Towards high-resolution large-scale multi-view stereo

Hoang-Hiep Vu, Renaud Keriven, Patrick Labatut, Jean-Philippe Pons

Towards high-resolution large-scale multi-view stereo

Vu Hoang Hiep[1,2] Renaud Keriven[1] Patrick Labatut[1] Jean-Philippe Pons[1]
IMAGINE
[1] Université Paris-Est, LIGM/ENPC/CSTB [2] ENSAM, Cluny
http://imagine.enpc.fr

Abstract

Boosted by the Middlebury challenge, the precision of dense multi-view stereovision methods has increased drastically in the past few years. Yet, most methods, although they perform well on this benchmark, are still inapplicable to large-scale data sets taken under uncontrolled conditions. In this paper, we propose a multi-view stereo pipeline able to deal at the same time with very large scenes while still producing highly detailed reconstructions within very reasonable time. The keys to these benefits are twofold: (i) a minimum s-t cut based global optimization that transforms a dense point cloud into a visibility consistent mesh, followed by (ii) a mesh-based variational refinement that captures small details, smartly handling photo-consistency, regularization and adaptive resolution. Our method has been tested on numerous large-scale outdoor scenes. The accuracy of our reconstructions is also measured on the recent dense multi-view benchmark proposed by Strecha et al.*, showing our results to compare more than favorably with the current state-of-the-art.*

1. Introduction

Motivation. Scene reconstruction from multiple images has always been an active field of research in computer vision. This classic problem finds many practical applications in the entertainment industry, in earth sciences and in cultural heritage digital archival for instance.

When high detail is needed, laser-based methods are usually applied successfully. However, these methods are rather complex to set for large-scale outdoor reconstructions, particularly when aerial acquisition is required. See for instance the recent detailed reconstruction of the Bayon temple in Angkor [2].

Our goal here is to replace these methods with image-based ones, yielding considerable savings both in time and money. We believe that recent advances in multi-view stereo methods made this goal closer than ever.

Multi-view stereo. Since the review of Seitz *et al*. [33] and the associated Middlebury evaluation, a lot of research has been focusing on multi-view reconstruction of small objects with tightly controlled imaging conditions. This has led to the development of many algorithms whose results are beginning to challenge the precision of laser-based reconstructions. However, as we will see, most of these algorithms are not directly suited to large-scale scenes.

A number of multi-view stereo algorithms have been proposed that exploit the visual hull [26]. They rely on it either as an initial guess for further optimization [8, 13, 15, 38, 43, 46, 50], as a soft constraint [21, 13] or even as a hard constraint [8, 35] to be fulfilled by the reconstructed shape.

While the unavailability of the visual hull discards many of the top-performing multi-view stereo algorithms of Middlebury challenge for our purpose, the requirement for the ability to handle large-scale scenes discards most of the others, in particular volumetric methods, *i.e*. methods based on a regular decomposition of the domain into elementary cells, typically voxels. Obviously, this approach is mainly suited to compact objects admitting a tight enclosing box, as its computational and memory costs quickly become prohibitive when the size of the domain increases. This includes space carving [4, 23, 34, 44, 49], level sets [19, 17, 32], and volumetric graph cuts [3, 15, 27, 38, 43, 47] (though [36, 14] propose photo-consistency adaptive grids to push the resolution limit further).

Finally, cluttered scenes disqualify variational methods [6, 7, 19, 13, 17, 28, 32] that get stuck into local minima, unless they provide a way of estimating a close and reliable initial guess that takes visibility into account.

Large-scale multi-view stereo. The multi-view stereo methods which have proved to be more adapted to large-scale scenes (*e.g*. outdoor architectural scenes) are those representing geometry by several depth maps [45, 12, 30, 11, 10, 22, 39, 40, 41]. However, their performance for complete reconstruction seems to be lower than previously discussed approaches, either as regards accuracy or completeness of the obtained model. This may be due to the merging process and to the difficulty to take visibility into

account globally and consistently. A notable exception could be Campbell's work [5] which is one of the most accurate method according to the Middlebury evaluation, but this method relies on a volumetric graph cut that cannot handle large-scale scenes.

Large-scale high-resolution multi-view stereo. In contrast to these depth maps based methods, Furukawa and Ponce proposed in [9] a very accurate reconstruction that generates and propagates a semi-dense set of patches. This method has shown impressive results but relies on heuristics and on a final Poisson surface reconstruction [18] that do not handle global consistency. The authors tested their method on the large-scale data set provided by Christoph Strecha *et al*. [42], the only available evaluation that, to our knowledge, fits our purpose. So far, their results were significantly more accurate and complete than the few other submitted ones. We will indeed rely on this quantitative challenge to demonstrate the superiority of our method. Results on other data sets will confirm this by visual qualitative evaluation.

Our multi-view stereo method consists in a pipeline that handles large-scale scenes while providing very accurate reconstructions. It takes the best of several previous methods both in multi-view stereo and in mesh processing. So in a way, one could argue for a lack of originality. Yet, we claim that this effort is valuable for at least two reasons: (i) assembling this pipeline from all the existing methods is not an obvious choice, and (ii) each existing method is modified in an essential way that makes it more robust, accurate or powerful. When concerning more commonly spread problems, comparable improvements are usually considered important advances! In the case of multi-view stereo, there is no reason to view them as only small increments. Our choices are justified by an analysis of the weak points of previous methods. Last but not least, our results show that these are more than details.

The pipeline consists in two main steps: (i) the generation of a dense point cloud from which a minimum s-t cut based optimization outputs a visibility consistent mesh close to the final reconstruction; and (ii) a variational optimization that photo-refines this mesh. The remainder of the paper is organized following these steps before presenting results and comparisons.

2. Globally optimal visibility consistent mesh

The first step consists in producing a mesh taking visibility into account and accurate enough to be then refined by a variational optimization. The hypothesis is then that the gradient descent of some energy will deform this mesh into a local minimum that will be considered as the final reconstruction. A method that comes up partially with this problem for cluttered large-scale scenes, is our previous work [24]. Our method consists in four steps: (i) a point cloud is generated from images, each point memorizing the two or more images from which it has been triangulated; (ii) the Delaunay triangulation of these points is performed; (iii) the Delaunay tetrahedra are labeled inside or outside the object so that the labeling minimizes some energy; and (iv) the surface is extracted as the set of triangles between inside and outside tetrahedra. The energy takes visibility into account: each ray from a vertex to the cameras from which it has been generated is enforced to intersect the oriented output surface as few times as possible. It is globally minimized with minimum s-t cut and the results in [24] show the ability of coping with cluttered scenes.

A denser and more accurate point cloud. However, it appears that the original method yields coarse meshes. This is contradictory with the goal of being an initial estimate that completely deals with visibility. To get a better mesh, we generate a much denser point cloud. The original method matches SIFT [29] points to produce it. Here, we increase the number of candidate feature points, extracting more DOGs, adding Harris points, or even using regular grids. Then, to counterbalance the high number of false matches, we improve the matching criterion. Indeed, relying on SIFT descriptors misses the fact that the camera geometry is known. Here, we simply use a robust photo-consistency criterion, namely the sum of normalized cross correlations (NCCs) for several fixed sizes of neighborhood [48]. This supposes that the object is locally fronto-parallel. In fact, we experimented with considering more than one orientation and keeping the best one without any appreciable improvement. Note also that, when using regular grids, this process boils down to a simple multiple-hypothesis plane sweeping algorithm.

As in the original method, close 3D points are then merged efficiently thanks to the Delaunay triangulation, so that a point of the final cloud originates from possibly more than two images. The overall process is fast, the NCCs (and, if necessary, the plane sweep) being easily implemented on graphical processors (GPUs).

A more adapted energy. The resulting point cloud is very dense and typically contains millions of points (see Fig. 2). The visibility term of the energy of [24] is very effective to filter out outliers from stereo point clouds. However, due to the high density of point clouds from regular grids, triangles lying near the surface are very small and the original complimentary photo-consistency term becomes almost useless. It is advantageously replaced with the surface quality term of [25] used for surface reconstruction from range scans. This term penalizes facets unlikely to appear on a densely sampled surface. As a result and contrarily to the original method, the minimum s-t cut step encodes discrete visibility and surface quality, saving an appreciable amount

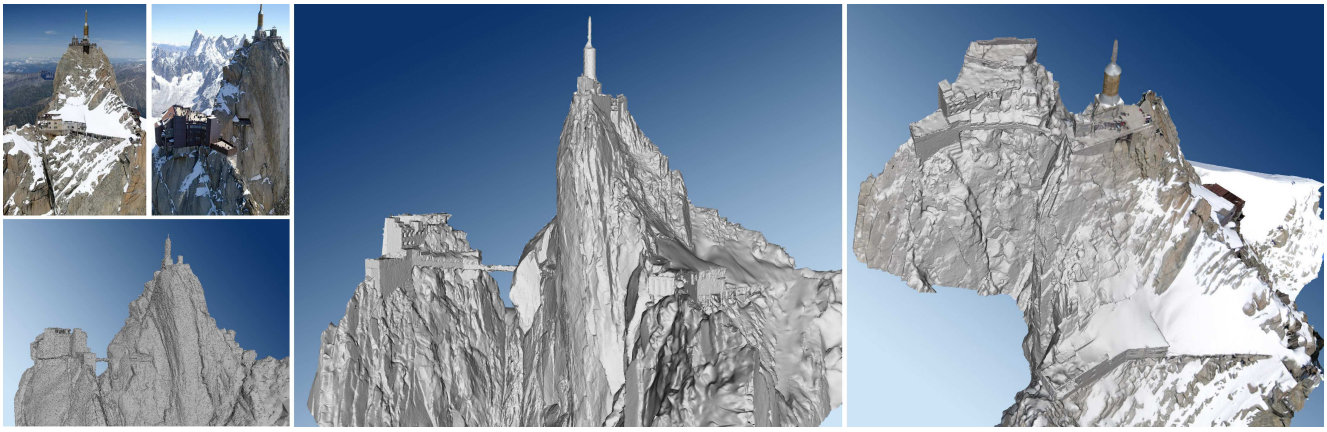

Figure 1. From left to right: two of the 51 images (5 Mpixels) of a mountain summit taken from a helicopter (© B.Vallet/IMAGINE) and the visibility consistent mesh M^0; final reconstruction without and with texture remapping using [1] (600, 000 triangles).

of time. Since [24], support for infinite tetrahedra was also added (tetrahedra with one facet on the convex hull and incident to the infinite vertex). This not only allows the observer to be "inside" the object, but also makes it possible to generate open meshes. This is an important aspect for outdoor scenes.

3. Variational refinement

The obtained mesh M^0 is noisy and does not capture small details. We refine it using the full image data, with a variational multi-view stereovision approach pioneered by [19]: we use M^0 as the initial condition of a gradient descent of an adequate energy function. As M^0 is close to the desired solution, this local optimization is very unlikely to get trapped in an irrelevant local minimum. Let us now justify our energy function and our optimization procedure, by presenting the numerous upgrades to the initial method. Again, these are not details. Or said another way: to get a detailed reconstruction, every detail is important.

The right summation. Let S be the object surface, X a point on S, N the normal to S at point X, $g_{kk'}(X, N)$ a positive decreasing function of a photo-consistency measure of patch (X, N) according to images k and k', and $v^S_{kk'}(X) \in \{0, 1\}$ the visibility of X in these images according to S. The original energy in [19] is

$$E_{KF}(S) = \sum_{k,k'} \int_S v^S_{kk'}(X)\, g_{kk'}(X, N)\, dX \ . \qquad (1)$$

To this energy, we prefer the reprojection error introduced by [32], namely:

$$E(S) = \sum_{k,k'} \int_{\Omega^S_{kk'}} h(I_k, I^S_{kk'})(x_k)\, dx_k \ , \qquad (2)$$

where $h(I, J)(x)$ is a positive decreasing function of a photo-consistency measure between images I and J at pixel x, $I^S_{kk'}$ the re-projection of image $I_{k'}$ into image I_k induced by S and $\Omega^S_{kk'}$ the domain of definition of this re-projection.

This summation has three major advantages over the original one: (i) reprojecting $I_{k'}$ into I_k according to S uses the exact geometry of S and does not use the tangent patch (X, N) approximation anymore, (ii) this re-projection can easily be computed with graphics hardware, and (iii) the less a surface element is viewed in a given image, the less it contributes to the energy. The first point is essential to an accurate reconstruction: in methods approximating the surface by planar patches, the choice of patch size is a difficult trade-off between robust and accurate photo-consistency.

The right surface representation. The level set representation used in [19, 32] has a prohibitive computational and memory cost for high resolution reconstructions. Unstructured polygonal meshes are much better at capturing extremely fine geometry. Moreover, both our global optimization step and the computation of the image reprojection $I^{kk'}$ on graphics hardware depend on a triangle mesh. Hence, the obvious choice for representing S is a deformable triangular mesh M with vertices V and triangles T.

Moreover, we assume that M^0 has the desired topology, which all our numerical experiments confirm. As a result, it is not necessary to resort to complex remeshing procedures [31, 51] to handle topology changes during deformation.

The right discretization. An overwhelming majority of methods in variational multi-view stereovision [7, 19, 13, 17, 28, 32], and more generally in computer vision, rely on an *optimize then discretize* approach: an energy functional depending on a continuous infinite-dimensional representation is considered, the gradient of this energy functional is

Figure 2. From left to right: three of the 27 images (8 Mpixels) of a sculpted calvary taken from the ground; the generated point cloud ($1,300,000$ points) and visibility consistent mesh M^0 ($1,200,000$ triangles); final reconstruction ($1,850,000$ triangles); close views of the final reconstruction with and without texture remapping. Note the high percentage of outliers in the point cloud and to what extent M^0 is noisy but close to the solution in position and topology.

computed analytically, then the obtained minimization flow is discretized.

In contrast, we adopt a *discretize then optimize* approach: we define an energy function depending on a discrete finite-dimensional surface representation, here a triangle mesh, and we use standard non-convex optimization tools. The benefits of this approach have long been recognized in mesh processing, but have seldom been demonstrated in computer vision [6, 37].

Following [6], we thus rewrite E as a function of M, and compute the velocity field as the partial derivatives of this energy with respect to vertex positions. First of all, it circumvents the difficult task of chosing a consistent discretization of differential quantities, such as normal and curvature, on a triangle mesh. Second, it is more faithful to the data, and it guarantees that the energy actually decreases: notably, the obtained gradient vector at a vertex involves integrals over the whole ring of triangular facets around it. This is in strong contrast with a pointwise, and thereby noise-sensitive, dependency on the input data that a late discretization typically causes. We note a crucial point here: this gradient flow may include a significant tangential component driving the vertices at the right places minimizing the energy. For instance, vertices naturally migrate to the object edges if any. This is illustrated by the crisp reconstruction of stair treads in Figure 3.

The right regularization. While the original intrinsic energy of Eq. 1 is self-regularizing due to the integration over the surface, this is not the case of Eq. 2. So we complement the energy function with a discrete analog of the thin-plate energy, described in [20]. This term penalizes strong bending, not large surface area. Consequently, the associated gradient flow is exempt from the classical shrinking bias. Moreover, beyond surface smoothness, it also redistributes vertices along the surface, and in particular it discourages degenerate triangles.

Mixing photo-consistency and regularization. A proper automatic balancing between data attachment and smoothing terms is a long-standing issue in variational methods. Designing a general solution to this problem is clearly out of the scope of this paper. Here we propose a specific strategy which allowed us to conduct all the following experiments *without adjusting parameters to each data set*. Our solution is twofold.

First, we observe that regularization is more important where photo-consistency is less reliable, in particular in textureless or low-textured image regions. Consequently, we weight the contribution of camera pair (k, k') at pixel x_k in Eq. 2 by a reliability factor $\min(\sigma^2, \sigma'^2)/[\min(\sigma^2, \sigma'^2) + \epsilon^2]$, where σ^2 and σ'^2 denote the local variance at x_k in images I_k and $I^S_{kk'}$, respectively, and ϵ is a constant.

Second, we homegenize the two terms of the energy function: while the data attachment term of Eq. 2 is homogeneous to an area in pixels, the discrete thin plate term is homogeneous to squared world units. After weighting the contribution of each image in Eq. 2 by the square of the ratio between the average depth of the scene and the focal length in pixels, we are able to define a scalar regularity weight whose optimal value is stable across very different datasets.

Adaptive mesh resolution. The resolution of the mesh is automatically and adaptively adjusted to image resolution: a triangular facet is subdivided if there exists one camera pair such that the visible facet projection exceeds a user-defined number of pixels in both images. We set this threshold to

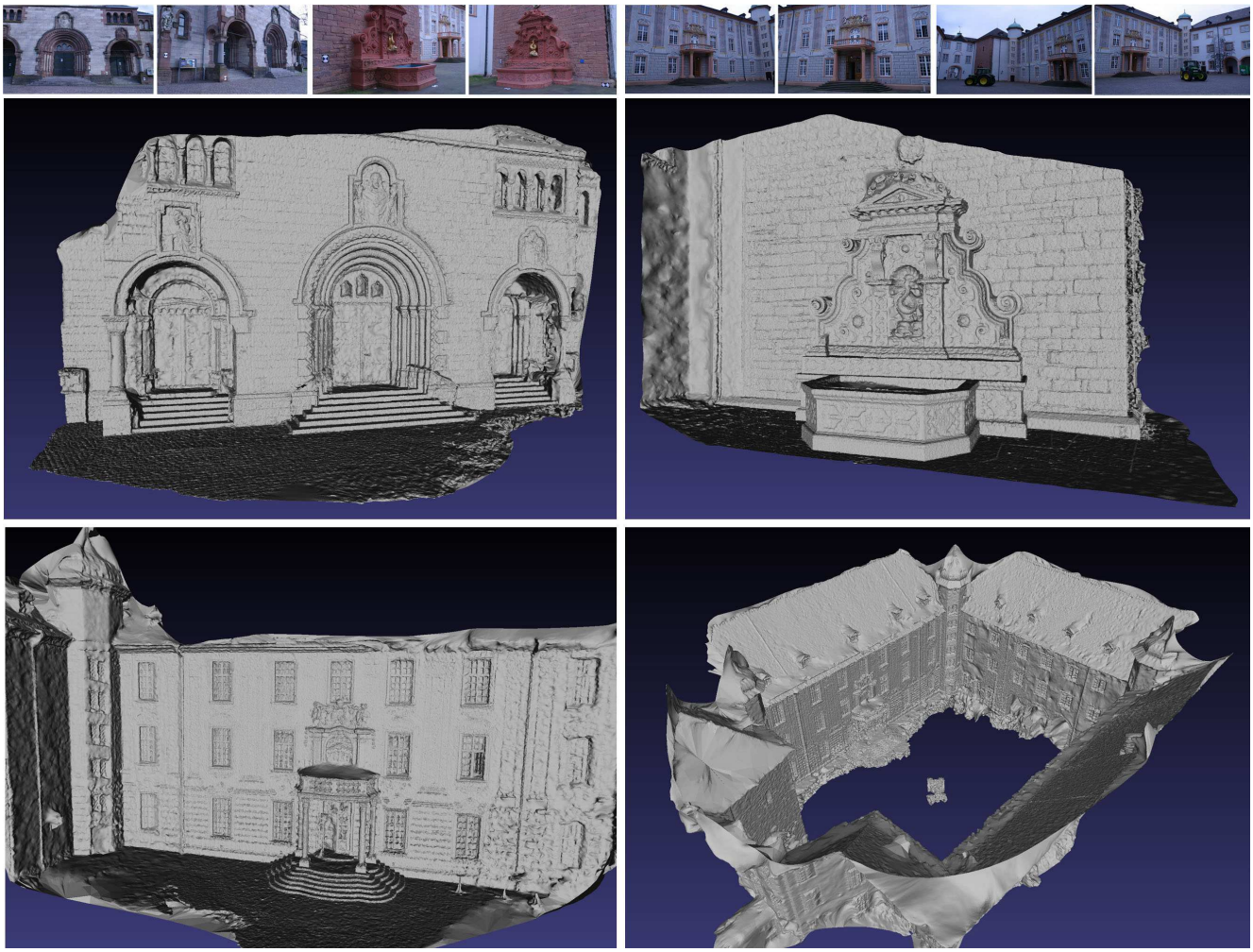

Figure 3. The four first data sets of [42]. From top to bottom, left to right: 2 images of each data set, namely *Herz-Jesu-P8* (8 images), *fountain-P11* (11 images), *entry-P10* (10 images) and *castle-P19* (19 images); our reconstructions, respectively $1,450,000$, $1,600,000$, $2,000,000$ and $3,000,000$ triangles. Note how details, topology (eg. columns) and edges (eg. stairs) are precisely recovered while regularization still handles as correctly as possible blurred or untextured parts.

16 pixels in all our experiments. We use a classical one-to-four triangle subdivision scheme, which has the advantage of preserving sharp edges.

4. Results

As already mentioned, all the following experiments have been conducted with the same parameters. Our photo-consistency h is NCC-based, although other more elaborate measures are feasible. Some operations are GPU implemented, mainly NCC estimations and image reprojections. Depending on the number of images, the running time of our pipeline ranges from fifteen to ninety minutes, on a 3.0 GHz CPU and an NVIDIA GeForce 8800 GTX GPU. Although we have obtained state-of-the-art results to the Middlebury challenge, we refer the reader to its web page [33] and we do not reproduce these results here. The sequel is devoted to experiments on large-scale scenes. For extensive, detailed and animated results, please visit our dedicated web page[1].

Original data sets. We tested our method our an aerial acquisition of the *Aiguille du Midi* summit. The data set consists in 51.5 Mpixels images. Figure 1 shows 2 of the images, the initial mesh M^0 and the final reconstruction. This experiment validates the whole pipeline and the ability to cope with uncontrolled imaging conditions (snow, sun, moving people from one image to another) and a mix of complex and smooth geometries. Note that the variational process is able to recover the top antenna although it is only partially present in M^0. Figure 2 shows extensively results

[1] `http://imagine.enpc.fr/demos/stereo/`

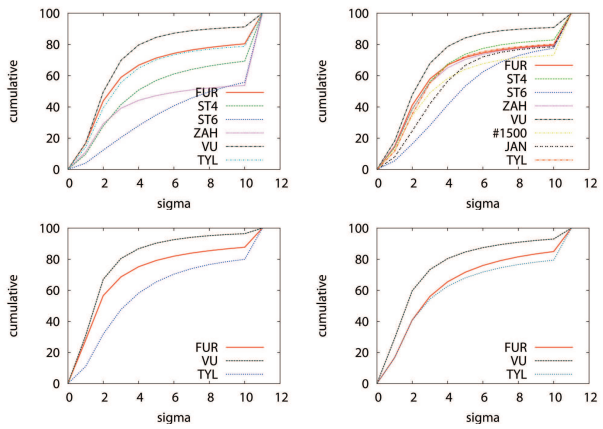

Figure 4. From top to bottom, left to right, relative error cumulated histograms, respectively for the *Herz-Jesu-P8*, *fountain-P11*, *entry-P10* and *castle-P19* data sets. Legend is the following: FUR for [9], ST4 for [39], ST6 for [40], ZAH for [51], TYL for [45], JAN for [16], VU for our work. On all data sets, measures confirm clearly our better results, both in accuracy and completeness.

on a 27.8 Mpixels images data set of a sculpted calvary taken from the ground. The cloud has $1,300,000$ points, with many outliers, mainly sky points (in white color) obtained matching clouds that have moved between shots. Only $660,000$ of these points are selected for the initial mesh M^0 ($1,200,000$ triangles). Note how this mesh if noisy, due the process of matching feature points that are just approximately view-point invariant. As the close views show, the final reconstruction ($1,850,000$ triangles) is sharp enough to capture meaningful details, while global visibility is still correct.

Dense multi-view stereo data sets. Provided by Strecha *et al*. [42], the already mentioned data sets consists in outdoor scenes acquired with 8 to 30 calibrated 6 Mpixels images. Ground truth has been acquired with a LIDAR system. Evaluation of the multi-view stereo reconstructions is quantified through relative error histograms counting the percentage of the scene recovered within a range of 1 to 10 times an estimated noise variance σ. We focus on the four first data sets, for which other groups have submitted results. Dedicated to large-scale objects and fitting perfectly our objective, these sets are particularly challenging, especially the *castle-P19* one, a complete courtyard acquired from the inside and where a tractor stays in the middle, disturbing reconstruction. So far, only Furukawa *et al*. [9] and Tylecek *et al*. [45] submitted for this particular data set. Figure 3 shows, for the four data sets, two of the images and a global view of our corresponding reconstruction. We output meshes going from $1,450,000$ to $3,000,000$ triangles, depending on the data set. Again, all the experiments are run with the same parameters. Comparison with the other methods are given in Fig. 4, where cumulated histograms show

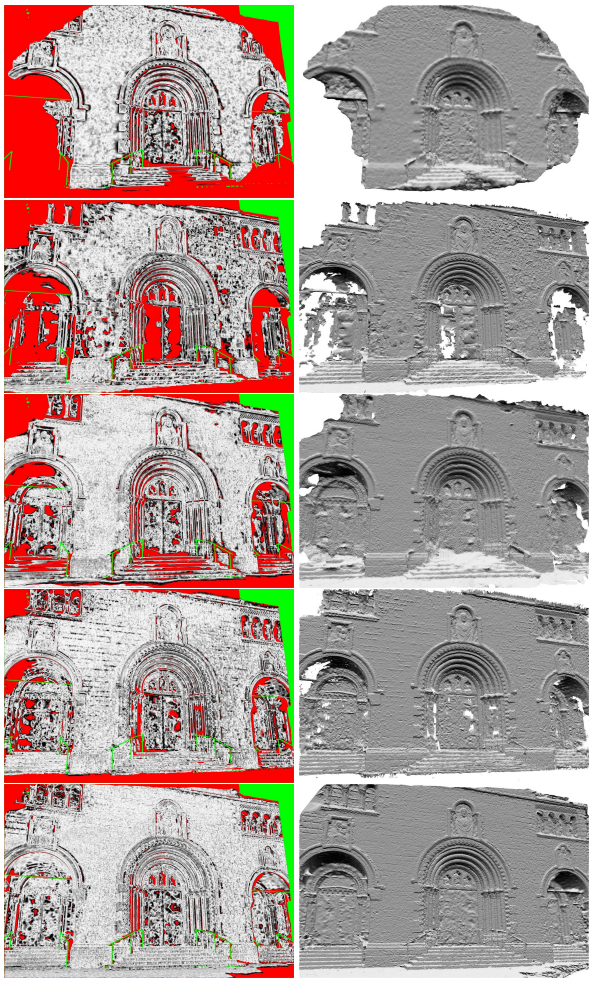

Figure 5. A visual comparison on the *Herz-Jesu-P8* data set. From top to bottom, results from ZAH [51], ST4 [39], TYL [45], FUR [9] and our work. Left: variance weighted depth difference (red pixels encode an error larger than 3σ; green pixels encode missing LIDAR data; the relative error between 0 and 3σ is encoded in gray. Right: diffuse renderings of the corresponding triangle meshes.

clearly that our method is both more accurate and complete. Focusing of the *Herz-Jesu-P8* data set, Fig. 5 gives a more visual comparison. For the four best methods so far and our method, the results are rendered and the corresponding error is color encoded. Note ground truth is not available everywhere (*e.g*. the metal bar under the left porch, which we actually partially recover). Finally, Fig. 6 compares, for the other three datasets, the rendering of our reconstruction with the one of the second best method. For extended results, we refer the reader to the challenge website.

5. Conclusion

We presented a method for multi-view reconstruction of cluttered large-scale scenes photographed under uncon-

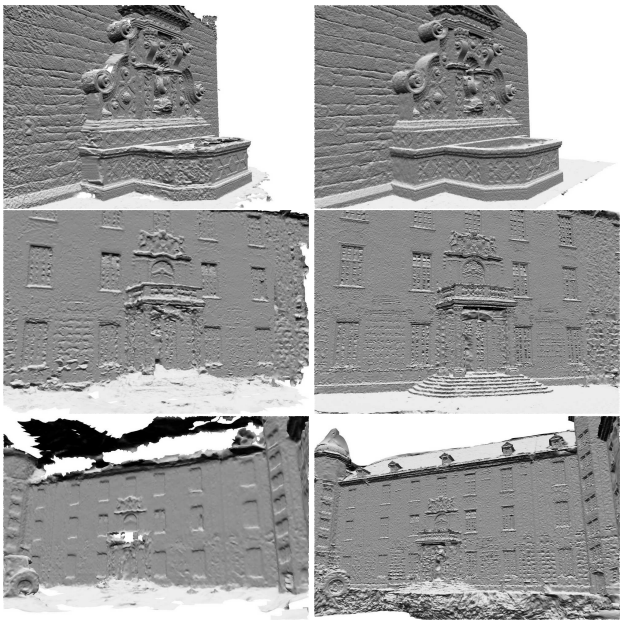

Figure 6. A short visual comparison for the three other data sets. Left: rendering of the second best method ([39] for *fountain-P11*, [9] for the *entry-P10* and *castle-P19*). Right: our method.

trolled conditions. Based on previous methods of stereovision and mesh processing, carefully analyzing their weak points and replacing them with robust or more adapted solutions, we obtain a complete pipeline outputting reconstructions visually and quantitatively more accurate and complete than state-of-the-art techniques. This is clearly just a first step toward high-resolution models that could compete with laser scans, but the road to them might not be so long. The availability of very high-resolution consumer-grade cameras will raise new issues like the problem of splitting the reconstruction problem into several smaller ones [52]. Above all, with the present paper, we hope to orient the competition on challenges like [42] and hope that the number of their participants will grow as rapidly as it happened to [33].

Acknowledgments

We thank Bernard Vallet for the *Aiguille du Midi* images, Marc Pierrot-Deseilligny for their calibration and Christoph Strecha for his challenge and his kind help.

References

[1] C. Allène, J.-P. Pons, and R. Keriven. Seamless image-based texture atlases using multi-band blending. In *International Conference on Pattern Recognition*, 2008.

[2] A. Banno, T. Masuda, T. Oishi, and K. Ikeuchi. Flying laser range sensor for large-scale site-modeling and its applications in bayon digital archival project. *International Journal of Computer Vision*, 78(2–3):207–222, 2008.

[3] Y. Boykov and V. Lempitsky. From photohulls to photoflux optimization. In *British Machine Vision Conference*, 2006.

[4] A. Broadhurst, T. W. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *IEEE International Conference on Computer Vision*, 2001.

[5] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, 2008.

[6] A. Delaunoy, E. Prados, P. Gargallo, J.-P. Pons, and P. Sturm. Minimizing the multi-view stereo reprojection error for triangular surface meshes. In *British Machine Vision Conference*, 2008.

[7] Y. Duan, L. Yang, H. Qin, and D. Samaras. Shape reconstruction from 3D and 2D data using PDE-based deformable surfaces. In *European Conference on Computer Vision*, 2004.

[8] Y. Furukawa and J. Ponce. Carved visual hulls for image-based modeling. In *European Conference on Computer Vision*, 2006.

[9] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[10] P. Gargallo and P. Sturm. Bayesian 3D modeling from images using multiple depth maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[11] M. Goesele, B. Curless, and S. M. Seitz. Multi-view stereo revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[12] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision*, 2007.

[13] C. Hernández and F. Schmitt. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[14] C. Hernández, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[15] A. Hornung and L. Kobbelt. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[16] M. Jancosek and T. Pajdla. Segmentation based multi-view stereo. In *Computer Vision Winter Workshop*, 2009.

[17] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 63(3):175–189, 2005.

[18] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing*, 2006.

[19] R. Keriven and O. Faugeras. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, 1998.

[20] L. Kobbelt, S. Campagna, J. Vorsatz, and H.-P. Seidel. Interactive multi-resolution modeling on arbitrary meshes. In *International Conference on Computer Graphics and Interactive Techniques*, 1998.

[21] K. Kolev and D. Cremers. Integration of multiview stereo and silhouettes via convex functionals on convex domains. In *European Conference on Computer Vision*, 2008.

[22] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *European Conference on Computer Vision*, 2002.

[23] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[24] P. Labatut, J.-P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *IEEE International Conference on Computer Vision*, 2007.

[25] P. Labatut, J.-P. Pons, and R. Keriven. Robust and efficient surface reconstruction from range data. *submitted to Computer Graphics Forum*, 2009.

[26] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[27] V. Lempitsky, Y. Boykov, and D. Ivanov. Oriented visibility for multiview reconstruction. In *European Conference on Computer Vision*, 2006.

[28] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):418–433, 2005.

[29] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[30] M. Pollefeys, D. Nistér, J.-M. Frahm, and A. Akbarzadeh. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2–3):143–167, 2008.

[31] J.-P. Pons and J.-D. Boissonnat. Delaunay deformable models: Topology-adaptive meshes based on the restricted delaunay triangulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[32] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

[33] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. `http://vision.middlebury.edu/mview/`.

[34] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.

[35] S. Sinha and M. Pollefeys. Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation. In *IEEE International Conference on Computer Vision*, 2005.

[36] S. N. Sinha, P. Mordohai, and M. Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *IEEE International Conference on Computer Vision*, 2007.

[37] G. G. Slabaugh and G. B. Unal. Active polyhedron: Surface evolution theory applied to deformable meshes. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[38] J. Starck, G. Miller, and A. Hilton. Volumetric stereo with silhouette and feature constraints. In *British Machine Vision Conference*, 2006.

[39] C. Strecha, R. Fransens, and L. V. Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[40] C. Strecha, R. Fransens, and L. V. Gool. Combined depth and outlier estimation in multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[41] C. Strecha, T. Tuytelaars, and L. V. Gool. Dense matching of multiple wide-baseline views. In *IEEE International Conference on Computer Vision*, 2003.

[42] C. Strecha, C. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. `http://cvlab.epfl.ch/~strecha/multiview/denseMVS.html`.

[43] S. Tran and L. Davis. 3D surface reconstruction using graph cuts with surface constraints. In *European Conference on Computer Vision*, 2006.

[44] A. Treuille, A. Hertzmann, and S. M. Seitz. Example-based stereo with general BRDFs. In *European Conference on Computer Vision*, 2004.

[45] R. Tylecek and R. Sara. Depth map fusion with camera position refinement. In *Computer Vision Winter Workshop*, 2009.

[46] G. Vogiatzis, C. Hernández, P. H. S. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2241–2246, 2007.

[47] G. Vogiatzis, P. H. S. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[48] R. Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[49] R. Yang, M. Pollefeys, and G. Welch. Dealing with textureless regions and specular highlights: A progressive space carving scheme using a novel photo-consistency measure. In *IEEE International Conference on Computer Vision*, 2003.

[50] T. Yu, N. Ahuja, and W.-C. Chen. SDG cut: 3D reconstruction of non-lambertian objects using graph cuts on surface distance grid. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[51] A. Zaharescu, E. Boyer, and R. P. Horaud. Transformesh: a topology-adaptive mesh-based approach to surface evolution. In *Asian Conference on Computer Vision*, 2007.

[52] A. Zaharescu, C. Cagniart, S. Ilic, E. Boyer, and R. P. Horaud. Camera clustering for multi-resolution 3-d surface reconstruction. In *ECCV 2008 Workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.