Photogrammetric Reconstruction of a Stolen Statue

Z. Liu[1,2] and E.L. Doubrovski[1] and J.M.P. Geraedts[1] and W. Wang[3,2] and Y. Yam[2] and C.C.L. Wang[4]

[1] Delft University of Technology, the Netherlands [2] Centre for Perceptual and Interactive Intelligence (CPII), Hong Kong, China
[3] The Chinese University of Hong Kong, Hong Kong, China [4] The University of Manchester, UK

Abstract

In this paper, we propose a method to reconstruct a digital 3D model of a stolen/damaged statue using photogrammetric methods. This task is challenging because the number of available photos for a stolen statue is in general very limited – especially the side/back view photos. Besides using standard structure-from-motion and multi-view stereo methods, we match image pairs with low overlap using sliding windows and maximize the normalized cross-correlation (NCC) based patch-consistency so that the image pairs can be well aligned into a complete model to build the 3D mesh surface. Our method is based on the prior of the planar side on the statue's pedestal, which can cover a large range of statues. We hope this work will motivate more research efforts for the reconstruction of those stolen/damaged statues and heritage preservation.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction; Mesh models;***

1. Introduction

The logo of TU Delft (Delft University of Technology, The Netherlands) is the flame of Prometheus. There was an over-life-sized bronze statue of Prometheus (see Fig. 1) created by Oswald Wenckebach (1895–1962) for the university's 100th anniversary and placed on the campus. However, the statue was stolen from its concrete pedestal on the night of 10/11 January 2012 and has not been found anymore.

Since there is very little chance to find the stolen statue, people try to reconstruct it digitally by using photos taken in the past years. The work of [Ren15] is an early effort to tackle this problem, which reconstructed an incomplete 3D point cloud (its Sec. 9.2). The difficulties of this reconstruction problem are summarized as:

- The number of photos is limited (e.g., only 34 photos are used in the Appendix J.1 of [Ren15]);
- The object's appearance changes because of the dirt (mainly tarnish) and the variation of lighting conditions in the photos taken at different times spanning over decades;
- The camera setup is unknown, without any precise information about the intrinsic or extrinsic parameters.

(a) (b) (c)

Figure 1: *The Prometheus statue was placed in two locations (a–b) on the campus of TU Delft during different periods. The pedestal (c) remained in the second location after the statue was stolen. All these photos are obtained from [vdKvdK10]. (a) and (b) are the #13th and #22nd photos in this paper's dataset.*

When using the state-of-the-art system based on photogrammetry (e.g., COLMAP [SF16, SZFP16]) and the dataset 1 of [Ren15] (listed in this paper's supplementary material), we found the 3D model of the Prometheus statue can be reconstructed as two sub-models (see Fig.2(b)), the main challenge is how to align them into the same coordinate frame. In a standard pipeline of photogrammetry, local features are extracted from images and matched in the first step. In this example, the matched photos are distributed into two clusters, where the matched features between clusters are too few to estimate a valid geometric relation. Photos from side-views cannot be well registered with those already matched photos shown in Fig.2(b) since there is a big angular gap between the front-view cluster and the back-view cluster with around 90°. Techniques employed in the subsequent steps of reconstruction, including structure-from-motion (SfM) [SF16] and multi-view stereo (MVS) [SZFP16], cannot align two sub-models easily.

Although there is little 3D overlap between the two reconstructed sub-models, we observe that the left and right sides of the pedestal are partially observed in both the front and the back view clusters. We can align two clusters by matching 2D points in the side re-

DOI: 10.2312/egs.20231000

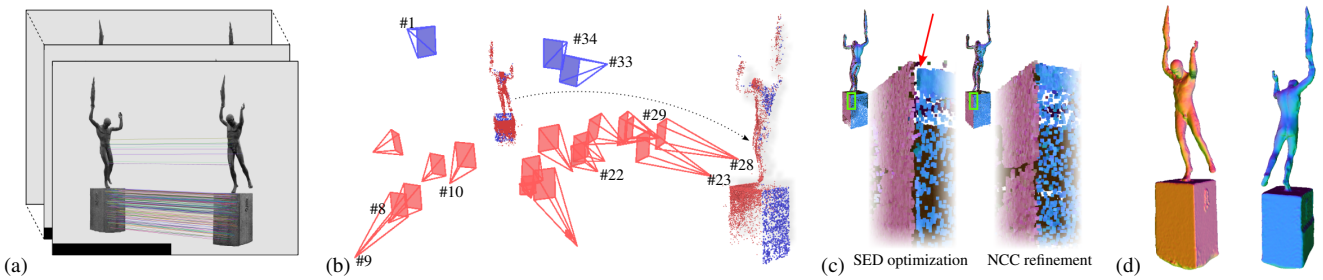


Figure 2: *Reconstruction pipeline for the Prometheus statue. (a) SIFT features are matched exhaustively between all image pairs. (b) Two sparse sub-models are reconstructed by SfM: the front sub-model (the red point cloud and cameras) and the back sub-model (the blue, manually aligned to the front model), between which there is little overlap, as shown in the zoom-in view. (c) Dense models can be reconstructed by MVS when two sub-models are aligned together. There is a seam (red arrow on left) since the alignment obtained by symmetric epipolar distance (SED) optimization is not accurate enough. The seam is eliminated after refinement with NCC (right, details are given in Sec. 2.3). (d) A mesh is reconstructed from the dense point cloud. The point clouds (c) and the mesh (d) are colored by their vertex normals.*

gions of the pedestal. Since the sides of the pedestal are basically planar, we transform the local patches from one image to another and find their optimal correspondences by evaluating the normalized cross-correlation (NCC) in sliding windows. Experiment results demonstrate the effectiveness of this method. It is unusual to use NCC in the image matching step for SfM as it is not invariant to scaling and orientation. However, we found it is very effective in the special case with a planar prior. This can be generalized to solve the reconstruction problem for other stolen/damaged statues with similar shapes – i.e., planar sides for the statue's pedestal.

2. Method

Our reconstruction mainly relies on the photogrammetry system COLMAP with the additional NCC-based alignment between the front-view and the back-view clusters (Sec. 2.3). The NCC-based alignment is the most critical step for a complete reconstruction.

2.1. Feature Extraction and Matching

Dataset 1 of [Ren15] has 34 photos of the statue. We remove #26 (i.e., a cropped #27) and use the rest 33 photos. As that work suggested, we manually segment the images and only keep the statue with the pedestal. SIFT features are extracted and matched (Fig. 2(a)) exhaustively considering the number of images is small.

2.2. Incomplete SfM

We run SfM using the 2D correspondences obtained in the previous step. Two sparse sub-models are then reconstructed. Each sub-model is a sparse 3D point cloud $\mathbf{P}$ and multiple images, each of which has estimated (intrinsic and extrinsic) camera parameters and inlier 2D feature points (2D projections of $\mathbf{P}$'s subset). The front sub-model contains 18 images (i.e., the red ones in Fig. 2(b)) and the back sub-model has 3 images (i.e., the blue ones in Fig. 2(b)).

It is not straightforward to align two sub-models because they do not share any image. Although there are side view photos (e.g., #2 and #30), these ones are not registered since their appearances are quite different. The result of feature matching with these views (e.g., #1 from the back cluster - #2 - #8 from the front cluster) is not good enough to estimate valid two-view geometries (fundamental matrices).

2.3. Front-Back Alignment by NCC

Now we have the front and back sparse sub-models. We need to estimate a similarity transformation to align the back model into the coordinate frame of the front model. There are two pairs of images that can help to construct the relation between the sub-models. One pair is image #8 (front) and #1 (back), both of which capture the right side of the Prometheus statue's pedestal. The other pair is image #29 (front) and #34 (back) which capture the left side. In the previous steps, SIFT features can not find sufficient correspondence between them. In this step, we hypothesize the pedestal sides are planar and guide the matching by measuring NCC-based consistency between sliding windows.

2.3.1. Rough Alignment Transformation

We start from an alignment transformation $\tilde{T}$ before guiding the NCC-based matching. The alignment transformation $\tilde{T}$ is a similarity consisting of a scalar scale factor $\tilde{s}$, a rotation matrix $\tilde{\mathbf{R}}$, and a translation vector $\tilde{\mathbf{t}}$. It transforms the back point cloud $\mathbf{P}_b$ to the coordinate frame of the front point cloud $\mathbf{P}_f$ by

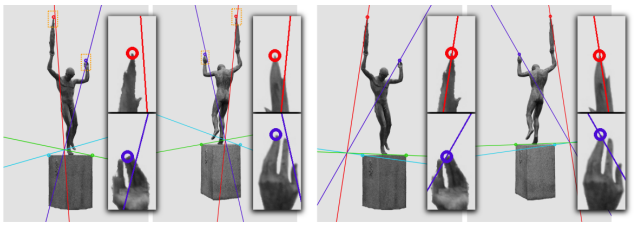
$$\mathbf{P}_{b2f} = \tilde{s}\tilde{\mathbf{R}}\mathbf{P}_b + \tilde{\mathbf{t}}. \tag{1}$$

If a camera's extrinsic matrix is $(\mathbf{R}_b|\mathbf{t}_b)$ in the back model, then it should be

$$(\mathbf{R}_{b2f} \mid \mathbf{t}_{b2f}) = \left(\tilde{\mathbf{R}}\mathbf{R}_b \;\middle|\; \tilde{s}\mathbf{t}_b - \mathbf{R}_b\tilde{\mathbf{t}}\right) \tag{2}$$

in the front model, transformed by the alignment similarity. The intrinsic matrix remains unchanged.

Such a similarity $\tilde{T}$ can be estimated, denoted as $\tilde{T}_0$, by manually picking at least three pairs of non-collinear 3D-3D correspondences between $\mathbf{P}_f$ and $\mathbf{P}_b$. However, the overlap between two SfM-reconstructed point clouds is too little to find enough 3D-3D matches for accurate estimation (Fig. 2(b)). We optimize $\tilde{T}_0$ by minimizing *symmetric epipolar distances* (Eq. (3)) on further picked 2D-2D correspondences (2D point-to-point distances can not be evaluated directly without depth information). That is to say, we



(a) $\tilde{T}_0$ before optimization (b) SED optimized $\tilde{T}_1$

Figure 3: *A rough alignment from the back model to the front is optimized by minimizing the symmetric epipolar distance (SED) on picked points (small circles here). We manually picked 9 pairs of points in total on 3 front/back image pairs: #8/#34 (this figure), #8/#1, and #29/#1. The distances from points to their epipolar lines are reduced after optimization (see zoom-in views in (a–b)).*

pick several 2D-2D correspondences on front-back image pairs. The 2D points are selected at unique corners such as the flame tip, fingertips, and pedestal corners, as shown in Fig. 3. A similarity $\tilde{T}$ defines a fundamental matrix $F_{f,b}$ between a front image and a back image (using Eq. (2)), which is sufficient to evaluate the distance $\mathrm{d}(\mathbf{p}, F\mathbf{p})$ from a 2D point $\mathbf{p}$ to the corresponding epipolar line $(F\mathbf{p})$. Therefore, the objective is to find

$$\tilde{T}_1 = \arg\min_{\tilde{T}} \sum_{f,b} \sum_{\mathbf{p}_f,\mathbf{p}_b} \mathrm{d}^2\left(\mathbf{p}_f, F_{f,b}\,\mathbf{p}_b\right) + \mathrm{d}^2\left(\mathbf{p}_b, F_{f,b}^{\top}\,\mathbf{p}_f\right). \quad (3)$$

To minimize the objective Eq. (3), we use CMA-ES (Covariance Matrix Adaptation Evolution Strategy), a gradient-free optimizer, as an ad hoc solution, which is acceptable to guide the NCC-based matching.

If we use the $\tilde{T}_1$ estimated to align the sparse models and take the next step MVS, there will be a visible seam on the reconstructed pedestal (Fig. 2(c) left), which encourages us to further refine the result.

2.3.2. Alignment Refinement by NCC

Now we have reconstructed the 3D geometry of the pedestal sides, which are (nearly) planar. Given a 3D region on a side (e.g., the right side), we can project it onto a front/back view pair (e.g., #8/#1 or #29/#34, denoted as #f/#b for generality) that both observe the region, then find a more accurate matching by sliding the projections. The consistency between projections is evaluated by NCC, which is robust to illumination variations. Though NCC is not invariant to geometric transformations (scaling, rotation, and translation), our pre-estimated $\tilde{T}_1$ can handle the transformation between projected patches.

We segment the pedestal side region on image #f and get a nearly planar sparse point cloud $\mathbf{P}$ from the front sub-model corresponding to the segmented region. The plane's equation is estimated by RANSAC. It is spanned by a pair of orthogonal vectors, denoted as $\mathbf{u}, \mathbf{v}$ (the decomposition is not unique but its choice is inconsequential).

For each reconstructed 3D point $\mathbf{p}$ on $\mathbf{P}$, a small square centered at $\mathbf{p}$ is constructed by four 3D vertices $\mathbf{p} \pm r\mathbf{u} \pm r\mathbf{v}$, in which r controls its size. We project the square onto image #f and #b, get-

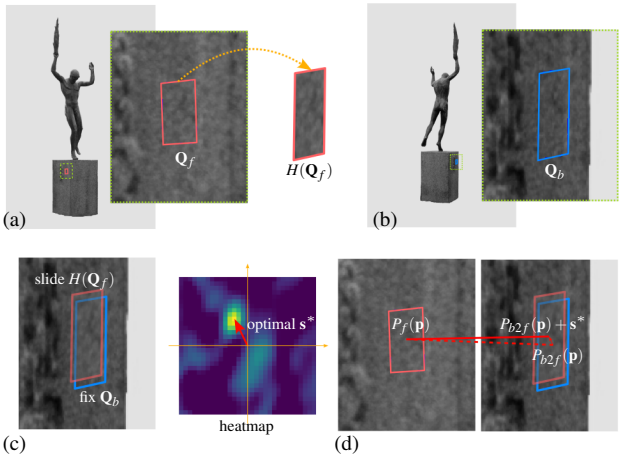


Figure 4: *NCC-based alignment refinement. The planar side of the pedestal is observed in front/back images #f/#b (e.g., #8/#1 in (a)/(b)). For a 3D point $\mathbf{p}$ on the side, a square patch around it is projected as quadrilaterals $\mathbf{Q}_f$ (a) and $\mathbf{Q}_b$ (b) on images #f/#b. Image patch in $\mathbf{Q}_f$ is transformed as $H(\mathbf{Q}_f)$ to match the shape of $\mathbf{Q}_b$ and compared to $\mathbf{Q}_b$. We fix $\mathbf{Q}_b$ and slide $H(\mathbf{Q}_f)$ in $\mathbf{Q}_b$'s $n \times n$ neighborhood. An $n \times n$ heatmap (c) is then obtained by evaluating the NCC scores on the sliding windows. If the maximum of the heatmap is not at the center, the 2D matching should be shifted (from the dashed line to the solid line in (d)). Then the correspondence between #f / #b is refined.*

ting two 2D quadrilaterals $\mathbf{Q}_f$ and $\mathbf{Q}_b$, between which there exists a homography transformation H.

We warp the image patch $\mathbf{Q}_f$ (converted as grayscale for simplicity) to $H(\mathbf{Q}_f)$, which has a pixel-wise correspondence to patch $\mathbf{Q}_b$ (Fig. 4(a–b)). When sliding $H(\mathbf{Q}_f)$ by a 2D shift vector $\mathbf{s} \in [-w, w] \times [-w, w]$, using one pixel as the step-length, the consistency between patches can be evaluated by NCC

$$\mathrm{NCC}_{f,b}(\mathbf{s}) = \frac{(\mathbf{f} - \bar{\mathbf{f}}) \cdot (\mathbf{g} - \bar{\mathbf{g}})}{\|\mathbf{f} - \bar{\mathbf{f}}\| \|\mathbf{g} - \bar{\mathbf{g}}\|}, \quad (4)$$

in which $\mathbf{f}$ is the flattened pixel array in $H(\mathbf{Q}_f)$, $\mathbf{g}$ is the flattened pixels in $\mathbf{Q}_b + \mathbf{s}$, and $\bar{\mathbf{f}}$ is the mean value of array $\mathbf{f}$.

For each 3D point $\mathbf{p}$, we compare the quadrilateral shaped sliding-windows and get a $(2w+1) \times (2w+1)$ heatmap (Fig. 4(c)). If $\mathbf{s}^*$ maximizes Eq. (4), then there is a triplet of 3D-2D-2D correspondence between the 3D point $\mathbf{p}$, 2D point $P_f(\mathbf{p})$ on image #f, and $P_{b2f}(\mathbf{p}) + \mathbf{s}^*$ on image #b (Fig. 4(d)). The projection $P_f(\cdot)$ is the camera projection of image #f and $P_{b2f}(\cdot)$ is the camera projection of image #b transformed by $\tilde{T}_1$ (see Eq. (2)).

We get a set of 3D-2D-2D correspondences after looping over all $\mathbf{p} \in \mathbf{P}$. We remove outlier correspondences by estimating the extrinsic parameter of image #b using RANSAC.

This process is repeated for front/back image pairs #8/#1 and #29/#34. The square size r is chosen to make the projected patches be around 50×50. The sliding range w is 15. The optimal shift vector $\mathbf{s}^*$ is estimated to have sub-pixel accuracy. On the right side of the pedestal, 64 out of 253 points are kept as inliers after RANSAC

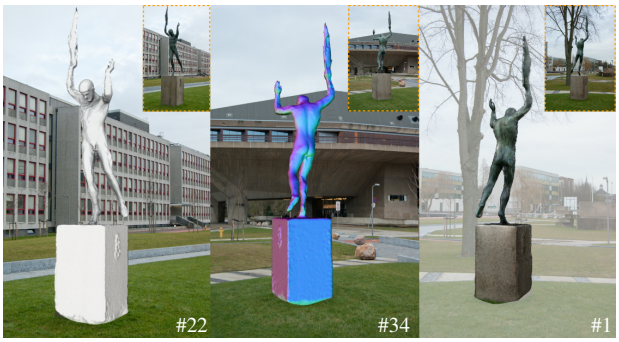


Figure 5: *The reconstructed mesh of the statue is rendered by ambient occlusion (#22), vertex normal (#34), or texture mapping (#1), overlaid with the original photos [vdKvdK10].*

extrinsic parameter estimation. On the left side, 42 out of 283 points are kept. The result helps to reconstruct a complete sparse model successfully and produces a visually 'seamless' dense point cloud, as shown in the complete SfM and MVS of the next step (Fig. 2(c) right).

2.4. Complete SfM, MVS, and Mesh Reconstruction

Now we have two sets of 2D correspondences between images. One is the inlier results of the incomplete SfM step. The other is obtained by NCC-based matching. Both sets of image matching results are used for SfM again. This time the SfM is complete since all images in either the front model or the back model are registered in the same coordinate system, though those images that are not registered in incomplete models still do not appear here.

Next, we use MVS to reconstruct a dense point cloud. In COLMAP, MVS estimates the depth and normal maps of each view by matching patches with other views, in which the patch consistency is evaluated by weighted NCC. We observe that it is not good to use all images to reconstruct a dense front model, because of the appearance inconsistency caused by a statue's cleaning in 2009 (compare Fig. 1 (a) v.s. (b)). Therefore, we only use the registered images taken after 2009 (10 images with indices in Fig. 2(b)) to reconstruct a dense model.

After obtaining a dense point cloud with MVS, we reconstruct the mesh (Figs. 2(d) and 5) using Poisson surface reconstruction.

3. Conclusion and Discussion

Reconstructing stolen/damaged statues is difficult because of lacking images, especially side-view images. We use an NCC-based matching method to align two partial reconstructions (front and back sub-models) into a complete one, after a manual initialization and optimization. The final reconstruction result on the TU Delft Prometheus statue (see Fig. 2(d) and Fig. 5) is visually reasonable. The average reprojection error of the complete sparse model is 0.277 pixels, evaluated on 10 images used for MVS. We are not able to quantitatively evaluate the error in 3D because the original shape is missing.

There are several directions to further improve the result. First of all, the pedestal can be more accurately reconstructed since it is still there, and then the photo can be registered onto it. Shape-from-shading techniques [XNSW19] can be used to enhance the geometric details using close-up photos #14 and #25. Meanwhile, recent deep learning advances can be used. The two most relevant topics are deep learning on MVS and neural reconstruction, both of which still rely on camera parameters estimated beforehand. Skilled 3D modelers can edit the mesh by taking the registered photos as references. The reconstructed model can be displayed by augmented reality (AR) [HLO*20] to place it virtually at its original spot. The proposed method can be generalized to reconstruct other shapes. If the 3D geometry of a region (whether planar or not) can be reconstructed from some views, then the region can be projected to a new view and the extrinsic parameters of the new view can be refined by maximizing NCC in sliding windows. The proposed method is more robust to large rotation angles than SIFT-like features when the surface is partially reconstructed.

We hope this work will motivate more research efforts for the reconstruction of those stolen/damaged statues and the problem of heritage preservation in general.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This project is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Limited in Hong Kong. We would like to acknowledge everyone who contributed to taking and collecting the photos of the Prometheus statue. Especially, thanks to René & Peter van der Krogt [vdKvdK10] for their written permission for using photos #1, 4, 7–10, 13, 15, 16, 22, 23, 25–29, and 34; and thanks to Ad Bercht for his written permission for using photos #24 and 33.

References

[HLO*20] HENNEMAN D., LI Y., OCHSENDORF J., BETKE M., WHITING E.: Augmented Reality for Sculpture Stability Analysis and Conservation. In *Eurographics Workshop on Graphics and Cultural Heritage* (2020), The Eurographics Association. doi:10.2312/gch.20201298. 4

[Ren15] RENKENS I. M.: *Prometheus: From 2D to 3D. A reconstruction based on photographs.* Master's thesis, Delft University of Technology, 2015. URL: http://resolver.tudelft.nl/uuid:7d2832e6-9921-4ad3-a925-0f17703a5894. 1, 2

[SF16] SCHÖNBERGER J. L., FRAHM J.-M.: Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4104–4113. doi:10.1109/CVPR.2016.445. 1

[SZFP16] SCHÖNBERGER J. L., ZHENG E., FRAHM J.-M., POLLEFEYS M.: Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision – ECCV 2016* (Cham, 2016), Springer International Publishing, pp. 501–518. doi:10.1007/978-3-319-46487-9_31. 1

[vdKvdK10] VAN DER KROGT R., VAN DER KROGT P.: Photos of the TU Delft Prometheus statue. https://standbeelden.vanderkrogt.net/object.php?record=ZH14ar, 2003–2010. 1, 4

[XNSW19] XIE W., NIE Y., SONG Z., WANG C. C. L.: Mesh-based computation for solving photometric stereo with near point lighting. *IEEE Computer Graphics and Applications 39*, 3 (2019), 73–85. doi:10.1109/MCG.2019.2909360. 4