will be labeled as candidate edge (as shown in Figure 2c, red dotted lines).

Given two sequential subgraphs G_α^s and G_β^s, we use the following greedy searching to find the optimal edge Θ efficiently. For all vertices $V_i \in G_\alpha^s$ and $V_j \in G_\beta^s$, we firstly select the candidate image pairs Θ_0 with maximal retrieval score $A_{i,j}$ by Equation (1), and then, we use feature matching and random sample consensus (RANSAC) [24] to verify the edge between V_i and V_j.

$$\Theta_0 = \arg\max_{<i,j>}\{A_{i,j} | V_i \in G_\alpha^s, V_j \in G_\beta^s, A_{i,j} \geq \frac{1}{n}\sum_{c=1}^{n} A_{i,c}\} \quad (1)$$

If sufficient inliers have been found, we keep this edge and remove all other retrieval edges between sequential subgraphs G_α^s and G_β^s. Otherwise, we select and check the second optimal edge from the rest of the candidate edges between the two subgraphs. This process is repeated until all existing candidate edges between two groups have been checked. Actually, in our experiments, we find that image pairs connected by the first selected optimal edge usually contains sufficient correct feature correspondences. As a result, our approach only needs to apply one feature matching to connect two sequential groups in most cases. This characteristic is essential for reducing the heavy computational cost of feature matching on redundant edges.

Here, we make a brief discussion about the relationship and advantage of cross-group retrieval compared to traditional single image retrieval.

For a given retrieval sampling with one image, we use S_w to represent the probability of finding correct overlapping images from the entire dataset. To extract panorama with enough confidence from a large-scale aerial dataset, the probability value should be as high as possible. However, due to the challenges of realistic aerial video with similar background or low texture regions, the probability S_w of a traditional single image retrieval is usually low.

Given S_w, the probability of retrieval failure for one image is $1 - S_w$. For a giving image group G^s with λ sample images, the probability of all retrieval results being wrong is $(1 - S_w)^\lambda$. Finally, the probability that at least one sample image of group G^s can find correct overlapping images is:

$$P_{group\ confidence} = 1 - (1 - S_w)^\lambda \tag{2}$$

Thus, through integrating the sequential grouping and crossing group retrieval in one framework, we can greatly increase the probability to find panoramas completely by Equation (2).

2.4. Graph-Based Global Panorama Rendering

After finding the cross-group edges, we have generated a undirected graph $G = (V, E)$ to represent the pairwise image relations of the dataset. With this undirected graph, it is convenient to identify image relations that have not been explicitly established by the pairwise image matching.

Since the number of panoramas is unknown in the original image set, we have to extract all of the connected subgraphs in the graph firstly and then compute the image transforms inside each component, respectively.

A connected component of a undirected graph is a subgraph in which any two vertices are connected to each other by paths. In this paper, we use the depth-first search [26] to compute the connected components of a graph in linear time. A search begins at a particular vertex V_j, and each new vertex reached is marked. When no more vertices can be reached along edges from marked vertices, a connected component has been found. An unmarked vertex is then selected, and the process is repeated until the entire graph is explored. For each connected component, we need to find a homography $H_{r,j}$ between reference image vertices V_r and other vertices $V_j, j = 1, ..., l$. In this work, we pick the image vertices with the maximal number of connected edges as the reference vertices V_r. Although the $H_{r,j}$ may be calculated by chaining together the homography on any path between vertices V_r and V_j, to reduce the accumulation error of long chains, we find a shortest path from V_j to V_r with the Dijkstra algorithm (as shown in Figure 2d, black solid lines).

Finally, images of the same group are warped together with the homography model, which is estimated by the shortest path in the previous graph. This strategy is very quick, but may retain seams in the scene with large depth variation, such as buildings or trees. In these cases, we can employ graph cuts [27] to minimize the total squared difference between two images, and the chosen seam is through those pixels where the image colors are similar. The pseudocode of the proposed hybrid stitching approach is shown in Algorithm 1.

Algorithm 1 Pseudocode for diverse scene stitching.

Input:

The large-scale aerial video dataset.

Algorithm:

Build an undirected graph $G = (V, E)$.
Extract SIFT features of all input images.
Generate sequential groups $\{G_1^s, ..., G_w^s\}$ by matching continuous images.
Build an HTK tree with all SIFT features.
Retrieve edges for each image vertices.
for each sequential group pairs $\{G_i^s, G_j^s\}_{i \neq j}$ **do**
    Select optimal edge (u, v) by Equation (1).
    Match candidate image pairs with SIFT features.
    Remove outliers with RANSAC.
    Estimate homography with correct inliers.
    If (u, v) is a connected edge with enough inliers, remove all other edges between G_i^s and G_j^s, and compare the next group pairs.
    Otherwise, remove (u, v), and repeat Step 7 until all existing edges between G_i^s and G_j^s have been checked.
end for
Extract all connected subgraphs $\{G_1^c, ..., G_h^c\}$ by depth-first search in global group G.
for each group $g \in \{G_1^c, ..., G_h^c\}$ **do**
    for each image vertices V_j of group g **do**
        Find the shortest path between image vertices V_j and reference image vertices V_r.
        Warp corresponding image I_j by homography on the shortest path.
        Seam cutting and stitching between downsampled warped image and previous panorama.
    end for
end for
Output complete panorama image set $\{P_1, P_2, ..., P_\tau\}$.

3. Experiments

In this section, we will present a comparative performance evaluation of the proposed method.

• Dataset

We use VIRAT benchmark aerial video dataset [17] with huge diversity in the scene. The VIRAT dataset contains 24 videos, which are selected from 25-h original videos recorded at various times with a CCD and IR camera on an aerial platform. Each selected video is 5 min long with an image size of 640 × 480. The VIRAT dataset is highly realistic, natural and challenging for aerial video stitching (as shown in Figure 1b,c). Given the input aerial images, we label images from the same scene and take the total number of image groups at the ground truth for evaluation (as shown in Table 1, third column). For instance, if the input aerial images are captured from two scenes, then the ground truth will be set as the scene Number 2 in this case.

• Implementation Details

We implement our algorithm with C++. The experiments in this paper are all performed on a laptop computer with Intel i5 1.6 GHz CPU and 12 G RAM. For sequential grouping, we apply a distinctive SIFT [15] descriptor to get the feature correspondence between images. For cross-group

retrieval, we build the hierarchical K-means tree [28] for feature indexing. The only input parameter of our method is the total number of correct feature correspondences, which determines whether a valid homography transformation can be found between two images or not. This parameter is set as 20 in all of our experiments.

• Evaluation Metrics

We use the standard ground truth (GT), true positive (TP), false positive (FP) and processing time to evaluate the robustness and efficiency of different algorithms.

Ground truth scene numbers are provided by our manually labeling results. For each output panorama, we define two criteria to identify whether it is a true positive or false positive. One is the label consistency, and another is scene integrity. Label consistency means all images in one panorama must have the same scene label. Scene integrity describes the matching ratio between the output panorama and the corresponding scene, which can be calculated by the image number of output panorama divided by the total image number of the corresponding scene. Only the panorama whose image label is consistent and the scene integrity is higher than 90% can be identified a true positive panorama. Otherwise, it is a false positive. Finally, the processing time is the total time from loading images to stitching all panorama results in our experiments.

• Quantitative Comparison Results

First, we compare our method to the sequential stitching and retrieval stitching. We choose them as the baseline results, since the sequential stitching model is widely used for aerial video stitching, and retrieval stitching shares the similar goal of finding overlapping images efficiently without brute-force matching. For a fair comparison, we use standard SIFT [15] for sequential stitching and combine SIFT [15] and bag-of-words (BOW) [6,29] for retrieval stitching. In this experiment, we extract 3 frames every second from VIRAT video. Thus, there are 932 images in each video.

Figure 3a shows an example of sequential stitching. The sequential stitching methods are based on the assumption that the input aerial images is continuous in time, and they usually require the input data to be taken on a stable platform without sudden changes between consecutive frames. However, in the realistic VIRAT aerial dataset, challenges, such as motion blur, sudden scene or view point changes, are common in the aerial images, and the continuity assumption of the sequential stitching method hence is not valid in this case. As a result, although SIFT is more robust for wide baseline image matching, it still cannot handle those image changes between consecutive frames. The panorama of sequential stitching is split into 4 fragments (as shown in Figure 3a). In contrast, our approach recovers the panorama completely (as shown in Figure 3b).

Figure 4 shows an example of retrieval stitching and our approach. The retrieval stitching method is based on feature indexing, and it requires sufficient distinctive image features. If the scene lacks enough unique features, the variance of different images is not enough for the method to reliably find correct matches. In Figure 4a, in the low texture and self-similar grassland scene, retrieval stitching outputs 12 small scene fragments from 3 scenes. Through exploring the sequential groups firstly, our method

significantly reduces the risk of retrieval failure by cross-group retrieval and completely generates three panoramas (as shown in Figure 4b).

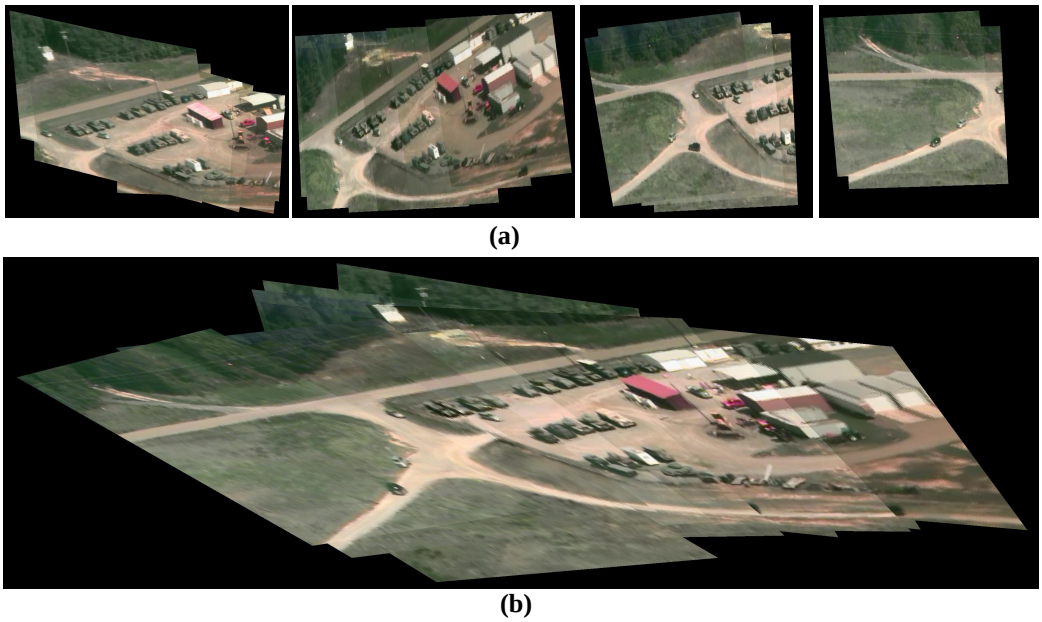

Figure 3. Example of sequential stitching *vs*. hybrid stitching. (**a**) Sequential stitching results over time with 4 panorama patches from 1 scene. (**b**) Our hybrid stitching result with a complete scene panorama

Further, we see that retrieval stitching achieves similar TP and FP as hybrid stitching only in two videos (Table 1, VIRAT 02 and 10). In all of the other 22 videos, hybrid stitching outperforms retrieval stitching in all metrics. In addition, the processing time of hybrid stitching is only half of retrieval stitching, and this advantage becomes more apparent as the increase of continuity inside the video (Table 1, VIRAT 10 and 17).

Figure 5 compares the statistic performance of 24 VIRAT videos. Our hybrid stitching achieves the highest TP and lowest FP, and its processing speed is even comparable with the fastest sequential stitching approach. On average, our method only costs 333.4 s to find 96% panoramas with 2 false positives. The significant improvements over retrieval stitching show the utility of combing temporal sequential grouping and cross-group retrieval in one hybrid stitching framework.

The previous experiments show two sample results between sequential stitching, retrieval stitching and our approach. Next, we compare the three methods on all 24 VIRAT videos.

Table 1 shows the quantitative comparison results on 24 VIRAT videos. The first column is the video indexing; the second column shows the image number of each VIRAT video. The manually-labeled ground truth (GT) is shown in the third column. The TP, FP and processing time of sequential stitching, retrieval stitching and our hybrid stitching are shown in the other columns.

Although the processing speed of sequential stitching is the the fastest, it is easily broken by diverse changes of aerial videos. As a result, we observe that both retrieval stitching and hybrid stitching significantly outperform sequential stitching in every video with higher TP and lower FP.

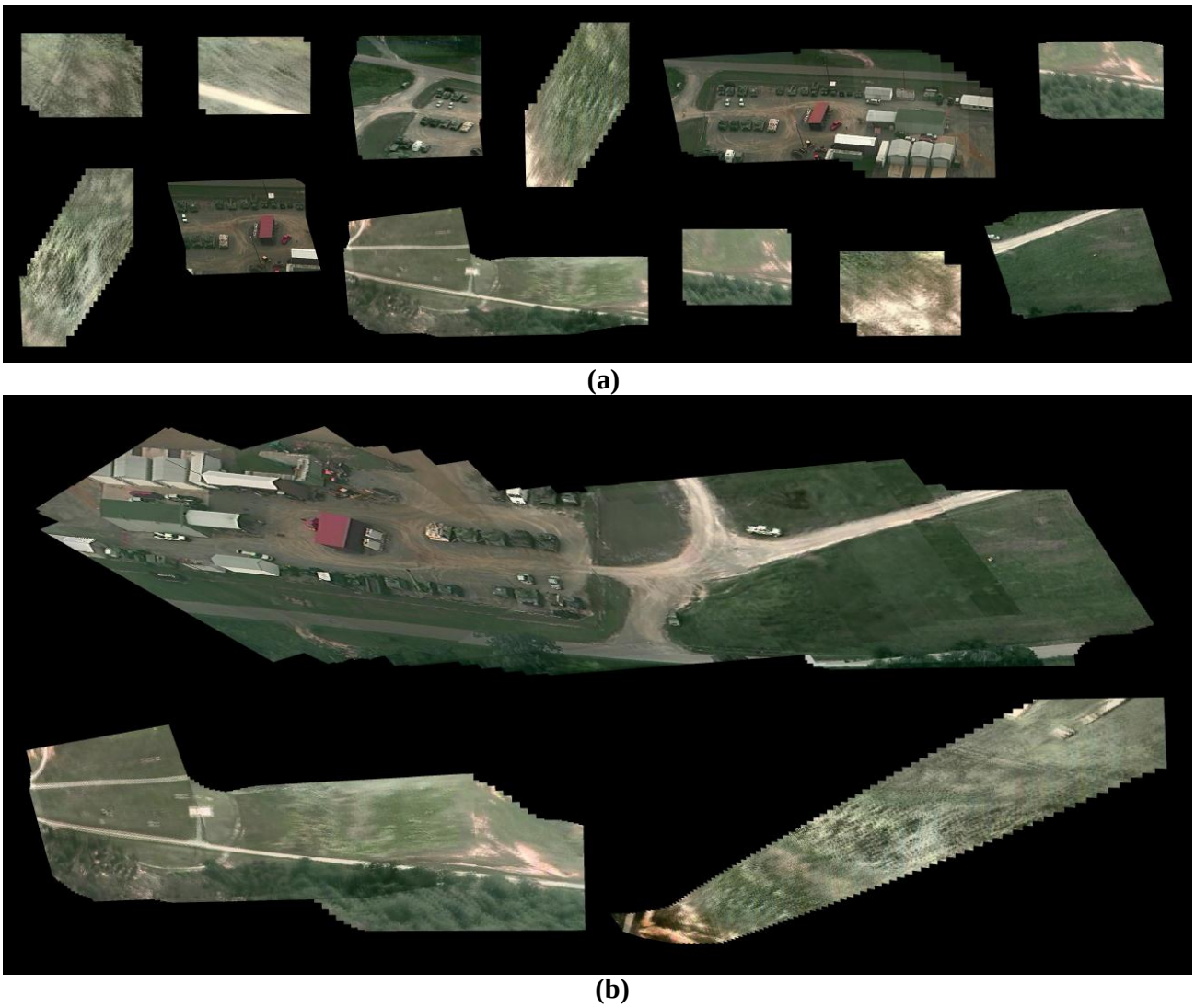

Figure 4. Example of retrieval stitching *vs.* hybrid stitching. (**a**) Retrieval stitching results with 12 panorama patches from 3 scenes. (**b**) Our hybrid stitching results with complete panorama of 3 scenes.

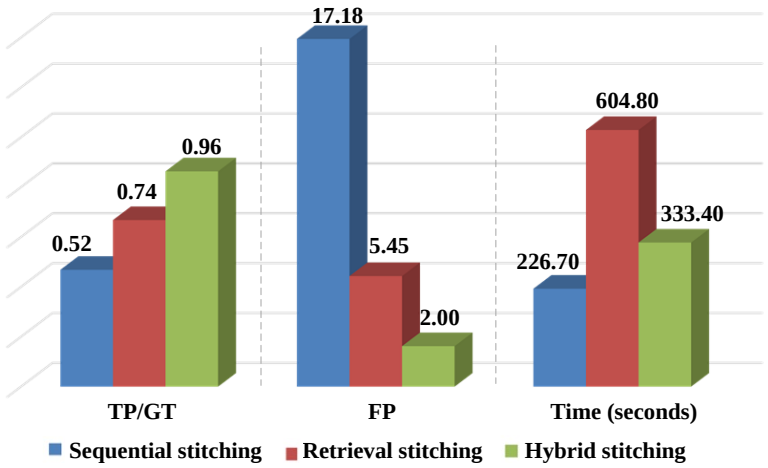

Figure 5. Comparison results of average performance on 24 VIRAT videos.

Table 1. Quantitative comparison of Sequential Stitching, Retrieval Stitching and Hybrid Stitching on 24 VIRAR videos.

Data set	Number of Images	GT	Sequential Stitching			Retrieval Stitching			Hybrid Stitching		
			TP	FP	Total Time(s)	TP	FP	Total Time(s)	TP	FP	Total Time(s)
VIRAT#01	932	17	6	17	247.6	12	4	780.1	17	0	365.9
VIRAT#02	932	15	8	21	264.4	14	1	946.2	14	1	425.8
VIRAT#03	932	20	7	25	269.4	11	9	668	19	1	416.3
VIRAT#04	932	25	9	25	196.3	16	2	613.8	25	0	323.4
VIRAT#05	932	20	10	22	249.4	16	4	743.3	19	2	371.8
VIRAT#06	932	23	16	14	208.9	21	1	546.4	22	1	337.5
VIRAT#07	932	19	6	27	250.3	14	7	585.4	16	6	370
VIRAT#08	932	19	9	19	241.2	17	3	623.6	18	2	312.7
VIRAT#09	932	18	9	19	209.3	12	4	590.7	18	3	320.4
VIRAT#10	932	2	1	10	224.2	2	0	722.7	2	0	302.0
VIRAT#11	932	13	12	5	222.4	10	3	774.0	13	1	336.0
VIRAT#12	932	22	8	29	253.2	14	13	701.7	21	2	440.8
VIRAT#13	932	10	3	25	238.5	7	9	668	9	3	253.2
VIRAT#14	932	12	6	14	240.0	10	3	624.9	12	1	306.3
VIRAT#15	932	14	6	17	234.8	10	7	540.3	14	1	311.2
VIRAT#16	932	19	7	16	216.6	14	1	485.3	18	2	342.5
VIRAT#17	932	9	6	2	196.7	7	3	461	9	0	228.5
VIRAT#18	932	14	10	16	223.4	9	10	526.4	13	5	281.6
VIRAT#19	932	11	8	14	176.6	9	6	382.7	10	3	211.4
VIRAT#20	932	18	13	10	198.7	13	7	490.7	18	3	349.5
VIRAT#21	932	12	8	12	228.4	9	4	460	10	2	392.7
VIRAT#22	932	9	5	14	231.0	4	11	481.2	9	4	297.6
VIRAT#23	932	13	8	12	192.2	11	3	547.9	13	0	316.8
VIRAT#24	932	16	6	27	227.9	7	16	550.9	16	5	389.1

• Qualitative Comparison Results

Second, we compare our system with state-of-the-art stitching systems, including PTGui [21], AutoStitch [19] and the very recent scheme by Autopano [22]. In order to compare the performance on the entire VIRAT dataset, we extract one image of every two seconds from the 24 VIRAT videos, and the total number of images is 2312.

Figure 6 shows the results of the state-of-the-art systems. Due to the lack of the ability for scene recognition, PTGui cannot automatically find overlapping image groups from the input video. It costs 2 h, 18 min and 52 s to generate one false panorama with enormous artifacts (as shown in Figure 6). Although AutoStitch [19] can recognize a scene from a small-scale image dataset, it cannot generate a panorama even after 10 h for the large-scale dataset of this experiment. The most recent Autopano [22] is the only system that can generate multiple panoramas from the large-scale input data. However, it costs over 10 h to create 38 panoramas, in which 18 panoramas contain significant artifacts (samples are shown in Figure 6, Autopano 1–5), and only 20 panoramas have reasonable visual effects (samples are shown in Figure 6, Autopano 6–8).

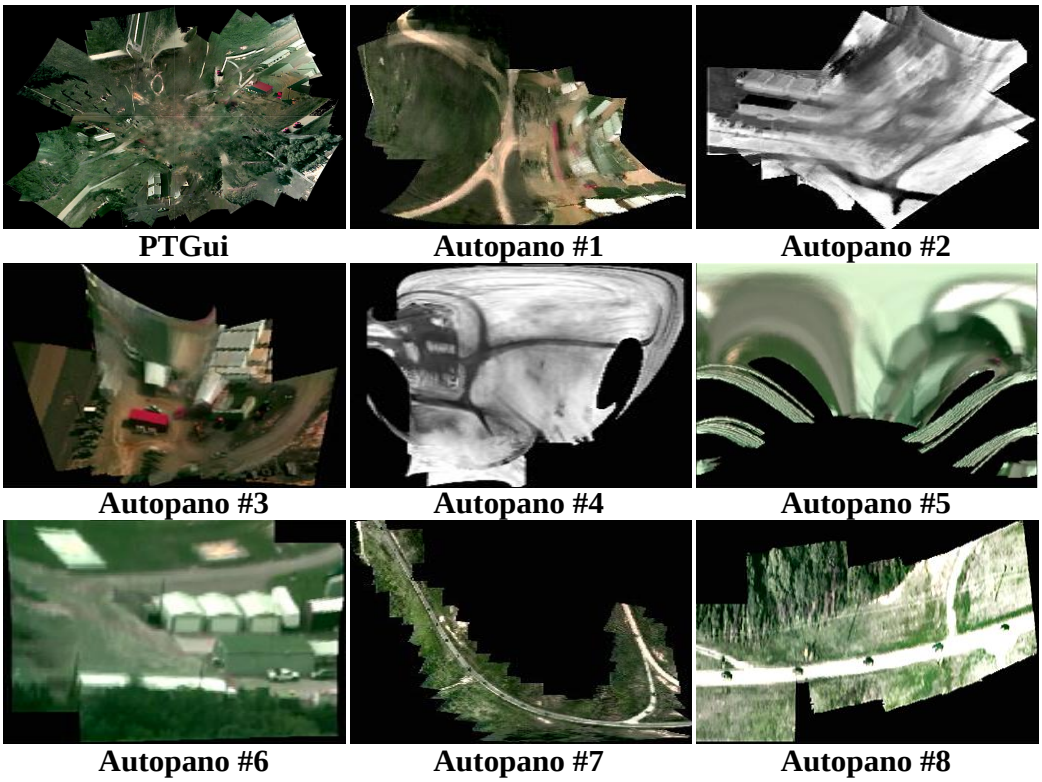

Figure 6. Example results of PTGui [21] and Autopano [22] from the VIRAT dataset with 2312 images.

In contrast, our approach costs only 15 min and 33 s to output panoramas of 48 scenes (as shown in Figure 7), which is much faster than the above state-of-the-art systems. Specifically, our approach firstly spends 136.112 s for feature detection and description and 160.049 s for feature indexing. Then, it uses 389.115 s for sequential grouping and cross-group retrieval. Finally, it costs 0.625 s for finding optimal paths in the entire graph, and 247.238 s for panorama stitching. As can be seen in Figure 7, the quality of our panoramas is much better than the state-of-the-art systems with less artifacts.

We also show a sample stitching result in Figure 8, which illustrates the dynamic stitching process of the overall site of VIRAT dataset (as shown in Figure 7, white bounding box). Our approach successfully extracts 22 revisits of this scene and generates a dynamic panorama with images captured at various times and viewpoints. We believe this reorganized scene panorama is particularly helpful for long-term surveillance and high level scene understanding. Experiments demonstrate the superiority of the proposed method.

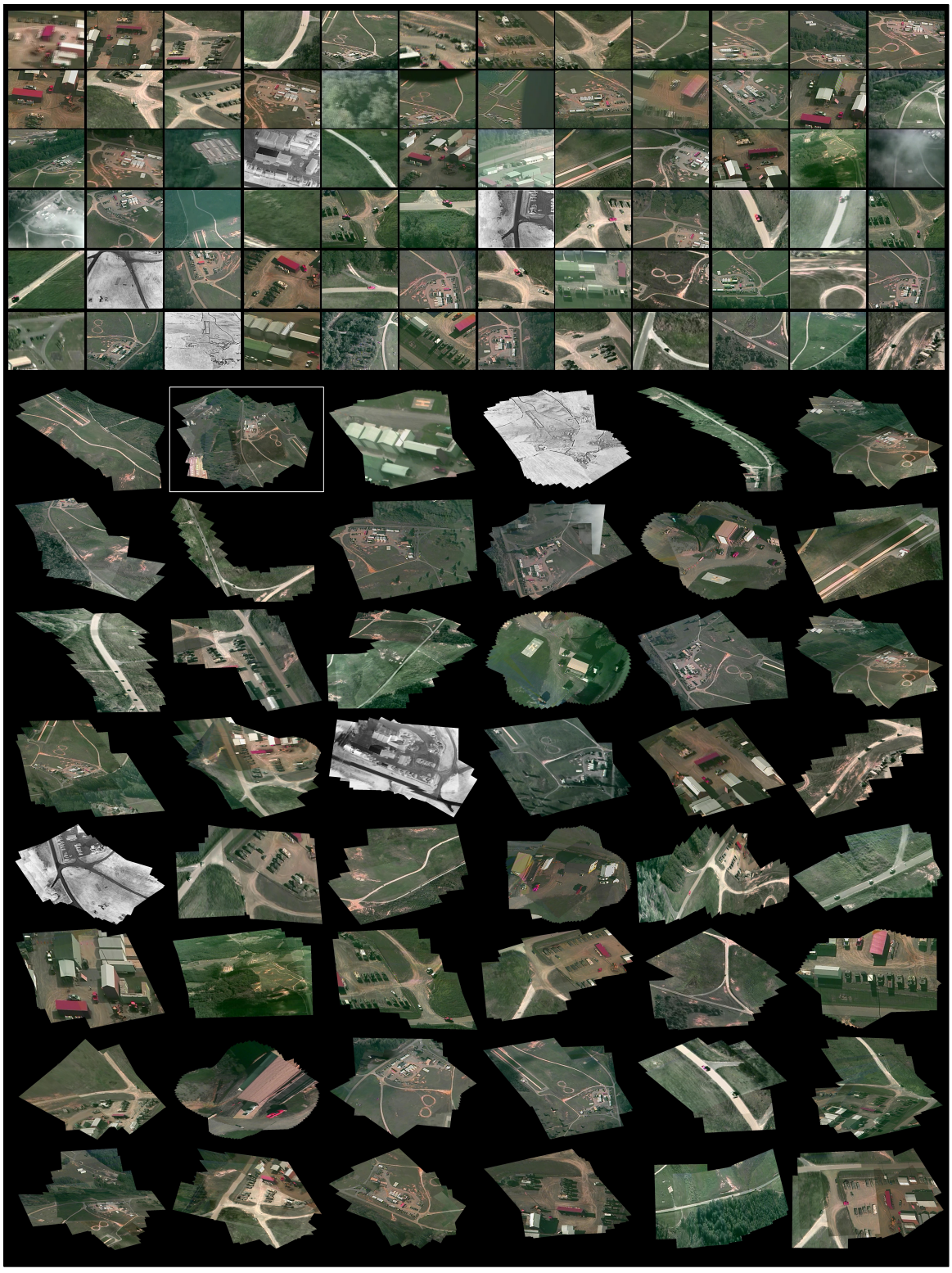

Figure 7. Our diverse scene stitching results from the VIRAT dataset with 2312 images. (Top) Examples of input VIRAT images. (Bottom) 48 panoramas of diverse scene by our approach after only 15 min and 33 s. The dynamic stitching results of panorama with the white bounding box are shown in Figure 8.

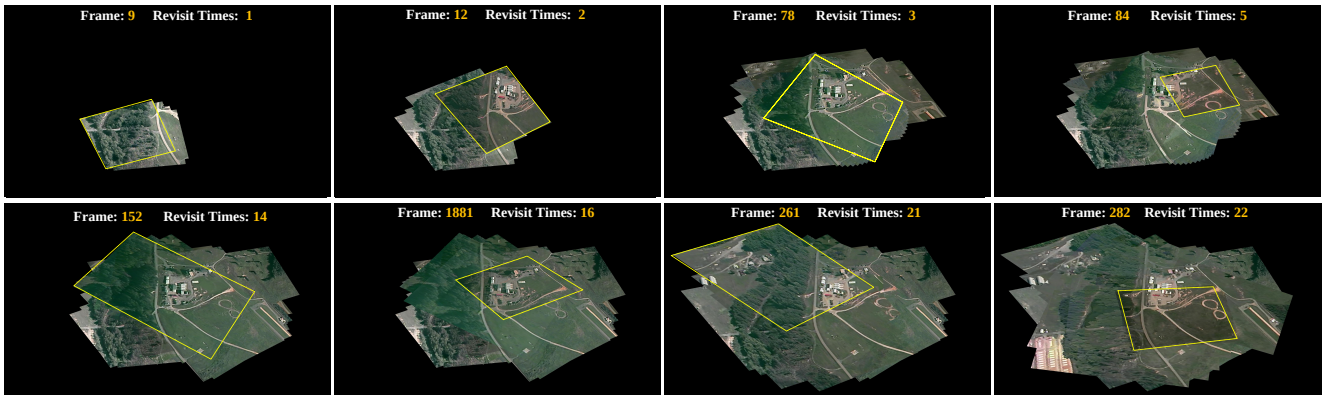

Figure 8. Example of dynamic stitching process of a surveillance scene with 22 revisits from 2312 VIRAT images; the yellow dotted line shows the first image from a new revisit.

4. Conclusions and Future Works

In this paper, we have proposed a powerful hybrid model for diverse scene stitching. The proposed method is based on the main findings that a large-scale aerial dataset has a close relationship to the human operators' monitoring behavior. For the same area of interest, human operators often apply continuous monitoring and repeated monitoring, which yield short, successive video clips that overlap in time or space.

Inspired by the temporal continuity and spatial repeatability of aerial surveillance video, our model integrates the sequential grouping and cross-group retrieval into a graph-based framework. We experiment with our method on the large-scale VIRAT aerial dataset [17], which is much more challenging than many other aerial datasets, due to its heterogeneity in image quality and diversity of the scene. To the best of our knowledge, this is the first stitching work on VIRAT. Experimental results show that our method can explicitly explore multiple panoramas from this challenging dataset. Moreover, our approach achieves a few orders of magnitude accelerations over the state-of-the-art stitching systems.

One limitation of the proposed method is that we use the homography model for vertical image stitching. Currently, most of the aerial images in our experiments are vertical images. However, For oblique aerial images, such as images taken parallel to the surface of the Earth, artifacts may appear with the homography model. For instance, for a three-dimensional building in an urban environment, the oblique images taken from different points of view are quite different. As a result, in order to generate a good panorama, highly accurate interior and exterior camera parameter extraction and 3D scene reconstruction may be required, which is more complicated compared with homography-based vertical image stitching. We would like to extend our work to diverse scene stitching with oblique aerial images in the future.

The current work is a pure image-based method for exploring diverse scenes from a large-scale aerial dataset. This image-based design makes our system very convenient to be used in many remote sensing tasks, even without any other sensor information. Considering that many aerial images may contain GPS/INS information, it is possible to embed the GPS information to further improve the robustness and efficiency of the diverse scene stitching. For instance, the GPS/INS information may be used to filter image pairs without overlapping area before feature matching or it can be used to split the entire aerial

images into several groups with a similar GPS location. We would also like to consider integrating the GPS/INS information in our future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61272288, No. 61231016), Northwestern Polytechnical University (NPU) New AoXiang Star (No. G2015KY0301), Fundamental Research Funds for the Central Universities (No.3102015AX007), Foundation of China Scholarship Council (No. 201206965020, No. 201303070083), NPU New People and Direction (No. 13GH014604) and NPU AoXiang Star (No. 12GH0311).

Author Contributions

Tao Yang and Jing Li designed the algorithm and wrote the source code and the manuscript together. Jingyi Yu provided suggestions on the algorithm and revised the entire manuscript. Sibing Wang and Yanning Zhang provided suggestions on the experiment and compared our work with other image stitching systems.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Kumar, R.; Sawhney, H.; Samarasekera, S.; Hsu, S.; Tao, H.; Guo, Y.L.; Hanna, K.; Pope, A.; Wildes, A.R.; Hirvonen, D.; Hansen, M.; Burt, P. Aerial video surveillance and exploitation. *Proc. IEEE* **2001**, *89*, 1518–1539.
2. Brown, M.; Lowe, D.G. Recognizing panoramas. In Proceedings of the IEEE Conference on International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1218–1225.
3. Agarwala, A.; Agrawala, M.; Cohen, M.; Salesin, D.; Szeliski, R. Photographing long scenes with multi-viewpoint. *ACM Trans. Graph.* **2006**, *25*, 853–861.
4. Brown, M.; Lowe, D.G. Automatic panoramic image stitching using invariant Features. *Int. J. Comput. Vis.* **2007**, *74*, 59–73.
5. Indelman, V.; Gurfil, P.; Rivlin, E.; Rotstein, H. Real-time mosaic-aided aerial navigation: I. motion estimation. In Proceedings of the AIAA Guidance, Navigation and Control Conference, Chicago, IL, USA, 10–13 August 2009; pp. 1–23.
6. Botterill, T.; Mills, S.; Green, R. Real-time aerial image mosaicing. In Proceedings of the 2010 25th International Conference of Image and Vision Computing New Zealand, IEEE, Queenstown, New Zealand, 8–9 November 2010; pp. 1–8.
7. Zaragoza, J.; Chin, T.J.; Brown, M.S.; Suter, D. As-projective-as-possible image stitching with moving DLT. In Proceedings of the IEEE Conferencce on Ccomputer Vision and Pattern Recognition, Portland OR, USA, 23–28 June 2013 ; pp. 2339–2346.

8. Li, J.; Yang, T.; Yu, J.Y.; Lu, Z.Y.; Lu, P.; Jia, X.; Chen, W.J. Fast aerial video stitching. *Int. J. Adv. Robot. Syst.* **2014**, *11*, doi:10.5772/59029.
9. Molina, E.; Zhu, Z.G. Persistent aerial video registration and fast multi-view mosaicing. *IEEE Trans. Image Proces.* **2014**, *23*, 2184–2192.
10. Zhang, F.; Liu, F. Parallax-tolerant Image Stitching. In Proceedings of the IEEE Conferencce on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3262–3269.
11. Szeliski, R. Image alignment and stitching: A tutorial. *Found. Trends Comput. Graph. Vis.* **2006**, *2*, 1–104.
12. Kekec, T.; Yildirim, A.; Unel, M. A new approach to real-time mosaicing of aerial images. *Robot. Auton. Syst.* **2014**, *62*, 1755–1767.
13. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In Proceedings of the International Joint Conference on ArtificialIntelligence, University of British Columbia Vancouver, BC, Canada, 24–28 August 1981; pp. 674–679.
14. Baker, S.; Matthews, I. Lucas-kanade 20 years on: A unifying framework. *Int. J. Comput. Vis.* **2004**, *56*, 221–255.
15. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
16. Bay, H.; Ess, A.; Tuytelaars, T.; Gool, L.V. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359.
17. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.C.; Lee, J.T; Mukherjee, S.; Aggarwal, J.K.; Lee, H.; Davis, L.; *et al.* A large-scale benchmark dataset for event recognition in surveillance video. In Proceedings of the Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3153–3160.
18. Schaffalitzky, F.; Zisserman, A. Multi-view matching for unordered image sets. In Proceedings of the European Conference on Computer Vision, Copenhagen, 27 May–2 June 2002; Springer: Berlin, Germany, 2002; pp. 414–431.
19. Brown, M. AutoStitch. Available online: http://www.cs.bath.ac.uk/brown/autostitch/autostitch.html (accessed on 25 May 2015).
20. Sibiryakov, A.; Bober, M. Graph-based multiple panorama extraction from unordered image sets. *Proc. SPIE* **2007**, doi:10.1117/12.704025.
21. New House Internet Services. PTGui Software. Available online: http://www.ptgui.com (accessed on 25 May 2015).
22. Kolor Company. Kolor autopano. Available online: http://www.kolor.com (accessed on 25 May 2015).
23. Kang, X.C.; Lin, X.G. Graph-based divide and conquer method for parallelizing spatial operations on vector data. *Remote Sens.* **2014**, *6*, 10107–10130.
24. Fischler, A.M.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395.
25. Muja, M.; Lowe, D.G. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In Proceedings of the International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, 5–8 February 2009; pp. 331–340.

26. Hopcroft, J.; Tarjan, R. Efficient algorithms for graph manipulation. *Commun. ACM* **1973**, *16*, 372–378.
27. Kwatra, V.; Schödl, A.; Essa, I.; Turk, G.; Bobick, A. Graphcut textures: Image and video synthesis using graph cuts. *ACM Trans. Graph.* **2003**, *22*, 277–286.
28. Muja, M.; Lowe, D.G. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2227–2240.
29. Nister, D.; Stewenius, H. Scalable recognition with a vocabulary tree. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2161–2168.