RECONSTRUCTION OF ORIENTATION AND GEOMETRY FROM LARGE UNORDERED IMAGE DATASETS FOR LOW COST APPLICATIONS

Mohammed Abdel-Wahab, Konrad Wenzel, Dieter Fritsch

ifp, Institute for Photogrammetry, University of Stuttgart
Geschwister-Scholl-Straße 24D, 70174 Stuttgart, Germany
{ mohammed.othman, konrad.wenzel, dieter.fritsch}@ifp.uni-stuttgart.de

KEY WORDS: Photogrammetry, Orientation, Reconstruction, Incremental, Adjustment, Close Range, Matching, High Resolution

ABSTRACT:

Reconstruction of camera orientations and structure from images is one of the basic tasks in photogrammetry and computer vision. A fully automated solution of this task from scratch in terrestrial applications is still pending in case of large unordered image datasets especially for close-range and low-cost applications. Current solutions require high computational efforts for image networks with high complexity and diversity regarding acquisition geometry. Unlike the methods suitable for landmark reconstruction from large-scale Internet image collections - we focus on datasets where one cannot reduce the number of images without losing geometric information of the dataset. Within the paper, an automated pipeline for the reconstruction of reliable and precise orientation and geometry from unordered image sets is presented. It was employed for several challenging large-scale datasets from different applications such as cultural heritage data recording or imagery from unmanned aerial vehicles (UAVs). However, results were also used for the derivation of initial values for commercial photogrammetric processing software such as Trimble Match-AT. Experimental results are shown to demonstrate the performance of the presented pipeline for applications with high accuracy requirements.

1. INTRODUCTION

In the past few years, low-cost photogrammetry has become a focus of research especially since cameras enable an efficient data acquisition at very low prices. For instance, recent work [Wenzel et al., 11] has shown that it is possible to use low-cost multi-camera systems (figure 1a) for efficient point cloud recording in close range applications with high accuracy requirements. Such applications lead to very large, unordered image networks with high complexity and diversity.

Furthermore, [Haala et al., 11] demonstrated that standard mapping products from airborne data acquisition like digital surface models (DSM) and ortho images could be generated well from low-cost UAV imagery, as can be seen in figure 1b. However, the imagery from such fixed-wing UAV systems has largely varying image overlaps due to the high flight dynamics and the relatively small footprint due to limitations of the currently employed consumer cameras.

The aim of this paper is to report a reliable and precise pipeline for fully automatic derivation of camera orientation from difficult imagery networks without initial orientation values. The following processing sequence is used: (1) Employ fast image indexing to avoid costly matching of all possible image pairs, which dominates computational complexity along with the multiple bundle adjustment steps. (2) Generate tie points by means of feature extraction and matching where the required automatic measurements are realized at maximum accuracy and reliability. (3) Identify reliable patches of images that have the mutual compatibility and optimize the geometry graph for each patch to ensure that the final tree is guaranteed to be optimal in minimizing the total edge cost. (4) Merge all patches and then finally adjust the full model with integrating the ground control points (if available).

The Structure and Motion (SaM) reconstruction approach, the core of this pipeline, was originally developed by the Computer Vision community to simultaneously estimate structure and camera motion from multiple images of a scene. SaM algorithms used for the derivation of exterior orientations for unorganized photo collections are used for the determination of initial values for the final bundle adjustment step in our pipeline.

Most SaM methods are starting with a small reconstruction, i.e. pair or triplet of images, and then expanding the bundle incrementally by adding new images and 3D points [Snavely et al., 07]. Here, each pose estimation and point triangulation is followed by an outlier rejection and a bundle adjustment. Other approaches increase the bundle hierarchically by merging smaller reconstructions [Farenzena et al. 09]. Unfortunately, both approaches require multiple intermediate bundle adjustment results and rounds of outlier removal to minimize error propagation as the reconstruction grows due to the incremental approach. This can be computationally expansive for large datasets. This issue is considered to be solved partially in [Farenzena et al. 09] by the introduction of a local bundle adjustment procedure and in [Snavely et al., 07] by optimizing the system over a graph to order the images and remove obsolete images from the dataset. However, we focus on datasets where one cannot reduce the number of images dramatically without losing a substantial part of the model. A third solution are so called partitioning methods [Gibson et al. 02] as used in [Nistér, D., 00, Klopschitz et al., 10], where the reconstruction problem is reduced to smaller and better conditioned sub-problems, represented by image triplets, which can be effectively optimized. The main advantage of these methods is the equalized error distribution on the entire dataset.

2. 3D RECONSTRUCTION PIPELINE OVERVIEW

Our 3D reconstruction pipeline is able to automatically process unordered sets of images to determine exterior camera orientations and a sparse point cloud of tie points without prior knowledge of the scene. The system mainly consists of four processing steps; starting with the initial network geometry analysis, followed by a pairwise matching step. After that, as shown in figure 2, the dataset is divided into optimal patches by using graphs. The reconstruction step is performed for each of

Figure 1: Low-cost sensors and its imagery. a) Five cameras rigidly mounted and protected by an aluminium frame. b) Fixed-wing UAV platform in flight, used consumer camera and mounting position on UAV belly. Right: Sparse point cloud and camera stations.

these patches separately. Finally, the results are stitched together and improved by a final common bundle adjustment.
A detailed description of the individual processing steps is given in the following sections. In general, calibrated camera settings are not strictly necessary for Euclidean 3D modelling, since self-calibration methods exist. However, if a stable camera is used with fixed focal length robustness and accuracy are usually greatly improved with values for the intrinsic orientation determined prior by standard calibration methods. Furthermore, also an increase in processing speed is achieved due to the lower dimensionality of the problem. Pursuant to that, we prefer to use intrinsic calibration parameters for high accuracy applications where these values can be considered to be stable.

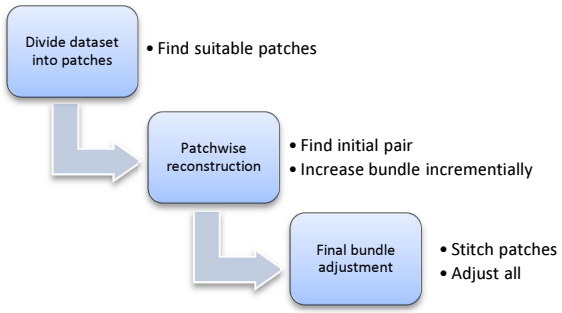

Figure 2: Flowchart of 3D reconstruction pipeline

2.1 Initial Network Geometry Analysis

This step is designed to accurately and quickly index unordered collections of photos. A connectivity matrix is the output of this step and is used as a heuristic about connections between the images. In addition, this connectivity matrix reveals singleton images and small subsets that should be excluded from the dataset. Finally, it is used to guide the process of pairwise matching (section 2.2) instead of trying to match every possible image pair.

Recent developments regarding this analysis can be distinguished into two major categories according to the type of image representation [Aly, M. et al. 10]. Local feature based approaches use quality measures of matched local descriptors while global feature based approaches utilize matching histograms of full images visual words. In fact, both categories represent the same approach with varying degrees of approximation to improve speed and/or storage requirements [Aly, M. et al. 10]. Generally, the first category provides superior recognition performance and the dimensionality is not an issue when only several thousands of images need to be processed. Consequently, we utilise a local feature based method in the pipeline presented in this paper. The first step is the extraction and description of local invariant features from each image by using the ***SIFT*** [Lo4we, 04] or ***SURF*** [Bay, H. et al., 05] operator on a downsampled image, e.g. using images with 2 Megapixels resolution.

For indexing, we follow an approach very close to the one presented in [Brown and Lowe 2003; Farenzena et al. 2009], where all the descriptors are stored in a randomized forest of kd-trees to improve the effectiveness of the representation in high dimensions. Then, each descriptor is matched to its k nearest neighbours in feature space. Therefore, we used the Fast Library for Approximate Nearest Neighbours FLANN [Muja, M. and Lowe, D., 09] and the kd-tree implementation in the VLFeat [Vedaldi, A., & Fulkerson, B., 08] library to find and analyse the 10 nearest neighbours. Afterwards, the weighted number of matches between each pair is stored in a 2D histogram where all matched features with a distance more than a certain threshold are deleted. We use 2δ as threshold where δ represents the standard deviation of the closest neighbours for each image. The inverse of the distances are used as weights. Furthermore, we introduce additional quality measures for possible connections between images such as the approximate image overlap derived from the convex hull of the matched feature points. The quality measures are normalized and summarized to one single quality value, which is stored in the index matrix (as shown in figure 3a). Finally, this index matrix is binarized using three thresholds to determine initial probable connections and disconnections (as shown in figure 3b).

2.2 Pairwise Feature Matching

Matching each connected image pair is accomplished using the connectivity matrix obtained during the previous step. Thus, corresponding 2D pixel measurements are determined between all connected image pairs. Afterwards, a weighted undirected geometry graph, $G\varepsilon = (V, E)$ where V is a set of vertices and E is a set of edges is constructed. Thus, two view relations are encoded such that each vertex refers to an image while each weighted edge presents the overlap between the corresponding image pair. The edges weights are stored according to the number of their shared matching points, w_{ij}^{p}, and the overlap area, w_{ij}^{a}, between view i & j. For the computation we follow the approach of [Farenzena et al. 09], where a set of candidate features are matched using a kd-tree procedure based on the approximate nearest neighbour algorithm. This step is followed by a refinement of correspondences using an outlier rejection procedure based on the noise statistics of correct/incorrect

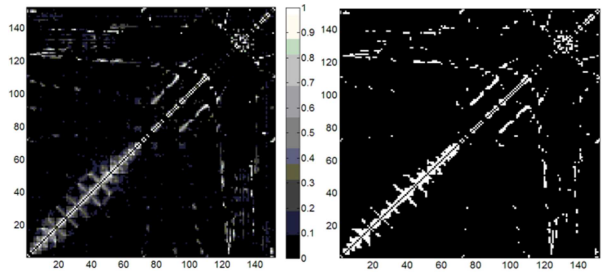

Figure 3: Top Patch of east tympanum dataset. Index matrix according to probabilistic model of relevance with 1457 edges, and adjacency connectivity matrix where the numbers of edges are reduced to 600.

matches. The results are then filtered by a standard RANSAC based geometric verification step, which robustly computes pairwise relations. Homography, ***H***, and fundamental, ***F***, matrices are used with an efficient outlier rejection rule called X84 [Hampel, F., et al. 1986] to increase reliability and accuracy. Finally, the best-fit (***H*** or ***F***) model is selected according to the Geometric Robust Information Criterion (GRIC) as initial model for the reconstruction. For an in-depth discussion see [Farenzena et al. 09; Snavely et al., 07] and references therein.

2.3 Graph of patches

In order to speed up the computation of the incremental reconstruction we address a fast local optimization instead of a global optimization approach. Therefore, we divide the dataset into n overlapping patches where each patch contains a manageable size of images. Thus, a parallelizable process replaces the process of reconstructing the whole scene at once where the large number of iterations with the growing number of unknowns can lead to very high computation times for complex datasets. The idea is to start from the most reliable part and use three images as the basic entity to extend each patch until a predefined size. In practice, we use the workflow as presented in ***algorithm 1*** to identify reliable patches with the highest mutual compatibility.

Algorithm 1: Building graph for patches

Input: geometry graph $G\varepsilon$
Output: collection of patches graph

1. Set new empty graph (patch) $G_p := \{\}$
2. Determine most reliable edge E_{ij} in $G\varepsilon$ which has max w_{ij}^p
3. Add $V_1, V_2,$ and E_{12} into G_p & set $E_{12} := 0$ in $G\varepsilon$
4. $\forall V_k$ in $G\varepsilon$ connected with two vertices $(V_i, V_j,)$ in G_p
   If w_{ik}^p & $w_{jk}^p \geq \max\left(\frac{1}{2}w_{ij}^p, 100\right)$ & w_{ik}^a & $w_{jk}^a \geq \frac{1}{2}w_{ij}^a$
   - Add V_k, E_{ik} & E_{jk} into G_p
   - Set E_{ik} & $E_{jk} := 0$ in $G\varepsilon$
5. Add edges in between inliers vertices in G_p & set all these edges $:= 0$ in $G\varepsilon$
6. Repeat steps 4,5 until $V_k = 0$ in step 4
7. Store G_p and repeat steps 1:6 until all edges in $G\varepsilon = 0$

3. PATCHWISE RECONSTRUCTION

Once the sub-graphs (patches) are calculated as described in the previous section, we can start the reconstruction process, as shown in figure 5.

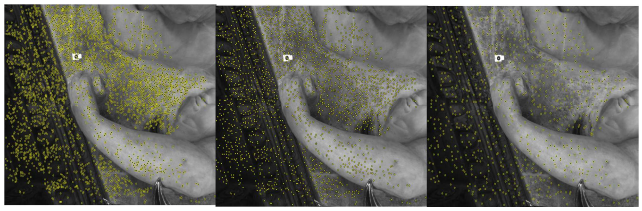

Figure 4: Point distribution in the image space before and after filtering (3395, 2007 and 819 points according to a filtering distance of 0, 20 and 40 pixels).

3.1 Optimize patch graph

For each patch we track the keypoints over all images in this patch and store the results in a visibility matrix, which depicts the appearance of points in the images. The results of this step will be the keypoints which have been correctly tracked in at least three images after rejecting those tracks as inconsistent in which more than one keypoint converges.

For more efficiency, we apply a homogeneous and dynamical filtering (see figure 4) approach for the tracked points to keep only the points with the highest connectivity. For each image we sort the keypoints in descending order according to their number of projections in other images. Then, the point with the greatest number of projections is visited, followed by an identification and rejection of all nearest neighbour points with a distance less than a certain threshold (e.g. 20 pixels). This step is repeated until the end of the points list. In order to maintain continuity, all points selected in an image must be considered as filtered (fixed) in the following filtering of other images. Filtering is done before the actual reconstruction step (section 3.1) in order to increase the accuracy but also to reduce the number of obsolete observations. Consequently, the geometric distribution of keypoints is improved, which reduces computational costs significantly without losing geometric stability.

Once correspondences have been tracked and filtered, we optimize the patch graph such that we construct a weighted undirected epipolar graph for each patch G_p containing common tracks. The weight w_{ij} of an edge represents the number of common points between the corresponding image pair. Then we build G_r, the edge dual graph of G_p, where every node in G_r corresponds to an edge in G_p. Two nodes in G_r are connected by an edge if and only if the corresponding image pairs share a camera and 3D points in common. Thus, each edge represents an image pair with sufficient overlap. Note that even when G_p is fully connected any spanning tree of G_r may be disconnected. This can happen if a particular pairwise reconstruction did not have 3D points in common with another pair. Thus, we use three images as basic geometric entity by using only points that were tracked in at least three images.

These points are used to build the graph in order to guarantee full connection for any sub sequential image. The *maximum spanning tree (MST)*, which minimizes the total edge cost of the final graph is then computed. The image relation retrieved as G_p^{max} graph is used for the bundle adjustment. For example, figure 7 presents the results of the top patch of east tympanum where the previous process reduced the pairwise connection from 600 edges (see figure 3) to 396 to orient 150 images.

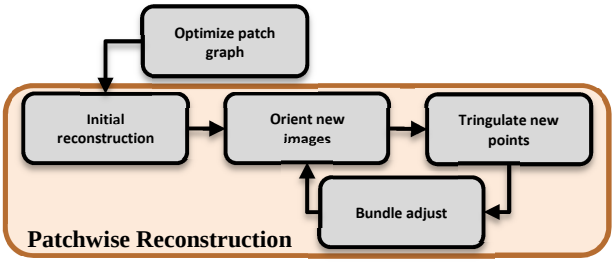

Figure 5: Block diagram of the optimized SaM pipeline.

3.2 Patchwise Reconstruction

As shown in figure 5, each patch is processed individually. Therefore an initial reconstruction is performed for the two, most suitable images of the patch. After this step orientations and tie points in object space are available for these two images where one image defines the local coordinate system. Within the incremental approach images are added to the existing bundle by triangulating new points, rejecting outliers and performing another iteration of the bundle adjustment. This incremental process is repeated until the maximum of stable imagery can be oriented.

Reconstruction of the initial pair

The incremental reconstruction step begins with the reconstruction of orientation and 3D points for an initial image pair. The choice of this initial pair is very important for the following reconstruction of the scene. The initial pair reconstruction can only be robustly estimated if the image pair has at the same time a reasonable large baseline for high geometric stability and a high number of common feature points. Furthermore, the matching feature points should be distributed well in the images in order to reconstruct a maximum of initial 3D structure of the scene and to be able to determine a strong relative orientation between the images.

Therefore, suitable image pairs should be selected according to the following conditions: the number of matching points is acceptable and the fundamental matrix must explain the matching points far better than homography models. In order to guarantee that GRIC scores are employed as used in [Pollefeys et al. 02 and Farenzena et al. 09).

After that, extrinsic orientation values are determined for this initial pair by factorizing the essential matrix and the tracks visible in the two images. A two-frame bundle adjustment starting from this initialization is performed to improve the reconstruction.

Adding new images and points

After reconstructing the initial pair additional images are added incrementally to the bundle. The most suitable image to be added is selected according to the maximum number of tracks from 3D points already being reconstructed. Within this step not only this image is added but also neighbouring images that have a sufficient number of tracks as mentioned in [Snavely et al., 07]. Adding multiple images at once reduces the number of required bundle adjustments and thus improves efficiency.

Next, the points observed by the new images are added into the optimization. A point is added if it is observed by at least two images, and if the triangulation gives a well-conditioned estimate of its location. This procedure follows the approach of [Snavely et al., 07].

Figure 6: East tympanum at the palace from distance and the sensor stabilized by an arm during the data acquisition.

Sparse bundle adjustment

Once the new points have been added, a bundle adjustment is performed on the entire model. This procedure of initializing a camera orientation, triangulating points, and running bundle adjustment is repeated, until no images observing a reasonable number of points remain. For the optimization we employ the sparse bundle adjustment implementation “SBA” [Lourakis, A., & Argyros, A., 09]. SBA is a non-linear optimization package that takes advantage of the special sparse structure of the Jacobian matrix used during the optimization step in order to provide a computation with reduced time and memory requirements.

4. STITCHING OF PATCHES AND GLOBAL ADJUSTMENT

After the reconstruction of points and orientations for the overlapping patches the results are merged. Since outlier rejection was performed within the previous processing the available 3D feature points are considered to be reliable and accurate. Due to the overlap the patches have a certain number of points and camera orientations in common which enable the determination of a seven-parameter transformation to align the patches into a common coordinate system. The transformed orientations and points are introduced into a common global bundle adjustment of the whole block. If ground control point measurements are available they can be used for the improvement of the bundle and to enable georeferencing.

5. EXPERIMENTAL RESULTS

During a cultural heritage data recording project 10k images were acquired using a multi-camera rig. The objective was the derivation of a point cloud with the resolution of 1mm and sub-mm accuracy of two large reliefs covering an area of about 125m² (see figure 6). Therefore, a rig with 5 industry cameras was used to record an image collection with high overlap efficiently. For the derivation of the point cloud the exterior orientations were derived using the presented method, followed by an additional dense image matching step relying on these orientations with very high accuracy requirements. Ground control point measurements provided the georeference. The interior orientation parameters were determined prior by a standard calibration method employing a calibration pattern and a free network adjustment.

The Structure and Motion reconstruction with such large number of images and high accuracy requirements is a computationally expensive step due to the high number of points and possible connections. Therefore, we used the proposed approach of splitting the dataset into patches. If all patches are reconstructed the dataset can be merged and adjusted at once. Furthermore, the initial network analysis is important to speed up the matching process of the feature

Figure 7: Adjacency $\mathbf{G_p^{max}}$ matrix for top part (Patch) of east tympanum with 396 edges and the reconstructed model – only the registration camera (150 images with 2 mega pixels resolution) used.

points. The point reduction technique employed within the pipeline was used to derive a homogeneous point distribution while saving overhead due to obsolete information.

By the presented pipeline the accurate orientations could be determined in a reasonable processing time (see figure 8). Finally, about 2 billion points were derived by the dense image matching step as shown in figure 8. A detailed report about this project and the applied methods is given by [Fritsch et al., 11 & Wenzel et al., 11].

6. CONCLUSION

The presented pipeline for the reconstruction of orientations and surface information is specifically designed for the efficient processing of large datasets with high accuracy requirements. An initial network analysis is used along other techniques to realize a reasonable processing time while adjusting a stable bundle containing information from a maximum number of images. Thus, it is specifically suitable for large scale photogrammetric applications at low costs.

7. REFERENCES

Aly, M., Munich, M., and Perona, P., 2011. Indexing in Large Scale Image Collections: Scaling Properties and Benchmark, 2011. IEEE Workshop on Applications of Computer Vision (WACV).

Bay, H., Ess, A., Tuytelaars, T., Luc, V., G., 2008. SURF: Speeded Up Robust Features, Computer Vision and Image Understanding (CVIU), pp. 346-359.

Brown, M. and Lowe, D., 2003. Recognizing panoramas. In Proceedings of the 9th International Conference on Computer Vision, Vol. 2, pp. 1218–1225.

Farenzena, M., Fusiello, A. and Gherardi, R., 2009. Structure and motion pipeline on a hierarchical cluster tree. ICCV Workshop on 3-D Digital Imaging and Modeling, pp. 1489–1496.

Fritsch, D., Khosravani, A. M., Cefalu, A. and Wenzel, K., 2011. Multi-Sensors and Multiray Reconstruction for Digital Preservation, Photogrammetric Week 2011, Wichmann Verlag, Berlin/Offenbach, pp. 305-323.

Gibson, S., Cook, J., Howard, T., Hubbold, R. and Oram, D., 2002. Accurate Camera Calibration for Off-Line, Video-Based Augmented Reality. IEEE and ACM Int'l Symp. Mixed and Augmented Reality, pp. 37-46.

Haala, N., Cramer, M., Weimer, F. and Trittler, M., 2011. Performance Test on UAV-based data collection. Proc. of the International Conf. on UAV in Geomatics. IAPRS, Volume XXXVIII-1/C22, 2011.

Hampel, F., Rousseeuw, P., Ronchetti, E. and Stahel, W., 1986. Robust Statistics: The Approach Based on Influence Functions, ser. Wiley Series in probability and mathematical statistics. New York: Wiley, 1986.

Irschara, A., Kaufmann, V., Klopschitz, M., Bischof, H., and Leberl, F., 2010. Towards fully automatic photogrammetric reconstruction using digital images taken from UAVs. Proc. of the ISPRS Symposium.

Klopschitz, M., Irschara, A., Reitmayr, G., and Schmalstieg, D., 2010. Robust incremental structure from motion. Proc. 3DPVT.

Lourakis, A., and Argyros, A., 2009. Sba: A software package for generic sparse bundle adjustment. ACM TOMS, 36(1):pp. 1–30.

Lowe, D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, Vol. 60, pp. 91-110

Muja, M. and Lowe, D., 2009. Fast approximate nearest neighbors with automatic algorithmic configuration. Proc. VISAPP.

Nistér, D., 2000. Reconstruction from Uncalibrated Sequences with a Hierarchy of Trifocal Tensor. Proc. European Conf. Computer Vision, pp. 649-663.

Schwartz, C., Klein, R., 2009. Improving initial estimations for structure from motion methods. In Proc. CESCG, April 2009.

Snavely, N., Seitz, S. and Szeliski, R., 2007. Modeling the world from internet photo collections. International Journal of Computer Vision.

Vedaldi, A., and Fulkerson, B., 2008. VLFeat: An open and portable library of computer vision algorithms.

Wenzel, K., Abdel-Wahab, M., Fritsch D., 2011. A Multi-Camera System for Efficient Point Cloud Recording in Close Range Applications. LC3D workshop, Berlin, December 2011.

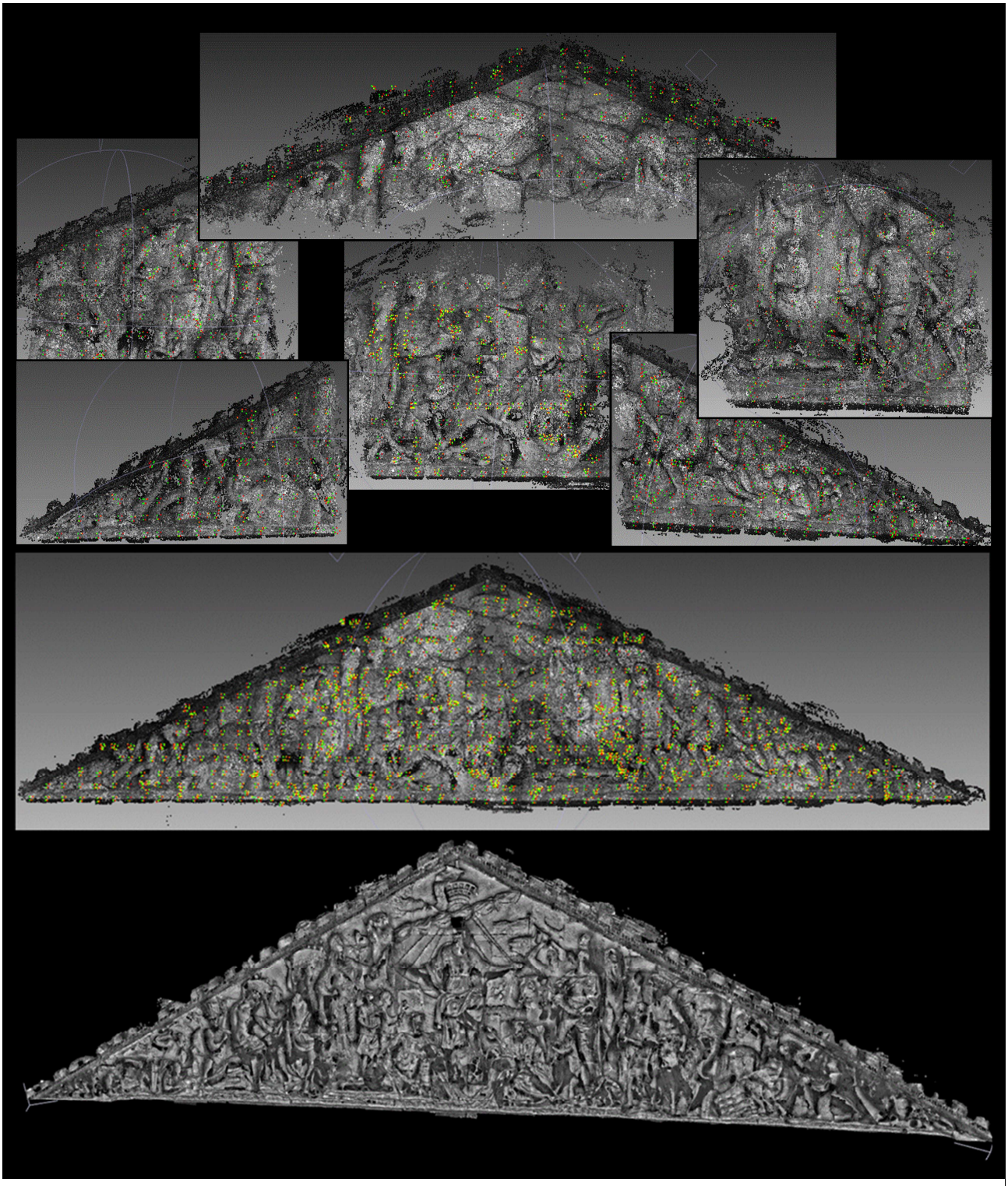

Figure 8: The reconstructed point clouds of the west tympanum. About $4k$ images from four different cameras of a rig - three for dense matching (5 mega pixels) and the fourth with larger field of view (2 mega pixels) for registration only. The first row shows the reconstructed 6 patches in a local coordinate system defined by the initial pair during the reconstruction step (mean reprojection errors around 1 pixel). The second row show the full sparse cloud of $\approx 1.1\ milion$ feature points in an object coordinates for all stitched patches after the final bundle adjustment step (mean reprojection errors reduced to 0.5 pixel). Finally, the dense point cloud derived by a subsequential dense image matching step with about 1.1 billion points.