On the improvement of three-dimensional reconstruction from large datasets

Guilherme Potje, Mario F. M. Campos, Erickson R. Nascimento
Computer Science Department
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
Email: {guipotje,mario,erickson}@dcc.ufmg.br

***Abstract*—The digital cameras granted many possibilities of structure and shape recovery from imagery that are quickly and inexpensively acquired by such devices. The state-of-art algorithms are now able to deliver 3D structure acquisition results from consumer-grade image sensors with quality and resolution comparable to industry standard systems such as LiDAR and photogrammetric equipments. Nonetheless, the processing time of the collected imagery to produce a 3D model quickly becomes prohibitive as the number of input images increases, demanding powerful hardware and days of processing time to generate full 3D models from these large datasets. In this work we propose an efficient approach based on Structure-from-Motion and multi-view stereo reconstruction techniques to automatically generate 3D models from images. The results from six large aerial datasets acquired by UAVs and four terrestrial datasets show that our approach [1] outperforms current strategies in processing time, and is also able to provide better or at least equivalent results in accuracy compared to three state-of-the-art SfM methods.**

***Keywords*-Structure-from-Motion; 3D Mapping;**

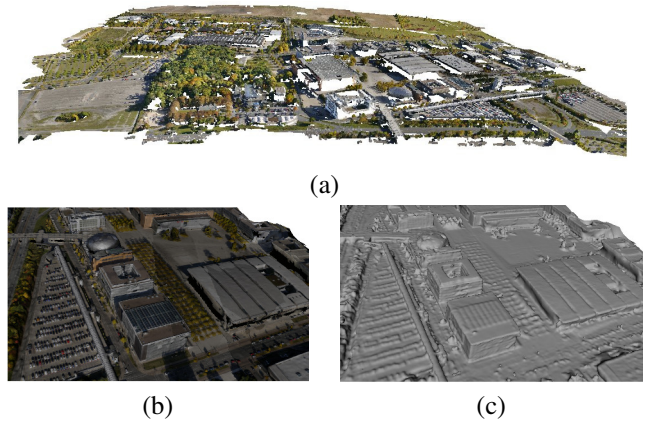


Fig. 1. Digital elevation model estimated by our methodology. (a) A dense 3D Model using $1,231$ high resolution images from expopark dataset estimated in only 17 hours (single core); (b) The projected textures into the mesh of a detailed region and (c) surface reconstruction of the same region.

I. Introduction

Geometric reconstruction of the world from a collection of images remains one of the key-challenges in Computer Vision. Three-dimensional recovery of the geometry of an object or a scene has several applications in Computer Vision and Robotics, such as scene understanding [1], object recognition and classification [2], digital elevation mapping and autonomous navigation, to name a few. For applications, such as aerial and terrestrial mapping, for instance, image-only based pipelines that incorporate recent SfM and multi-view stereo techniques are strong competitors to LiDAR based surface measurements [3].

Although Structure-from-Motion (SfM) techniques have significantly advanced in accuracy and scalability in the past few years, in general, their processing time increases non-linearly with respect to the number of pictures, which makes the processing time for most real outdoor scenes, such as open-pit mines and large areas of cities, undesired, or even prohibitive, specially on consumer-grade computers.

Contributions: The general contribution of our work is the improvement of the incremental SfM pipeline by combining existing ideas and proposing new ones. We propose a new pipeline based on: I) The use of GPS information to avoid matching distant image pairs [4], II) Consideration of only the most discriminant features of the images to make a coarse estimation of the pair geometry [5], III) The use of approximate nearest neighbor search of the corresponding features of the valid pairs, and IV) The use of the vocabulary tree search to speed up the matching phase to $O(n \log(n))$ [6]. These steps combined outperform the previously proposed approaches individually, and reduces even more the time required to perform the matching step when compared to the individual approaches. Also, we propose a modified maximum spanning tree algorithm used in the epipolar graph that carefully selects high quality image pairs to be used in the reconstruction, consequently improving accuracy, but also ensures the completeness of the results by avoiding the graph to be disconnected. Finally, we introduce a modified local bundle adjustment (LBA) window approach that targets local consistency by overlapping the BA windows.

As shown in the experiments, our approach is capable of computing the DEM faster than the other methods used in the experiments in all the tested datasets while preserving the quality of the results.

II. Structure-from-Motion Review

The incremental approach [7] [8] [4] [5] gained attention in the past years, because of its robustness to outliers (wrong

[1]This work relates to a Master dissertation.

correspondences and relative motion estimation) and missing data, such as the absence or wrong intrinsic parameters. Other methods such as factorization-based [9] and global SfM [10] can process datasets faster than incremental SfM, because they do not need to constantly optimize the model, since they solve the whole problem at once.

The epipolar graph is widely used to represent the geometric relation between each pair of image in the scene [7][10] [4] and can be defined as follows: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $v \in \mathcal{V}$ represents an image and there is an edge $e \in \mathcal{E}$ between two vertices if there is a valid geometry relation between the images which is encoded by the fundamental, essential or homography matrix (the last in case of planar scenes). The naïve approach for constructing the epipolar graph is time consuming, requiring $O(n^2)$ pairs verification to optimally build the graph. However, efficient methods like [6] can be used to reduce the time complexity to $O(n \log n)$ with loss of weak edges, however the results have been shown to be a good alternative for large datasets. Techniques like SIFT [11] and many others can be used to detect and match points across images, generally using a RANSAC scheme in order to handle outlier correspondences

Besides the feature matching and geometric validation step, another bottleneck of SfM approaches is the optimization of the camera parameters and 3D points, that needs to be constantly done during the reconstruction to reduce error drifting. Optimizing large SfM problems demands highly specialized algorithms that need to be efficient and well-implemented. In spite of the improvements already made, specially in exploiting the sparse block structure that arises in bundle adjustment to speed up the computation [12] [8] [13], the problem is still costly to solve for large datasets.

III. Methodology

In this section we detail the main steps of our methodology. It is a novel pipeline that provides two new features: An efficient epipolar graph building procedure improved by an additional filtering step, and a local bundle adjustment adapted to large-scale reconstructions.

First, the GPS constraint in addition to the vocabulary tree score are used to efficiently prune non-overlapping pairs (Fig. 2– 1) followed by a coarse to fine geometry validation to save even more processing time in the feature matching phase (Fig. 2– 2). The epipolar graph's edges are then updated by the modified maximum spanning tree algorithm (Algorithm 1) that carefully selects the best ones to be used in estimation of the camera parameters and the scene structure while enforcing the completeness of the graph (Fig. 2– 3). The camera motion and intrinsics, as well the 3D structure parameters are incrementally recovered and locally optimized by an overlapping window containing the most recent cameras (Fig. 2– 4,5). In the final step, our pipeline computes the dense model using a patch-based multi-view-stereo technique and Poisson reconstruction to obtain the final mesh (Fig. 2– 6).

Pruning: In general, in the aerial image acquisition process, GPS data (even if noisy) will be available. It is fair to assume that if the Euclidean distance between the position of image pairs is large, they do not share any portion of view. By considering that, we generate an initial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $v \in \mathcal{V}$ represents an image. We connect the $d\_nearest$ images according to the distance obtained by compairing each pairs' GPS coordinates. We used $d\_nearest = 40$ in our experiments, which is a sufficient value for all datasets in our experiments, considering the standard aerial mapping acquisition pattern. For terrestrial datasets, we raise this value to 60, since their mean overlap are generally more redundant.

The constraint increases the time performance and reduces the time complexity of matching n images from $O(n^2)$ to $O(n)$ considering aerial and large datasets. Additionally, this avoids comparing ambiguous pairs, which makes the approach more robust to wrong reconstructions due to views that are actually geometrically consistent but are not viewing the same portion of the scene (*e.g.* symmetric building facades).

In some cases, the GPS tags are missing for some images, and it can become a problem when a dataset has most of its images without GPS information. To overcome this problem, we use a vocabulary tree approach similar to [6] to avoid the $O(n^2)$ time complexity in the matching step. Vocabulary trees are used in scalable image recognition to retrieve similar images. We used SIFT [11] features to build the vocabulary tree. In our experiments, we search for the top 60 most similar images for each image, and we prune the edges in the epipolar graph from the query vertice to those vertices that are not among the highest scores of this query, excepting the edges that were validated by the GPS distance.

Registration: In general, SfM techniques look for the correspondence tracks between images to estimate the camera extrinsic parameters and a sparse point cloud. We use SIFT to extract and match the keypoints. In order to to minimize processing costs, we sort the found keypoints by descending order of scale and remove the small keypoints so that we keep the features with large scale attribute up to 9.000 features per image, which is a sufficient amount of keypoints for the most scenarios, as suggested by [5]. The reason we select the features with large scale attribute in many steps in the approach is because they are more robust.

After the graph construction, we can efficiently match image pairs in a reduced space search, which initially had $O(n^2)$ and now has $O(n)$ pairs. For each edge of the graph, the matching step procedure first attempts to match the descriptors of two small sets containing the biggest (most discriminant) keypoints of their respective images, selected according to the scale attribute. We consider a pair as valid if the number of inlier correspondences returned by the fundamental matrix estimation using the normalized 8-point algorithm [14] in a RANSAC scheme is higher than at least 15% of the number of cross-validated descriptor matches between each pair, which we call $coarse\_inlier\_rf$. The 15% value was chosen by performing tests on image pairs and we concluded when there is less than 15% of inliers in the correspondences using the top 600 features in scale, the likelihood of overlap

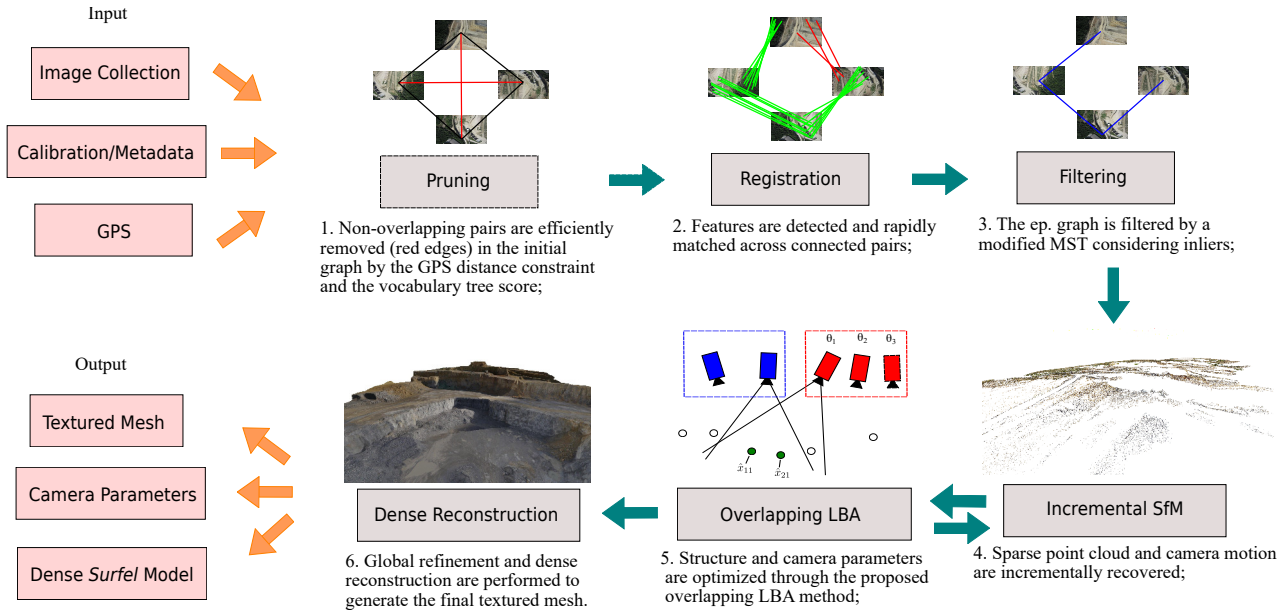


Fig. 2. Main steps of our methodology. We initialize the epipolar graph by connecting images with a large chance of having overlap, according to GPS data and a vocabulary tree search. In this example, the black edges are below the threshold distance, and the vocabulary tree query of at least one of the images are among the highest score matches of the other. After the optimized pairwise registration, we update the epipolar graph by selecting high quality matches enforcing completeness (blue edges in step III). The camera motion is incrementally recovered for each image and a sparse point cloud generated from the matching points and optimized through robust and fast local bundle adjustment. At the end, we compute the dense model.

between them is minimal. If the correspondences are able to minimally satisfy the epipolar geometry constraint, a full pairwise matching considering all keypoints is performed to obtain a fine registration using the Fast Approximate Nearest Neighbour search (FLANN) [15].

A $threshold\_fm = 0.07\%$ of the image width is used to determine if the correspondence is an inlier or not [7], depending on the distance that it is from the respective epipolar lines.

Filtering: We set the weights in the epipolar graph using the number of *inliers* returned by RANSAC for each estimated pair. A naïve approach would consider removing the edges with a small number of inliers using a hard threshold and perform the triangulation by using only the remaining pairs. However, this may remove edges that keep the graph connected, which results in missing parts in the final 3D model, specially because it is difficult to define a hard threshold for this purpose, depending on many factors. Therefore, we propose applying a maximum spanning tree approach (MST) to remove only the edges with small number of inliers but enforcing the connectivity of the graph, since the MST avoids us breaking the epipolar graph into smaller connected components when we try to remove an edge with low number of inliers.

Algorithm 1 Epipolar graph filtering.

procedure EPIPOLARFILTERING(EG, τ_i)
  MAXSPANNINGTREE(EG, $FilteredEG$)
  for each edge e in EG **do**
    if $weight > \tau_i$ & $e \notin FilteredEG$ **then**
      ADD($FilteredEG$, e)
  return $FilteredEG$
▷ The $FilteredEG$ contains the maximum spanning tree plus all edges higher than a threshold.

The last step of the epipolar filtering consists in extracting the sub-graph that contains the edges from the maximum spanning tree and the edges with the number of inliers larger than a defined threshold τ_i (we use a value of 60 inliers in our experiments, a standard value used by Bundler [7] and VisualSFM [5]). This procedure is described by the Algorithm 1.

Incremental SfM: The incremental reconstruction algorithm begins by selecting a pair of images and then incrementally estimate the points and cameras parameters. The camera motion estimation happens in a greedy manner with respect to the number of 2D-3D correspondences.

Choosing the initial pair is crucial to the quality of the reconstruction. If we choose a pair not having enough overlap, the reconstruction can fail immediately. To avoid that, we sort the edges of the graph and keep a percentile of 0.4 of the most valued edges (this value is arbitrary and is not sensitive when it is not set on the extremes like ≤ 0.10 or ≥ 0.90 according to our experiments), which contains consistent geometric pairs that undoubtedly overlap. Then, we sort this subset considering the ratio between the essential matrix inliers and the homography inliers and use a percentile of 0.25 (again, the percentile value is not sensitive and is arbitrary) of the subset containing the highest ratio between the fundamental matrix inliers and homography inliers ($F_{inliers}/H_{inliers}$), which is

useful to avoid the use of small-baseline pairs in the seed reconstruction. We then finally select the pair which provide the lowest mean re-projection error after triangulation in this small subset of candidates.

From the initial point cloud, we find the image with the largest 2D correspondences with 3D points already estimated and we calculate the extrinsic parameters from the camera. To do that, we use an a contrario camera resectioning approach [16], which estimates an inlier threshold for each camera estimation, differently from approaches that only uses the standard RANSAC resectioning with fixed threshold, thus improving accuracy. The camera resectioning step is repeated iteratively for all cameras, and after a certain amount of camera estimations, we call a local bundle adjustment to minimize the re-projection error. Once there is no more cameras to be added, we run a global bundle adjustment.

Local Bundle Adjustment and Global Refinement: Finding the optimal solution for the global bundle adjustment problem (BA) time consuming when considering thousands of cameras and millions of 3D points. To tackle with this problem, we propose an overlapping local bundle adjustment (LBA) window approach that optimizes the camera poses and points locally, but it overlaps with already optimized 3D points to hold the consistency and avoid fast propagation of error (drift).

Let $\mathbf{P}$ be a vector containing all parameters describing the m projection matrices and the n three-dimensional points, where $\mathbf{P} = (\theta_1, \ldots, \theta_m, X_1, \ldots, X_n)^T$, and $\mathbf{X}$ the measurement vector composed of the measured image point coordinates across all cameras. By using the parameter vector, we can create the estimated measure matrix as $\hat{\mathbf{X}} = (\hat{x}_{11}^T, \ldots, \hat{x}_{1m}^T, \ldots, \hat{x}_{n1}^T, \ldots, \hat{x}_{nm}^T)^T$, where $\hat{x}_{ij}^T$ is the projection of the 3D point i in the camera j.

We can rewrite the BA as the optimization problem of finding the values of $\mathbf{P}$ and $\mathbf{X}$ that minimize $(\mathbf{X} - \hat{\mathbf{X}})^T \Sigma_{\mathbf{X}}^{-1} (\mathbf{X} - \hat{\mathbf{X}})$ over the projection matrices $\mathbf{P}$. $\Sigma_{\mathbf{X}}$ is the norm matrix. The minimization can be performed by using the Levenberg-Marquardt algorithm [17] to solve the augmented weighted normal equations $(\mathbf{J}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{J} + \mu \mathbf{J})\delta = \mathbf{J}^T \Sigma_{\mathbf{X}}^{-1} (\mathbf{X} - \hat{\mathbf{X}})$, where $\mathbf{J}$ represents the Jacobian of $\hat{\mathbf{X}}$, δ the update parameter of $\mathbf{P}$ that we are estimating and μ is the damping term which is used to change the diagonal elements of the Jacobian.

Incremental approaches optimize camera motion and scene structure calling BA multiple times during reconstruction. As the number of parameters of the model incrementally increases, the time to perform an iteration rapidly grows. To tackle this problem, we fasten the parameters of the 3D points that have already been bundle adjusted and only adjust the parameters of the newest estimated cameras and points.

The time complexity of bundle adjustment considering the sparse block structure is $O(m^3)$ [18], where m is the number of cameras. In the the incremental approach, $O(m)$ global BA calls are required to avoid the propagation of drifting, which makes the complexity raise to $O(m^4)$. By using LBA, we are able to reduce the complexity to $O(m^3)$ again.

The window contains the most recent estimated cameras and all the 3D points that projects onto them. When the window achieves the limit of cameras, we call a BA that will optimize all cameras in the set and the points in their field of view. It is important to notice that points that have been already optimized contributes to the minimized re-projection error, although their parameters remain fixed, to maintain the local consistency and prevent the fast propagation of drift. Fig. 2 (5.) shows two sets of cameras (blue and red). The blue set was optimized and the current iteration is trying to adjust the three new cameras (in red). The green points should not be modified in the optimization process. Global BA can be performed sometimes during reconstruction to obtain the optimal parameters as we do in our experiments, but much fewer global optimization calls are required (bound to a constant value), and it is optional depending on the size of the dataset and the desired accuracy. We used Ceres Solver [19] as the optimization engine in our implementation.

Dense Reconstruction: Once we have the complete set of projection matrices and undistorted images estimated by our approach, we use them as input to a MvS dense reconstruction technique [20]. The quality of the camera parameters provided by the SfM algorithm as well as the quality of the images (*e.g.* resolution, texture and image sharpness) strongly influence on the density and quality of the estimated *quasi-dense surfel* model. Finally, by using the Poisson Surface Reconstruction method [21], we convert the set of oriented points into a textured mesh model.

IV. Experimental Evaluation

In this section, we show the obtained results of our pipeline in ten different datasets and compare them against three state-of-the-art implementations for solving moderate and large scale SfM problems, namely, Bundler [7], VisualSFM [5] and OpenMVG [16].

Both aerial and terrestrial datasets were used to evaluate our method, each one from a different scene. Challenging aspects are present in many of these datasets: Low-textured regions, reflective surfaces such as lakes, occlusions caused by moving objects and strong illumination and perspective changes.

We used six large scale aerial datasets composed of high resolution overlapping images acquired by unmanned aerial vehicles with large baseline, obtained from publicly available drone websites. We also used four terrestrial datasets. NotreDame was obtained from the internet (Flickr), while the other three were made using a *Samsung S4* smartphone.

A virtual machine hosted by a computer equipped with two Intel(R) Xeon(R) CPU E5-2620 @ 2.00GHz processors and 132 GB of RAM were used in the experiments. For a fair comparison we computed the total time time each approach used in CPU. Our focus was to test the scalability of the SfM approaches *per se*.

We also tested VisualSFM on a Xeon E3-1200 v2/3rd Gen 8-core processor and a GeForce GTX 560 Ti GPU to evaluate the impact the parallel optimization has on timings. In this experiment (VisualSFM-GPU), we computed the wall-clock time.

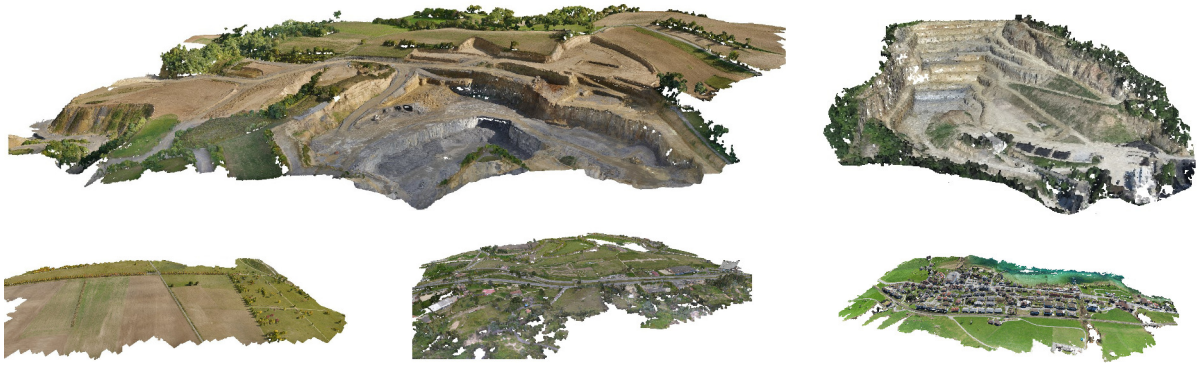
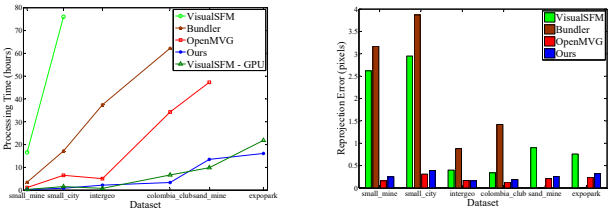


Fig. 3. Dense *surfel* models estimated for the aerial datasets. In clockwise order from the top-left to bottom-right: sand_mine, small_mine, intergeo, colombia_club and small_city.



(a) Processing time (b) Re-projection error

Fig. 4. Time performance considering the entire pipeline (a) and the median re-projection error results (b) of each approach for the large aerial datasets. Our approach was the only able to provide the results for the expopark dataset within the time-out value of 120 hours. VisualSFM-GPU re-projection error is equivalent to VisualSFM without GPU.

	images	speedup	local error	global error
small_mine	127	2.33	0.60	0.45
small_city	297	2.51	0.56	0.53
intergeo	479	2.45	0.38	0.41
colombia_club	795	1.83	0.39	0.39
sand_mine	978	1.28	0.37	0.37
expopark	1,231	2.04	0.38	0.23

TABLE I
SPEEDUP GAIN AND MEAN RE-PROJECTION ERROR IN PIXELS OF THE LOCAL APPROACH COMPARED TO GLOBAL BA. THERE IS A SIGNIFICANT SPEEDUP IN EXCHANGE OF A SMALL OSCILLATION IN THE RESIDUALS.

A. Evaluation Methodology and Parameter Tuning

To quantitatively evaluate the output of all the SfM techniques in our experiments, we use the mean and median residual re-projection error values in pixels. Due to the lack of ground-truth data, we were only able to evaluate the estimation quality achieved by measuring the re-projection error values.

To choose the window size in the local bundle adjustment step, we ran several experiments with multiple window size values on 3 moderate-sized datasets. We chose the size equal to 80, since it provided the best time performance gains with small fluctuations in the re-projection error. We tested several combinations of detectors and descriptors and we found that SIFT holds the best results, providing more robust and accurate correspondences.

B. Results and Discussion

The Fig. 4 (b) shows the median re-projection error for each dataset and each method. Fig. 4 (a) shows the time performance. We set a time-out of 120 hours for the single core experiment. Bundler and VisualSFM were unable to generate the results for some datasets in the established time-out.

In Fig. 4 (a), we can clearly see an expressive increase in processing time by all implementations but ours, as the number of images increases. Our method shows a smoothed growth and it leads time performance, reflecting the efficient steps adopted in our method. We are even able to outperform VisualSFM with all parallel optimizations enabled including GPU (Fig. 4 (a) dark green curve). Also, Fig. 4 (b), shows that our approach and OpenMVG (which is an implementation exclusively focused in accuracy) lead accuracy performance. This is the result of a careful selection of pairs to be matched through the filtering performed by Algorithm 1, which avoids false positive matches that can lead to an increase of the re-projection error of the cloud compromising the model's accuracy, besides assuring the completeness in the model estimation. As can be seen in Fig. 3, Fig. 1 and Fig. 5 the final models do not present any visible drift or abnormalities on the mesh.

We used a collection of 715 unorganized images from the challenging Notre Dame dataset [7] to show the capability of our approach to deal with unordered collection of images in the wild. Most of GPS tags are missing from images. The dense 3D model generated by our method is shown in Fig. 5 (top-left model). For this experiment, our method estimated the model with a re-projection error of 0.43 pixels in 27.4 hours. Bundler method, for its turn, spent 86.7 hours and got a larger error (0.47 pixels). VisualSFM was not able to provide the results within the established time-out value of 96 hours, and OpenMVG could not handle the missing focal lengths of a good portion of the images, not returning any results.

Quantitative results for the terrestrial datasets except for Notredame were not considered, because of their small size (below 125 images) and there were little gain in time and accuracy compared to the other approaches. Qualitative results can be seen in Fig. 5.

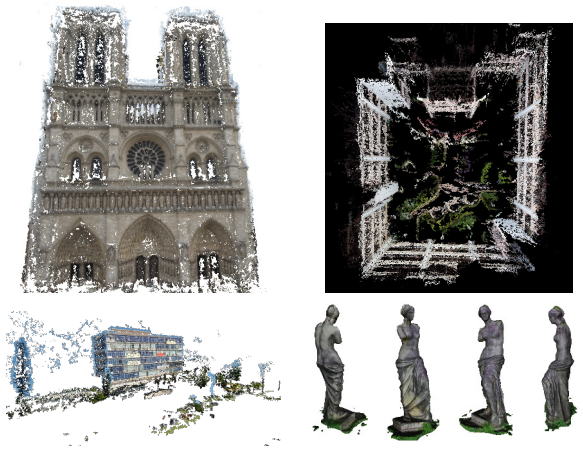


Fig. 5. Dense models estimated for the terrestrial datasets. In clockwise order from the top-left to bottom-right: notre_dame, ICEx_square, UFMG_rectory and UFMG_statue. All the last three datasets were estimated in less than 30 minutes with our approach.

The speedup provided by using the LBA method proposed can be verified in Table I. We compare the total time used to generate the DEM with the LBA against the classic approach of globally optimizing the model multiple times. After the reconstruction using LBA finishes, we run a final global BA to obtain the optimal solution. We can see that even running a global BA in the end, the speedup gain is considerable, and it is able to achieve global minima. It means that the multiple local BAs are able to maintain the necessary consistency and avoid the final minimization to fail.

We also evaluate the benefits of using the Algorithm 1 and the GPS information to filter the epipolar graph. We performed 3 experiments with the largest datasets. A gain up to 62% in accuracy using Algorithm 1 is obtained in our implementation. Qualitative results from the aerial datasets can be seen in Fig. 3, and for the terrestrial datasets, in Fig. 5.

V. Conclusion

In this work, we proposed and implemented a new SfM pipeline adapted to high resolution aerial image (but not only limited to this kind) datasets which incorporates and improves previously used methods in the literature aiming at time efficiency. It is important to mention that most of these methods were used separately in previous works and we explore and adapt them into a single approach, in addition to the maximum spanning tree that ensures the graph's completeness and also contributes to an improved accuracy. The speed-up achieved as well as the low re-projection error can be seen in the experiments performed to evaluate the time efficiency and point cloud quality.

Limitations and future improvements: In our approach, besides the GPS pruning, we do not treat geometric ambiguity in the scene, thus, in some scenarios the reconstruction can fail for this reason. Some works in SfM aim at solving this specific problem [22]. Using both aerial and ground images to generate more complete models is also a subject of our interest. Currently, SfM approaches struggle to merge these views due to strong change of perspective from aerial to ground. Another issue is that if a significant drift occurs in SfM, mainly because the lack of tracks on images, the reconstruction may also fail. In the future, we plan to extend this work to consider these problems.

Acknowledgment

The authors would like to thank the agencies CAPES, CNPq, FAPEMIG and ITV (Vale Institute of Technology) for funding different parts of this work.

References

[1] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *CVPR*. IEEE, 2009, pp. 2036–2043.

[2] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *ICCV*. IEEE, 2009, pp. 221–228.

[3] F. Leberl, A. Irschara, T. Pock, P. Meixner, M. Gruber, S. Scholz, and A. Wiechert, "Point clouds," *Photogrammetric Engineering & Remote Sensing*, vol. 76, no. 10, pp. 1123–1134, 2010.

[4] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik *et al.*, "Building rome on a cloudless day," in *ECCV*, 2010, pp. 368–381.

[5] C. Wu, "Towards linear-time incremental structure from motion," in *3DV*, 2013, pp. 127–134.

[6] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, vol. 2. IEEE, 2006, pp. 2161–2168.

[7] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *IJCV*, vol. 80, no. 2, pp. 189–210, Nov. 2008.

[8] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *ICCV*, 2009, pp. 72–79.

[9] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *IJCV*, vol. 9, no. 2, pp. 137–154, 1992.

[10] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *CVPR*, June 2011, pp. 3001–3008.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[12] M. A. Lourakis and A. Argyros, "SBA: A Software Package for Generic Sparse Bundle Adjustment," *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 1–30, 2009.

[13] Y. Jeong, D. Nister, D. Steedly, R. Szeliski, and I.-S. Kweon, "Pushing the envelope of modern methods for bundle adjustment," *PAMI*, vol. 34, no. 8, pp. 1605–1617, 2012.

[14] R. I. Hartley, "In defense of the eight-point algorithm," *PAMI*, vol. 19, no. 6, pp. 580–593, 1997.

[15] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009, pp. 331–340.

[16] P. Moulon, P. Monasse, and R. Marlet, "Adaptive structure from motion with a contrario model estimation," in *ACCV*, 2012, pp. 257–270.

[17] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial & Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[18] K. Mitra and R. Chellappa, "A scalable projective bundle adjustment algorithm using the l infinity norm," in *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*. IEEE, 2008, pp. 79–86.

[19] S. Agarwal, K. Mierle, and Others, "Ceres solver," http://ceres-solver.org, 2015.

[20] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *PAMI*, vol. 32, no. 8, pp. 1362–1376, 2010.

[21] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, 2006.

[22] K. Wilson and N. Snavely, "Network principles for sfm: Disambiguating repeated structures with local context," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 513–520.