High Accuracy and Visibility-Consistent Dense Multiview Stereo

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

Abstract—Since the initial comparison of Seitz et al. [48], the accuracy of dense multiview stereovision methods has been increasing steadily. A number of limitations, however, make most of these methods not suitable to outdoor scenes taken under uncontrolled imaging conditions. The present work consists of a complete dense multiview stereo pipeline which circumvents these limitations, being able to handle large-scale scenes without sacrificing accuracy. Highly detailed reconstructions are produced within very reasonable time thanks to two key stages in our pipeline: a minimum s-t cut optimization over an adaptive domain that robustly and efficiently filters a quasidense point cloud from outliers and reconstructs an initial surface by integrating visibility constraints, followed by a mesh-based variational refinement that captures small details, smartly handling photo-consistency, regularization, and adaptive resolution. The pipeline has been tested over a wide range of scenes: from classic compact objects taken in a laboratory setting, to outdoor architectural scenes, landscapes, and cultural heritage sites. The accuracy of its reconstructions has also been measured on the dense multiview benchmark proposed by Strecha et al. [59], showing the results to compare more than favorably with the current state-of-the-art methods.

Index Terms—Dense multiview stereo, surface reconstruction, large-scale scenes, minimum s-t cut, deformable mesh.

1 Introduction

1.1 Motivation

The classic problem of scene reconstruction from multiple images finds many practical applications in reverse engineering, in the game and entertainment industry, and in the digital archives of cultural heritage. However, when high-accuracy reconstructions are required, the reconstruction of outdoor scenes has traditionally been done using range scanning and a combination of surface reconstruction from point clouds and geometry processing techniques. These methods and the acquisition process are rather complex to set for large-scale outdoor reconstructions, and this often proves to be time consuming, expensive, and dependent on the scene, particularly when aerial acquisition is required (see, for instance, the reconstruction of the Bayon temple in Angkor [2], which used range finders attached to flying balloons). Providing an image-based reconstruction solution would certainly eliminate most if not all of these drawbacks. This problem has thus always been one of the main goals and an active field of research in computer vision. Recent advances in multiview stereo methods have made this goal closer than ever. In this paper, the focus is on the dense multiview stereo problem, i.e., the reconstruction of a surface model from a set of calibrated images where camera calibration is assumed to be accurately known.

1.2 Previous Work on Dense Multiview Stereo for Compact Objects

Since the review of [48] and the associated Middlebury evaluation, a lot of research has been focusing on multiview reconstruction of small objects taken under 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 will be explained, most of these algorithms are not directly suited to large-scale outdoor scenes. A number of multiview stereo algorithms have been proposed that exploit the *visual hull* [41]. Many dense multiview methods rely on this information either as an initial guess for further optimization [26], [18], [29], [28], [55], [61], [64], [69], as a soft constraint [26], [35], or even as a hard constraint [51], [18] to be fulfilled by the reconstructed shape.

While the unavailability of the visual hull discards many of the top-performing multiview stereo algorithms of the Middlebury challenge [48], the requirement for the ability to handle large 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 [49], [37], [8], [68], level sets [14], [32],[46], and volumetric graph cuts [65], [6], [28], [43], [61] (though [50], [27] propose regular volumetric grid adaptive to photo-consistency measures to push the resolution limit further).

- *H.-H. Vu is with IMAGINE/CSTB, École des Ponts ParisTech, Université Paris-Est, 19, rue Alfred Nobel-Cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée Cedex 2, France.*
  E-mail: hoang.vu@polytechnique.org.
- *P. Labatut is with Nokia, 5, rue des Feuillantines, Paris 75005, France.*
  E-mail: patrick.labatut@normalesup.org.
- *J.-P. Pons and R. Keriven are with Acute3D-Center International de Communication, Avancée, 2229 route des Cretes, 06560 Sophia Antipolis, France. E-mail: {jean-philippe.pons, renaud.keriven}@acute3D.com.*

Manuscript received 25 Oct. 2010; revised 6 July 2011; accepted 21 July 2011; published online 8 Aug. 2011.
Recommended for acceptance by I. Reid.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-10-0813.
Digital Object Identifier no. 10.1109/TPAMI.2011.172.

Finally, cluttered scenes disqualify variational methods [14], [26], [12], [32], [44], [46], [11] that can easily get stuck into local minima, unless a way of estimating a close and reliable initial guess that takes visibility into account is provided.

1.3 Previous Work on Dense Multiview Stereo for Outdoor Scenes

Multiview stereo methods that have been proven to be more adapted to larger scenes, e.g., outdoor architectural scenes, usually initialize the scenes with sparser measurements such as depth maps or point clouds to reconstruct a surface.

The performance of some depth maps-based methods [36], [58], [56], [22], [57], [23], [24] for complete reconstruction, however, seems to be lower than previously discussed approaches, as regards either 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. While visibility is taken into account to fuse depth maps in [45], the focus on high performance prevents the use of a global optimization. Zach et al. [70] proposed a globally optimal variational merging of truncated signed distance maps using a volumetric grid. Another exception could be the work of [9], currently one of the most accurate methods according to the Middlebury evaluation, but this method relies on a volumetric graph cut [27] that cannot handle large-scale scenes.

Furukawa and Ponce [20] proposed a very accurate reconstruction that generates and propagates a semidense set of patches. This method has shown impressive results, but relies on filtering and expansion heuristics to process a set of oriented patches. The surface reconstruction step that converts the set oriented patches into a mesh is done by applying the well-known Poisson surface reconstruction [33], which requires dense and uniformly sampled point clouds and does not handle visibility issues. Finally, the obtained mesh has to be refined using a mesh evolution. This method has been tested on the data sets provided by Christoph Strecha et al. [59], the only available evaluation that allows comparison on large outdoor scenes (to our knowledge) and which obtained the best results at the moment of its publication. More recently, Tylecek and Sara [63] used depth map fusion, then refined camera center and mesh refinement, which obtained high accuracy but still lacked completeness. Salman and Yvinec [47], using our point clouds in [66], achieved nice completeness of the scenes.

1.3.1 3D Reconstruction on Internet Scale

Recent progress of Structure from Motion (SfM) and multiview methods allow researchers to handle larger collections of images of a given site available on the Internet. The challenge is how to calibrate thousands, even millions of images and how to reconstruct a 3D scene from these images within reasonable time. The standard way of calibration is to use bundler adjustment to estimate SfM of all these images [53], or its skeletal graph to reduce the computing cost [54]. Preprocessing to remove redundant images and fast matching before calibration could be useful to accelerate the calibration [16]. Implementations are also taken into account to exploit parallel computing on a cluster [1] or on a single computer with many CPU and GPU cores [16]. For 3D reconstruction, Goesele et al. [24] compute depth maps that form a point clouds and water-tight mesh using Poisson surface reconstruction. Furukawa et al. [17] partitioned the cameras in view clusters to run a multiview stereo of choice for each cluster. Taking into acount the known vertical of urban scenes, Frahm et al. [16] perform GPU-accelerated plane sweeping, and then extract the polygonal mesh. However, there is no qualitative benchmark on Internet scale to our knowledge and it may be difficult to create one.

From the above methods, the 3D reconstruction from an extremely large collection of images should take care of scability: removal of redundancy, streaming, division of data, parallelized computing, and perhaps a combination of partial results. In this paper, we are not targeting this challenge, and focus on a global reconstruction of a smaller scale (up to hundreds of high-resolution images) to result in a highly complete and accurate watertight mesh for a large outdoor scene from calibrated images.

1.4 Contributions

Our multiview stereo method consists of a pipeline that naturally handles large-scale open scenes while providing very accurate reconstructions within a very reasonable time. The whole pipeline is designed to not sacrifice accuracy for scalability. Several design choices are made and justified by an analysis of the weak points of other methods. The pipeline contains three main steps:

1. The generation of a quasidense point cloud with standard passive multiview stereo techniques.
2. The extraction of a mesh that respects visibility constraints and is close to the final reconstruction, with a minimum s-t cut-based optimization to fit a surface over the Delaunay triangulation of the points.
3. The variational refinement of this initial mesh to optimize its photo-consistency.

The present paper is an extended version of our recent conference paper [66], which builds on our previous work in this field. Compared to our preliminary work [38] on robust surface reconstruction from semidense point clouds from multiview stereo matching, the initial point cloud is generated in a denser and more accurate fashion; the surface reconstruction has been adapted to use a more suitable energy similar to [39]. Finally, the variational refinement uses an energy inspired from our previous work [46], but in a lightweight and scalable Lagrangian framework. Our experiments clearly demonstrate the its competitiveness on large data sets.

The rest of this paper is organized as follows: Section 2 gives some background on the different techniques needed in our approach: Delaunay triangulations, minimum s-t cuts for optimal binary labelings, and surface mesh optimization. In Section 3, the different steps of our multiview stereo reconstruction pipeline are described in details. Implementation aspects are discussed in Section 4 and, finally, Section 5 presents experiments on a variety of real data sets to demonstrate the potential of our pipeline for reconstructing complex large-scale scenes.

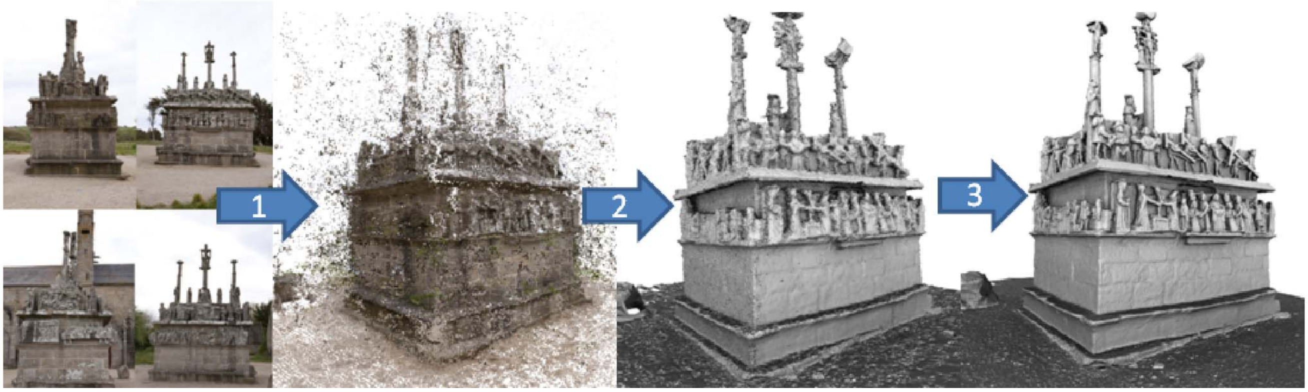


Fig. 1. Reconstruction pipeline. 1) Generate a points cloud. 2) Extract a visibility-consistency mesh. 3) Refine the mesh with photo-consistency optimization and regularization.

2 Background

In this section, some background and notations are provided: first, on the Delaunay triangulation upon which our surface reconstruction is based and, second, on variational methods applied to mesh deformation.

2.1 Delaunay Triangulation

A triangulation of a point set P in $\mathbb{R}^d$ is a partition of its convex hull into simplices of dimension d. In three dimensions, it is also called *tetrahedralization*. A Delaunay triangulation of a point set P is a triangulation in which no point in P is inside the circumcircle of any simplex of this triangulation. In the general position, where there are no $d+2$ points on the same sphere, the Delaunay triangulation is unique. Delaunay triangulation is a classical tool in the field of mesh generation and mesh processing due to its optimality properties [10].

2.2 Surface Optimization with Minimum s-t Cut

Given a finite directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$ and edges $\mathcal{E}$ with nonnegative weights (capacities) w_{pq}, and two special vertices, the source s and the sink t, an s-t-cut $\mathcal{C} = (\mathcal{S}, \mathcal{T})$ is a partition of $\mathcal{V}$ into two disjoints sets $\mathcal{S}$ and $\mathcal{T}$ such that $s \in \mathcal{S}$ and $t \in \mathcal{T}$.

The cost of the cut is the sum of the capacities of all the edges going from $\mathcal{S}$ to $\mathcal{T}$:

$$c(\mathcal{S}, \mathcal{T}) = \sum_{\substack{v_p \in \mathcal{S}\setminus\{s\} \\ v_q \in \mathcal{T}\setminus\{t\}}} w_{pq} + \sum_{v_p \in \mathcal{S}\setminus\{s\}} w_{pt} + \sum_{v_p \in \mathcal{T}\setminus\{t\}} w_{sp}. \qquad (1)$$

The minimum s-t-cut problem consists of finding a cut $\mathcal{C}$ with the smallest cost: The Ford-Fulkerson theorem [15] states that this problem is equivalent to computing the maximum flow from the source s to the sink t and many classical algorithms exist to efficiently solve this problem.

2.3 Dynamic Meshes: From Continuous to Discrete Gradient Flow

The last part of our pipeline evolves a triangular mesh to minimize a photo-consistency energy *w.r.t.* provided input images. In what follows, we describe how to compute a discrete gradient flow from a continous one. The variations of an energy E attached to a surface S can be analyzed with a functional gradient defined as the vector field ∇E such that for all vector fields v on S we have

$$D\mathrm{E}(S)[v] = \left.\frac{\partial \mathrm{E}(S + \epsilon\, v)}{\partial \epsilon}\right|_{\epsilon=0} = \int_S \nabla \mathrm{E}(x) v(x) dx. \qquad (2)$$

If the S is a the triangulated mesh, consisting of n vertices $X_i \in \mathbb{R}^3$, $i \in [1, n]$, a discrete vector field is defined at the vertices of this mesh by a sequence of vectors $v_i \in \mathbb{R}^3$, $i \in [1, n]$. Such a vector field is interpolated between the vertices over the whole mesh: $v(x) = \sum_i v_i \phi_i$ with $\sum_i \phi_i(x) = 1$ for all $x \in S$ (in the case of triangular facet, $\phi_i(x)$ is the barycentric coordinate corresponding to vertex i if i is one of vertices of a triangle containing x and 0 otherwise). Equation (2) becomes

$$D\mathrm{E}(S)[v] = \sum_i v_i \int_S \phi_i(x) \nabla \mathrm{E}(x) \mathrm{d}x. \qquad (3)$$

This equation naturally shows how to formulate a discrete gradient from a continous one:

$$\frac{\mathrm{dE}(S)}{\mathrm{d}X_i} = \int_S \phi_i(x) \nabla \mathrm{E}(x) \mathrm{d}x \qquad i \in [1, n]. \qquad (4)$$

3 Multiview Reconstruction Pipeline

As shown in Fig. 1 and previously announced, our dense multiview stereo pipeline is composed of three successive stages. Given calibrated cameras associated with the input images, a quasidense set of points is first extracted from the images. These points are matched pairwise between different views: From these matches, a quasidense 3D point cloud is generated by reconstructing and optionally merging the triangulated 3D points. This point cloud is then fed to the second stage, which builds a Delaunay triangulation from it and then robustly extracts an initial surface from the facets of this triangulation, filtering out most of the outliers. Finally, the last step improves the quality of the recovered surface by refining it using a criterion mixing photo-consistency and fairness.

3.1 Quasidense Point Cloud

In order to apply the surface fitting of the next step of our reconstruction pipeline, a slightly nonconventional way to generate point clouds from passive stereo is used that favors density over matching robustness. We describe two different but related point cloud generation strategies, one matching interest points in the input images and another using plane sweeping to compute sparse depth maps. We prefer the latter strategy, which is used in all our recent experiments because it generates more points.

3.1.1 Match of Interest Points

First, interest points are located in all the input images. For this purpose, and to capture most of the geometry of the sampled shape, two complementary kinds of interest points are considered: Harris corners, which typically lie on "corners" in images, and Laplacian-of-Gaussian, located at the center of blob-like structures in images. LoG blobs and Harris corners are extracted at some fixed scale[1] in all the input images. Then, for each potential camera pair (i, j) and for each interest point m_i (of the same type) in the first image I_i of this pair, its best matching point $m_j^\star$ is sought within a small band around the corresponding epipolar line in the other image I_j. The width of this band is fixed and should partially depend on the accuracy of the calibration.[2]

The best matching point $m_j^\star$ is the point with the highest matching score against the reference interest point m_i. The neighborhood of a potential match m_j in the image I_j is reprojected in the reference image I_i through a plane parallel to the focal plane of the camera i and passing through the potential reconstructed 3D point (the underlying assumption is that the surface is locally fronto-parallel to the camera i). The matching score can then be estimated in a window around the reference point. Since the choice of an appropriate matching window size is difficult, multilevel matching is used, and the matching criterion is the sum of normalized cross correlations (NCC) for several fixed window sizes[3] (or scale σ) as in [67].

Furthermore, this best matching interest point $m_j^\star$ is kept only if its matching score is above some threshold and if it is also successfully validated: The original interest point has to be the best matching interest point of its best matching interest point. An initial 3D point can then be reconstructed from the calibration by using standard triangulation optimization [25].

The final step aggregates the different 3D points. In each image, the 2D Delaunay triangulation of the interest points (of the same type) is computed. This geometric data structure allows to efficiently locate the nearest interest points of a given 2D point. Now, a pair of matched interest points in two different views has given rise to a 3D point by triangulation. By projecting this initial 3D point in the other views, potential other unmatched interest points that are close enough (within a tolerance similar to the half-width of the epipolar band) are located. Closest unmatched interest points are merged with the original pair and a new 3D point (replacing the previous one) is reestimated from all the interest points. The final result is a set of points each carrying a tuple of views where they were seen. In addition, a confidence value has been assigned to each 3D point, cumulating the photo-consistency scores of all its originating pairs. Obviously, as the whole technique relies on simple greedy or winner-take-all "optimization," it possibly generates a noisy point cloud with a decent amount of outliers.

3.1.2 Sparse Depth Maps

While the previous passive stereo approach is general and copes with scenes that have enough texture, it tends to generate lots of outliers and the 3D points are often poorly located. A different passive stereo technique can be devised when strong planar structures are observed, as is often the case in architectural scenes.

Initial sparse depth maps are computed between pairs of input images. These depth maps have a downscaled resolution[4] *w.r.t.* the images and are filled using a simple geometric plane sweep with the same thresholded multilevel NCC matching score and winner-takes-all optimization as above. A plane is swept in the reference camera frustum and its offset follows a geometric sequence between the near and far planes of the camera.

These initial depth maps are merged and clusters of points are formed according to their position in the different camera frustums. These clusters are hierarchically split until the bounding boxes of their projections in the images is small enough. A 3D k-D tree [3] of this clustered initial point set is then build to efficiently find the k nearest neighbors of each point using a large neighborhood.[5] A plane is tentatively fitted to each point's neighborhood with least squares. Provided the fit is good enough, the point is retained and its position is iteratively refined using the same matching score as above. The final result is the same as what was obtained from interest points: a set of points, each carrying a tuple of views where they were seen and an associated confidence. Again, this step still generates a noisy point cloud with a decent amount of outliers, but tends to yield better results on architectural scenes (fewer outliers and noise).

The advantage of the two passive stereo techniques presented lies in the fact that the reprojection and multilevel matching process can leverage the computational resources of common graphics hardware allowing the overall process to be reasonably fast (a few minutes in the data sets of [59] featuring from 8 to 30 images of 6 Mpixel, on an Intel Xeon 3.0 GHz CPU with a NVIDIA 260 GTX GPU).

As the reconstruction involves matching points in different images, the corresponding 3D error distribution is complex and cannot be modeled as simply as in the range scanning case. Mismatches are also almost inevitable, leading to gross outliers. Depending on the geometry of the cameras and the repetitiveness of texture patterns, these mismatches may even aggregate in structured clusters of outliers producing phantom structures in the point cloud. Another limitation of passive stereo is the highly nonuniform density of samples that depends on the amount of texture on the scene and object. While visibility filtering and expansion techniques combining heuristic-based optimizations have been able to improve the quality of point clouds from stereo, as in [19] and [24], standard point clouds from

[1] In practice, a scale of 2 pixels is used for 6 Mpixel images.

[2] A 3 pixel-wide band is typically chosen for 6 Mpixel images.

[3] Five levels are used on 6 Mpixel images.

[4] By a factor 4 × 4 for 6 Mpixel images.

[5] $k = 25$.

multiview such as the two acquisition methods described have notoriously higher levels of noise and higher ratio of outliers that point clouds acquired with laser range finding.

However, in our case, relying on thresholds and possibly generating numerous outliers is not a serious concern. The only goal of this point cloud from the passive stereo step is to generate enough points so that the following global optimization finds a close enough surface from the tetrahedra facets.

3.2 Visibility-Based Surface Reconstruction

The second step of our multiview pipeline consists of filtering gross outliers from the point cloud and reconstructing an initial surface. These two goals are achieved at once by relying on the Delaunay triangulation described in Section 2.1 and using a visibility-based formulation to build a surface and discard outliers.

3.2.1 Optimal Tetrahedron Binary Labeling

From the image-based point cloud $\mathcal{P}$ where each point memorizes the two or more images from which it has been triangulated v (as described in the previous section), the 3D Delaunay triangulation of these points is built. Then, the Delaunay tetrahedra are labeled inside or outside the object so that this binary labeling minimizes some energy and, finally, the surface is extracted as the set of triangles between inside and outside tetrahedra (called a pseudosurface in what follows).

3.2.2 Surface Visibility

A pseudosurface S^* is sought so as to minimize visibility constraints imposed by the line of sight of the acquired points: $S^* = \arg\min_S \mathrm{E}_{\mathrm{vis}}(S, \mathcal{P}, v)$.

A surface should never cross the empty space traversed by the various lines of sight attached to the points. Ideally, one would like to minimize the conflicts of the lines of sight with the surface S induced by the tetrahedron labeling l. This corresponds to the following term:

$$\sum_{P\in\mathcal{P}}\sum_{Q\in v_P} V_{\mathrm{conflict}}\left(l_{T_1^{Q\to P}}, \ldots, l_{T_{N_{[QP]}}^{Q\to P}}\right),$$

where $T_1^{Q\to P}, \ldots, T_{N_{[QP]}}^{Q\to P}$ is the ordered sequence of the $N = N_{[QP]}$ tetrahedra crossed from the camera center position Q to the point P (see Fig. 2). Since P is a vertex of the Delaunay triangulation, the sequence is terminated before the tetrahedron lying behind P as shown in the upper part of Fig. 2. Each oriented facet $F = (T_i^{Q\to P} \cap T_{i+1}^{Q\to P})$ for $i \in [1, N-1]$ is intersected by the line segment $[QP]$. To cast as a minimum s-t cut problem, we penalize the number of misalignments of the tetrahedra's label and define V_{conflict} as (we drop the notation $Q \to P$)

$$V_{\mathrm{conflict}}(l_{T_1}, \ldots, l_{T_N}) = \sum_{i=1}^{N-1} V_{\mathrm{align}}(l_{T_i}, l_{T_{i+1}}),$$

where V_{align} is a simple pairwise subterm defined for two adjacent cells of the complex (since in the above equation the cells are crossed in that order, they are adjacent to each other) $V_{\mathrm{align}}(l_{T_i}, l_{T_j}) = \alpha_{\mathrm{vis}}\ \mathbb{1}[\ l_{T_i} = 0\ \wedge\ l_{T_j} = 1]$ with α_{vis} is a constant *w.r.t.* the labeling but depends on the point or line of sight considered: It is a confidence measure of the point or line of sight. α_{vis} can be linked to the photo-consistency score of the triangulated 3D point.

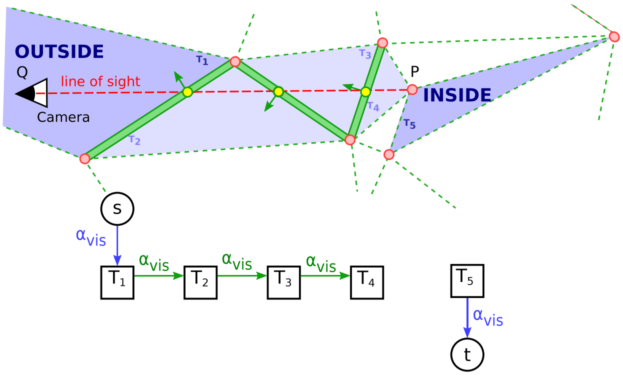


Fig. 2. Visibility and graph construction. A line of sight from a reconstructed 3D point traverses a sequence of tetrahedra, the graph construction, and the assignment of weights to the tetrahedra and oriented facets.

Since the trivial labeling $l^0 : t \in \mathcal{T} \to 0$ marking all tetrahedra as outside and to which an empty pseudosurface corresponds, satisfying these constraints, the facts that the point is assumed to lie near the surface and the camera centers have to be outside have to be considered. $T_1^{Q\to P}$ is the tetrahedron containing the camera and it should be marked as outside. We denote by $T_{N+1}^{Q\to P}$ the tetrahedron behind the point P in the direction of the line of sight and this tetrahedron should be favored as inside. Therefore, we add two more terms: $D_{\mathrm{out}}(l_T) = \alpha_{\mathrm{vis}} \mathbb{1}[l_T = 1]$ and $D_{\mathrm{in}}(l_T) = \alpha_{\mathrm{vis}} \mathbb{1}[l_T = 0]$.

To this end, $\mathrm{E}_{\mathrm{vis}}$ is the following expression:

$$\mathrm{E}_{\mathrm{vis}}(S, \mathcal{P}, v) = \sum_{P\in\mathcal{P}}\sum_{Q\in v_P} D_{\mathrm{out}}\left(l_{T_1^{Q\to P}}\right) \tag{5}$$

$$+ \sum_{i=1}^{N_{[QP]}-1} V_{\mathrm{align}}\left(l_{T_i^{Q\to P}}, l_{T_{i+1}^{Q\to P}}\right) \tag{6}$$

$$+ D_{\mathrm{in}}\left(l_{T_{N_{[QP]}+1}^{Q\to P}}\right). \tag{7}$$

The corresponding weight construction is shown in Fig. 2: The s-link of the vertex representing the tetrahedron T_1 is assigned α_{vis}, the t-link of vertex representing the tetrahedron T_{N+1} ($N = 4$ in Fig. 2) behind the point P is assigned α_{vis}, and each oriented facet crossed by the line of sight from P to Q is also assigned α_{vis}. These weight assignments are accumulated over all lines of sight, and computing a minimum s-t on this graph yields a globally optimal labeling.

One might wonder if alternatives would not be better suited to this problem, e.g., using the D_{out} subterm for all crossed tetrahedra, which leads to a guided ballooning force [42], [27]. Without an appropriate regularization term, that energy tends to minimize the number of "inside" tetrahedra in a light of sight no matter whether these tetrahedra are adjacent or not. It might lead to a fragmented surface. On the other hand, our visibity term minimizes the

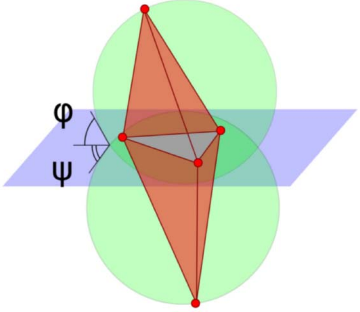


Fig. 3. Surface quality. A facet of the triangulation, its two adjacent tetrahedra (red), and their circumspheres (green). Their angles ϕ and ψ with the facet influence the weight this facet will get.

number of time the surface cut the light-of-sights, which favors a more regularized surface.

3.2.3 *Surface Quality*

As input images are available, an additional photo-consistency term E_{photo} may be used to favor surfaces with the best matching reprojections in the different views. This can also be implemented within the minimum s-t cut framework [38]. However, the resulting point cloud typically might contain millions of points (see Fig. 9); the photo-consistency term is quite expensive. Moreover, the visibility term of our energy is very effective to filter out outliers from stereo point clouds. Since the output surface is only used as an initialization for a variational photometric refinement, the photo-consistency term is advantageously replaced with the simple surface quality term E_{qual} of [40] for surface reconstruction from range scans: $E_{\text{qual}}(S) = \sum_f w_f \mathbb{1}[l_{T_1^f} \neq l_{T_2^f}]$. This sum is over every facet f in the triangulation, T_1^f and T_2^f the two tetrahedra incident to f, $w_f = 1 - \min\{\cos(\phi), \cos(\psi)\}$, where ϕ, ψ are the angles of the facet f with the circumspheres of T_1^f, T_2^f, respectively, (Fig. 3).

This term penalizes facets unlikely to appear on a densely sampled surface by using a geometric criterion related to the size of the empty circumspheres of a triangle. Support for infinite tetrahedra is 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 of outdoor scenes.

The energy to label tetrahedra, which can be globally minimized with minimum s-t cut, is thus

$$E(S) = \mathrm{E}_{\text{vis}}(S, \mathcal{P}, v) + \lambda_{\text{qual}}\, \mathrm{E}_{\text{qual}}(S), \tag{8}$$

where $\mathcal{P}$ is the generated point cloud and v the associated visibility sets of the points.

3.3 Photometric Robust Variational Refinement

As the initial surface reconstruction method is interpolatory and the point cloud still contains a decent amount of noise, the obtained initial mesh, noted as M^0, is noisy and fails to capture fine details. By using all the image data, this mesh is refined with a variational multiview stereovision approach pioneered by Faugeras and Keriven [14]: M^0 is used as the initial condition of a gradient descent of an adequate energy function. As the mesh M^0 is already close to the desired solution, this local optimization is very unlikely to get trapped in an irrelevant local minimum. The details of the energy function and the optimization procedure are now presented and the improvements over the initial method justified. This collection of improvements should not be considered as mere implementation details and all have a strong impact on the accuracy of the final reconstruction.

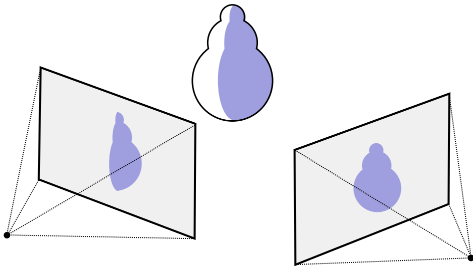


Fig. 4. Reprojection induced by the surface.

The inital mesh M^0, as the surface between interior and exterior tetrahedra, may still contain isolated triangles respecting visibility constraint (for example, from false points in the sky, background of scene) or big-size triangles (due to lack of density of points or lack of images in some area). Moreover, it might capture the landscape far from our scene, which we do not need to reconstruct in detail (plus it is impossible to refine this part accurately because of inexact calibration for scenes very far from cameras). For these reasons, we remove these triangles by some thresold of triangle size or number of triangles in an isolated piece, and manually cut unnecessary far landscape background.

3.3.1 *Photo-Consistency Refinement*

Let S be the object surface x a point on S, $\vec{n}$ the normal to S at point x, $g_{ij}(I_i, I_j)(x, \vec{n})$ a positive decreasing function of a photo-consistency measure of the patch $P = (x, \vec{n})$ according to images I_i and I_j, and $v_{ij}^S(x) \in \{0, 1\}$ the visibility of x in these images according to S. The original energy in [14] is

$$\mathrm{E}_{\text{photo}}(S) = \sum_{i,j} \int_S v_{ij}^S(x)\, g_{ij}(x, \vec{n})\, \mathrm{d}S. \tag{9}$$

Instead of this energy, the reprojection error introduced by [46] is preferred, namely,

$$\mathrm{E}_{\text{error}}(S) = \sum_{i,j} \int_{\Omega_{ij}^S} h(I_i, I_{ij}^S)(x_i)\, \mathrm{d}x_i, \tag{10}$$

where $h(I, J)(x)$ is a decreasing function of a photo-consistency measure between images I and J at pixel x (typically the opposite of normalized cross correlation), $I_{ij}^S = I_j \circ \Pi_j \circ \Pi_i^{-1}$ is the reprojection of image I_j into image I_i induced by S, and Ω_{ij}^S is the domain of definition of this reprojection (see Fig. 4), Π_i and Π_i^{-1} are the projection and back projection from an image i to the surface. This energy measures, for each considered camera pair, the dissimilarity between the portion of a reference image corresponding to the projected surface and a portion of another image reprojected via the surface into the reference image.

This summation has several major advantages over the original one:

1. Reprojecting I_j into I_i according to S uses the exact geometry of S and does not rely on any approximation of the tangent patch $(x, \vec{n})$.
2. The less a surface element is viewed in a given image, the less it contributes to the energy.
3. This reprojection can easily and efficiently be computed on graphics hardware with projective texture mapping.

The first point is essential to get an accurate reconstruction: In methods approximating the surface by planar patches, the choice of patch size is a difficult tradeoff between robust and accurate photo-consistency. In practice, we set the photo-consistency measure as the opposite of normalized cross correlation. This measure has the advantage of robustness to noise and light change, which occurs frequently for outdoor images, due to real lighting change and internal image processing inside cameras.

3.3.2 Regularization

The original intrinsic energy $\mathrm{E}_{\text{photo}}$ of (9) is self-regularizing due to the area-weighted integration over the surface. This is, however, not the case of (10). The energy function $\mathrm{E}_{\text{error}}$ is thus complemented with a surface fairing term E_{fair}, thin-plate energy that measures the total curvature of the surface. This term penalizes strong bending, not large surface area:

$$\mathrm{E}_{\text{fair}}(S) = \int_S (\kappa_1^2 + \kappa_2^2)\,\mathrm{d}S, \tag{11}$$

where κ_1 and κ_2 are the principal curvatures of the surface at the considered point. Consequently, the associated gradient flow is exempt from the classical shrinking bias.

3.4 Discretization

Many methods in variational multiview stereovision [12], [14], [32], [44], [46], 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 computed analytically, then the obtained minimization flow is discretized.

In contrast, a *discretize then optimize* approach is adopted: An energy function that depends on a discrete finite-dimensional surface representation, here a triangle mesh is considered, and standard nonconvex optimization tools are used. The benefits of this approach have long been recognized in mesh processing, but have seldom been demonstrated in computer vision [11], [26], [52].

As (4) shows, the obtained gradient vector at a vertex involves integrals over the ring of triangular facets around it (see also [13, Sec. 2.2]). This is in strong contrast with a point wise, and thereby noise sensitive, dependency on the input data that a late discretization typically causes. A crucial point has to be noted here: This discrete 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 Fig. 7.

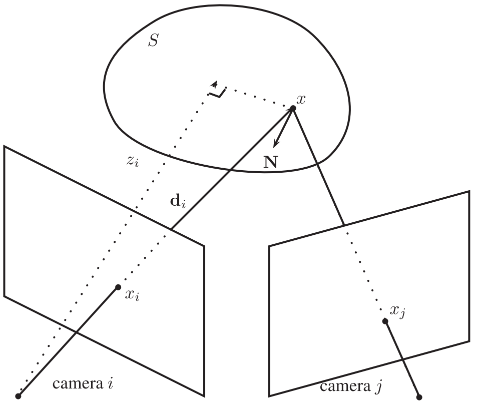


Fig. 5. Some notations in reprojection.

In what follows, we recall some definitions and results of [46] that are the base of our discretization.

Given two images: $I, J : \Omega \to \mathbb{R}^d$, let us consider $M(I,J) = \int_\Omega h(I,J)(x)\mathrm{d}x$ as a function of similarity of two images. $\partial_2 M(I,J)$ is defined as the derivative of $M(I,J)$ with respect to the second image, in the sense that, for any image variation δJ,

$$\lim_{\epsilon \to 0} \frac{M(I, J + \epsilon\delta J)}{\epsilon} = \int_\Omega \partial_2 M(I,J)(x)\delta J(x) dx. \tag{12}$$

We note $\mathcal{M}_{ij}(S) = M(I_i, I_{ij}^S)$, thus $\mathrm{E}_{\text{error}}(S) = \sum_{i,j} \mathcal{M}_{ij}(S)$ and $\nabla \mathrm{E}_{\text{error}}(S) = \sum_{i,j} \nabla \mathcal{M}_{ij}(S)$.

With a point $x \in S$ visible for cameras i and j, we note: $x_i = \Pi_i(x)$, $x_j = \Pi_j(x)$ the projection in image i, j, $\mathbf{d}_i$ the vector joining the center of camera i, and x, z_i the depth of x in camera i, $\mathbf{N}$ the outward surface normal at x (see Fig. 5). From [46, p. 10], we have

$$\mathrm{d}x_i = -\mathbf{N}^T \mathbf{d}_i \mathrm{d}x / z_i^3, \tag{13}$$

$$\nabla \mathcal{M}_{ij}(x) = -\left[\partial_2 M(x_i) DI_j(x_j) D\Pi_j(x) \frac{\mathbf{d}_i}{z_i^3}\right]\mathbf{N}, \tag{14}$$

with M the abbreviation for $M(I_i, I_{ij}^S)$, D denotes the Jacobian matrix of a function. The term between square brackets line is a scalar quantity. We note $f_{ij}(x_i) = \partial_2 M(x_i) DI_j(x_j) D\Pi_j(x)\mathbf{d}_i$, then $\nabla \mathcal{M}_{ij}(x) = -f_{ij}(x_i)\,\mathbf{N}/z_i^3$.

We rewrite (4) in dropping the index i of X_i and ϕ_i:

$$\frac{\mathrm{d}\mathrm{E}_{\text{error}}(S)}{\mathrm{d}X} = \int_S \phi(x) \sum_{i,j} \nabla \mathcal{M}_{ij}(x)\mathrm{d}x \tag{15}$$

$$= -\int_S \phi(x) \sum_{i,j} f_{ij}(x_i)\mathbf{N}/z_i^3 \mathrm{d}x \tag{16}$$

$$= -\sum_{i,j} \int_S \phi(x) f_{ij}(x_i)\mathbf{N}/z_i^3 \mathrm{d}x, \tag{17}$$

$$= \sum_{i,j} \int_{\Omega_{ij}} \phi(x) f_{ij}(x_i)\mathbf{N}/z_i^3 \frac{z_i^3}{\mathbf{N}^T\mathbf{d}_i}\mathbf{N}\mathrm{d}x_i \tag{18}$$

$$= \sum_{i,j} \int_{\Omega_{ij}} \phi(x) f_{ij}(x_i) / (\mathbf{N}^T \mathbf{d}_i) \mathbf{N} \mathrm{d}x_i, \qquad (19)$$

where Ω_{ij} is the map of the reprojection from image j to image i via the surface. Therefore, the gradient of each vertex equals the summation weighted (with barycentric coordinate) of contribution of all pixels lying in the projection of all the triangles containing this vertex for all pairs of images (i, j).

When the mesh parameterization is close to isometric, the gradient from the complementary thin-plate energy reduces to a simple bi-Laplacian Δ^2. A discrete analog of such simplified thin-plate energy and associated flow, described in [34], is used by applying the umbrella operator of [60] to approximate the Laplace-Beltrami operator. This particular choice has a convenient property of redistributing vertices along the surface, and in particular discourages degenerate triangles.

3.4.1 Balance between Photo-Consistency and Regularization

A long-standing issue in variational methods is the proper and automatic balancing between data attachment and smoothing terms. Designing a general solution to this problem is clearly beyond the scope of this paper. A specific strategy is instead proposed that allows to conduct all the following experiments *without adjusting parameters to each data set*. The solution is twofold.

First, the fact that regularization has to be more important where photo-consistency is less reliable is observed, in particular in textureless or low-textured image regions. Consequently, the contribution of camera pair (i, j) at pixel $\boldsymbol{x}_i$ in (19) is weighted by a reliability factor $r(x_i) = \min(\sigma_i^2, \sigma_j^2)/(\min(\sigma_i^2, \sigma_j^2) + \epsilon^2)$, where σ_i^2 and σ_j^2 denote the local variance at $\boldsymbol{x}_i$ in images I_i and I_{ij}^S, respectively, and ϵ is a constant.

Second, the two terms of the energy function are homogenized: While the data attachment term of (10) is homogeneous in an area in pixels, the discrete thin-plate term is homogeneous in squared world units. After weighting the contribution of each image in (10) by the square of the ratio between the average depth of the scene and the focal length in pixels, a scalar regularity weight can be defined whose optimal value is stable across very different data sets. As we previously mentioned, this thin-plate term not only plays an a priori knowledge of the model (Bayesian arguments), but stabilizes the mesh during the refinement by redistributing vertices along the surface.

3.4.2 Mesh Resolution

The resolution of the mesh is automatically and adaptively adjusted to image resolution: A triangular facet is subdivided if there is one camera pair such that the visible facet projection exceeds a user-defined number of pixels in both images. This threshold is set to 16 pixels in the experiments. A classical one-to-four triangle subdivision scheme is used, which has the advantage of preserving sharp edges. Nevertheless, we believe that other subdivision methods can be used.

4 Implementation Aspects

Parts of our reconstruction pipeline take advantage of the cheap parallel processing resources available in many consumer-grade graphics card: namely, the computation of the initial quasidense point cloud, the computation of the mesh velocity field (the normalized cross-correlation and the image reprojections), and also its evolution, which are mostly done with a custom combination of vertex, geometry, and fragment shaders. The independence of pixels of images in our computation helps our pipeline adapt very well to the graphics card. Our approach heavily relies on geometric data structures and queries: from the 2D and 3D Delaunay triangulations and its corresponding queries to dynamic meshes. Fortunately the computational geometry algorithms library (CGAL)[6] [4] defines robust and efficient implementations of all the geometric data structures, primitives, queries, and traversals needed for our different algorithms. Finally, the max-flow algorithm described in [5][7] is used to compute a minimum *s-t*-cut of our specifically designed network graphs.

With these implementation advantages, the overall running time is quite reasonable; for example, it takes 45 minutes for the whole pipeline in the data set Herz-Jesu-P25 provided by Strecha et al. [59], consisting of 25 images of resolution $3{,}072 \times 2{,}048$, most of the time being spent either in computing and selecting points when generating the initial point cloud or in the final photometric refinement.

5 Experimental Results

5.1 Compact Objects

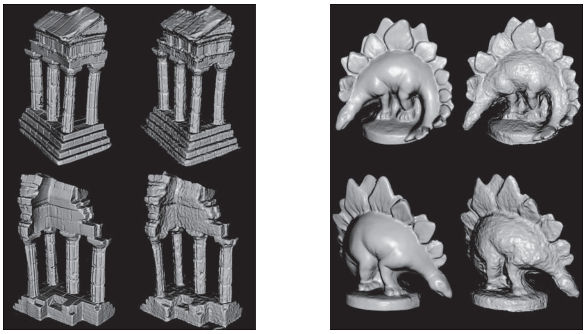
As mentioned in the introduction, our reconstruction pipeline does not target small-scale data sets for which the acquisition conditions can typically be easily modified to allow a foreground/background segmentation. Nevertheless, Fig. 6 shows the results of our final variational refinement step (from a mesh approximating the visual hull) and evaluation on the Middlebury dense multiview stereo benchmark of [48]. For the sake of comparison, we have included the results of other methods, including the results of our previous level set-based method [46] and another mesh-based variational approach based on the same energy function as our previous work [71]. Our results on the *templeRing* are currently the best both in completeness and accuracy. However, on the *dinoRing*, while a highly complete reconstruction is indeed achieved, our results are less competitive in terms of accuracy. This may be explained in the strong lack of texture on this particular data set that makes our photo-consistency measurement less peaked near the ground-truth surface.

5.2 Outdoor Architectural Scenes

Provided by Strecha et al. [59], the already mentioned data sets consists of outdoor scenes acquired with 8 to 30 calibrated 6 Mpixel images. Ground truth has been acquired with a LIDAR system. The evaluation of the

[6] http://www.cgal.org/.

[7] And implemented in http://www.adastral.ucl.ac.uk/~vladkolm/software.html.



	accuracy (at 90%)	completeness (at 1.25mm)
Us	0.45mm	99.8%
[9]	0.48mm	99.4%
[21]	0.47mm	99.6%
[26]	0.52mm	99.5%
[46]	0.60mm	99.5%
[71]	0.55mm	99.2%

	accuracy (at 90%)	completeness (at 1.25mm)
Us	0.53mm	99.7%
[7]	0.39mm	97.6%
[21]	0.28mm	99.8%
[35]	0.43mm	99.4%
[26]	0.45mm	97.9%
[46]	0.55mm	99.0%
[71]	0.42mm	98.6%

Fig. 6. Comparison to ground truth (top images: left column is ground truth, right column is our result) and evaluation results (bottom tables) on the *dinoRing* and *templeRing* data sets of [48].

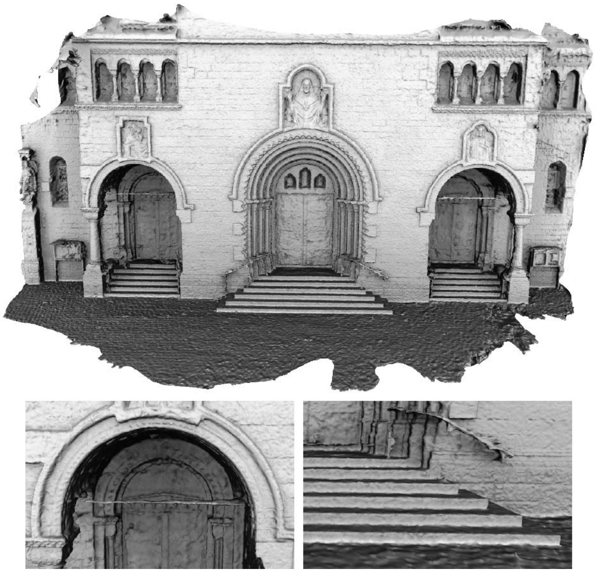


Fig. 7. Top: Overview of our reconstruction of *Herz-Jesu-P25*. Bottom: Close-ups on reconstruction details such as the thin metal bars, the facade relief or the staircases.

multiview stereo reconstructions is quantified through relative error histograms counting the percentage of the scene recovered within a range of 1 to 10 times the estimated LIDAR depth standard deviation σ. Dedicated to large-scale objects and fitting our objective perfectly, these sets are particularly challenging, especially the *castle-P19*, a complete courtyard acquired from the inside and where a tractor is placed in the middle, disturbing reconstruction. So far, [19], [31], [47], [62], [63] submitted for all these particular data sets. Some of them appeared after our conference version of this paper [66], yet our results still achieve the best at accuracy and completeness in most data sets of this benchmark. Comparisons with the other methods are given in Fig. 8, where cumulated histograms clearly show that the proposed pipeline is both more accurate (thanks to the final variational refinement) and complete (thanks to the initial visibility-consistent mesh). More detailed views of our reconstruction of the *Herz-Jesu-P25* data set are shown in Fig. 7. Note how details, topology (e.g., columns), and edges (e.g., stairs) are precisely recovered, while regularization still handles as correctly as possible blurred or untextured parts. Further results are available on the challenge website.[8]

5.3 Landscape and Cultural Heritage Scenes

The method was tested on an aerial acquisition of the *Aiguille du Midi* summit (data and calibration courtesy Bernard Vallet and Marc Pierrot-Deseilligny, respectively). The data set consists of 53 images of 5 Mpixel. Fig. 9 shows two of the images, the generated point cloud, 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. The variational process is able to recover the top antenna, although it is only partially present in M^0. Fig. 1 shows results on a data set of 27 images of 10 Mpixel of a sculpted calvary taken from the ground. The cloud has 802K points, with many outliers, mainly sky points obtained by matching clouds that have moved between shots; 539K of these points are selected for the initial mesh. This mesh is noisy due to the process of matching interest points that are just approximately viewpoint invariant. The closer views in Fig. 10 show the final reconstruction (2,331K triangles) is very sharp, to capture meaningful details. Fig. 11 shows results on a data set of 30 images of 14 Mpixel of Cluny Abbey in France, taken from a balloon in front. Lacking different views, the total scene is not complete, but the final reconstruction proves its great details from the direction of input images. We also tested on an aerial acquisiton of *Entrevaux* (Fig. 12), consisting of 109 images of 3.1 Mpixel. The final mesh is very complete, capturing small details of buildings and cliffs with trees. Note that trees are not suitable for multiview or mesh representation because of their complex and changing shape in time. Nevertheless, our method is robust enough to give them a reasonable form.

6 Conclusion and Future Work

A novel dense multiview stereo reconstruction pipeline has been presented. The whole method is designed to handle the reconstruction of large-scale cluttered scenes taken under uncontrolled imaging conditions, a scenario where traditional multiview stereo methods are either not applicable or have completeness and accuracy issues in part due to a lack of a correct treatment of visibility issues. The initial surface reconstruction problem is cast as the recovery of a visibility-consistent surface from the Delaunay triangulation of a quasidense point generated from the input images. This problem is reduced to a binary labeling of tetrahedron

[8] http://cvlab.epfl.ch/~strecha/multiview/denseMVS.html.

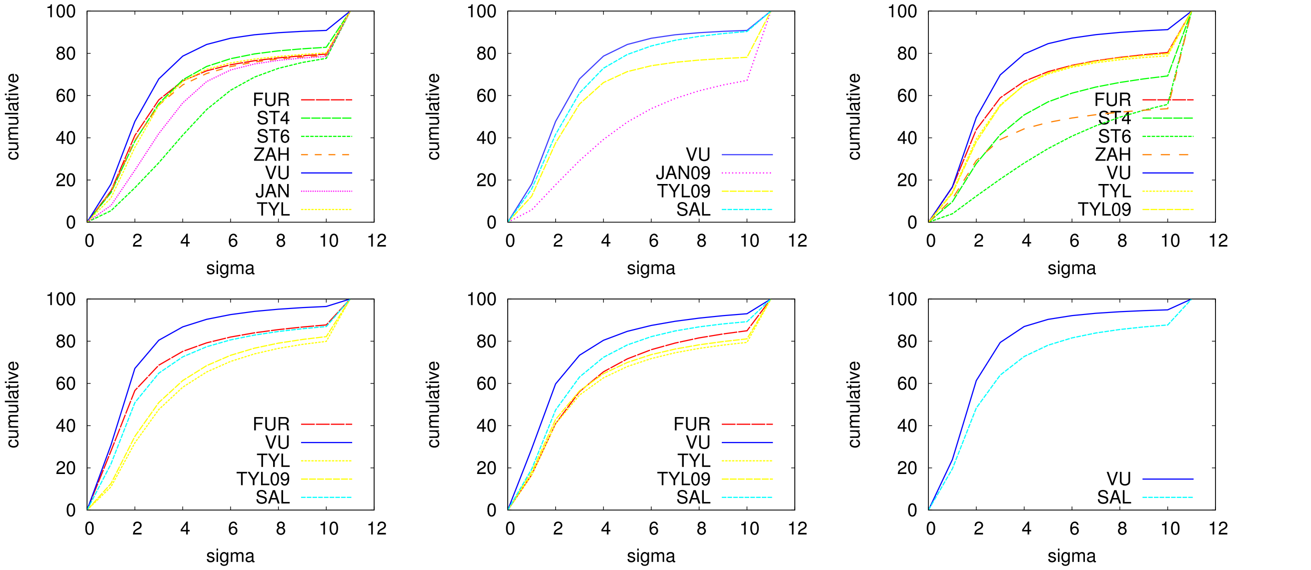


Fig. 8. Relative error cumulated histograms. From left to right, up to down, the relative error cumulated histograms, respectively, for the *fountain-P11* (two first histograms), *Herz-Jesu-P8*, *entry-P10*, *castle-P19*, *Herz-Jesu-P25* data set. The legend is the following: FUR for [19], ST4 for [56], ST6 for [57], ZAH for [71], TYL for [62], TYL09 for [63], JAN for [30], JAN09 for [31], SAL for [47], and VU for our work. On all data sets, the measurements clearly confirm our better results, both in accuracy and completeness.

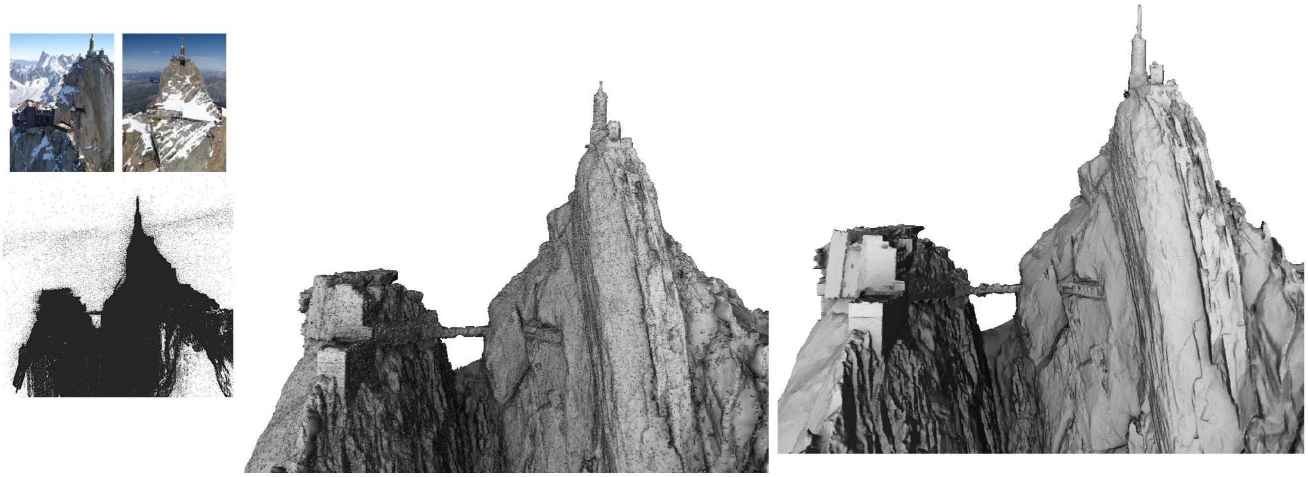


Fig. 9. Results on an Aiguille-du-Midi data set. From left to right: Two sample images taken from a helicopter (© B.Vallet/IMAGINE), point cloud from interest points, initial surface, and our final reconstruction.

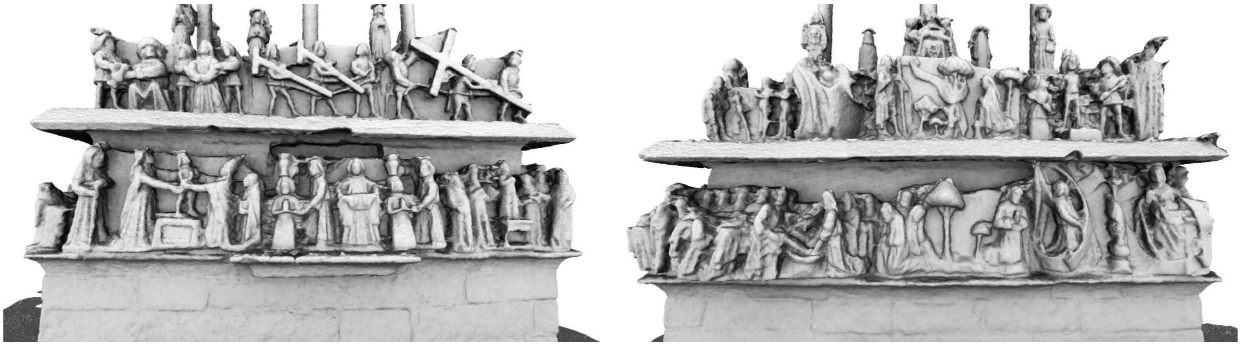


Fig. 10. Refined mesh on ground-level scene calvary.

that can be efficiently computed with a minimum s-t cut: The obtained surface is both complete and close to the ground truth and serves as a coarse initial estimate of the scene or object of interest. Its accuracy is then improved by a carefully designed and also scalable variational refinement. The full multiview stereo pipeline has been demonstrated on a number of large-scale scenes. Its output reconstructions are visually and quantitatively more accurate and complete than state-of-the-art techniques. Regarding future work, we will adapt the whole pipeline with parallel computing to reduce running time and to cope with larger data sets.

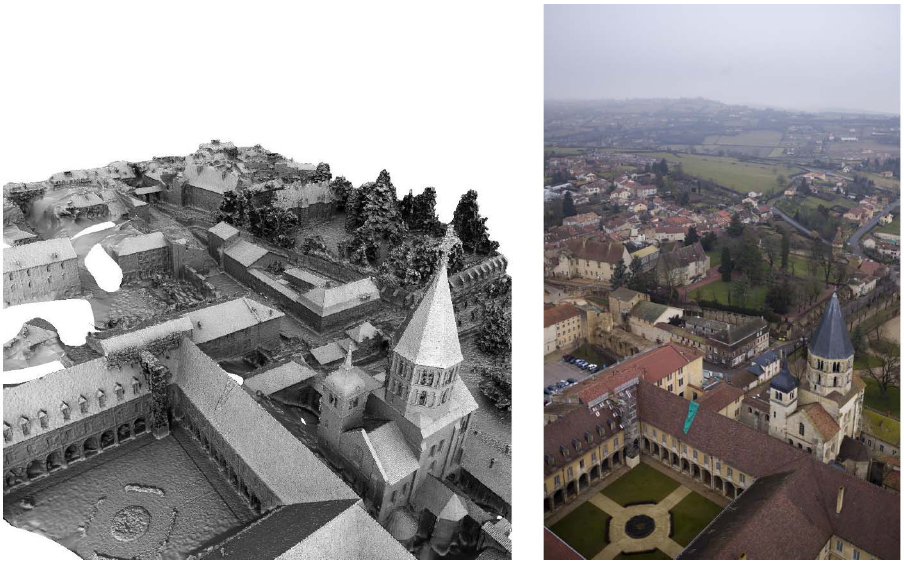


Fig. 11. Results on Cluny data set. Cluny Abbey built from 30 images taken from a balloon (© B.Vallet/IMAGINE) and an image of data set.

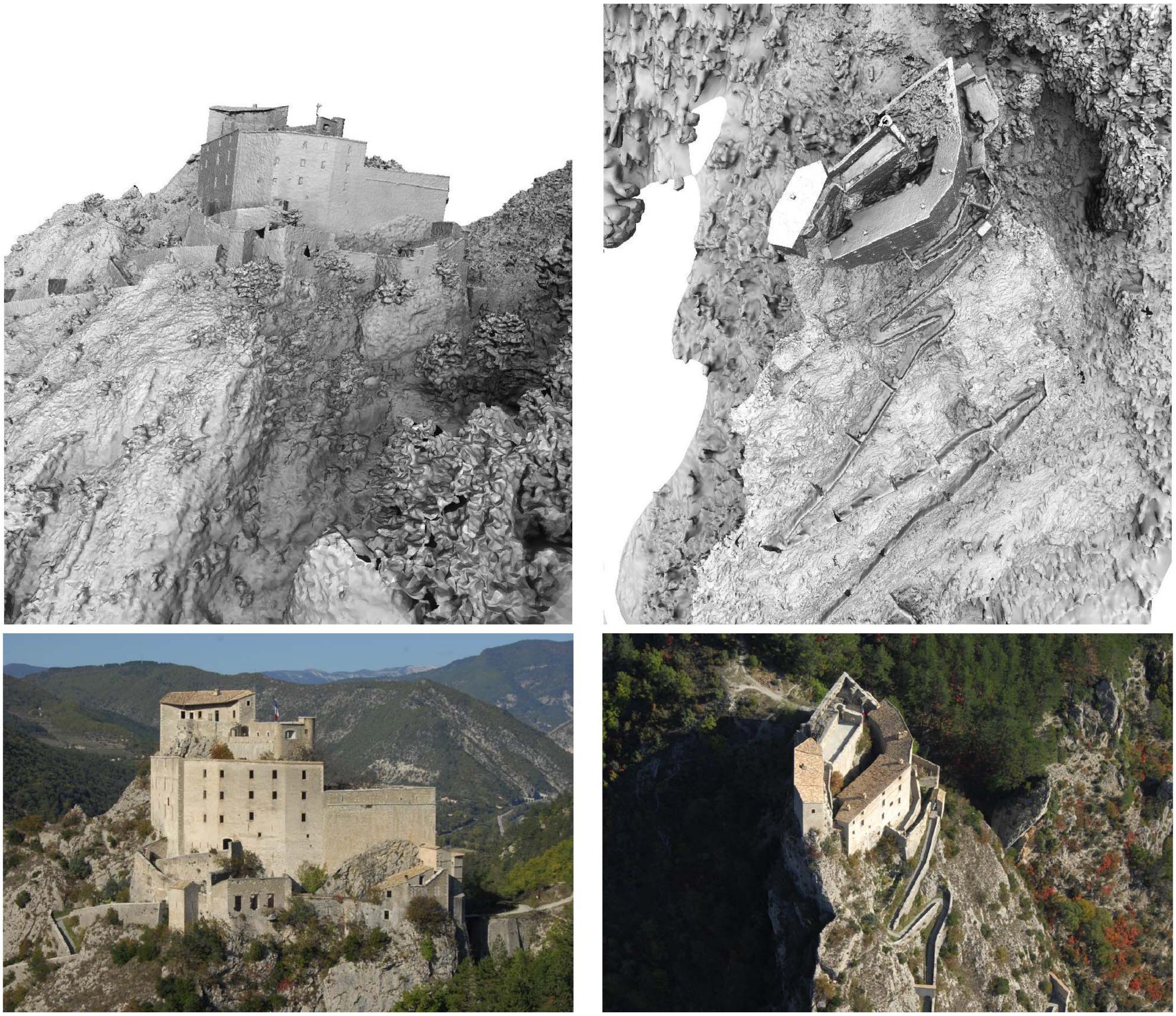


Fig. 12. Results on Entrevaux data set. Nontexture mesh from 109 images taken from a helicopter (© IMAGINE/CSTB), seen in two views associated with similar images.

Acknowledgments

The authors of the Middlebury challenge [48] are warmly thanked for stimulating research on the dense multiview stereo reconstruction problem within the computer vision and graphics communities. The authors would like to express their gratitude to the authors of [59] for making large-scale and fully calibrated data sets publicly available to the community along with a corresponding benchmark. The authors are especially thankful to Christoph Strecha for his time and swift answers when submitting the presented results. The Aiguille du midi mountain summit and the Cluny abbey data sets are courtesy of Bernard Vallet. Marc Pierrot-Deseilligny provided the calibration data for the Aiguille du midi mountain summit data set.

References

[1] S. Agarwal, N. Snavely, I. Simon, S.M. Seitz, and R. Szeliski, "Building Rome in a Day," *Proc. 12th IEEE Int'l Conf. Computer Vision*, 2009.

[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," *Int'l J. Computer Vision*, vol. 78, nos. 2/3, pp. 207-222, 2008.

[3] J.L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching," *Comm. ACM*, vol. 18, no. 9, pp. 509-517, 1975.

[4] J.-D. Boissonnat, O. Devillers, M. Teillaud, and M. Yvinec, "Triangulations in CGAL," *Proc. 16th Ann. Symp. Computational Geometry*, pp. 11-18, 2000.

[5] Y. Boykov and V. Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124-1137, Sept. 2004.

[6] Y. Boykov and V. Lempitsky, "From Photohulls to Photoflux Optimization," *Proc. British Machine Vision Conf.*, vol. 3, pp. 1149-1158, 2006.

[7] D. Bradley, T. Boubekeur, and W. Heidrich, "Accurate Multi-View Reconstruction Using Robust Binocular Stereo and Surface Meshing," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[8] A. Broadhurst, T.W. Drummond, and R. Cipolla, "A Probabilistic Framework for Space Carving," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 388-393, 2001.

[9] N.D.F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo," *Proc. European Conf. Computer Vision*, pp. 766-779, 2008.

[10] F. Cazals and J. Giesen, "Delaunay Triangulation Based Surface Reconstruction, Mathematics and Visualization," *Effective Computational Geometry for Curves and Surfaces*, pp. 231-276, Springer, 2006.

[11] A. Delaunoy, E. Prados, P. Gargallo, J.-P. Pons, and P. Sturm, "Minimizing the Multi-View Stereo Reprojection Error for Triangular Surface Meshes," *Proc. 19th British Machine Vision Conf.*, 2008.

[12] Y. Duan, L. Yang, H. Qin, and D. Samaras, "Shape Reconstruction from 3D and 2D Data Using PDE-Based Deformable Surfaces," *Proc. European Conf. Computer Vision*, vol. 3, pp. 238-251, 2004.

[13] I. Eckstein, J.-P. Pons, Y. Tong, C.C.J. Kuo, and M. Desbrun, "Generalized Surface Flows for Mesh Processing," *Proc. Fifth Symp. Geometry Processing*, pp. 183-192, 2007.

[14] O. Faugeras and R. Keriven, "Variational Principles, Surface Evolution, PDE's, Level Set Methods and the Stereo Problem," *IEEE Trans. Image Processing*, vol. 7, no. 3, pp. 336-344, Mar. 1998.

[15] L.R. Ford and D.R. Fulkerson, *Flows in Networks.* Princeton Univ. Press, 1962.

[16] J.-M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dumn, B. Clipp, S. Lazebnik, and M. Pollefeys, "Building Rome on a Cloudless Day," *Proc. European Conf. Computer Vision*, 2010.

[17] Y. Furukawa, B. Curless, S.M. Seitz, and R. Szeliski, "Towards Internet-Scale Multi-View Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[18] Y. Furukawa and J. Ponce, "Carved Visual Hulls for Image-Based Modeling," *Proc. European Conf. Computer Vision*, pp. 564-577, May 2006.

[19] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2007.

[20] Y. Furukawa and J. Ponce, "Dense 3D Motion Capture from Synchronized Video Streams," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2008.

[21] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362-1376, Aug. 2009.

[22] P. Gargallo and P. Sturm, "Bayesian 3D Modeling from Images Using Multiple Depth Maps," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 885-891, June 2005.

[23] M. Goesele, B. Curless, and S.M. Seitz, "Multi-View Stereo Revisited," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 2402-2409, 2006.

[24] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S.M. Seitz, "Multi-View Stereo for Community Photo Collections," *Proc. IEEE Int'l Conf. Computer Vision*, 2007.

[25] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, second ed. Cambridge Univ. Press, 2004.

[26] C. Hernández and F. Schmitt, "Silhouette and Stereo Fusion for 3D Object Modeling," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 367-392, Dec. 2004.

[27] C. Hernández, G. Vogiatzis, and R. Cipolla, "Probabilistic Visibility for Multi-View Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2007.

[28] A. Hornung and L. Kobbelt, "Hierarchical Volumetric Multi-View Stereo Reconstruction of Manifold Surfaces Based on Dual Graph Embedding," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 503-510, 2006.

[29] A. Hornung and L. Kobbelt, "Robust and Efficient Photo-Consistency Estimation for Volumetric 3D Reconstruction," *Proc. European Conf. Computer Vision*, 2006.

[30] M. Jancosek and T. Pajdla, "Segmentation Based Multi-View Stereo," *Proc. Computer Vision Winter Workshop*, 2009.

[31] M. Jancosek, A. Shekhovtsov, and T. Pajdla, "Scalable Multi-View Stereo," *Proc. IEEE Int'l Workshop 3D Digital Imaging and Modeling*, 2009.

[32] H. Jin, S. Soatto, and A.J. Yezzi, "Multi-View Stereo Reconstruction of Dense Shape and Complex Appearance," *Int'l J. Computer Vision*, vol. 63, no. 3, pp. 175-189, 2005.

[33] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson Surface Reconstruction," *Proc. Fourth Eurographics Symp. Geometry Processing*, pp. 61-70, June 2006.

[34] L. Kobbelt, S. Campagna, J. Vorsatz, and H.-P. Seidel, "Interactive Multi-Resolution Modeling on Arbitrary Meshes," *Proc. Int'l Conf. Computer Graphics and Interactive Techniques*, pp. 105-114, 1998.

[35] K. Kolev, M. Klodt, T. Brox, and D. Cremers, "Continuous Global Optimization in Multiview 3D Reconstruction," *Int'l J. Computer Vision*, vol. 84, pp. 80-96, 2009.

[36] V. Kolmogorov and R. Zabih, "Multi-Camera Scene Reconstruction via Graph Cuts," *Proc. European Conf. Computer Vision*, vol. 3, pp. 82-96, May 2002.

[37] K.N. Kutulakos and S.M. Seitz, "A Theory of Shape by Space Carving," *Int'l J. Computer Vision*, vol. 38, no. 3, pp. 199-218, 2000.

[38] P. Labatut, J.-P. Pons, and R. Keriven, "Efficient Multi-View Reconstruction of Large-Scale Scenes Using Interest Points, Delaunay Triangulation and Graph Cuts," *Proc. IEEE Int'l Conf. Computer Vision*, Oct. 2007.

[39] P. Labatut, J.-P. Pons, and R. Keriven, "Hierarchical Shape-Based Surface Reconstruction for Dense Multi-View Stereo," *Proc. IEEE Int'l Workshop 3D Digital Imaging and Modeling*, 2009.

[40] P. Labatut, J.-P. Pons, and R. Keriven, "Robust and Efficient Surface Reconstruction from Range Data," *Computer Graphics Forum*, vol. 28, no. 8, pp. 2275-2290, 2009.

[41] A. Laurentini, "The Visual Hull Concept for Silhouette-Based Image Understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150-162, Feb. 1994.

[42] V. Lempitsky and Y. Boykov, "Global Optimization for Shape Fitting," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2007.

[43] V. Lempitsky, Y. Boykov, and D. Ivanov, "Oriented Visibility for Multiview Reconstruction," *Proc. European Conf. Computer Vision*, vol. 3, pp. 226-238, May 2006.

[44] M. Lhuillier and L. Quan, "A Quasi-Dense Approach to Surface Reconstruction from Uncalibrated Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 418-433, Mar. 2005.

[45] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-Time Visibility-Based Fusion of Depth Maps," *Proc. IEEE Int'l Conf. Computer Vision,* Oct. 2007.

[46] J.-P. Pons, R. Keriven, and O. Faugeras, "Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score," *Int'l J. Computer Vision,* vol. 72, no. 2, pp. 179-193, 2007.

[47] N. Salman and M. Yvinec, "Surface Reconstruction from Multi-View Stereo," *Proc. Int'l Workshop Representation and Modeling of Large-Scale 3D Environments,* 2009.

[48] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 519-526, June 2006.

[49] S.M. Seitz and C.R. Dyer, "Photorealistic Scene Reconstruction by Voxel Coloring," *Int'l J. Computer Vision,* vol. 35, no. 2, pp. 151-173, Nov. 1999.

[50] S.N. Sinha, P. Mordohai, and M. Pollefeys, "Multi-View Stereo via Graph Cuts on the Dual of an Adaptive Tetrahedral Mesh," *Proc. IEEE Int'l Conf. Computer Vision,* Oct. 2007.

[51] S.N. Sinha and M. Pollefeys, "Multi-View Reconstruction Using Photo-Consistency and Exact Silhouette Constraints: A Maximum-Flow Formulation," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 349-356, Oct. 2005.

[52] G. Slabaugh and G. Unal, "Active Polyhedron: Surface Evolution Theory Applied to Deformable Meshes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 84-91, 2005.

[53] N. Snavely, S.M. Seitz, and R. Szeliski, "Modeling the World from Internet Photo Collections," *Int'l J. Computer Vision,* vol. 80, pp. 189-210, 2008.

[54] N. Snavely, S.M. Seitz, and R. Szeliski, "Skeletal Sets for Efficient Structure from Motion," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[55] J. Starck, G. Miller, and A. Hilton, "Volumetric Stereo with Silhouette and Feature Constraints," *Proc. British Machine Vision Conf.,* vol. 3, pp. 1189-1198, 2006.

[56] C. Strecha, R. Fransens, and L.V. Gool, "Wide-Baseline Stereo from Multiple Views: A Probabilistic Account," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 552-559, 2004.

[57] C. Strecha, R. Fransens, and L.V. Gool, "Combined Depth and Outlier Estimation in Multi-View Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2394-2401, 2006.

[58] C. Strecha, T. Tuytelaars, and L.V. Gool, "Dense Matching of Multiple Wide-Baseline Views," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 1194-1201, 2003.

[59] C. Strecha, W. von Hansen, L.V. Gool, P. Fua, and U. Thoennessen, "On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[60] G. Taubin, "A Signal Processing Approach to Fair Surface Design," *Proc. ACM Siggraph,* pp. 351-358, 1995.

[61] S. Tran and L. Davis, "3D Surface Reconstruction Using Graph Cuts with Surface Constraints," *Proc. European Conf. Computer Vision,* vol. 2, pp. 219-231, 2006.

[62] R. Tylecek and R. Sara, "Depth Map Fusion with Camera Position Refinement," *Proc. Computer Vision Winter Workshop,* 2009.

[63] R. Tylecek and R. Sara, "Refinement of Surface Mesh for Accurate Multi-View Reconstruction," *Int'l J. Virtual Reality,* vol. 9, pp. 45-54, 2010.

[64] G. Vogiatzis, C. Hernández, P.H.S. Torr, and R. Cipolla, "Multi-View Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 12, pp. 2241-2246, Dec. 2007.

[65] G. Vogiatzis, P.H.S. Torr, and R. Cipolla, "Multi-View Stereo via Volumetric Graph-Cuts," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 391-398, 2005.

[66] H.H. Vu, R. Keriven, P. Labatut, and J.-P. Pons, "Towards High-Resolution Large-Scale Multi-View Stereo," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2009.

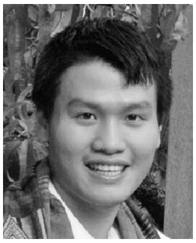
[67] R. Yang and M. Pollefeys, "Multi-Resolution Real-Time Stereo on Commodity Graphics Hardware," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 211-220, 2003.

[68] 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," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 576-584, 2003.

[69] T. Yu, N. Ahuja, and W.-C. Chen, "SDG Cut: 3D Reconstruction of Non-Lambertian Objects Using Graph Cuts on Surface Distance Grid," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2269-2276, 2006.

[70] C. Zach, T. Pock, and H. Bischof, "A Globally Optimal Algorithm for Robust TV-L1 Range Image Integration," *Proc. IEEE Int'l Conf. Computer Vision,* Oct. 2007.

[71] A. Zaharescu, E. Boyer, and R.P. Horaud, "TransforMesh: A Topology-Adaptive Mesh-Based Approach to Surface Evolution," *Proc. Asian Conf. Computer Vision,* pp. 166-175, Nov. 2007.



Hoang-Hiep Vu received the MS degree of computer science from the École Polytechnique, France, in 2008. Currently, he is working toward the graduate degree in the IMAGINE group at the École des Ponts ParisTech, under the supervision of Renaud Keriven. His research interests include multiview stereovision, mesh processing, and GPGPU.

Patrick Labatut is an alumnus of the École Normale Supérieure, Paris, where he received the BS and MS degrees in 2003 and 2005, respectively. He received the PhD degree in computer science from the Université Paris VII in 2009. Currently, he is working as a principal engineer at Nokia, working on GIS rendering. His research interests include image-based modeling and surface reconstruction.

Jean-Philippe Pons received the MS degree from the École Polytechnique, France, in 1999, and the PhD degree from the École des Ponts ParisTech in 2005. Currently, he is working as a general manager of Acute3D, a start-up dedicated to providing the 3D industry with software components on automatic 3D reality capture. From 2006 to 2010, he was a permanent researcher in the IMAGINE group, directed by Renaud Keriven, at the École des Ponts ParisTech and CSTB. In 2006, he was a postdoctoral researcher in the GEOMETRICA Team at INRIA Sophia Antipolis, under the supervision of Jean-Daniel Boissonnat, and in the Applied Geometry Laboratory at Caltech, under the supervision of Mathieu Desbrun.

Renaud Keriven received the MS degree from the École Polytechnique, France, in 1988, and the PhD degree from the École des Ponts ParisTech in 1997 on level sets method in stereo and computer vision. Currently, he is working as a general manager of Acute3D. Previously, he was a professor of computer science at the École des Ponts ParisTech, where he headed the IMAGINE group, and an associate professor at the École Polytechnique, France. From 2002 to 2007, he was an assistant director of the INRIA Odyss'ee team (leader Professor O. Faugeras) at the École Normale Supérieure, Paris. His research interests include 3D photography, multiview stereovision, shapes and shape priors in computer vision, discrete and continuous optimization in computer vision, and generic programming on graphics processing units.