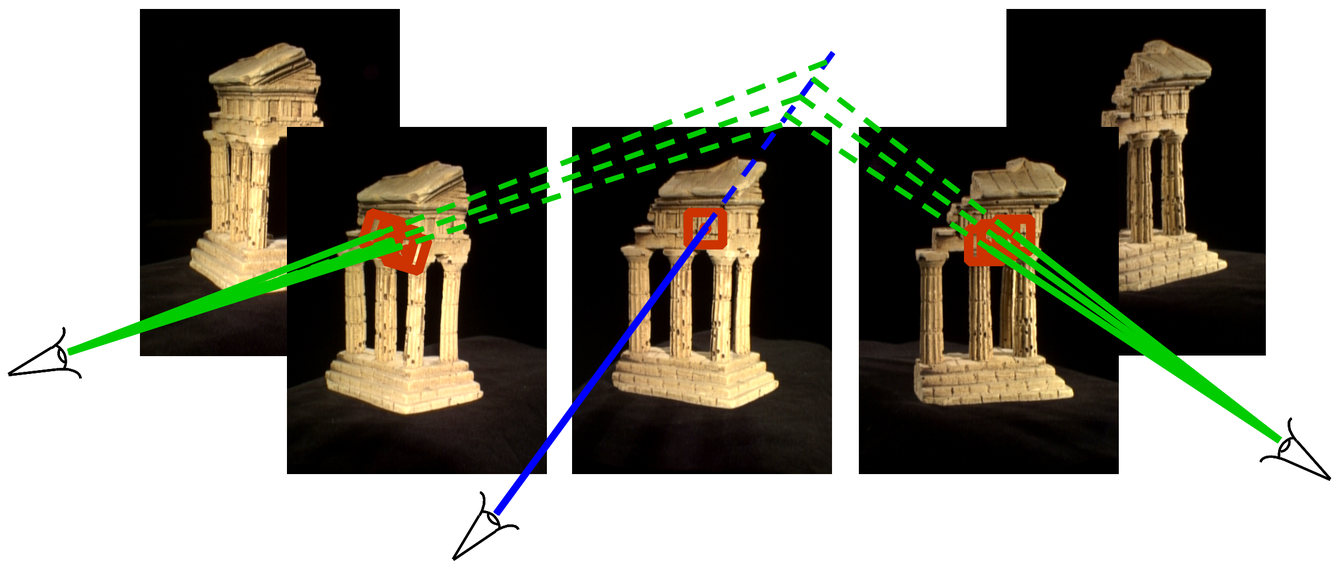
Multi-View Stereo Revisited

Michael Goesele, Brian Curless, and Steven M. Seitz
Department of Computer Science and Engineering
University of Washington, Seattle, WA

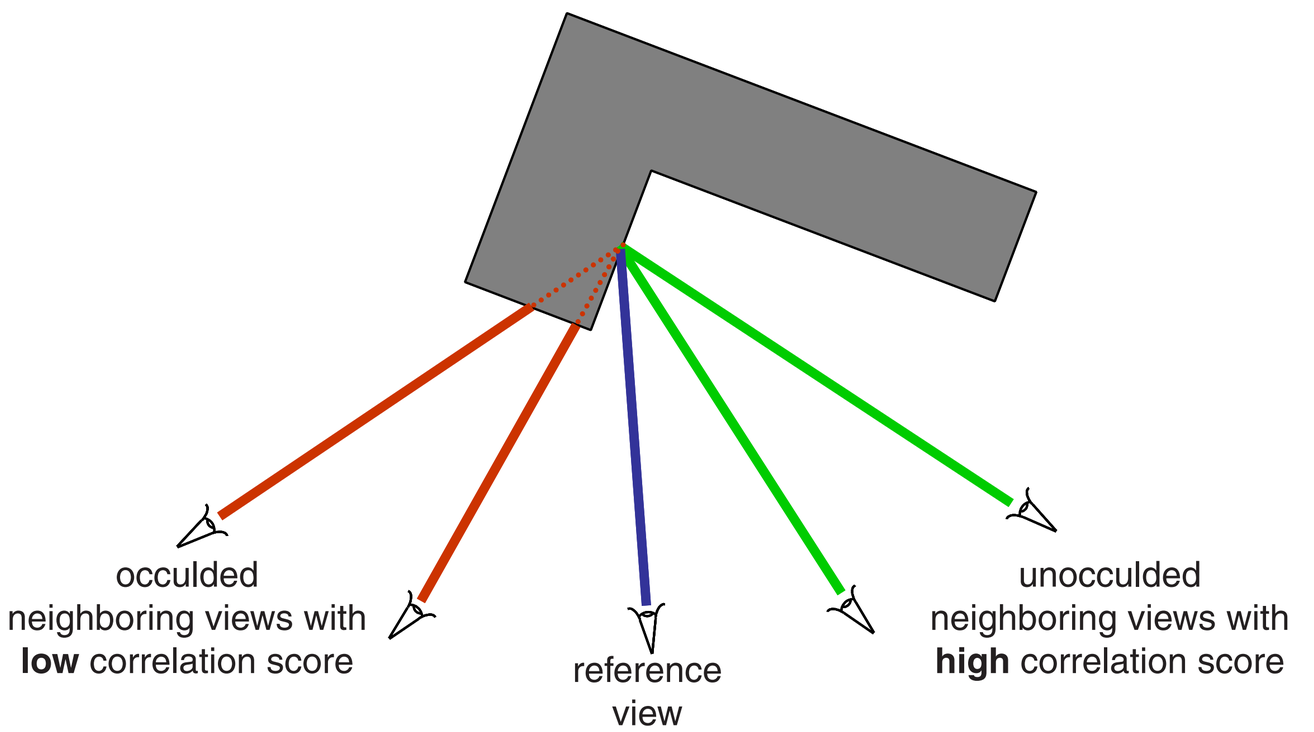
Abstract
We present an extremely simple yet robust multi-view stereo algorithm and analyze its properties. The algorithm first computes individual depth maps using a window-based voting approach that returns only good matches. The depth maps are then merged into a single mesh using a straightforward volumetric approach. We show results for several data sets, showing accuracy comparable to the best of the current state of the art techniques and rivaling more complex algorithms.



Principle of window matching. A fixed window around a pixel *p* in a reference view *R* (blue) is compared to its projection in the neighboring images at various depths *d* using the NCC score.

Key Ideas

1. Implicit view selection: Compute overall correlation score only based on views with high pairwise correlation with the reference view (similar to Hernandez and Schmitt [3] and Pollefeys et al. [5]). Occluded views are automatically rejected based on their low correlation score.
2. Reconstruct only scene parts that are matched with high confidence. This leads to holes at or near silhouettes, oblique surfaces, occlusions, highlights, and low-textured areas but avoids problems due to outliers as in Narayanan et al. [4]. The complete object or scene geometry can be recovered by combining information from all input views.



Implicit view selection: Occluded views have typically a low correlation score and do not contribute to the correlation score *corr(d)*.

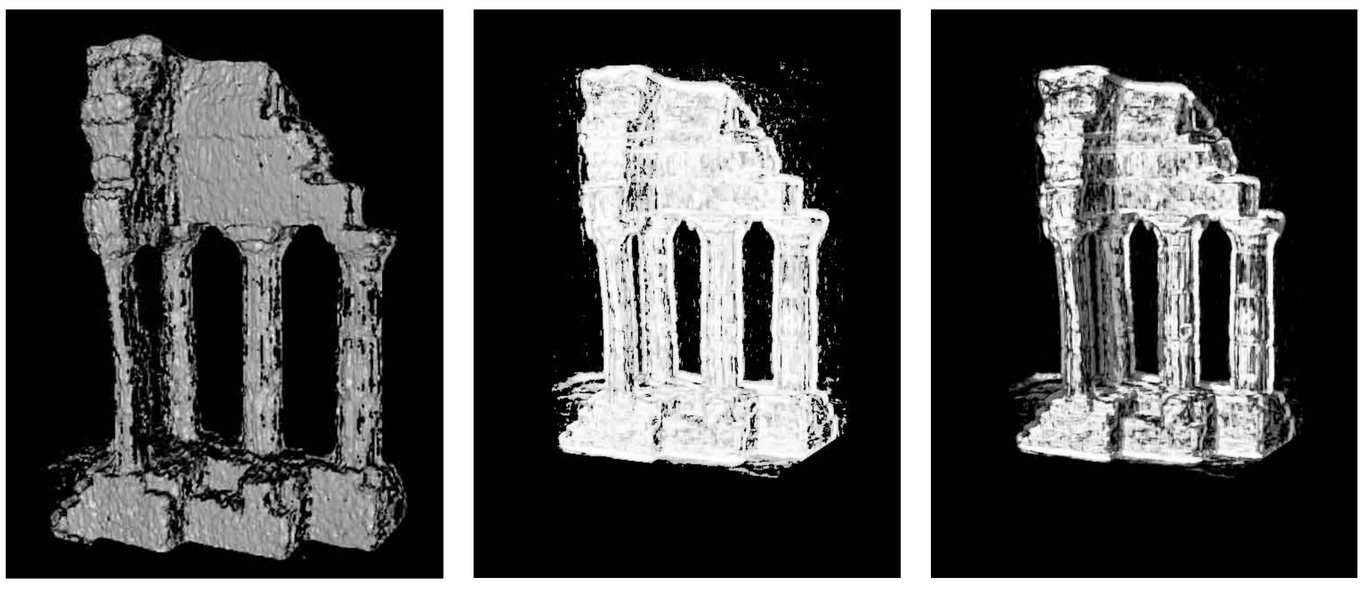
Step 1: Depth Map Generation
For each input view *R* (reference view), we select a set of *k* neighboring images against which we correlate *R* using robust window matching based on normalized cross correlation (NCC). Using a plane sweep approach, we compute a correlation score *corr(d)* for each pixel *p* in *R* and each candidate depth *d*.

$$corr(d) = \frac{\sum_{C_j \in \mathbf{C}_v(d)} NCC(R, C_j, d)}{\| \mathbf{C}_v(d) \|}$$

We exclude views with an NCC score below 0.6 and mark *d* as invalid if less than two views contribute to the correlation score. We store for each pixel *p* the depth value *d* that maximizes *corr(d)* in a depth map and compute a confidence value *conf(d)*.

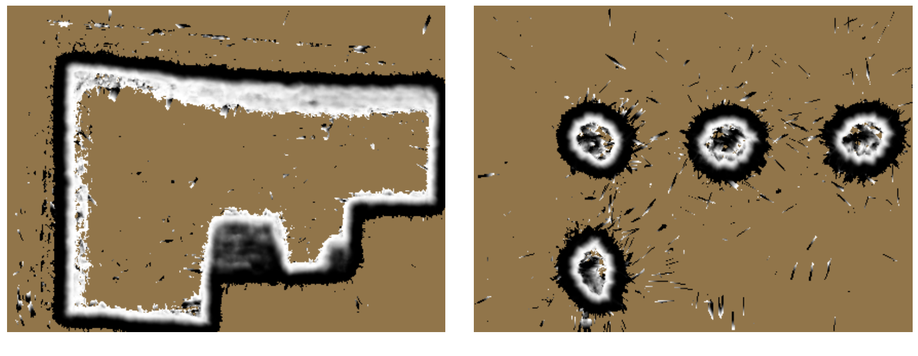
$$conf(d) = \frac{\sum_{C_j \in \mathbf{C}_v(d)} (NCC(R, C_j, d) - thresh)}{\| \mathbf{C} \| (1 - thresh)}$$

p is marked as invalid if no valid *d* is found.



Reconstructed depth map (rendered), correlation values and confidence values for a view from the temple data set.

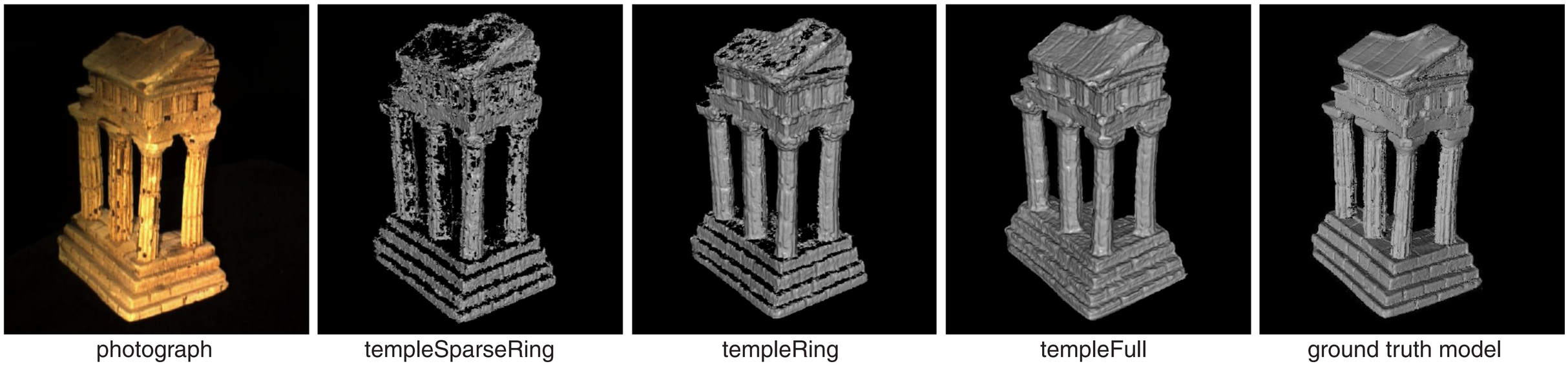
Step 2: Merging Depth Maps
We merge the set of incomplete depth maps with confidence values from the previous step into a single surface mesh using the volumetric method by Curless and Levoy [1]. This approach converts each depth map into a weighted signed distance volume, takes a sum of these volumes, and extracts a surface at the zero level set using a marching cubes algorithm.



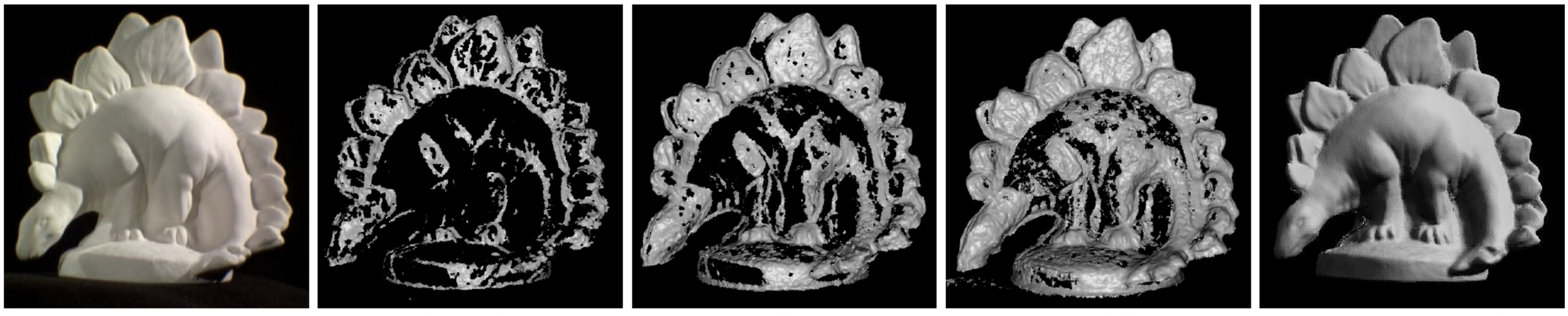
Cross sections through the weighted signed distance volume at the base of the temple (left) and through the columns (right).

Multi-View Stereo Revisited

Michael Goesele, Brian Curless, and Steven M. Seitz
Department of Computer Science and Engineering
University of Washington, Seattle, WA



photograph | templeSparseRing | templeRing | templeFull | ground truth model

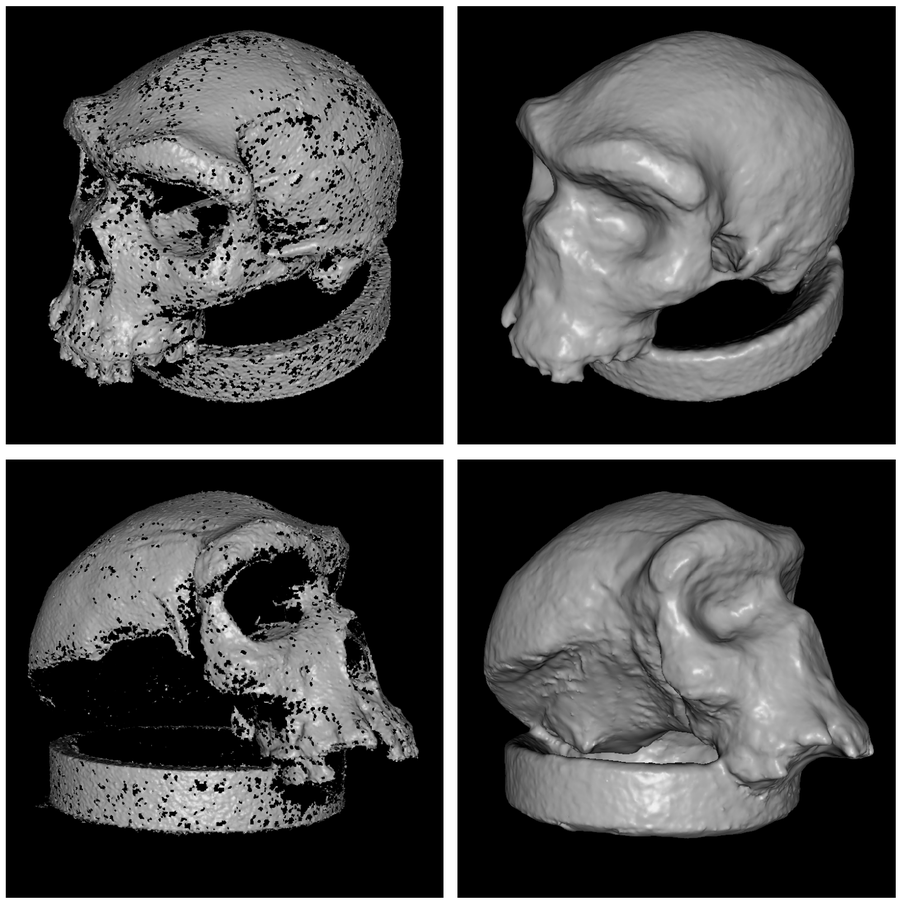


photograph | dinoSparseRing | dinoRing | dinoFull | ground truth model

Validation
Reconstructed surface meshes of the temple and the dino plaster models were submitted to a multi-view stereo evaluation study [6]. The reconstructed models were evaluated regarding their accuracy and completeness. In addition, we reconstructed a model of a plaster cast of a human skull and compared it to the reconstruction using one of the currently top-performing multi-view stereo methods [2].

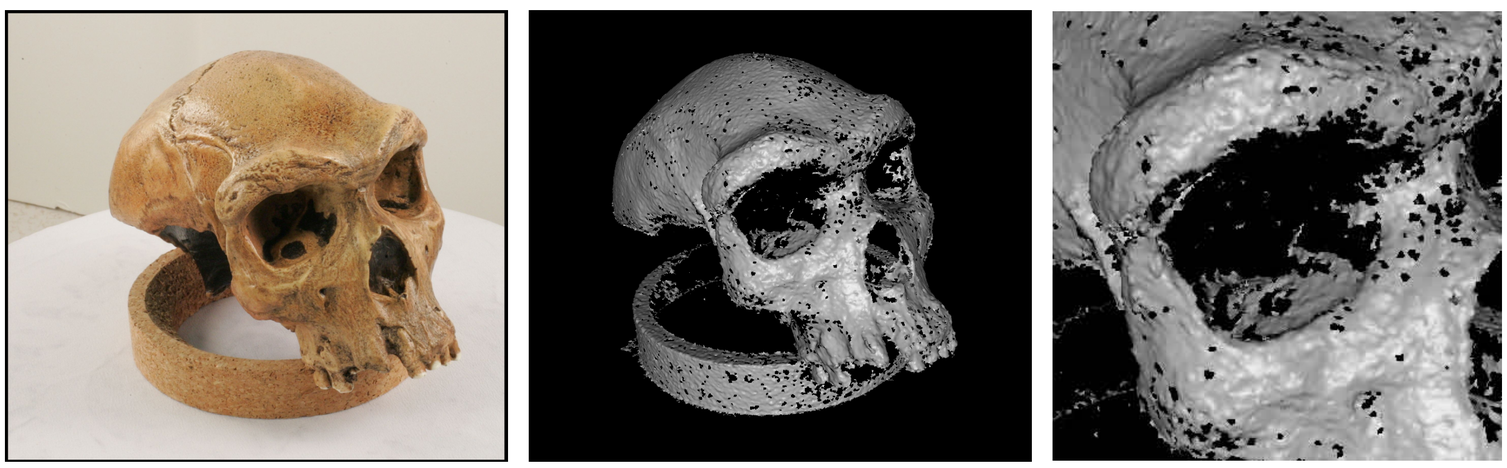
dataset	views	accuracy	difference to best method	completeness	run time (hours:minutes)
templeFull	317	0.42 mm	0.06 mm	98.0 %	*200:00*
templeRing	47	0.61 mm	0.09 mm	86.2 %	*30:00*
templeSparseRing	16	0.87 mm	0.12 mm	56.5 %	10:06
dinoFull	363	0.56 mm	0.14 mm	80.0 %	*281:00*
dinoRing	48	0.46 mm	0.04 mm	57.8 %	*37:00*
dinoSparseRing	16	0.56 mm	-	26.0 %	12:24

Results of the evaluation regarding accuracy, completeness, and run time. The third column lists the difference between the accuracy of our method and the accuracy of the best performing method in [6]. The heights of the objects are 159.6 mm (temple) and 87.1 mm (dino). Run times are given for a 3.4 GHz Pentium 4 processor. Models whose run times are given in italics were computed on a PC cluster and timings were not directly available. The run times were therefore estimated based on the run times for the sparseRing datasets.



our method | Furukawa and Ponce

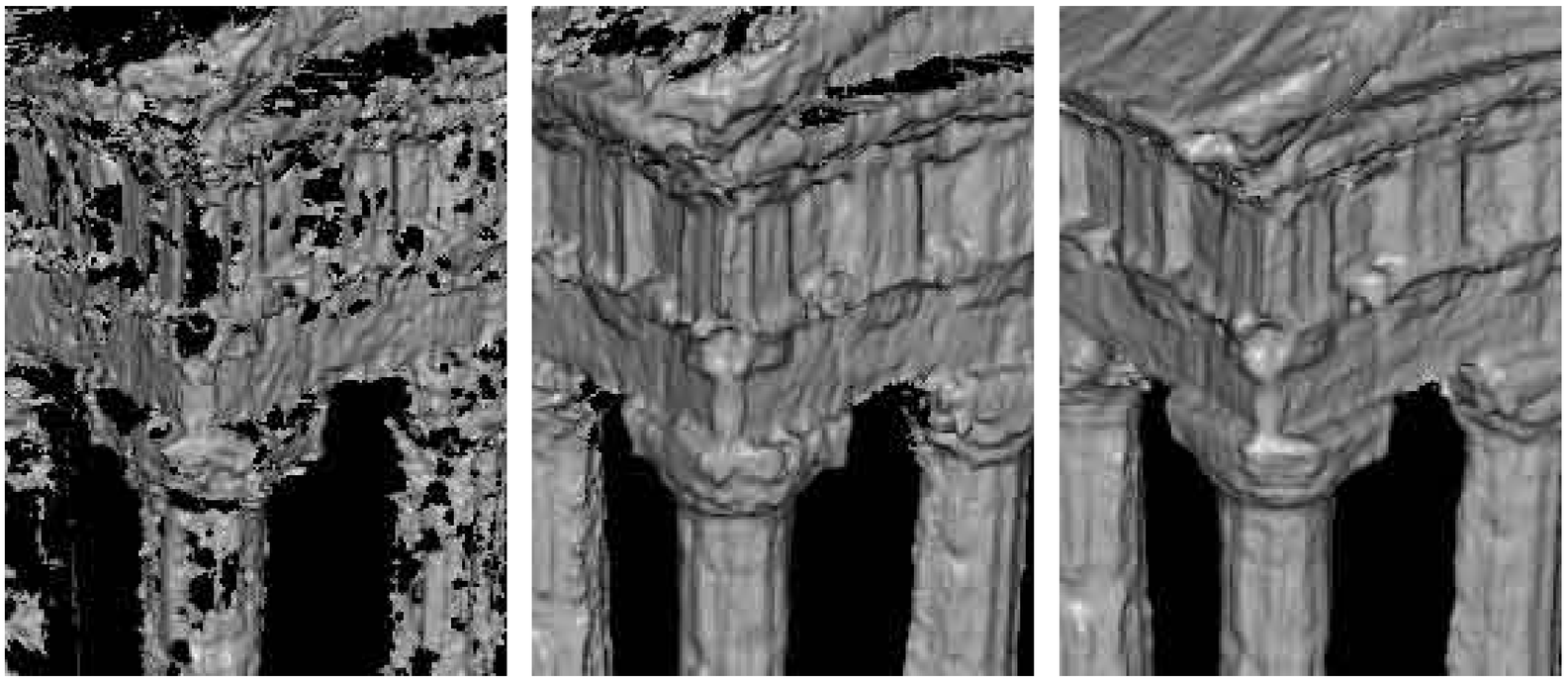
Comparison of the nskulla-half dataset reconstructed with our method (left) and with one of the currently top-per-forming multi-view stereo methods [2].



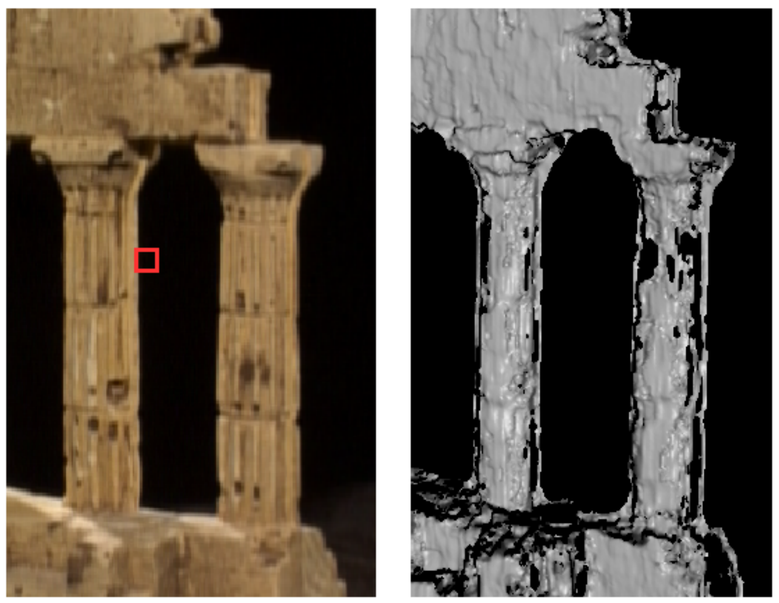
Example view from the nskulla-half dataset and the reconstructed mesh. Note the specular reflection on the skull surface and reconstructed geometry in the eye socket.

Multi-View Stereo Revisited

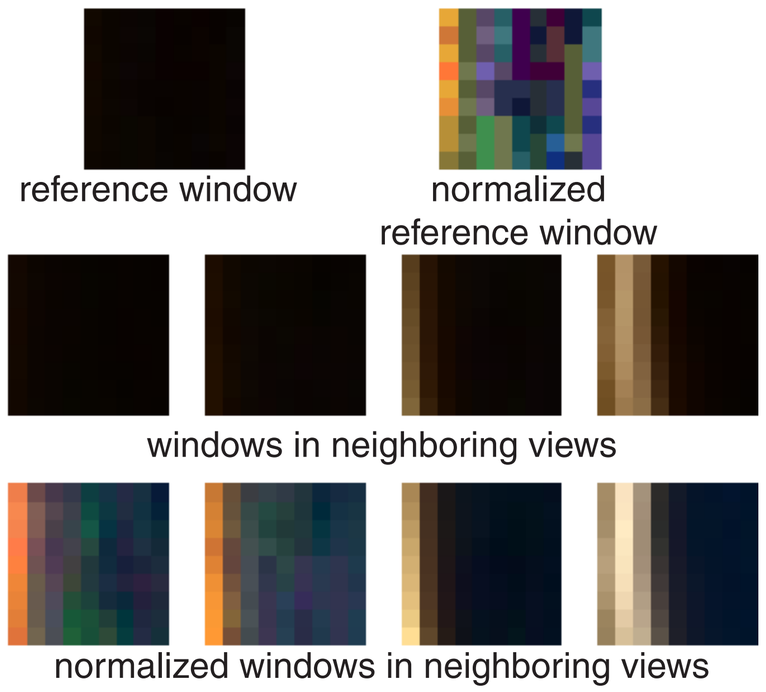
Michael Goesele, Brian Curless, and Steven M. Seitz
Department of Computer Science and Engineering
University of Washington, Seattle, WA



Influence of the number of input images on the reconstruction quality.
Left to right: templeSparseRing (16 views), templeRing (47 views), templeFull (317 views).



Cropped reference view with approximate location of reference window (left) and rendering of the corresponding depth map (right).



Spurious geometry can occur at silhouettes with low image contrast. The seemingly structureless 9x9 window of the reference view actually contains structure which is revealed by the normalization used in the NCC. The reference window matches the two windows from neighboring views shown on the left and spurious geometry is created along the edges of the columns.

References

[1] B. Curless and M. Levoy. A Volumetric Method for Building Complex Models from Range Images. In Proc. SIGGRAPH 1996, pages 303-312.
[2] Y. Furukawa and J. Ponce. Carved Visual Hull for Image-Based Modeling. In Proc. ECCV 2006.
[3] C. Hernandez and F. Schmitt. Silhouette and Stereo Fusion for 3D Object Modeling. Computer Vision and Image Understanding, 96(3):367-392, 2004.
[4] P. J. Narayanan, P. Rander, and T. Kanade. A Multiple-Baseline Stereo. TPAMI, 15(4):353-363, 1993.
[5] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual Modeling with a Hand-held Camera. International Journal of Computer Vision, 59(3):207-232, 2004.
[6] S. M. Seitz, B. Curless, J. Diebel, D. Scharenstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In Proc. CVPR 2006.