2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Efficient Covisibility-based Image Matching for Large-Scale SfM

Zhichao Ye Guofeng Zhang* Hujun Bao*
State Key Lab of CAD&CG, Zhejiang University

Abstract— **Obtaining accurate and sufficient feature matches is crucial for robust large-scale Structure-from-Motion. For unordered image collections, a traditional feature matching method with geometric verification requires a huge cost to find sufficient feature matches. Although several methods have been proposed to speed up this stage, none of them makes full use of existing matches. In this paper, we propose a novel efficient image matching method by using the transitivity of region covisibility. The overlapping image pairs can be efficiently found in an iterative matching strategy even only with few inlier feauture matches. The experimental results on unordered image datasets demonstrate that the proposed method is three times faster than the state-of-the-art and the matching result is high-quality enough for robust SfM.**

I. INTRODUCTION

Over the past decades, Structure-from-Motion (SfM) has made significant progress in both efficiency and robustness [1], [2], [3], [4]. Although different SfM systems may have different strategies [4], [5], [6], most of them can be divided into two common stages in essence. In the first stage, SfM systems find overlapping images through feature matching and geometric verification, and then generate scene graph [7], [8] with images as nodes and verified pairs of images as edges. In the second stage, the main work is to recover camera parameters and 3D points based on the inlier matches from the first stage.

As the camera registration and the point cloud reconstruction depends heavily on the results of matching stage, finding as many inlier feature matches as possible is crucial in SfM. A naive method is to use a brute-force matching strategy by testing every image pairs to find all feature associations. However, for a large dataset, especially the dataset collected from the Internet, the brute-force matching is intractable due to the $O(n^2)$ computation complexity.

The matching effectiveness can be boosted given the covisibility, which means that two images see common landmarks. A lot of unnecessary computation in uncovisibility image pairs can be omitted. For example, when dealing with an image sequence or a video stream, thanks to strong covisibility among consecutive frames, a simple sequential matching strategy can be used to match the consecutive frames, whose time complexity is linear to the number of frames. However, it is not straightforward to predict the overlapping image pairs in an unordered image set.

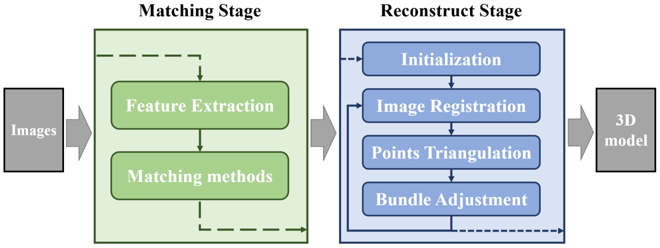

Fig. 1. A general pipeline of incremental SfM.

The previous work [1], [9] predict the covisibility by image similarity. However, the approximate is not reliable enough, resulting in useless geometric verification. By contrast, the verified feature matches are more trustworthy evidences of covisibility, especially when there are multiple feature tracks in a local region. In this paper, we propose a novel method to reliably construct an image covisibility graph based on the observation that the covisibility of regions has transitivity. In addition, to omit unnecessary matching, we propose an iterative matching strategy which can efficiently find overlapping pairs for unordered image datasets. We found that poor quality pictures were isolated, while most of the registered images are connected with others. According to the feature associations, we divide the images into two parts: images of possible registration and the others. Searching covisible image pairs from the former part can effectively reduce the matching with poor quality images. We show that the proposed matching method is 3 times faster than the state-of-the-art method and the achieved matching results are sufficient for accurate and complete 3D reconstruction.

II. RELATED WORK

SfM technique has achieved great success in the past decade [1], [10], [2], [3], [4]. As shown in Fig. 1, a general pipeline of increamental SfM contains two major stages: matching stage and reconstruction stage. The process of reconstruction stage can be divided into the following four steps:

1) Initializing camera poses and 3D points from a selected image pair.
2) Finding the most suitable image from the unregistered ones, and then recover the camera pose of this image with a number of newly triangulated 3D points.
3) If the registration is successful, extending the triangulated point set with the new inlier matches.

*Corresponding authors: {zhangguofeng, bao}@cad.zju.edu.cn
The authors are also affiliated with ZJU-SenseTime Joint Lab of 3D Vision. This work was partially supported by NSF of China (Nos. 61822310 and 61672457), and the Fundamental Research Funds for the Central Universities (No. 2019XZZX004-09).

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

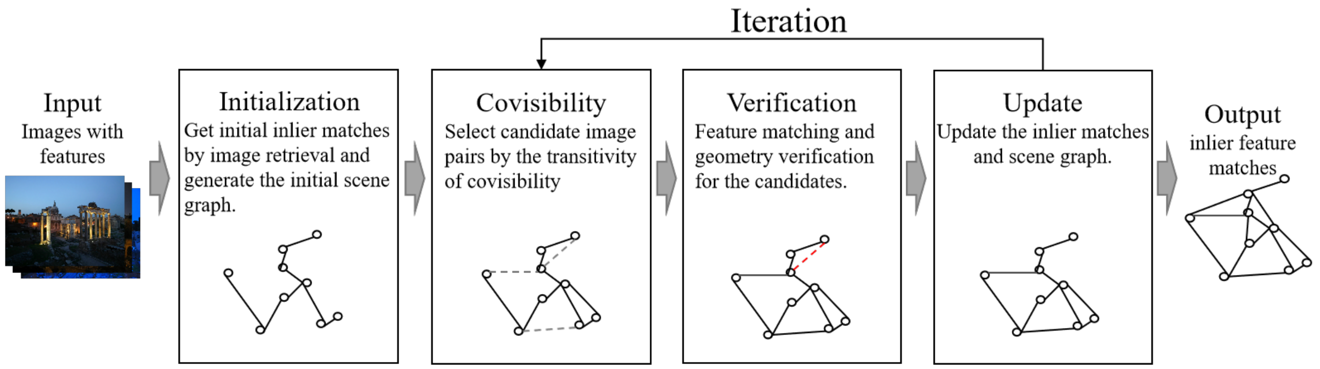
Iteration
Input
Images with features
Initialization
Get initial inlier matches by image retrieval and generate the initial scene graph.
Covisibility
Select candidate image pairs by the transitivity of covisibility
Verification
Feature matching and geometry verification for the candidates.
Update
Update the inlier matches and scene graph.
Output
inlier feature matches

Fig. 2. The pipeline of our iterative matching strategy

4) Optimizing camera poses and 3D points via bundle adjustment.

Generally, the reconstruction stage needs to iterate the above 2-4 steps until there is no more suitable frame to be registered.

While the above steps of the reconstruction stage are relatively fixed, there are a variety of speed-up matching techniques for unordered images in matching stage. Generally speaking, these methods can be basically summed up as the following two ways: reducing the number of image pair candidates or the cost of feature matching.

The most representative of the first kind of methods is image retrieval [11]. For each image, this method retrieves a fixed number N_R of most similar images to match features, so the time complexity of feature matching is $O(N_R * n)$, where n is the number of images. If we employ this method in the matching stage, the integrity of the reconstruction is affected by both N_R and the used image retrieval algorithm.

However, there are some inherent defects in retrieval methods with a fixed param N_R. Some images ought to have many potential matches, while some others only have few overlapping images. For the former, a small N_R leads to the lack of inlier matches that is difficult to guarantee the completeness of reconstruction. For the latter, a big N_R resulting in a computational waste of many false matching pairs.

Vocabulary tree [9] is a common image retrieval method, which is widely employed in various SfM systems and loop detection [12]. This method heavily relies on the pre-training of the vocabulary tree and the descriptors. Other kinds of image retrieval methods generate global image descriptors, such as GIST [13]. Recently, with the great success of depth learning in computer vision, image retrieval methods based on Convolutional Neural Networks (CNNs) [14] have emerged and they outperform traditional methods. Because the CNN based methods have a strong image representing ability, they are more robust to illumination and view angle changes.

Some researchers propose to extend the existing image retrieval results to improve robustness and enhance matching. A simple query expansion method [15] is to match the query results of neighbor frames. This method can find some extra inlier matches, but it costs a lot for images with rich matching relations. Another method is MatchMiner [16], which iteratively updates the weights of vocabularies through matching results of existing queries, distinguishes valuable vocabularies from noisy vocabularies to achieve good performance. Moreover, a vote-and-verify strategy [17] of vocabulary tree was proposed for fast spatial verification. However, all of these methods do not make full use of the feature correspondences of existing inlier matches.

Furthermore, many methods are aiming at reducing the cost of feature matching. The preemptive matching method [10] predicts image pairs with overlapping by matching a few representative features. Hartmann et al. [18] proposed to control the number of features to achieve sufficient but not too many matches for acceleration. VocMatch [19] proposed to use the vocabulary tree further. The features indexed to the same visual word are considered as potential matches to skip the descriptor matching. Alternatively, another method PAIGE [20] uses the change of feature location and rotation to predict the image pairs with overlapping.

Different from the previous methods, we propose to use the covisibility to predict overlapping images, which can significantly reduce the number of image pairs to be matched. Covisibility refers to the fact that two images observe the same landmarks. Covisibility graph is a graph that regards images as the nodes and covisibility image pairs as edges, which is widely used in 3D vision problems [21], [22]. In ORB-SLAM2 [21], feature inlier matches provide information to build a covisibility graph which simplifies data association and facilitates data maintenance. Some researchers [22] proposed another kind of covisibility graph constructed by vocabulary association to search correspondences among unmatched images to handle place recognition and loop closure. We can judge the covisibility is reliable if two pictures share enough feature points or vocabularies. However, in order to discover the overlapping image pairs as many as possible, a good algorithm needs to perform well with a small amount of support from the existing information.

III. METHODOLOGY

This section presents a new effective matching strategy for unordered image sets. We introduce the matching strategy in two parts in detail. First, we propose a region-based algorithm to find the potential overlapping image pairs. Second, we introduce an iterative matching strategy that efficiently extends inlier feature matches by mitigating unuseful matchings.

A. Transitivity of the covisibility

The matching stage searches image correspondences in the input images $U = \{I_i \mid i = 1...N_I\}$. The feature extraction part detects the features $F_i = \{f_i^k \mid k = 1...N_{F_i}\}$ where F_i denotes the set of features in image I_i, N_{F_i} is the number of features in F_i, and f_i^k denotes the k-th feature in the i-th image.

Features detected in images are the projections of 3D scene points. A correct feature match connects a feature pair which corresponde to the same 3D scene point. The features in a feature track $\{f_{i_1}^{k_1}, f_{i_2}^{k_2}, ..., f_{i_n}^{k_n}\}$ derived from successive correct feature matches $\{(f_{i_1}^{k_1}, f_{i_2}^{k_2}), ..., (f_{i_{n-1}}^{k_{n-1}}, f_{i_n}^{k_n})\}$ are belong to the same scene point. The image pairs sharing feature tracks can be considered to be covisible. Nevertheless, the projections of a 3D scene points sometimes are not detected so that some covisible image pairs are ignored due to missing matches. In addition, there are still some mismatches in the inlier feature matches even after the geometric verification, which makes the covisibility supported by few feature tracks unreliable.

We propose a co-visibility construction method to address the problems of both missing matches and mismatches. A 3D point is a part of a scene structure, so we extend the covisibility of features to the local regions where the features lie. The projections of a small 3D structure adjacent to a 3D point in an image pair are called a covisible region pair. As shown in Fig. 3, there is no feature track shared by r_{i1} and r_{i3} by missing matches, but the potential overlapping region pair $(r_{i1}^{k1}, r_{i3}^{k3})$ still can be found by the transitivity of covisibility. Moreover, the number of shared feature tracks measure the confidence of co-visibility, so that a single feature mismatch does not cause a false positive co-visible region pair. The covisibility of the region pairs, e.g., $(r_{i1}^{k1}, r_{i2}^{k2})$ and $(r_{i2}^{k2}, r_{i3}^{k3})$, are relatively reliable because there are multiple feature tracks. By contrast, the region pair, $(r_{i1}^{k1'}, r_{i2}^{k2})$, shares only one feature track which is identified as unreliable because they are potential mismatches.

It is hard to determine the boundary of the regions without the scene structure, so we uniformly divide each image I_i into N_p^2 patches $P_i = \{p_i^k | k = 1...N_p^2\}$ as an approximation. A covisible patch pair (p_i^{k1}, p_j^{k2}) is confirmed once there are at least T feature tracks shared by patch p_i^{k1} and p_j^{k2}, and all the covisible patch pairs make up a patch covisibility graph with patches as nodes and covisible patch pairs as edges. If a patch pair is connected in the patch covisibility graph, there is a potential covisibility of the patch pair. We transfer the covisibility of patches to the covisibility of images by the following equation:

$$dist(I_i, I_j) = \min\{dist(p_i^{k1}, p_j^{k2}) | k1 = 1...N_p^2, k2 = 1...N_p^2\}, \quad (1)$$

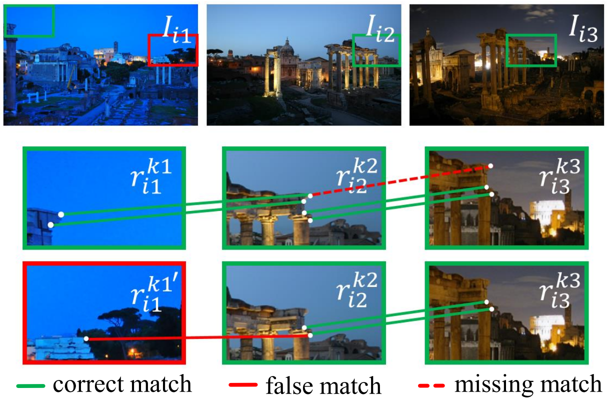

Fig. 3. The top row contains three covisible images; the second row shows the correct transitive covisibility $(r_{i1}^{k1}, r_{i3}^{k3})$; In the thrid row, $(r_{i1}^{k1'}, r_{i3}^{k3})$ shares a feature track, but it is not a covisible region pair.

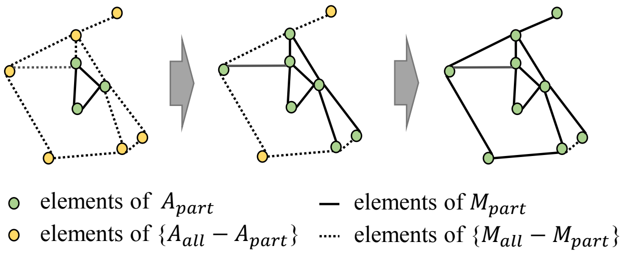
elements of A_{part}
elements of $\{A_{all} - A_{part}\}$
elements of M_{part}
elements of $\{M_{all} - M_{part}\}$

Fig. 4. Iteratively extending the registered images.

where $dist(p_i^{k1}, p_j^{k2})$ is the distance of p_i^{k1} and p_j^{k2} in the patch covisibility graph. If there is no path connecting the two patches, $dist(p_i^{k1}, p_j^{k2})$ is defined as infinite. $dist(I_i, I_j)$ is the minimum distance between patches in I_i and that in I_j. As a patch is an approximation to a region, the overlapping ratio of a patch pair may decrease with the distance at patch covisibility graph growing. We therefore build a covisibility relation between I_i and I_j by the following equation:

$$covisible(I_i, I_j) = \begin{cases} True & dist(I_i, I_j) \leq \sigma \\ False & dist(I_i, I_j) > \sigma \end{cases}, \quad (2)$$

where σ is the distance threshold. I_i and I_j are covisible if $dist(I_i, I_j)$ does not exceed the threshold σ.

B. Iterative matching strategy

We propose an iterative algorithm that extends the inlier feature matches step by step to make full use of existing matches, as illustrated in Fig. 2. Firstly, we verify high-ranking image pairs in the retrieval results to get the initial inlier matches represented by M_{init}. In each iteration, we select some image pairs as candidates according to the potential covisibility supported by the existing inlier feature matches. Then, we perform candidate verification, and add new inlier matches to update the possibility of covisibility. Finally, most of the inlier matches can be found. The key problem of the proposed algorithm is how to select the candidates.

For Internet photo collections, e.g. images collected via Flickr keyword search API, there are many poor quality pictures and irrelevant picutres in U, so some images still can not be registered successfully even if all the inlier feature matches M_{all} has been established. A_{reg} denotes the set of images which can be registered with the support of M_{all}. A_{rest} denotes the rest of the photos in U, and all the elements in A_{rest} can not be registered. The Brute-Force matching can find all the inlier feature matches M_{all}, but the matching procedures related to A_{rest} are usually noneffective. Given the registrable image set A_{reg}, we only need to verify image pairs constituted by the images in A_{reg}, which reduces the useless verifications. However, A_{reg} is not maintainable at the beginning.

In each iteration of the proposed method, we have partial inlier feature matches M_{part} (a subset of M_{all}), so partial images A_{part} can be registered. We only select the candidate image pairs composed of at least an element in A_{part}. Because A_{part} is a subset of A_{reg}, this strategy is very effective by avoiding the unnecessary verification for the image pairs composed of two elements of A_{rest}.

To get the set A_{part}, a naive method is running the reconstruction with existing inlier feature matches M_{inlier}, but it is quite time-consuming. We exploit a fast algorithm to approximate the registration process. Before diving into the details, we explain some definitions used in the Algorithm 1.

$$Tri(I_i, I_j) = \{f_i^a, f_j^b | (f_i^a, f_j^b) \in M_{inlier}\} \tag{3}$$

$$Match(I_i, S) = \{(f_i^a, f_j^b) | (f_i^a, f_j^b) \in M_{inlier}, f_j^b \in S\} \tag{4}$$

The matched features f_i^a and f_j^b makes up the set $Tri(I_i, I_j)$ for simulating the progress of triangulation in reconstruction. Given a feature set S, $Match(I_i, S)$ is the set of inlier feature matches between features of I_i and S.

A_{appr} is an approximation of A_{part}. As show in Algorithm 1, Our registration approximation algorithm includes an initial stage(step 1 to 4 in Algorithm 1) and an iteration extending stage(step 5 to 17 in Algorithm 1), which corresponds to the reconstruction stage of SfM. The SfM systems select the next candidates to register from the images whose triangulated points are larger than a certain threshold t_{reg}. It is obviously that A_{appr} is a subset of A_{part} when the threshold t is equal to t_{reg}.We use the (2) to select candidates C with A_{appr} in Algorithm 2.

To control the time of covisibility test, we select the covisible image pairs from the retrieval results with a big N_R (step 5 in Algorithm 2). $Retrieval(I_i, k)$ denotes the retrieval results of I_i with the retrieval number $N_R = k$. In our implementation, we use NetVLAD [23] to get retrieval results, but there is no limitation to use other image retrieval methods. If there are few initial inlier feature matches, the candidates are too few to be extended. We add some extra candidates by the votes from the retrieval results to alleviate the lack of initial connections. With more image matches to A_{part}, an image is more likely to be registered. Since image retrieval can be regarded as an approximation of image matching, an image whose retrieval result contains more images belonging to A_{part} is more likely to be registered. When we take A_{appr} as the approximation of A_{part}, the proportion of images belonging to A_{appr} in the retrieval results determines the possibility of registration. What's more, we limit I_j in $Retrieval(I_i, k)$. In the worst case, the proposed method will match each image with its k closest neighbors.

Algorithm 1: Registration approximation algorithm

```
   Input: inlier matches $M_{inlier}$, thresold $t$
   Output: $A_{appr}$
 1 $A_{appr} = \emptyset$, $S = \emptyset$;
 2 Select a matched image pair $(I_i, I_j)$
 3 $A_{appr} = A_{appr} \cup \{I_i, I_j\}$;
 4 $S = S \cup Tri(I_i, I_j)$;
 5 $finded \leftarrow True$;
 6 while $finded$ do
 7     $finded \leftarrow False$;
 8     for $I_i \in U - A_{appr}$ do
 9         if $|Match(I_i, S)| > t$ then
10             $A_{appr} \leftarrow A_{appr} \cup \{I_i\}$;
11             $finded \leftarrow True$;
12             for $I_j \in A_{appr}$ do
13                 $S \leftarrow S \cup Tri(I_i, I_j)$;
14             end
15         end
16     end
17 end
```

Algorithm 2: Iterative matching strategy

```
   Input: initial inlier matches $M_{init}$, retrieval param $k$
   Output: inlier matches $M_{inlier}$
 1 $M_{inlier} = M_{init}$;
 2 $C = \emptyset$;
 3 Compute $A_{appr}$ from $M_{inlier}$ by Algorithm 1
 4 for $I_i \in A_{appr}$ do
 5     for $I_j \in Retrieval(I_i, k)$ do
 6         if $covisible(I_i, I_j)$ then
 7             $C \leftarrow C \cup (I_i, I_j)$;
 8         end
 9     end
10 end
11 Verify the candidate pairs in $C$ and update $M_{inlier}$
12 Repeat from line 2 until there are suitable candidates
     or the maximum number of iterations is reached.
```

IV. EXPERIMENTS

We conduct the experiments on 14 large unordered datasets [24] to evaluate both our methods and the state-of-the-art image retrieval method NetVLAD [23]. These 14 datasets contain a total of 58k unordered Internet photos, covering a wide variety of scenes. We use the same features

TABLE I

Evaluation results on 14 large-scale unordered Internet photo collections

	#Size	#Registered				#Time[s]				#Precision[%]			
		IR_5	IR_{25}	IR_{50}	$Ours$	IR_5	IR_{25}	IR_{50}	$Ours$	IR_5	IR_{25}	IR_{50}	$Ours$
Alamo	2,915	683	810	862	760	144	722	1405	233	40.74	27.61	23.68	37.29
Ellis Island	2,587	295	344	351	331	104	584	988	177	49.25	33.27	26.94	52.09
Gendarmenmarkt	1,463	702	984	1020	923	62	346	596	230	59.16	46.25	39.86	52.09
Madrid Metropolis	1,344	245	409	435	406	47	244	430	113	42.31	27.76	23.1	37.35
Montreal Notre Dame	2,298	475	554	564	552	99	523	972	171	54.79	41.97	36.31	48.54
NYC Library	2,550	385	614	574	592	102	544	975	170	44.32	28.60	21.74	43.58
Piazza del Popolo	2,251	332	901	951	865	99	468	872	234	47.10	34.15	28.53	43.23
Piccadilly	7,351	2213	2871	2988	2838	406	1508	2717	995	40.95	29.17	24.73	40.63
Roman Forum	2,364	1291	1500	1599	1546	164	587	1226	473	59.45	43.15	35.57	33.87
Tower of London	1,576	477	651	699	632	84	386	732	199	42.65	28.08	22.41	35.58
Trafalgar	15,685	4397	7048	7725	7122	713	3474	6396	2819	41.41	30.76	26.54	37.98
Union Square	5,961	536	985	1070	971	311	1436	2313	449	22.63	13.48	10.29	30.07
Vienna Cathedral	6,288	924	1060	1119	1033	533	1657	3328	707	40.25	20.69	20.69	39.28
Yorkminster	3,368	452	655	1060	927	165	1182	1620	382	48.79	32.09	24.89	37.16
Average	4,142	957	1384	1501	1399	216	975	1754	525	45.27	31.22	26.09	40.62

and the same implementation in both feature matching and geometric verification supported by the open-source project COLMAP [4], which shows the highly improved efficiency of our proposed matching method. All the experiments are conducted on a desktop PC with an Intel i7-9700K 3.6GHz CPU, 64GB of memory and a NVidia GTX 2070 graphic card. In this section, we first exhibit the comparison experiment to verify the superiority of our proposed method and then present the ablation studies to examine the effectiveness of each component.

We set the NetVLAD as our baseline, which retrieves N_R images as IR_{N_R}. For NetVLAD, higher N_R offers more registered images but requires more computation time. $Registered$ is the number of registered frames, $Time$ denotes the time consumed in candidate verification, and $Precision$ is computed from N_T/N_{All} (N_T is the number of found overlapping image pair and N_{All} is the number of the candidate image pairs). For comparison, we show these three metrics on both NetVLAD and our method in Table I. The NetVLAD is tested with IR_5, IR_{25}, and IR_{50}. The M_{init} in Algorithm 2 of our proposed method is set as the result of IR_5 and the k is 50. It shows that IR_5 is the fastest method, but can not guarantee the integrity of reconstruction The time of our method is the second fastest because ours takes the result of IR_5 as the initial inlier matches. Comparing with the experiments on NetVLAD that use larger param N_R, our method is much faster than IR_{25} and IR_{50}, and the number of registered frames of our method approaches to that of these two implementations. In addition, the precision of our method (verified image pairs) is higher than IR_{25} and IR_{50} on average, which means that our method can accurately predict overlapping images.

In order to reflect the advantages of our method more intuitively, we list the total time and the sum of registered number of Table I. As shown in Fig. 5, the speed of our method is much faster than the NetVLAD with the premise of complete reconstruction. Fig 6 shows the reconstruction results of the different methods in Madrid Metropolis (the first row), Union Square (the second row). Compared with the reconstruction result of IR_5, the reconstruction result of our method is more complete without the absence of walls.

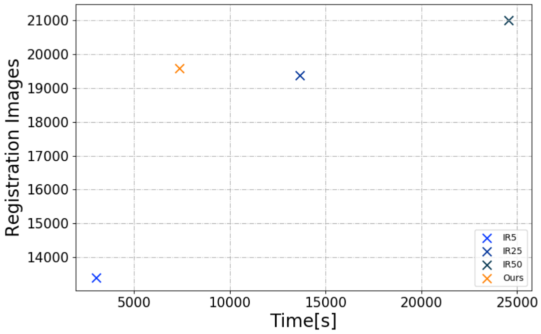

Fig. 5. Visualization of the overall evaluation results

Generally speaking, our method is quite robust and mines covisible image pairs well on these datasets.

A. Transitivity of the covisibility

TABLE II

The results of comparison experiments with different N_p

	#Registered	#Time	#Precision
$N_p = 5, T = 2$	953	439	0.1947
$N_p = 10, T = 2$	943	405	0.2233
$N_p = 20, T = 2$	927	382	0.2360

N_p controls the number of patches the image divided into and the threshold T denotes the minimum sharing tracks between a covisible patch pair. To show the effects of the parameters N_p and T, we present the ablation studies on Yorkminster in Table II and Table III. It is noteworthy that the $Precision$ in these two tables removes N_T and N_{all}

TABLE III

The results of comparison experiments with different T

	#Registered	#Time	#Precision
$N_p = 5, T = 1$	962	486	0.1716
$N_p = 5, T = 2$	953	440	0.1947
$N_p = 5, T = 4$	921	372	0.2407
$N_p = 5, T = 6$	573	228	0.3057

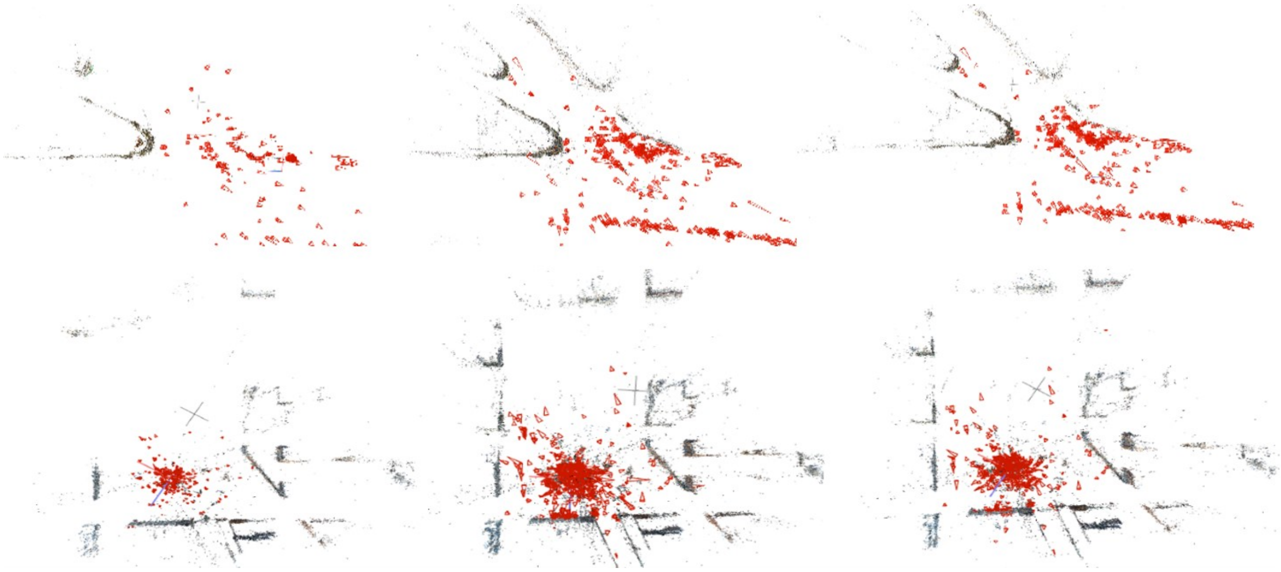

Fig. 6. The top view of reconstruction results : IR_5 (left) , IR_{50} (mid) , $Ours$ (right).

TABLE IV

THE FOUND PAIRS RATIO OF DIFFERENT METHODS

	found pairs ratio		
	IR_5	IR_{25}	$Ours$
Alamo	0.8776	0.9591	0.9597
Ellis Island	0.5933	0.9333	0.9510
Gendarmenmarkt	0.5708	0.9180	0.9441
Madrid Metropolis	0.7308	0.9203	0.9457
Montreal Notre Dame	0.7642	0.9681	0.9783
NYC Library	0.6495	0.9300	0.9409
Piazza del Popolo	0.6551	0.8990	0.9156
Piccadilly	0.5497	0.7902	0.8665
Roman Forum	0.7506	0.9402	0.9425
Tower of London	0.7096	0.9496	0.9465
Trafalgar	0.5243	0.8074	0.8870
Union Square	0.4989	0.8469	0.8758
Vienna Cathedral	0.7943	0.9602	0.9615

TABLE V

THE RESULTS OF EXPERIMENTS FOR A_{appr}

	#Registered	**#Time**	**#Precision**
with A_{appr}	331	177	0.5980
without A_{appr}	335	264	0.5214

in M_{inlier} for better observation of change. Increasing N_p reduces the time consumption and raises the precision but decreases the number of registered images. And a larger threshold T results in fewer registered images, less time consumption, and higher precision. The possibility that false feature tracks existing in a small area is small. When a large N_p divides image into several small cell, it skips many mismatched image pairs, so the precision is high. Similarly, a strict requirement ($T = 6$) also has a high precision.

As we set $k = 50$ in our methods, the candidate pairs of our method are the subset of IR_{50}. In order to indicate that our method finds most of the co-visibility images pairs in IR_{50}, we define a new metric called found pairs ratio. We take the results of IR_{50} as the groundtruth, and A_{gt} is the registered image set of IR_{50}. S_{gt} is the set of verified image pairs of IR_{50} between A_{gt}. S_{found} is the image pairs which sharing at least 30 feature tracks between A_{gt}. The found pairs ratio is computed from $|S_{found}|/|S_{gt}|$. Compared with other methods IR_5 and IR_{25}, the found pairs ratio of our methods is the highest in all the datasets as shown in Table IV.

B. Iterative matching strategy

To evaluate the improvement of our proposed iterative matching strategy on poor quality images, we present the experiments with and w/o A_{appr} on the dataset Ellis Island where there are many poor quality images. As listed in Table. V, the registered image number of the method with A_{appr} is almost the same as that of the method without A_{appr}, but the running time is reduced to the half.

V. CONCLUSIONS

This paper proposes an efficient iterative matching strategy and a reliable method to predict covisible image pairs by the transitivity of region covisibility. The comprehensive evaluation shows that our proposed method is three times faster than the state-of-the-art and the matching result of this method is good enough for complete reconstruction. In addition, the covisibility of image pairs can be efficiently found by our method by mining inlier feature matches fully. The direction of our future work is to explore how to integrate our work with other modules in SfM well to achieve a very efficient and robust SfM system.

ACKNOWLEDGEMENTS

The authors would like to thank Youji Feng, Bangbang Yang, Hanchao Li, Zhaoyang Huang, and Yijin Li for helping to proofread and format the paper. Thanks Kai Huang for video narration.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[2] C. Wu *et al.*, "Visualsfm: A visual structure from motion system," 2011.

[3] G. Zhang, H. Liu, Z. Dong, J. Jia, T. Wong, and H. Bao, "Efficient non-consecutive feature tracking for robust structure-from-motion," *IEEE Trans. Image Processing*, vol. 25, no. 12, pp. 5957–5970, 2016.

[4] J. L. Schönberger and J. Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[5] Z. Cui and P. Tan, "Global structure-from-motion by similarity averaging," in *IEEE International Conference on Computer Vision*, 2015, pp. 864–872.

[6] M. Farenzena, A. Fusiello, and R. Gherardi, "Structure-and-motion pipeline on a hierarchical cluster tree," in *IEEE International Conference on Computer Vision Workshops*, 2009, pp. 1489–1496.

[7] N. Snavely, S. M. Seitz, and R. Szeliski, "Skeletal graphs for efficient structure from motion," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[8] R. Raguram, C. Wu, J. Frahm, and S. Lazebnik, "Modeling and recognition of landmark image collections using iconic scene graphs," *International Journal of Computer Vision*, vol. 95, no. 3, pp. 213–239, 2011.

[9] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

[10] C. Wu, "Towards linear-time incremental structure from motion," in *International Conference on 3D Vision*, 2013, pp. 127–134.

[11] Y. Rui, T. S. Huang, and S. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39–62, 1999.

[12] A. Angeli, D. Filliat, S. Doncieux, and J. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[13] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[14] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision*, 2014, pp. 584–599.

[15] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, "Bundle adjustment in the large," in *European Conference on Computer Vision*, 2010, pp. 29–42.

[16] K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, J. N. Weinstein, *et al.*, "MatchMiner: a tool for batch navigation among gene and gene product identifiers," *Genome biology*, vol. 4, no. 4, p. R27, 2003.

[17] J. L. Schönberger, T. Price, T. Sattler, J. Frahm, and M. Pollefeys, "A vote-and-verify strategy for fast spatial verification in image retrieval," in *Asian Conference on Computer Vision*, 2016, pp. 321–337.

[18] C. Mei, G. Sibley, and P. Newman, "Closing loops without places," in *IEEE/RSJInternational Conference on Intelligent Robots and Systems*, 2010, pp. 3738–3744.

[19] M. Havlena and K. Schindler, "VocMatch: Efficient multiview correspondence for structure from motion," in *European Conference on Computer Vision*, 2014, pp. 46–60.

[20] J. L. Schönberger, A. C. Berg, and J. Frahm, "PAIGE: pairwise image geometry encoding for improved efficiency in structure-from-motion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1009–1018.

[21] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[22] C. Mei, G. Sibley, and P. Newman, "Closing loops without places," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 3738–3744.

[23] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[24] K. Wilson and N. Snavely, "Robust global translations with 1DSfM," in *European Conference on Computer Vision*, 2014, pp. 61–75.