Randomized Structure from Motion Based on Atomic 3D Models from Camera Triplets

Michal Havlena Akihiko Torii Jan Knopp Tomáš Pajdla
Center for Machine Perception, Department of Cybernetics
Faculty of Elec. Eng., Czech Technical University in Prague
{havlem1,torii,knoppj1,pajdla}@cmp.felk.cvut.cz

Abstract

This paper presents a new efficient technique for large-scale structure from motion from unordered data sets. We avoid costly computation of all pairwise matches and geometries by sampling pairs of images using the pairwise similarity scores based on the detected occurrences of visual words leading to a significant speedup. Furthermore, atomic 3D models reconstructed from camera triplets are used as the seeds which form the final large-scale 3D model when merged together. Using three views instead of two allows us to reveal most of the outliers of pairwise geometries at an early stage of the process hindering them from derogating the quality of the resulting 3D structure at later stages. The accuracy of the proposed technique is shown on a set of 64 images where the result of the exhaustive technique is known. Scalability is demonstrated on a landmark reconstruction from hundreds of images.

1. Introduction

Despite recent advancements of techniques for 3D reconstruction from unorganized image data sets [27, 3, 36, 16, 30, 31, 19], real scalability has not been yet reached.

When thinking of thousands of images, exhaustive computation of pairwise matches and epipolar geometries between all image pairs becomes infeasible. We propose a novel technique based on image pair similarity scores computed from the detected occurrences of visual words [25, 29] allowing us to perform a costly pairwise image matching only when it is likely to be successful. As the detection of visual words is very fast, this leads to a significant speedup while having only a small influence on the quality of the resulting model.

Speeding up the SfM computation has been a topic of many papers, real-time systems reconstructing urban scenes from video were presented in [1] and [7]. These techniques rely on the temporal order of the frames.

Photo Tourism [30], one of the most known 3D modeling systems from unordered image sets, uses exhaustive pairwise image feature matching and global bundle adjustment after connecting each new image to obtain an accurate model of the reconstructed object. The approach becomes very inefficient when images do not share a common view.

Recently, an advancement of this technique finding a skeletal subset giving almost optimal reconstruction has been presented [31]. An image graph with vertices being images and edges weighted by the uncertainty of pairwise relative position estimations is constructed. Its augmentation into a pair graph avoids traversing paths leading to undetermined scale between partial reconstructions by testing image connectivity. The skeletal set is found as a subgraph of the image graph having as few internal nodes as possible while keeping high number of leaves and having at most constant times longer shortest paths. Reconstructing from the skeletal set only and connecting the rest of the cameras later yields a great speedup without a significant loss of quality. On the other hand, the construction of the image graph is still very slow as one needs to compute epipolar geometries between all pair of images to evaluate its edges.

Unlike the methods suitable for landmark reconstruction from large-scale contaminated Internet image collections, we focus on datasets containing also evenly distributed cameras, where one cannot reduce the number of images dramatically without losing a substantial part of the model. On the other hand, a simple pre-filtering step based on the GIST descriptor [26] together with geometric verification according to [12] would allow us to work with datasets containing dense "hot spots" too. "Iconic image selection" and "iconic scene graph construction" concepts described in [12] are close to our technique, the main difference being the purpose of constructed partial 3D models. 3D models constructed in [12] may fully represent the reconstructed object when viewed from a certain viewpoint and should model the whole object when merged. Our 3D models are primarily intended for the geometrical verification of tentative image feature matches.

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

We use atomic 3D models reconstructed from camera triplets that share at least 100 points as the seeds which form the final large-scale 3D model when merged together. Using three views instead of two allows us to reveal most of the outliers of pairwise geometries at an early stage of the process hindering them from derogating the quality of the resulting 3D structure at later stages. Global optimization is replaced by faster locally suboptimal optimization of partial reconstructions which turns into the global technique when all parts are merged together. Cameras sharing fewer points are glued to the largest partial reconstruction during the final stage of the process.

Our pipeline is operating in the "easy first, difficult later" manner where pairwise matching and other computations are performed on demand. Therefore, it is possible to get a result close to the optimality in a given time available. Particular threshold values present at several places of the paper are the proposed values for obtaining a model whose quality is comparable to the results of the state of the art techniques using all pairwise matches. For easy data, there always exist many subsets of all pairwise matches that are sufficient for computing a reconstruction of a reasonable quality there but using just a subset of pairwise matches instead of the whole set yields a much faster reconstruction. Our method can be viewed at as a random selection of one of these subsets guided by the image similarity scores. Furthermore, unlike the aforementioned techniques, our pipeline is able to work both with calibrated perspective and calibrated omnidirectional images which is broadening its usability.

2. The Pipeline

Our pipeline consists of four consecutive steps, which are executed one after another: (1) Computing image similarity matrix (2) reconstructing atomic 3D models from camera triplets, (3) merging partial reconstructions, and (4) gluing single cameras to the best partial reconstruction (see Figure 1). The input of the pipeline is an unordered set of images acquired by cameras with known calibration. For perspective cameras, EXIF information can be used to obtain the focal length and we can assume principal point in the middle of the image. Omnidirectional cameras have to be pre-calibrated according to the appropriate lens or mirror model [20].

2.1. Computing Image Similarity Matrix

First, up to thousands of Speeded Up Robust Features (SURF) [2] are detected and described on each of the input images. Image feature descriptors are quantized into visual words according to a vocabulary containing 130,000 words computed from urban area images [9]. Assignment is done by Fast Library for Approximate Nearest Neighbors (FLANN) [22] searching for approximate nearest neighbours using a hierarchical k-means tree with branching factor 32 and 15 iterations. The parameters were obtained by FLANN automatic algorithm configuration finding the best settings for obtaining nearest neighbours with precision 90% in the shortest time possible. Next, term frequency–inverse document frequency (tf-idf) vectors [29], which weight words occurring often in a particular document and downweight words that appear often in the database, are computed for each image with more than 50 detected visual words and finally, pairwise image similarity matrix S_{II} containing cosines of angles between normalized tf-idf vectors t_a, t_b of images I_a, I_b is computed as

$$S_{II}(a,b) = t_a \cdot t_b. \tag{1}$$

Images with less than 50 detected visual words are excluded from further computation.

Figure 1. Overview of the pipeline. Input images are described by SURF and an image similarity matrix is computed. Atomic 3D models are reconstructed from camera triplets, merged together into partial reconstructions, and finally single cameras are glued to the largest partial reconstruction.

2.2. Reconstructing Atomic 3D Models from Camera Triplets

Image similarity matrix S_{II} is used as a heuristics telling us which triplets of cameras are suitable for reconstructing atomic 3D models. As S_{II} is symmetric with units on the diagonal, we take the upper triangular part of S_{II}, exclude the diagonal, and search for the maximum score. This gives us a pair of cameras with indices i and j. Then, we find three "third camera" candidates with indices k_1, k_2, and k_3 such that $\min(S_{II}(i,k_1), S_{II}(j,k_1))$ is maximal, $\min(S_{II}(i,k_2), S_{II}(j,k_2))$ is the second greatest and $\min(S_{II}(i,k_3), S_{II}(j,k_3))$ is the third greatest among all possible choices of the third camera. Atomic 3D models are reconstructed for each of the candidates as described below. The resulting models are ranked by the quality score and the model with the highest quality score is selected and passed to the next step of the pipeline.

Denoting the index of the third camera corresponding to the selected atomic 3D model as k, cameras with indices i, j, and k are removed from future selections by zeroing rows and columns i, j, and k of S_{II}. If the quality of all three 3D models is 0, no 3D model is selected and $S_{II}(i,j)$ is zeroed preventing further selection of this pair of cameras. The whole procedure is repeated until the maximum score in S_{II} is lower than 0.1.

Quality score. Each 3D point X reconstructed from a triplet of cameras has associated three apical angles [33], one apical angle per each camera pair $\tau_{ij}(X)$, $\tau_{ik}(X)$, and $\tau_{jk}(X)$. The formula giving us the 3D model quality q is the following:

$$\tau(X) = \min(\tau_{ij}(X), \tau_{ik}(X), \tau_{jk}(X)) \quad (2)$$

$$P_1 = \{X : \tau(X) \geq 5^\circ\} \quad q_1 = \begin{cases} |P_1| & |P_1| \geq 10 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$P_2 = \{X : \tau(X) \geq 10^\circ\} \quad q_2 = \begin{cases} |P_2| & |P_2| \geq 10 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$P_3 = \{X : \tau(X) \geq 15^\circ\} \quad q_3 = \begin{cases} |P_3| & |P_3| \geq 10 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$q = q_1 + 4q_2 + 20q_3 \quad (6)$$

Our q formula checks whether there is a sufficient number of 3D points with large apical angles, as they ensure good relative camera pose estimation [33]. Constants 4 and 20 favour atomic models with 3D points having really large apical angles, as we seek for atomic models with distant cameras, while threshold value 10 ensures that the quality is not overestimated when only few points have sufficiently large apical angles. As $P_1 \supseteq P_2 \supseteq P_3$, points with $\tau(X) \in \langle 10^\circ, 15^\circ)$ have five times bigger weight than those with $\tau(X) \in \langle 5^\circ, 10^\circ)$ and the same applies to points with $\tau(X) \in \langle 15^\circ, \infty)$ against $\tau(X) \in \langle 10^\circ, 15^\circ)$.

Atomic 3D model reconstruction. The atomic 3D model from a triplet of cameras is reconstructed in several steps. After each step, the reconstruction is terminated if the number of reconstructed 3D points falls under 100 and the model quality score set to 0. All intermediate results of the computation are stored into separate files and can be reused if needed which speeds up the computation. The procedure is the following:

1. Image features, namely Maximally Stable Extremal Regions (MSER) [17] on intensity and saturation channels and Affine Invariant Interest Points (APTS) [21] Laplacian-Affine and Hessian-Affine, are detected on three input images (denoted as I_i, I_j, and I_k) and the assigned Local Affine Frames (LAF) [18] are described by Discrete Cosine Transform (DCT).

2. Tentative matches between the three image pairs ($I_i\, I_j$, $I_i\, I_k$, and $I_j\, I_k$) are computed using FLANN [22] searching for approximate nearest neighbours using 4 random kd-trees, filtered to keep only the mutually best matches, and then further filtered into tentative matches among triplets by chaining matches in all three images.

3. Homogeneous image coordinate vectors of filtered tentative matches are normalized to unit direction vectors using the known camera calibration. Pairwise relative camera poses are obtained by softvoting for the epipole positions [11] using 5 votes from independent Progressive Sample Consensus (PROSAC) [6] runs with the 5-point algorithm [23].

4. Shared inliers of these geometries, i.e. final matches, together with three pairwise triangulations [8] are computed. The relative positions of the cameras and the common scale of all three reconstructions is found using one 3D point correspondence (with RANSAC).

5. 3D points are reconstructed from the pair with the largest baseline for omnidirectional cameras, or by optimal triangulation from three views [5] for perspective cameras.

6. Very distant points, i.e. likely outliers, are filtered out and sparse bundle adjustment [13] modified similarly as in [10], regarding non-perspective central cameras as a kind of a generalized camera, refines both points and cameras.

Detecting multiple types of image features (ad. 1.) is favourable as they are usually located in different parts of an image: MSER features are found on uniform regions while APTS features fire on corners. To achieve high computation speed, tentative matches are found using an approximate technique with 80% precision and subsequent two-step filtering of the computed tentative matches (ad. 2.) decreases their contamination by mismatches leading to a speedup of the epipolar geometry estimation.

As the ordered randomized sampling in PROSAC still has the randomness of selecting matches, each epipolar geometry resulted by a single run of PROSAC may be different, especially when the tentative matches are strongly contaminated by mismatches. To increase the chance of finding the correct model, we cast the epipole positions, i.e. relative motion directions, of the best epipolar geometries recovered by several independent runs of PROSAC (ad. 3.). The best model is selected as the one with the epipole position closest to the maximum in the accumulator space. This strategy works when the correct, or almost correct, models provide consistent motions while the incorrect models with high support generate different ones, which is often the case. More details can be found in [34].

It is a natural property of evaluating matches by epipolar geometry that incorrect matches lying on epipolar lines or in the vicinity to epipoles are often regarded as inliers. However, they can be easily filtered out by finding shared inliers of three views as 3D points successfully verified in three views are unlikely to be incorrect. Therefore, RANSAC obtaining the common scale of the three reconstructions

(ad. 4.) is a good test of the quality of pairwise geometries. To find, which triplets of final matches generate a consistent 3D point, we use a "cone test" checking the existence of a 3D point that would project to desired positions in all three matches after the scales were unified. During the cone test, four pixels wide cones (two pixels to each side) formed by four planes (up, down, left, and right) are casted around the final matches and we test whether the intersection of the cones is empty or not using the LP feasibility test [15].

The definitive advantage of the cone test over the standard technique checking the reprojection errors [8] lies in the fact that inaccurately reconstructed 3D points, e.g. those with small apical angles which have large uncertainties in depth estimates, do not affect the error measure. If one used the reprojection errors instead, which is equivalent to testing whether or not a given reconstructed 3D point lies in the intersection of the casted cones, some correct matches could be rejected due to corresponding inaccurately reconstructed 3D points. Inaccurate 3D points triangulated from accepted matches do not cause any harm as they are later re-triangulated (ad. 5.) and bundled (ad. 6.).

As an exhaustive test is faster than LP for three cones, LP is used only when intersecting a higher number of cones during merging and gluing and not in this particular case. The exhaustive test constructs all candidates for the vertices of the convex polyhedron comprising the intersection of the cones as the intersections of triplets of planes. The intersection of the cones is empty iff none of these candidates lies in all 12 positive halfspaces formed by the planes. To reject atomic 3D models with low-quality pairwise geometries, the quality score is set to 0 if the inlier ratio of the cone test is under 80%.

To ensure an uniform image coverage by the projections of reconstructed 3D points, a unit sphere surrounding the camera center representing different unit vector directions is tessellated into 980 triangles $\mathcal{T}$ using [4]. A triangle T is non-empty if there exists a reconstructed 3D point projecting into it, empty otherwise. The image coverage measurement c_I of image I is defined as

$$\mathcal{T}_o = \{T \in \mathcal{T} : T \text{ is non-empty}\} \quad c_I = \frac{|\mathcal{T}_o|}{|\mathcal{T}|}. \tag{7}$$

If more than one image from the triplet has $c_I < 0.01$, the quality score of the atomic 3D reconstruction is set to 0.

2.3. Merging Partial Reconstructions

First, we construct a new similarity matrix S_{TT} containing similarity scores between selected atomic 3D models. Having two atomic 3D models each constructed from camera sets $\mathcal{C}_a = \{i, j, k\}$ and $\mathcal{C}_b = \{i', j', k'\}$ respectively, there are always nine pairs of cameras such that the cameras are contained in different models. The similarity score between two atomic 3D models is computed as the mean of the similarity scores of those nine pairs as

$$S_{TT}(a, b) = \frac{1}{9} \sum_{a_x \in \mathcal{C}_a} \sum_{b_y \in \mathcal{C}_b} S_{II}(a_x, b_y). \tag{8}$$

The matrix is again used as the heuristics telling us which pairs of atomic 3D models are suitable for merging. At the beginning, we have one partial reconstruction per accepted 3D model, each of them containing three cameras and 3D points triangulated from them. Partial reconstructions will be connected together during the merging step forming bigger partial reconstructions containing the union of cameras and 3D points of the connected reconstructions.

We take the upper triangular part of S_{TT}, exclude the diagonal, and search for the maximum score. This gives us a pair of atomic 3D models with indices m and n. Next, we try to merge the two partial reconstructions R_p and R_q containing the models with indices m and n respectively. After a successful merge, elements $S_{TT}(p', q')$ are zeroed for all indices of models p' contained in partial reconstruction R_p and all indices of models q' contained in partial reconstruction R_q in order to prevent further merging between atomic 3D models which are both contained in the same partial reconstructions. If the merge is not considered to be successful, partial reconstructions are not connected and $S_{TT}(m, n)$ is zeroed preventing further selection of this pair of atomic models. Notice however, that this is not a strict decision on the mergeability of partial reconstructions R_p and R_q as they can be connected later using a different pair of atomic models contained in them. The whole procedure is repeated until the maximum score in S_{TT} is lower than 0.05.

Merging two atomic 3D models. The actual merge is performed in several steps. Given two atomic 3D models with indices m and n, first, tentative 3D point matches are found. Each 3D point X reconstructed from a triplet of cameras with indices i, j, and k has three LAF+DCT descriptors D_i^X, D_j^X, and D_k^X connected with it. Having six sets of descriptors (D_i, D_j, and D_k for 3D points from model m and $D_{i'}$, $D_{j'}$, and $D_{k'}$ for 3D points from model n), we find the mutually best matches between all nine pairs of descriptors (D_i $D_{i'}$, D_i $D_{j'}$, etc.) independently. As particular descriptors of a single 3D point from model m can be matched to descriptors of different 3D points in model n in individual matchings, unique 3D point matches need to be constructed. The nine lists of the 3D point matches output from the individual matchings are concatenated and sorted by the distance of the descriptors in the feature space. A unique matching is obtained in a greedy way by going through the sorted list and accepting only those 3D point matches whose 3D points are not contained in any of the 3D point matches accepted before.

If there are less than 10 tentative 3D point matches, the merge is not successful, otherwise we try to find a similarity transform bringing model m to the coordinate system of model n. As three 3D point matches are needed to compute the similarity transform parameters [35], RANSAC with samples of length three is used. A 3D point match is an inlier if the intersection of the three cones from cameras contained in model n and the three cones from the transformed cameras contained in model m is non-empty. Local optimization is performed by repeating the similarity transform computation from all inliers.

If the inlier ratio is higher than 60%, the merge is considered successful and the whole partial reconstructions R_p and R_q are merged according to this similarity transform computed from atomic 3D models m and n only. R_q remains fixed and the 3D points and cameras of R_p are transformed, 3D point matches which were inliers are merged into a single point with the position being the mean of the former positions after transformation.

Sparse bundle adjustment [13] is used to refine the whole partial reconstruction after a successful merge. The resulting partial reconstruction is then transformed to a normalized scale to allow easy visualization and to ease the next step of the pipeline.

2.4. Gluing Single Cameras to the Best Partial Reconstruction

The best partial reconstruction R_r is selected as the one containing the highest number of cameras. In this step, we are trying to find the poses of the cameras which are not contained in R_r. Another similarity matrix S_{TI}, which contains similarity scores between atomic 3D models contained in R_r and cameras not contained in R_r, is constructed. The similarity score between the atomic 3D model constructed from cameras $\mathcal{C}_a$ and a camera with index b is computed as the mean of similarity scores of three pairs of cameras as

$$S_{TI}(a,b) = \frac{1}{3} \sum_{a_x \in \mathcal{C}_a} S_{II}(a_x, b). \tag{9}$$

We search for the maximum score in S_{TI} and obtain the atomic 3D model with index o and the camera with index l. During the gluing step, we compute the pose of the camera l using 3D points contained in the atomic model o. The gluing being successful, we zero the column l of S_{TI} in order to prevent further selection of already glued single cameras, otherwise only element $S_{TI}(o,l)$ is zeroed. The whole procedure is repeated until the maximum score in S_{TI} is lower than 0.025.

Gluing a single camera. When performing the actual gluing, we find mutually best tentative matches between three pairs of descriptors ($D_i\, D_l$, $D_j\, D_l$, and $D_k\, D_l$) independently. Unique 2D-3D matches are obtained using the same greedy approach as when performing a merge. If the number of tentative matches is smaller than 20, the gluing is not successful. Otherwise, RANSAC sampling triplets of 2D-3D matches is used to find the camera pose [24] having the largest support evaluated by the cone test again. Local optimization is achieved by repeated camera pose computation from all inliers [28] via SDP and SeDuMi [32].

If the inlier ratio is higher than 80%, the gluing is considered successful and the camera with index l is added into the partial reconstruction R_r. Sparse bundle adjustment is used to refine the whole partial reconstruction and the reconstruction is transformed to a normalized scale again because improper scale of the reconstruction can influence the convergence of the SDP program.

Figure 2. Example input image data. Top row: Perspective images from data set DALIB. Bottom row: Omnidirectional images from data set CASTLE.

3. Results

We present results on two data sets. The first one consists of 64 images and the camera poses obtained by the exhaustive method computing matches between all pairs of cameras [16] are known. We consider them being near the ground truth as their accuracy has been proven by a successful dense reconstruction. For the second experiment, we use a set of 4,472 omnidirectional images captured while walking through Prague. Our method was able to find images sharing the views and reconstruct several landmarks present in them.

DALIB data set. The data set DALIB consists of 64 perspective images capturing a paper model of a house acquired by a camera with known calibration (see Figure 2).

The pipeline selected 13 atomic 3D models out of 132 candidates (S_{II} was sampled only 44 times for the best pair). It was sufficient to compute just 199 pairwise image matches compared to 2,016 computed by the exhaustive method. All atomic models were successfully merged into a single partial reconstruction and the poses of 25 missing

Figure 3. Complete reconstruction of data set DALIB. Partial reconstruction containing all 39 cameras from selected atomic 3D models was extended with 25 missing cameras during gluing.

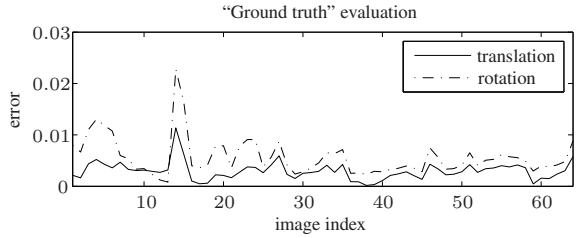

Figure 4. Measured errors of the camera pose estimation of data set DALIB. Translational error is the fraction of the diameter of a sphere containing all cameras, rotational error is in radians. Note that all cameras but camera number 14 were estimated with translational error smaller than 0.7%.

Name	Similarity	Atomic 3D	Merge	Gluing
DALIB	2 min	37 min	2 min	2 min
CASTLE	6 hrs	257 hrs	18 hrs	19 hrs

Table 1. Time spent in different steps of the pipeline while reconstructing data sets DALIB and CASTLE.

Method	Features	Matching	Geometry
MATLAB+MEX	10 min	65 min	15 min
Photo Tourism	→	8 min	←

Table 2. Time spent in different steps of our exhaustive method and Photo Tourism [30] for data set DALIB. Photo Tourism time is the total time spent by the method as one cannot measure the times of the individual steps.

cameras were obtained during gluing resulting in the model shown in Figure 3. The time spent in different steps of the pipeline having a MATLAB+MEX implementation running on a standard Core2Duo PC can be found in Table 1. The total computation time was less than 45 minutes. Sparse bundle adjustment takes less than a second in average to run for an atomic 3D model and at most several seconds when

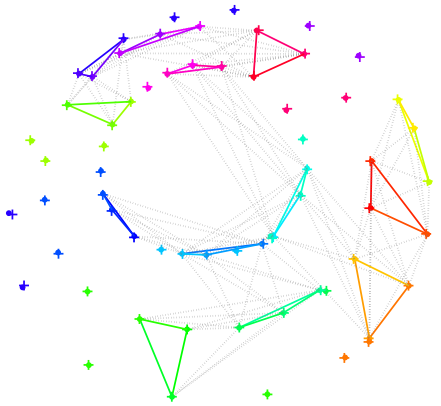

Figure 5. Visualization of the selected atomic 3D models and their merging of data set DALIB. Cameras computed by our method (denoted as •) contained in the same atomic 3D model are connected by a coloured line, cameras glued to a given model are sharing its colour. Merging is shown by dashed grey lines. Cameras obtained by the exhaustive method are denoted as +.

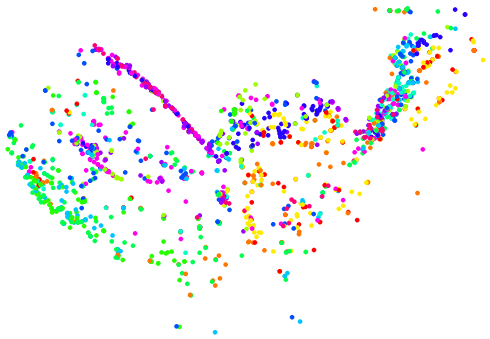

Figure 6. The partitioning of the resulting 3D point cloud among 13 selected atomic 3D models of data set DALIB. Colour coding is the same as for Figure 5.

applying for refining larger partial reconstructions because there are not so many constraints as we do not match all image pairs.

Computation time of the exhaustive method using a similar MATLAB+MEX implementation on the same hardware was around 90 minutes, most of the time being spent on computing pairwise image feature matches (see Table 2). When reconstructing the same data set using Photo Tourism [30] which also uses exhaustive pairwise image feature matching, the computation time went down to 8 minutes but the resulting camera poses were less accurate than those obtained by our method due to the lack of SIFT [14] image features on the paper model. Photo Tourism is faster mainly because of implementation reasons as it contains a more optimized C/C++ code.

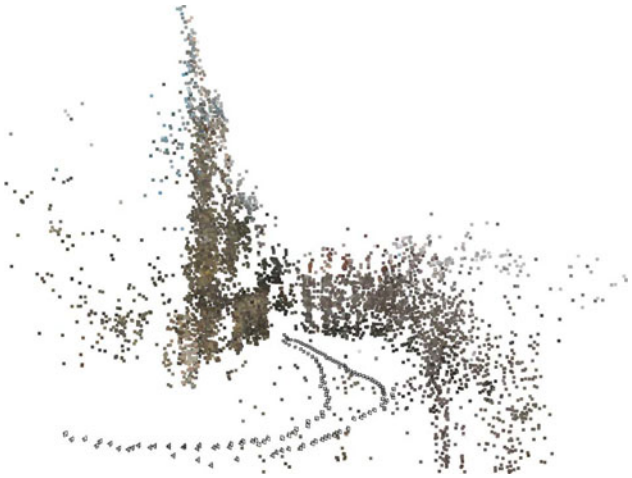

Figure 7. Partial reconstruction #486 of data set CASTLE. Right part of the St.Vitus Cathedral and other buildings surrounding the square were reconstructed from 90 cameras, another 49 cameras were connected during gluing.

After finding the similarity transform between the camera poses computed by our method and those computed by the exhaustive one, we were able to measure the error of the camera pose estimation. It has shown that there is no significant loss of quality (see Figure 4). Both sets of cameras can be seen in Figure 5 together with the visualization of atomic 3D models and their merging. Figure 6 shows the partitioning of the resulting 3D point cloud among 13 selected atomic 3D models.

CASTLE data set. Our second data set CASTLE consists of 4,472 omnidirectional images captured by a 180° fisheye lens camera with known calibration.

The images were acquired in several sequences while walking in the center of Prague and around the Prague Castle but they were input into the pipeline as an unordered set. The pipeline selected 652 atomic 3D reconstructions out of 100,410 candidates and only 58,961 pairwise image matches were computed while the number of all possible image pairs is 9,997,156. Several partial reconstructions containing remarkable landmarks were obtained (see Figures 7, 8, and 9). The total computation time was around 12.5 days. As Photo Tourism works with perspective images only, we could not compare its performance with the performance of the proposed method on this data set directly but if we linearly extrapolated the computation time of Photo Tourism using the number of all possible image pairs, it would be around 27.5 days.

Minor merging and gluing errors caused by repetitive image structures and matching clouds can be found in some of the resulting partial reconstructions. As our current "winner takes all" approach is unable to recover from such errors, our future work lies in introducing alternative ways of merging and a method evaluating their quality in order to bound incorrect ones.

Figure 8. Partial reconstruction #407 of data set CASTLE. Part of the Old Town Square with the clock tower was reconstructed from 69 cameras, another 39 cameras were connected during gluing.

Figure 9. Partial reconstruction #471 of data set CASTLE. Entrance to the Prague Castle was reconstructed from 60 cameras, another 49 cameras were connected during gluing.

4. Conclusions

We have presented a new efficient technique for large-scale structure from motion from unordered data sets. Pairwise image similarity scores are used to reduce the number of computed image feature matchings drastically, yielding a significant speedup compared to techniques based on exhaustive pairwise matching. Using atomic 3D models instead of reconstructions from camera pairs as the seeds, the quality of the triangulated 3D points is higher as they are verified in three views. Merging, which connects the atomic 3D models into partial reconstructions, both extends and improves accuracy of the model because the number of image projections of merged points is increased. Finally, poses of the cameras not contained in the given partial reconstruction are estimated using 2D-3D matches during gluing.

The method is fully scalable storing all results of the computation on a hard drive instead of in RAM. Performance could be improved by using a fast SSD drive instead of a standard SATA drive.

Acknowledgements

This work was supported by EC project FP6-IST-027787 DIRAC and by Czech Science Foundation under Project 201/07/1136. T. Pajdla was supported by Czech Government under the research program MSM6840770038. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

References

[1] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, H. Towles, D. Nistér, and M. Pollefeys. Towards urban 3d reconstruction from video. In *3DPVT*, May 2006.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *CVIU*, 110(3):346–359, June 2008.

[3] M. Brown and D. Lowe. Unsupervised 3d object recognition and reconstruction in unordered datasets. In *3DIM05*, pages 56–63, 2005.

[4] J. Burkardt. Sphere grid: Points, lines, faces on a sphere - http://people.scs.fsu.edu/~burkardt/datasets/sphere_grid, 2007.

[5] M. Byrod, K. Josephson, and K. Astrom. Fast optimal three view triangulation. In *ACCV07*, pages II: 549–559, 2007.

[6] O. Chum and J. Matas. Matching with prosac: Progressive sample consensus. In *CVPR05*, pages I: 220–226, 2005.

[7] N. Cornelis, K. Cornelis, and L. Van Gool. Fast compact city modeling for navigation pre-visualization. In *CVPR06*, pages II:1339–1344, 2006.

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2003.

[9] J. Knopp, J. Šivic, and T. Pajdla. Location recognition using large vocabularies and fast spatial matching. Research Report CTU–CMP–2009–01, CMP Prague, January 2009.

[10] M. Lhuillier. Effective and generic structure from motion using angular error. In *ICPR06*, pages I: 67–70, 2006.

[11] H. Li and R. Hartley. A non-iterative method for correcting lens distortion from nine point correspondences. In *OMNIVIS 2005*, 2005.

[12] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV08*, pages I: 427–440, 2008.

[13] M. Lourakis and A. Argyros. 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, August 2004.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.

[15] A. Makhorin. Glpk: Gnu linear programming kit - http://www.gnu.org/software/glpk, 2000.

[16] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR07*, 2007.

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC02*, pages 384–393, 2002.

[18] J. Matas, S. Obdrzalek, and O. Chum. Local affine frames for wide-baseline stereo. In *ICPR02*, pages IV: 363–366, 2002.

[19] Microsoft. Photosynth - http://livelabs.com/photosynth, 2008.

[20] B. Mičušík and T. Pajdla. Structure from motion with wide circular field of view cameras. *PAMI*, 28(7):1135–1149, July 2006.

[21] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, October 2004.

[22] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP09*, 2009.

[23] D. Nistér. An efficient solution to the five-point relative pose problem. *PAMI*, 26(6):756–770, June 2004.

[24] D. Nister. A minimal solution to the generalized 3-point pose problem. In *CVPR04*, pages I: 560–567, 2004.

[25] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR06*, pages II: 2161–2168, 2006.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, May 2001.

[27] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or 'how do i organize my holiday snaps?'. In *ECCV02*, pages I: 414–431, 2002.

[28] G. Schweighofer and A. Pinz. Globally optimal o(n) solution to the pnp problem for general camera models. In *BMVC08*, 2008.

[29] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *CLOR06*, pages 127–144, 2006.

[30] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3d. In *SigGraph06*, pages 835–846, 2006.

[31] N. Snavely, S. Seitz, and R. Szeliski. Skeletal graphs for efficient structure from motion. In *CVPR08*, 2008.

[32] J. Sturm. Sedumi: A software package to solve optimization problems - http://sedumi.ie.lehigh.edu, 2006.

[33] A. Torii, M. Havlena, T. Pajdla, and B. Leibe. Measuring camera translation by the dominant apical angle. In *CVPR08*, 2008.

[34] A. Torii and T. Pajdla. Omnidirectional camera motion estimation. In *VISAPP08*, pages II: 577–584, 2008.

[35] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *PAMI*, 13(4):376–380, April 1991.

[36] M. Vergauwen and L. Van Gool. Web-based 3d reconstruction service. *MVA*, 17(6):411–426, December 2006.