Efficient Structure from Motion by Graph Optimization

Michal Havlena[1], Akihiko Torii[1,2], and Tomáš Pajdla[1]

[1] Center for Machine Perception, Department of Cybernetics, Faculty of Elec. Eng.,
Czech Technical University in Prague, Technická 2, 166 27 Prague 6, Czech Republic
{havlem1,pajdla}@cmp.felk.cvut.cz
[2] Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, Japan
torii@ctrl.titech.ac.jp

Abstract. We present an efficient structure from motion algorithm that can deal with large image collections in a fraction of time and effort of previous approaches while providing comparable quality of the scene and camera reconstruction. First, we employ fast image indexing using large image vocabularies to measure visual overlap of images without running actual image matching. Then, we select a small subset from the set of input images by computing its approximate minimal connected dominating set by a fast polynomial algorithm. Finally, we use task prioritization to avoid spending too much time in a few difficult matching problems instead of exploring other easier options. Thus we avoid wasting time on image pairs with low chance of success and avoid matching of highly redundant images of landmarks. We present results for several challenging sets of thousands of perspective as well as omnidirectional images.

Key words: Structure from motion, Image set reduction, Task prioritization, Omnidirectional vision

1 Introduction

We seek to reconstruct 3D scene structure and camera poses from a large collection of images downloaded from the web or taken by a camera mounted on a moving vehicle as in the Google Street View. This is a challenging task because unstructured web collections often contain a large number of very similar images of landmarks while, on the other hand, image sequences often have very limited overlap between images. Computation effort of large scale structure from motion is dominated by image matching, which is often done only to find that matched images actually do not have visual overlap.

Most of the state-of-the-art techniques for 3D reconstruction from unorganized image sets [1–4] start the computation by performing exhaustive pairwise image matching which becomes infeasible for image sets comprising thousands of images. Even Photo Tourism [5], one of the most known 3D modeling systems from unordered image sets, uses exhaustive pairwise image feature matching and exhaustive pairwise epipolar geometry computation to create the image graph with vertices being images and edges weighted by the uncertainty of pairwise relative position estimations which is later used to lead the reconstruction. By finding the skeletal set [6] as a subgraph of the image graph having as few internal nodes as possible while keeping a high number of leaves and the shortest paths being at most constant times longer, the reconstruction time improves significantly but the time spent on image matching remains the same. Recent advancement of the aforementioned technique [7] abandons exhaustive pairwise image matching by using shared occurrences of visual words [8, 9] to match only the ten most promising images per each input image. On the other hand, the number of computed image matchings still remains rather high for huge image sets. The presented computational speed is achieved also thanks to massive parallelization which demands grid computing on 496 cores.

We aim at reducing the number of image matchings by reducing the size of the image set, because it may be highly redundant. Opposed to the technique presented in [10], we do not cluster the input images using GIST [11] but we select a subset of input images in such a way that all the remaining images have a significant visual overlap with at least one image from the selected ones (Section 2). As this visual overlap is measured by shared occurrences of visual words [9], the method is more robust to viewpoint changes because it seeks for images capturing the same 3D structure rather than for images acquired from the same viewpoint, as demonstrated in [12]. Furthermore, the method works also for omnidirectional images where GIST often fails. For selecting the subset of input images, the approximate minimal connected dominating set is computed by a fast polynomial algorithm [13] on the graph constructed according to the visual overlap. The algorithm used is closely related to the maximum leaf spanning tree algorithm employed in [6] but the composition of the graph is quite different and less computationally demanding in our case.

The actual SfM pipeline uses the atomic 3D models reconstructed from camera triplets introduced by [14] as the basic elements of the reconstruction but the strict division of the computation into steps is relaxed by introducing a priority queue which interleaves different reconstruction tasks in order to get a good scene covering reconstruction in limited time (Section 3). Our aim here is to avoid spending too much time in a few difficult matching problems by exploring other easier options which lead to a comparable resulting 3D model in shorter computational time. We also introduce model growing by constructing new 3D points when connecting an image which allows for sparser image sets than those which could be reconstructed by [14].

2 Image Set Reduction

When performing sparse 3D reconstruction from user-input images, the input image set may often be highly redundant, such as photographs acquired by tourists at landmark sites. As it is not needed to use all such input images in order to get a 3D model covering the scene captured in them, it is possible to speed the reconstruction up by using only a suitable subset of input images.

Algorithm 1 Approximate minimum CDS computation [13]

```
Input  Unweighted undirected graph $G = (V, E)$.
Output  List $S$ of vertices belonging to the minimum CDS of $G$.

  I. Label all vertices $v \in V$ white.
 II. Set $D := \{\}$ and repeat until no white vertices are left:
     1: For all black vertices $v \in V$ set $c(v) := 0$.
     2: For all gray and white vertices $v$ set $c(v) :=$ number of white neighbours of $v$.
     3: Set $v^* := \arg\max_v c(v)$.
     4: Label $v^*$ black and add it into $D$.
     5: Label all neighbours of $v^*$ gray.
III. Set $S := D$ and connect components of the subgraph of $G$ induced by $D$ by adding
     at most 2 vertices per component into $S$ in a greedy way.
 IV. Return $S$.
```

We seek for a method that would remove the unnecessary images from the input image set while affecting neither the quality nor the connectivity of the resulting 3D model much. The concept of visual words, which first appeared in [9], has been used successfully for matching images and scenes [8]. It proved its usefulness also for near duplicate image detection [12] when the scene is captured from different viewpoints or under different lighting conditions. Our aim is to (i) evaluate pairwise image similarity efficiently following [15, 7] and (ii) formulate the selection of the desired subset of input images as finding a suitable subgraph of the graph constructed according to image similarity.

2.1 Image Similarity

We use the bag-of-words approach to evaluate image similarity. In particular, we follow the method proposed in [15] to create the pairwise image similarity matrix M_{II} containing the cosines of the angles between the normalized tf-idf vectors computed from the numbers of occurrences of the quantized SURF [16] image feature descriptors in individual images. Next, we create an unweighted undirected graph G_{II} expressing image similarity. Vertices of G_{II} are the input images and we add five edges per vertex connecting it with the five most similar images according to the values of M_{II}, which is close to the approach used in [7]. Edges are not added if the measured similarity falls under 0.05. Notice that there may (and often will) exist vertices with degree higher than five in the resulting graph as some images may be similar to many other images.

2.2 Minimum Connected Dominating Set

According to [13], the minimum connected dominating set (CDS) problem is defined as follows. Given a graph $G = (V, E)$, find a minimum size subset S of vertices, such that the subgraph induced by S is connected and S forms a

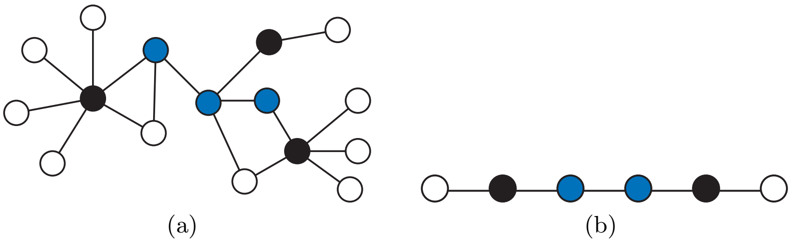


Fig. 1. Minimum CDS computation. Vertices belonging to the minimum dominating set D are labeled black, vertices added when connecting the components in order to get S are labeled blue. (a) General graph. (b) Graph being a singly connected line.

dominating set in G. In a graph with a dominating set, each vertex is either in the dominating set or adjacent to some vertex in the dominating set. The problem of finding the minimum CDS is known to be *NP*-hard [17] but [13] presents a fast polynomial algorithm with an approximation ratio of $\ln \triangle + 3$, $\triangle$ being the maximum vertex degree in the graph, see Algorithm 1.

We use the aforementioned algorithm to find the minimum connected dominating set S_{II} of the graph G_{II}, see Figure 1(a), and *only the images corresponding to the vertices in S_{II} are further used for the sparse 3D model reconstruction.* Edges of the subgraph of G_{II} induced by D (Algorithm 1, Step III.) together with the edges connecting the components of this subgraph in order to get S_{II} are used as the seeds of the reconstruction.

The usage of the dominating set provides for connecting the removed images to the resulting 3D model reconstructed from the selected ones using camera resectioning [18] if required, as an image is removed only if it is similar to at least one image which remains in the selected subset, i.e. there exists visual overlap between the resulting model and each of the removed images. Furthermore, the connectivity of the resulting 3D model is preserved by using the connected dominating set which does not allow for splitting the originally connected graph into components. For non-redundant image sets, e.g. when the graph expressing image similarity is a singly connected line, the method removes only the first and the very final images because removing more images would affect model connectivity, see Figure 1(b). On the other hand, the reduction of highly redundant image sets is drastic, as shown in Section 4.1.

3 3D Model Construction Using Tasks Ordered by a Priority Queue

The reduced image set is input into our 3D reconstruction pipeline which grows the resulting 3D model from several atomic 3D models. The computation is divided into tasks, each of them can either try to create a new atomic 3D model from three images, or try to connect one image to a given 3D model, see Figure 2. The order of the execution of different tasks is determined by task priority

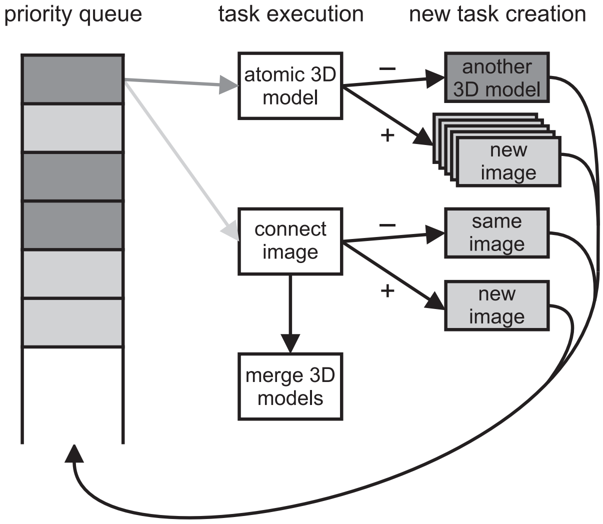


Fig. 2. Schematic visualization of the computation. The task retrieved from the head of the priority queue can be either an atomic 3D model construction task (dark gray) or an image connection task (light gray). Unsuccessful atomic 3D model reconstruction (–) inserts another atomic 3D model reconstruction task with the priority key doubled into the queue, a successful one (+) inserts five image connection tasks. Unsuccessful image connection (–) inserts the same task again with the priority key doubled, a successful one (+) inserts a new image connection task. Merging of overlapping 3D models is called implicitly after every successful image connection if the overlap is sufficient.

keys set when adding them to the priority queue being the essential underlying data structure. Note that *the task with the smallest priority key has the highest priority*, i.e. it is always in the head of the queue, in our implementation of the priority queue. Our aim is to set task priority keys in such a way that stopping the computation at any time would give a good scene covering sparse 3D model for the time given which is demanded e.g. by online SfM services. The state-of-the-art SfM approaches [5–7] implement this priority queue implicitly in such a way that they may get stuck by solving a difficult part of the reconstruction even when an easier path to the goal exists, as they are greedily growing from a single seed. Using our approach, several seeds are grown in parallel so the easiest path is actively searched for.

First, the queue is filled with one candidate camera triplet for atomic 3D model reconstruction per seed. The triplet is constructed from the two cameras C_1, C_2 being the endpoints of the edge corresponding to the seed. The third camera C_3^* is selected as

$$C_3^* = \underset{C_3}{\arg\max}\ \min(M_{II}(C_1, C_3), M_{II}(C_2, C_3)) \tag{1}$$

and the priority key of this task is set to $1 - M_{II}(C_1, C_2)$.

Next, the task from the head of the priority queue is taken and executed. As we are just starting the computation, it will be an atomic 3D model creation task. If the atomic 3D model reconstruction from a given candidate camera triplet is not successful, the camera triplet is rejected and another candidate camera triplet for the same seed is input into the queue with the priority key doubled. The new third camera accompanying cameras C_1 and C_2 is selected similarly as in Equation 1 by taking the camera C_3^* with the n-th largest value of $\min(M_{II}(C_1, C_3^*), M_{II}(C_2, C_3^*))$ and increasing n. After a successful atomic 3D model creation, the vicinity of the respective seed is searched for camera candidates suitable for connecting with the newly created atomic 3D model and tasks connecting the five most suitable cameras are input into the queue. We put the cameras contained in the atomic 3D model into the set $\mathcal{C}^c$ and the rest of the cameras into $\mathcal{C}^n$. Then, we search for a candidate camera C_r^* to be connected to the atomic 3D model using

$$(C_r^*, C_s^*) = \underset{(C_r, C_s) \,\in\, \mathcal{C}^n \times \mathcal{C}^c}{\arg\max} \; M_{II}(C_r, C_s). \tag{2}$$

The priority key of this task is set to $1 - M_{II}(C_r^*, C_s^*)$. Other four candidate cameras are selected similarly using the second, third, fourth, and fifth largest value of $M_{II}(C_r^*, C_s^*)$.

Alternatively, the head of the priority queue may contain an image connection task. After a successful image connection, a task connecting another camera to the same partial 3D model is created using Equation 2 again with a larger set $\mathcal{C}^c$ and input into the queue in order to keep the number of image connection tasks at five per a partial model. When the connection of an image to a given 3D model is unsuccessful, the task is input into the queue again with the priority key doubled because it may be successful if tried again after other images are successfully connected. In order to keep the resulting reconstruction consistent and connected, grown 3D models are implicitly merged together when they share at least five images. If the merge is not successful, it will be tried again when the number of shared images increases again.

The whole procedure is repeated until the priority queue is empty or the available time runs out. The following paragraphs describe particular parts of the pipeline in deeper detail.

3.1 Creation of Atomic 3D Models

Atomic 3D model reconstruction introduced in [14] has been improved and extended in several ways:

1. SIFT [19] and SURF [16] image feature detectors and descriptors have been added as it shows out that a combination of many different detectors is needed for difficult image sets. On the other hand, for easy image sets, it is possible to use only the fastest of them, which is SURF in our case.
2. Camera calibration does not need to be the same for all images in the set and can be obtained from the EXIF info of JPEG images.

3. The formula computing the quality score q has been simplified into:

$$q = |\{X : \tau(X) \geq \ 5^\circ\}|, \tag{3}$$

$\tau(X)$ being the apical angle measured at the 3D point X. In contrast with the original formula, 3D points with even larger apical angles do not contribute more to the quality score as we found out that it does not bring any significant improvement over the simple formula.

We require the quality score of at least 20 to accept a given candidate camera triplet as being suitable for reconstructing. Together with the remaining triplet quality pre-tests, the decision rule is the following: A given candidate camera triplet is accepted if and only if the results of pairwise epipolar geometries are consistent (the inlier ratio of the RANSAC finding the common scale is higher than 0.7), at least fifty 3D points have been reconstructed, at least twenty of them have apical angles larger than 5 degrees, and their projections cover a sufficiently large portion of the three respective viewfields.

3.2 Model Growing by Connecting Images

Connection of a new image to a given partial 3D model proceeds in two stages. First, the pose of the corresponding camera C_g with respect to the 3D model is estimated. Secondly, promising cameras from the vicinity of the newly connected one are used to create new 3D points.

Every 3D point already contained in the model has a descriptor which is transferred from one of the corresponding images during its triangulation. Thus it is easy to find 2D-3D matches between the reconstructed 3D points and the feature points detected and described in the candidate image being connected. To ensure reasonable speed even for large models with millions of points, we do one-way matching only with strict criteria on the first/second nearest neighbour distance ratio, setting it to 0.7 [19]. If the number of tentative matches is smaller than 20, the connection is not successful. Otherwise, RANSAC sampling triplets of 2D-3D matches is used to find the camera pose [18] having the largest support evaluated by the cone test [14]. Local optimization is achieved by repeated camera pose computation from all inliers [20] via SDP and SeDuMi [21]. We require the inlier ratio to be higher than 60% to consider the connection as successful and continue.

Next, we find the cameras already contained in the partial model, which have some viewfield overlap with the newly connected camera, by examining the projections of the inlier 3D points from the previous stage. We take a set $\mathcal{C}^p$ of all cameras, which contain projections of at least 20 inlier 3D points, and try to triangulate 3D points from camera pairs $(C_g, C_i) : C_i \in \mathcal{C}^p$. Newly triangulated 3D points with apical angles larger than 5 degrees are accepted if they are projected to at least three cameras after being merged based on the shared 2D feature points in C_g. Cone test can further reject a 3D point if those projections are not consistent with any possible 3D point position. Finally, sparse bundle adjustment [22] is used to refine the whole partial reconstruction after adding new 3D points and their projections.

3.3 Merging Overlapping Partial Models

When two partial 3D models share images, they usually share also 2D feature points which are the projections of some already triangulated 3D points. Therefore, we can avoid costly descriptor matching and create tentative 3D point matches between the two partial 3D models from pairs of 3D points which project to the same 2D feature points in both models.

As the 2D-3D matching used when connecting new images is rather strict, it often fails to find correspondences between not so distinctive regions, e.g. regions corresponding to the repetitive scene structures, which leads into triangulating the same scene 3D point once more at the latter stage. After connecting many images, scene 3D points may have several triangulated copies in the model, that is why the tentative 3D point matches created for merging often form large connected components, each of them corresponding to a single scene 3D point. After splitting all of these components into two parts, one per each partial model being merged, we use the cone test for each of those parts to verify that given 3D points can be merged into one. When this "internal merge" consolidating the partial models is finished, we continue with merging the two models using the collapsed tentative 3D point matches.

If there are less than 10 tentative 3D point matches, the merge is not successful, otherwise we try to find a similarity transform between the coordinate systems of the models. As three 3D point matches are needed to compute the similarity transform parameters [23], RANSAC with samples of length three is used. Inliers are evaluated by the cone test using image projections from both partial models and local optimization is performed by repeating the similarity transform computation from all inliers. Camera poses corresponding to the images shared by the models are averaged (rotation and position separately) inside the RANSAC loop before the cone test, so the similarity transforms which would lead into incorrectly averaged cameras would not be accepted. We require the inlier ratio to be higher than 60% to consider the merge as successful.

Finally, the smaller model is transformed to the coordinate system of the larger one because transforming the smaller model is faster. 3D point matches which were inliers are merged into a single point with the position being the mean of the former positions after transformation and duplicate image projections are removed. Sparse bundle adjustment [22] is used to refine the whole partial reconstruction after a successful merge.

4 Experiments

We demonstrate the proposed method in three experiments. The first one shows the efficient reduction of a highly redundant image set using the approximate minimum connected dominating set of a graph constructed using the image similarity matrix, the latter ones present the output of our 3D model reconstruction pipeline after 6 hours of computation for an omnidirectional and a perspective image set. All measured times are achieved by running a MATLAB+MEX implementation on a 2.83GHz Core2Quad PC.

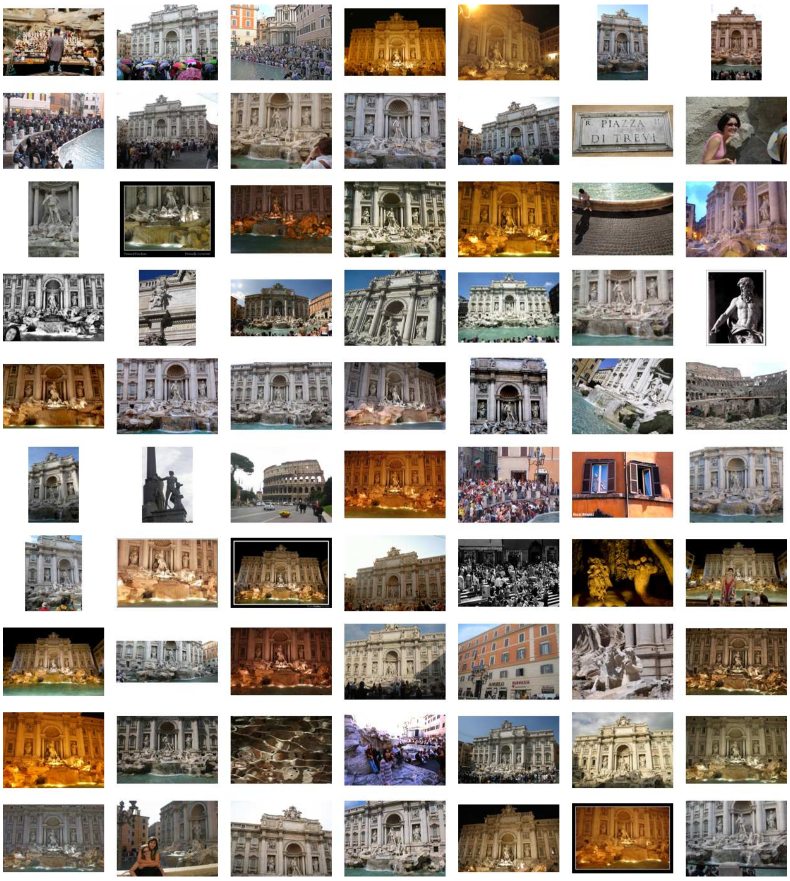


Fig. 3. Images corresponding to the approximate minimum connected dominating set computed for image set DiTrevi. Image set size has been reduced by 97% from 2,545 to 70 and the contamination ratio of the image set decreased from 17% to 7%.

4.1 Image Set Reduction

Image set DiTrevi consists of 2,545 images resulting from a Flickr Photo Sharing site [24] search for “di trevi” (April 2009). The image set is highly redundant and contaminated with images not capturing Di Trevi Fountain as it comprises pictures uploaded by hundreds of tourists visiting Rome. After detecting SURF image features and computing the image similarity matrix in 2 hours, the algorithm finding the approximate minimum connected dominating set of the corresponding graph returned 70 images in 5 seconds, see Figure 3. Selected images reasonably cover different scene viewpoints while the image set size was reduced by more than 97%. Furthermore, the contamination ratio of the image set decreased from 17% to 7% after the reduction.

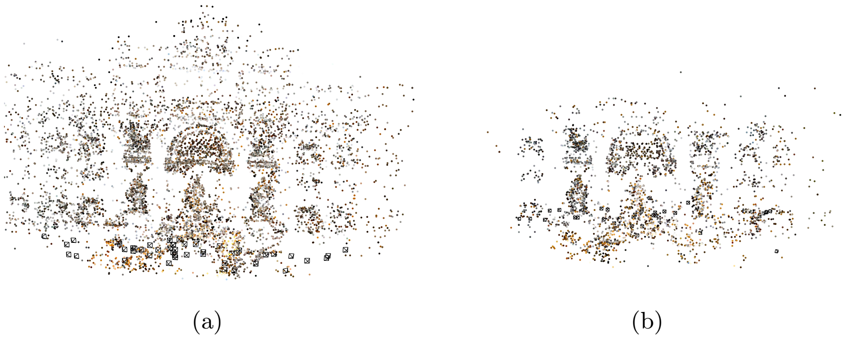


Fig. 4. (a) 3D model computed by Bundler [5] from the 70 images selected from image set DiTrevi by CDS. (b) The best from the 3D models returned by the five runs of Bundler on different random selections of 70 images from image set DiTrevi.

We used Bundler [5], a publicly available SfM tool, to evaluate the suitability of the image selection done by CDS for 3D reconstruction. The model returned in 44 minutes contains 47 camera poses and 8,489 3D points, see Figure 4(a). We ran Bundler also on five randomly selected sets of 70 images out of 2,545. Two of the runs did not return any result, two returned small fragments of the model with fewer than 5 camera poses, and one returned an incomplete 3D model having 32 camera poses and 3,355 3D points, as can be seen in Figure 4(b).

4.2 Sparse 3D Model Reconstruction

Two city sequences with landmark areas visited several times are used to demonstrate sparse 3D model reconstruction, see Figure 5. Nevertheless, they were input into the pipeline as unordered image sets.

Castle image set. Omnidirectional image set Castle [14] captured by a 180° fisheye lens camera with known calibration [25] consists of 4,472 omnidirectional images captured while walking in the center of Prague and around the Prague Castle. The obtained approximate minimum connected dominating set comprises 1,063 vertices and 1,359 edges are used as the seeds of the reconstruction. Image set reduction is not as drastic as for image set DiTrevi because the images are more evenly distributed. We use MSER [26], SIFT, and SURF image features in order to create sufficiently many 3D points even when image resolution is low. Several 3D models showing the important landmarks captured in the image set were obtained when the reconstruction time was limited to 6 hours, see Figure 6.

The resulting sparse 3D models are very similar to those presented in [14] but the speed of the reconstruction differs significantly as the authors of the aforementioned paper needed 12.5 days to obtain those results. Using our approach, the models are obtained in 10 hours, including 4 hours for image similarity matrix computation, which shows proper task priority key assignment.

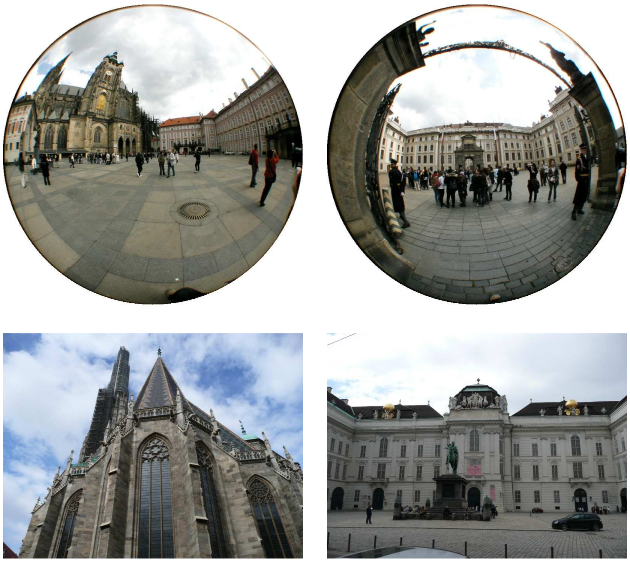


Fig. 5. Sample input images from image sets Castle and Vienna respectively.

Vienna image set. Image set Vienna [27] consists of 2,448 radially undistorted perspective images captured by a pre-calibrated camera while walking in the center of Vienna. After computing the image similarity matrix in 90 minutes, 1,008 vertices and 1,900 edges being the seeds of the reconstruction are obtained in 10 seconds as the result of the search for the approximate minimum connected dominating set of the corresponding graph. As image resolution is sufficient, only SURF image features are used for 3D model reconstruction. The 3D models showing several important landmarks captured in the image set, received after 6 hours of reconstruction, can be seen in Figure 7.

Compared to the omnidirectional image set Castle, only parts of the landmarks are reconstructed in the 6 hour limit because more images are needed to capture the whole landmark as the field of view of the perspective camera is limited. Partial 3D models become larger and connected gradually when the reconstruction continues, see Table 1 for different quantitative results of the reconstruction process at given times. Notice that the number of active seeds drops (86 → 77 → 65) after some time as the overlapping models are merged and also the sub-quadratic number of computed pairwise matchings w.r.t. the number of images contained in the partial models being far behind the quadratic number which would be achieved by methods using exhaustive pairwise image matching.

Note that when running Bundler on the reduced image set, 3 hours are spent on detecting and describing SIFT image features and 1,922 out of 15,753 tested image pairs are accepted after additional 6 hours of computation. No partial 3D models are output at this time as bundling starts later, after all 507,528 possible image pairs are tested.

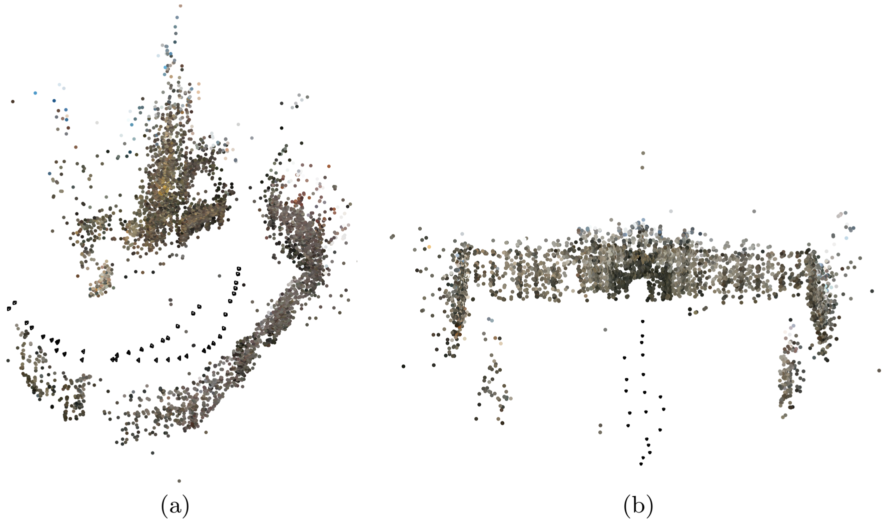


Fig. 6. Two largest partial 3D models reconstructed from the reduced image set Castle (1,063 images) after 6 hours of computation.

If one modified Bundler according to [7] so that it would test only the ten most promising image pairs per image based on image similarity and ran it on the non-reduced image set comprising 2,448 images, the whole 6 hour limit would still be spent on testing 16,762 obtained image pairs. This demonstrates the need for a prioritized structure from motion pipeline for large image sets.

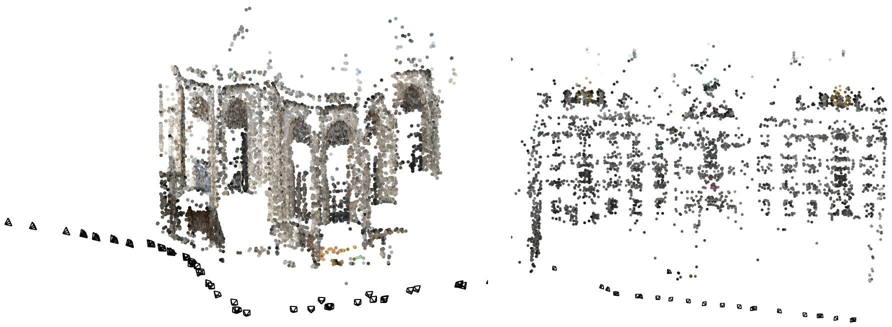
5 Conclusions

We presented a pipeline for efficient sparse 3D model reconstruction from highly redundant unordered image sets, such as those acquired by tourists at landmark sites as well as image sequences. The approximate minimum connected dominating set of a graph constructed according to the image similarity matrix computed from tf-idf vectors over SURF image features is used both for (i) reducing the size of the image set by removing nearly duplicate images and (ii) setting priority keys of the reconstruction tasks stored in a priority queue. The proposed interlacing of different reconstruction tasks allows for obtaining either a good scene covering sparse 3D model in limited time or a complete sparse 3D model when time is not limited.

Based on our experiments, image similarity works very well for the presented image sets and the number of the edges which were kept after the reduction was sufficient for 3D reconstruction. On the other hand, revisiting the reduction step may be necessary for difficult image sets. This is in principle possible and is a part of our future work together with fine tuning of the priority keys assigned to different tasks.

Table 1. The number of computed pairwise matchings, the number of active seeds, and the number of images contained in at least one partial model for the reduced image set Vienna (1,008 images) at given times of the reconstruction process.

Time	1h	2h	3h	4h	5h	6h	7h	8h	9h	10h	11h	12h
# pairs	548	991	1432	1773	2100	2360	2624	2882	3172	3437	3679	4030
# seeds	44	57	66	77	86	80	79	77	73	71	71	65
# images	153	244	313	368	411	438	466	496	521	546	572	600



(a) (b)

Fig. 7. Two largest partial 3D models reconstructed from the reduced image set Vienna (1,008 images) after 6 hours of computation.

Acknowledgements

This research was supported by the EC under Project HUMAVIPS FP7-ICT-247525 and by Czech Government under the research program MSM6840770038.

References

1. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or 'How Do I Organize My Holiday Snaps?'. In: ECCV'02. (2002) I: 414–431
2. Brown, M., Lowe, D.: Unsupervised 3D object recognition and reconstruction in unordered datasets. In: 3-D Digital Imaging and Modeling (3DIM). (2005) 56–63
3. Vergauwen, M., Van Gool, L.: Web-based 3D reconstruction service. Machine Vision and Applications (MVA) **17** (2006) 411–426
4. Martinec, D., Pajdla, T.: Robust rotation and translation estimation in multiview reconstruction. In: CVPR'07. (2007)
5. Snavely, N., Seitz, S., Szeliski, R.: Modeling the world from internet photo collections. IJCV **80** (2008) 189–210
6. Snavely, N., Seitz, S., Szeliski, R.: Skeletal graphs for efficient structure from motion. In: CVPR'08. (2008)

7. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building Rome in a day. In: ICCV'09. (2009) 72–79
8. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR'06. (2006) II: 2161–2168
9. Sivic, J., Zisserman, A.: Video Google: Efficient visual search of videos. In: Toward Category-Level Object Recognition (CLOR). (2006) 127–144
10. Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.: Modeling and recognition of landmark image collections using iconic scene graphs. In: ECCV'08. (2008) I: 427–440
11. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
12. Chum, O., Philbin, J., Isard, M., Zisserman, A.: Scalable near identical image and shot detection. In: Conference on Image and Video Retrieval (CIVR). (2007) 549–556
13. Guha, S., Khuller, S.: Approximation algorithms for connected dominating sets. Algorithmica **20** (1998) 374–387
14. Havlena, M., Torii, A., Knopp, J., Pajdla, T.: Randomized structure from motion based on atomic 3D models from camera triplets. In: CVPR'09. (2009) 2874–2881
15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR'07. (2007)
16. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). CVIU **110** (2008) 346–359
17. Garey, M., Johnson, D.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman (1979)
18. Nister, D.: A minimal solution to the generalized 3-point pose problem. In: CVPR'04. (2004) I: 560–567
19. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
20. Schweighofer, G., Pinz, A.: Globally optimal O(n) solution to the PnP problem for general camera models. In: BMVC'08. (2008)
21. Sturm, J.: SeDuMi: A software package to solve optimization problems - `http://sedumi.ie.lehigh.edu` (2006)
22. Lourakis, M., Argyros, A.: The design and implementation of a generic sparse bundle adjustment software package based on the Levenberg-Marquardt algorithm. Tech. Report 340, Institute of Computer Science – FORTH (2004)
23. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. PAMI **13** (1991) 376–380
24. Yahoo!: Flickr: Online photo management and photo sharing application - `http://www.flickr.com` (2005)
25. Mičušík, B., Pajdla, T.: Structure from motion with wide circular field of view cameras. PAMI **28** (2006) 1135–1149
26. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC'02. (2002) 384–393
27. Irschara, A., Zach, C., Bischof, H.: Towards wiki-based dense city modeling. In: Virtual Representations and Modeling of Large-scale environments (VRML). (2007)