A hybrid multi-view stereo algorithm for modeling urban scenes

Florent Lafarge, Renaud Keriven, Mathieu Brédif, Hoang-Hiep Vu

A Hybrid Multi-View Stereo Algorithm for Modeling Urban Scenes

Florent Lafarge, Renaud Keriven, Mathieu Brédif, Hoang-Hiep Vu

Abstract—We present an original multi-view stereo reconstruction algorithm which allows the 3D-modeling of urban scenes as a combination of meshes and geometric primitives. The method provides a compact model while preserving details: irregular elements such as statues and ornaments are described by meshes whereas regular structures such as columns and walls are described by primitives (*planes*, *spheres*, *cylinders*, *cones* and *tori*). We adopt a two-step strategy consisting first in segmenting the initial mesh-based surface using a multi-label Markov Random Field based model and second, in sampling primitive and mesh components simultaneously on the obtained partition by a Jump-Diffusion process. The quality of a reconstruction is measured by a multi-object energy model which takes into account both photo-consistency and semantic considerations (i.e. geometry and shape layout). The segmentation and sampling steps are embedded into an iterative refinement procedure which provides an increasingly accurate hybrid representation. Experimental results on complex urban structures and large scenes are presented and compared to the state-of-the-art multi-view stereo meshing algorithms.

Index Terms—3D modeling, multi-view stereo, urban scenes, hybrid representation, jump-diffusion.

1 Introduction

The 3D reconstruction of urban environments from multi-view stereo images is a well known computer vision problem which has been addressed by various approaches but remains a challenging issue [1], [2]. Urban scenes are difficult to analyze mainly because the structures significantly differ in terms of complexity, diversity, and density within the same scene. Created according to man-made rules, this kind of scenes has some common characteristics. In particular, the high occurrence of piecewise planar structures and the shape repetition are increasingly exploited in the recent methods.

1.1 Related works

In the literature, urban scene reconstruction from multi-view stereo images has been addressed by different families of approaches which can be ranked according to the nature of the 3D-models.

Meshes constitute a collection of vertices, edges and faces that defines the shape of polyhedral objects. Several types of methods are used to generate a mesh from multi-view stereo images. Region growing based methods consist in generating a cloud of 3D points from the images, and then expanding small flat patches from these points using photo-consistency considerations. These methods provide good results in practice, as demonstrated by Furukawa *et al.* [3] or Goesele *et al.* [4], but usually suffer from a large amount of tunable parameters which makes difficult to reach the optimal results. Others works, *e.g.* Campbell *et al.* [5] or Kolev *et al.* [6], tackle the problem by fusing a set of depth maps, *i.e.* view-dependent $2.5D$ representations where each point of the map is associated with a single depth value. These methods can be easily parallelizable. Nevertheless the depth map quality is usually low when the images are either poorly textured or of low resolution, as underlined in various works and comparative studies [7], [8], [9]. Some variational methods, *e.g.* Pons *et al.* [10], make the meshes evolve towards a local solution minimizing image prediction errors. Less common but particularly efficient for occlusions, some methods, *e.g.* Vu *et al.* [11] or Labatut *et al.* [12], use 3D Delaunay triangulations generated from a cloud of 3D points to extract a photo-consistent surface. Note also that the mesh-based representations of urban scenes can be obtained from laser scanning [13], [14], video sequences [15] and even Internet-based photo collections [16], [17], even if, in this last case, the results are still particularly sparse.

The efficiency of these different methods is evaluated on common datasets, in particular the Middlebury [18] and Strecha facade [19] benchmarks. Generally speaking, the mesh-based models generated from multi-view stereo images provide highly detailed descriptions of urban structures featuring ornaments, statues and other irregular shapes. However, man made objects contain many regular structures which are not optimally modeled by meshes in practice. For

- *F. Lafarge is with the Geometrica Research Group, INRIA Sophia Antipolis, 2004 route des Lucioles 06902 Sophia Antipolis, France. E-mail: Florent.Lafarge@inria.fr*
- *R. Keriven is with Acute3D, 2229 route des Crêtes - 06560 Sophia Antipolis, France. E-mail: Renaud.Keriven@acute3D.com*
- *M. Brédif is with the Matis laboratory, IGN/Université Paris Est, 73 avenue de Paris, 94165 Saint-Mandé, France. E-mail: Mathieu.Bredif@ign.fr*
- *H.-H. Vu is with the Imagine group, LIGM/Université Paris Est, 6 avenue Blaise Pascal, 77455 Marne la Vallée, France. E-mail: vhh@imagine.enpc.fr .*

example, modeling a planar wall of a facade by several thousand of triangular faces is both less accurate and less compact than by using a 3D-plane. In addition, no semantic consideration is taken into account in these approaches: the 3D models cannot be consolidated by the understanding of the scenes.

3D-primitive arrangements are also frequently used for reconstructing urban scenes from multi-view stereo images. Piecewise planar models are particularly well adapted to describe urban environments. Planes are usually detected from the data, and then arranged in 3D using photo-consistency but also model complexity considerations, *e.g.* Baillard *et al.* [20] who reconstruct buildings from aerial images or Chauve *et al.* [21] in a more general multi-view stereo context. In the works proposed by Sinha *et al.* [22] or Xiao *et al.* [23], the arrangement of the planes is not directly realized in 3D but is performed in the images or estimated depth maps. This simplifies the arrangement procedure but makes the 3D-models sparse and view dependent. Strong urban assumptions can also be done to simplify the plane arrangement procedure and to consolidate the 3D model. For example, Furukawa *et al.* [24] use the *Manhattan-world* assumption which states the predominance of three mutually orthogonal directions in the scenes [25]. Block assembling based models usually provide a more general description of scenes than piecewise planar models: the different components in the scenes are identified via a library of parametric urban objects, and assembled together according to strong urban assumptions related to man-made rules. Dick *et al.* [26] assemble facade components using a Monte Carlo sampler from terrestrial images whereas Lafarge *et al.* [27] combine various types of 3D roof blocks from aerial data. Grammar-based models also constitute elegant methods which are usually well-designed to insert urban knowledge driving the reconstruction. It is particularly adapted to facade modeling [28], [29], [30] and also to building reconstruction [31]. Note also that many works, *e.g.* [32], [33], [34], have been proposed to introduce semantic information in 3D building representations by detecting and inserting various urban objects such as windows, doors or roof superstructures.
The primitive arrangement based methods produce efficient regularized 3D-models driven by semantic considerations, and have an appealing storage and rendering capacities. However, they remain simplistic representations and fail to model fine details and irregular shapes. In particular, evaluating these models on multi-view stereo benchmarks [18], [19] usually produces average results which cannot compete against the best mesh-based models.

To a lesser extent, other types of approaches have also been proposed. However they offer less possibilities and are usually restricted to specific contexts. In particular, Level set methods [35], [36], [37] have interesting properties to deform surfaces. They are efficient for organ modeling, face tracking and, more generally, variational shape representation but are not well adapted to describe urban objects and scenes containing piecewise planar structures.

1.2 Motivations

The two main families of approaches mentioned above have complementary advantages: semantic knowledge and compaction for primitive arrangement based models, reconstruction of details and non-restricted use for mesh-based models. A natural but still lightly explored idea is the combination of 3D-primitives for representing regular elements and meshes for describing irregular structures (see Fig. 1). Such **hybrid models** are of interest especially with the new perspectives offered to navigation aids by general public softwares such as *Street View* (Google) or *GeoSynth* (Microsoft) where the 3D representation systems must be both compact and detailed.

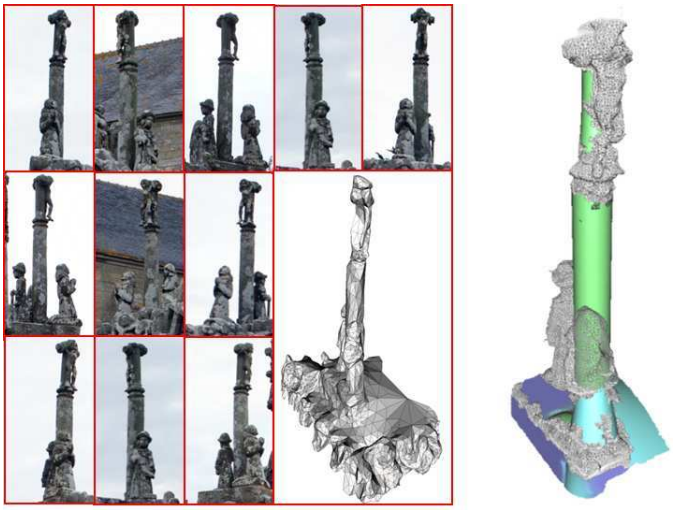

Fig. 1. Our hybrid reconstruction (right) with the inputs composed of multi-view images and an initial rough mesh (left). Irregular elements such as statues are described by mesh-based surfaces whereas regular structures such as columns or walls are modeled by 3D-primitives.

This idea has been partially addressed by several works in specific contexts. In former works [38], we proposed to detect and inject geometrical objects in dense meshes of urban scenes according to curvature attributes. However, the method solely relies on the input mesh and no photo-consistency information is used for detecting primitives and controlling their quality. This drawback limits both the accuracy and the consolidation of the models. Gallup *et al.* [39] reconstruct facades from video frames by hybrid depth maps composed of planar and non-planar components. The method is formulated as a segmentation problem from the images which takes into account

photo-consistency and texture considerations. Labatut *et al.* [40] generate a cloud of 3D points from multi-view stereo images in order to extract 3D-primitives by Ransac. Primitives are then arranged and completed by Delaunay triangulations. The resulted models are consistent but fail at describing fine details due to outliers contained in the point clouds.

1.3 Contributions

In this paper, we develop the concept of model hybridness for reconstructing complex urban scenes from multi-view images. Our method presents several contributions to the field which are explained below.

Mesh and 3D-primitive joint coordination - As mentioned above, hybrid modeling has been explored lightly in multi-view stereo reconstruction either by generating meshes where primitives are then inserted [38] or by detecting primitives and then meshing the unfitted parts of the scene using Delaunay triangulation [40]. These approaches, based on successive heuristics, cannot make two different types of 3D representation tools (*i.e.* 3D-primitives and meshes) evolve and interact in a common framework. A rigorous mathematical formulation is proposed to address this problem through an original multi-object energy model which is based on stochastic sampling techniques especially adapted for exploring complex configuration spaces efficiently.

Shape layout prior in urban scenes - The lack of information contained in the images is compensated by the introduction of urban knowledge in the stochastic model we propose. These priors allow the consolidation of the 3D-models by favoring regularized primitive layouts according to shape parallelism/perpendicularity and repetitiveness properties. In particular, some structures which are partially occluded in the images can be fully reconstructed by 3D-primitives.

Efficient global optimization - The sampling of both 3D-primitives and meshes is performed by a Jump-Diffusion based algorithm [41] which combines probabilistic and variational mechanisms. Such an algorithm is particularly interesting in our case because it allows the escape from local minima thanks to the stochastic relaxation, while guaranteeing fast local explorations with gradient descent based dynamics.

1.4 Overview

Starting from multi-view stereo images and a rough initial surface, we aim to obtain a hybrid model with a very good accuracy to compaction ratio. A two-step strategy is adopted, consisting first in segmenting the initial mesh-based surface and second in sampling primitive and mesh components simultaneously on the obtained partition. A preliminary segmentation is important because it allows us to significantly reduce the complexity of the problem. These two stages are embedded into a general iterative procedure which provides, at each iteration, an increasingly more refined hybrid model: the extracted 3D-primitives are collected along iterations whereas the mesh patches are subdivided and used as the initialization of the next iteration (See Fig. 2). The procedure stops when the mesh subdivision generates facets which are too small to be accurately matched with the images.

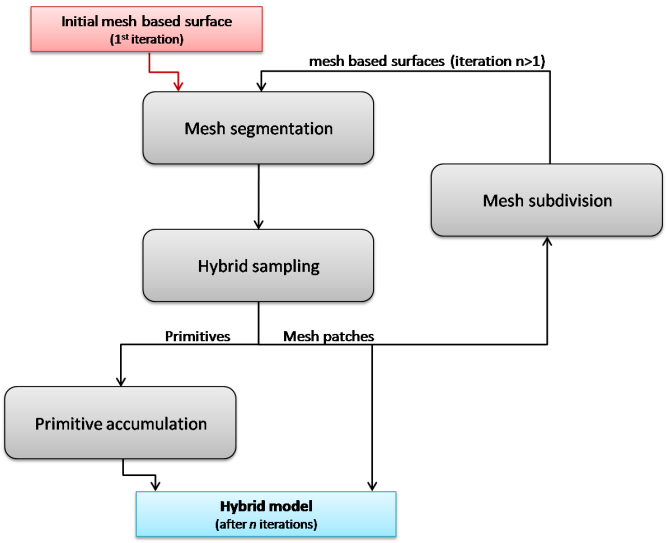

Fig. 2. Overview of the proposed approach.

This paper extends the work presented in [42] by detailing the hybrid reconstruction model and its implementation, by presenting new results and comparisons as well as commenting the mesh segmentation impact on the general procedure. The paper is organized as follows. The segmentation method is briefly presented in Section 2. Section 3 details the multi-object energy model for sampling simultaneously mesh patches and 3D-primitives. The general iterative refinement procedure is proposed in Section 4. Finally, experimental results are presented and commented in Section 5.

2 Mesh-based surface segmentation

The sparse initial mesh-based surface is first segmented by using an algorithm proposed in former works [38]. A Markov Random Field (MRF) whose sites are specified by the vertex set of the mesh and whose neighborhood relationship is given by the edge adjacency, is used to formulate the segmentation problem. Each vertex is labeled by a local curvature attribute, *i.e. planar*, *developable convex*, *developable concave* and *non developable*. The quality of a label configuration over the surface is measured by an energy composed of both a data term based on a combination of principal curvature distributions, and some propagation constraints taking into account

the label homogeneity and the edge preservation. These two components are balanced by a parameter β. The α-expansion algorithm [43] is used to quickly reach an approximate solution close to the global optimum.
The segmentation process has several interesting properties. First, it allows the extraction of the scene components with few errors even when the surface is strongly corrupted by both noise and degenerated facets. Many approaches have been proposed in the literature for segmenting synthetic mesh-based surfaces, as detailed in comparative studies [44], [45]. However, most of these methods cannot be easily extended to the surfaces generated by multi-view stereo processes and fail to correctly segment them. An efficient segmentation of such non-synthetic surfaces cannot be restricted to the use of local descriptors but must take into account propagation constraints in order to both be robust to the noise and meshing degeneracies, and introduce urban knowledge on the desired partition. Also of interest, the algorithm is adapted to different mesh densities. The energy takes into account a scale factor which is fixed proportionately to the mean edge length of the global surface. The main structures of urban scenes are detected even from sparse meshes whereas urban details can be located from dense meshes, as shown in Fig. 3. This constitutes a key point for performing our general iterative refinement procedure detailed in Section 4.

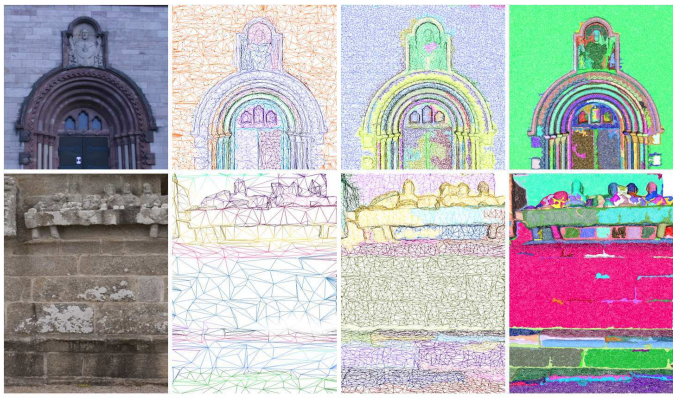

Fig. 3. Mesh segmentation. From left to right: one of the input images and results from approximated mesh-based surfaces with increasing triangle densities. Each cluster is drawn with some random color. (NB: on the right column, triangles are so small that the mesh seems plain)

In the sequel, a *cluster* is defined as a connected region of identical labels extracted by region growing. Note that the partitioning errors/approximations do not have critical consequences on the final result due to the general iterative refinement procedure. The irrelevant clusters, *e.g.* clusters composed of several regular shapes, will not be associated to 3D-primitives but will take advantage of a more accurate re-meshing to be more correctly segmented at the next iteration of the refinement procedure (see Fig. 2).

3 Stochastic Hybrid Reconstruction

In the second stage, 3D-primitives and mesh patches are simultaneously sampled from the previously segmented mesh-based surface.

3.1 Definitions

Let us first define some mathematical notations.

- $\boldsymbol{x}^{(0)}$ is the initial rough mesh-based surface segmented in N clusters by using the algorithm presented in Section 2. Each cluster, denoted by C_i with $i \in [1, N]$, is composed of a set of vertices and edges defining triangular facets, as illustrated on Fig. 4, left column. The bordering vertices of the cluster C_i are denoted by ∂C_i, and the set of edges connecting two clusters C_i and C_j by $E_{\partial C_i, \partial C_j}$.
- m_i is a mesh patch associated with the cluster C_i. m_i has the same number of vertices and the same edge adjacency than C_i, but with different vertex positions.
- p_i is a primitive associated with the cluster C_i, and defined by the couple (r_i, θ_i) where r_i is the primitive type chosen among a set of basic geometric shapes (*plane, cylinder, cone, sphere* and *torus*) and θ_i specifies its parameter set. The border of p_i corresponds to the projection of ∂C_i on the primitive surface.
- x is a hybrid model defined as a set of N_m mesh patches and N_p primitives, each of them associated with a cluster C_i such that we have

$$\boldsymbol{x} = \{x_i\}_{i\in[1,N]} = \{m_i\}_{i\in[1,N_m]} \cup \{p_i\}_{i\in[N_m+1,N]} \quad (1)$$

  with $N_m + N_p = N$. Note that a hybrid model $\boldsymbol{x}$ preserves the topology induced by the initial surface $\boldsymbol{x}^{(0)}$, as illustrated in Fig. 4. In practice, each bordering vertex of ∂C_i cannot be displaced by more than half of the length of its shortest adjacent edge in $\boldsymbol{x}^{(0)}$. This condition allows the preservation of a non-degenerated cluster adjacency, and excludes degenerated models from the possible solutions, in particular the empty set.
- $\mathcal{H}$ is the configuration space of the hybrid models given the segmented initial mesh-based surface $\boldsymbol{x}^{(0)}$, and be defined as a union of 6^N continuous subspaces $\mathcal{H}_n$, each subspace containing a predefined object type per cluster (*i.e.* 5 primitive types and 1 mesh-based structure). Note that $\mathcal{H}$ is a variable dimension space because the object types are defined by a different number of parameters.
- $U(\boldsymbol{x})$ denotes an energy measuring the quality of a hybrid model $\boldsymbol{x} \in \mathcal{H}$.

In the next two parts, we propose a formulation of $U(\boldsymbol{x})$ and a sampler allowing to find the optimal hybrid model $\widehat{\boldsymbol{x}}$ minimizing U:

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\mathcal{H}}{\arg\min}\, U(\boldsymbol{x}) \quad (2)$$

In the sequel, we call an *object*, an element x_i of a hybrid model $\boldsymbol{x}$, which is associated with the cluster C_i and can be either a primitive or a mesh patch.

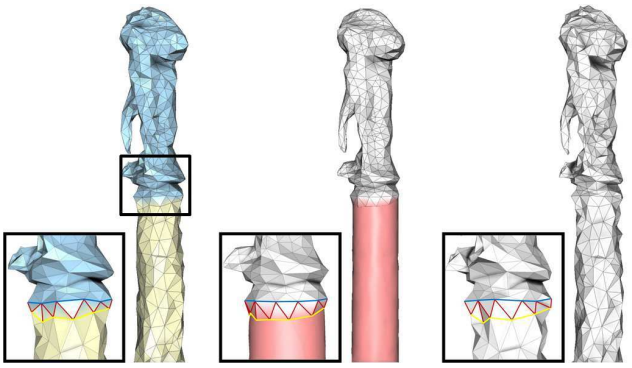

Fig. 4. Hybrid models. From left to right: An initial rough mesh $\boldsymbol{x}^{(0)}$ segmented into 2 clusters C_1 and C_2, and two hybrid models. The first hybrid model (center) is more relevant: the upper cluster is associated to a refined mesh-patch whereas the lower cluster is described by a cylindrical primitive. On the contrary, the second hybrid model (right) is associated to two mesh-patches of poor quality. Note that the cluster adjacency $E_{\partial C_1,\partial C_2}$ (red segments on the crops) is preserved with respect to the bordering edges of the clusters (yellow and blue segments) such that the hybrid models maintain the topology of $\boldsymbol{x}^{(0)}$.

3.2 Multi-object energy model

Formulating an energy U from the configuration space $\mathcal{H}$ is not a conventional problem because several kinds of objects (i.e. mesh and 3D-primitives) must be simultaneously taken into account. In addition, U must verify certain requirements. In particular, U must be differentiable in order to perform efficient gradient descent based optimization methods. The proposed energy is expressed as an association of three terms by:

$$U(\boldsymbol{x}) = \sum_{i=1}^{N} U_{pc}(x_i) + \beta_1 \sum_{i=1}^{N_m} U_s(m_i) + \beta_2 \sum_{i \bowtie i'} U_a(p_i, p_{i'}) \quad (3)$$

where U_{pc} measures the coherence of an object surface with respect to the images, U_s imposes some smoothness constraints on the mesh-based objects, U_a introduces semantic knowledge on urban scenes for positioning the 3D-primitives, $i \bowtie i'$ represents the primitive pairwise set and (β_1, β_2) are parameters balancing these three terms.

3.2.1 Photo-consistency U_{pc}

This term, based on the work of Pons *et al.* [10], computes the image back-projection error with respect to the object surface.

$$U_{pc}(x_i) = A(x_i) \sum_{\tau,\tau'} \int_{\Omega^{\mathcal{S}}_{\tau\tau'}} f(I_\tau, I^{\mathcal{S}}_{\tau\tau'})(s)\, ds \;, \quad (4)$$

where $\mathcal{S}$ is the surface of the object x_i, $f(I,J)(s)$ a positive decreasing affine function of a photo-consistency measure between images I and J at pixel s, $I^{\mathcal{S}}_{\tau\tau'}$ the re-projection of image $I_{\tau'}$ into image I_τ induced by $\mathcal{S}$, $\Omega^{\mathcal{S}}_{\tau\tau'}$ the domain of definition of the back-projection and $A(x_i)$ a function tuning the occurrence of 3D-primitive-based/mesh-based surfaces. $A(x_i) = 1$ if the object x_i is a mesh patch and $A(x_i) = \lambda$ otherwise. The λ parameter is typically set slightly inferior to 1 so that 3D-primitives are favored relatively to the meshes. The lower the value of λ, the higher the primitive to mesh ratio in the hybrid representation. The photo-consistency measure used for computing $f(I,J)(s)$ corresponds to the zero-mean normalized cross-correlation function between images I and J at pixel s.

This term has several interesting properties. Such an image back-projection with respect to the surface $\mathcal{S}$ first does not use an approximated geometry of $\mathcal{S}$. Secondly, it is easily computed with graphics hardware. And thirdly, it intrinsically takes into account visibility information as occluded reprojected surface patches do not contribute to the energy.

3.2.2 Mesh smoothness U_s

This term allows the regularization of mesh patches through the introduction of smoothness constraints. We use the thin plate energy E_{TP} proposed by Kobbelt *et al.* [46] which penalizes strong bending. In particular, this local bending energy is efficient for discouraging degenerate triangles. Our term U_s is then given by:

$$U_s(m_i) = \sum_{v \in V_{m_i}} E_{TP}(v, \{\overline{v}\}) \quad (5)$$

where $\{\overline{v}\}$ represents the set of adjacent vertices to the vertex v and V_{m_i}, the vertex set of the mesh patch m_i.

3.2.3 Priors on shape layout U_a

This term allows us to both improve the visual representation by realistic layouts of 3D-primitives, and consolidate the 3D-model, *i.e.* compensate for the lack of information contained in the images by urban assumptions. The term is inspired from the *Manhattan-world* assumption [25], and is expressed through a pairwise interaction potential which favors both perpendicular and parallel primitive layouts and object repetition in a scene. For instance, an urban environment composed of perpendicular and parallel planar structures is more probable than multiple randomly oriented structures. By considering two primitives p_i and $p_{i'}$, we introduce

$$U_a(p_i, p_{i'}) = w_{ii'}(1 - cos(2\gamma_{ii'})^{2\alpha}) \quad (6)$$

where $\gamma_{ii'}$ is the angle between the direction of revolution of the two primitives. In the case of a spherical shape, we have infinitely many axes of revolution and thus the angle is considered as null. In the case of a plane, its normal is considered as direction of revolution.

α is a coefficient fixed to 5 which allows a quasi-constant penalization of non perpendicular and non-parallel primitive layouts while keeping the existence of the derivative of U_a. $w_{ii'}$ is a weight which tends to favor the repetitiveness of similar primitive types in the scene. The higher the number of primitive types t_i and $t_{i'}$ in the scene in terms of surface, the lower the weight $w_{ii'}$.

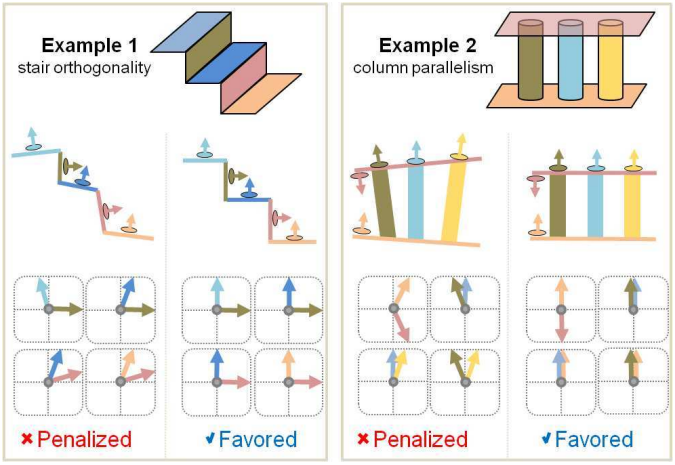

Fig. 5. Shape layout prior. The piecewise interaction U_a compares the orientation of the normals between two primitives in order to favor their mutual orthogonality (stair case on the left) and parallelism (column case on the right).

Fig. 5 illustrates the impact of this prior on the shape layout. On the stair case for example, the photo-consistency term usually does not provide a robust estimation on all the components of the stairs due to visibility problems in the images. Without this prior, some primitives are detected with a wrong orientation and a wrong type. The prior corrects this problem by favoring an ideal perpendicular/parallel structure arrangement of similar primitive types.

3.3 Jump-Diffusion based sampling

Finding the optimal hybrid model $\widehat{\boldsymbol{x}}$ requires non convex optimization techniques which are able to sample two different types of 3D-structures (*i.e.* primitives and meshes) in a high and variable dimension space $\mathcal{H}$. This complex optimization problem is addressed using a Jump-Diffusion based algorithm [41]. This type of sampler has shown potential in various applications such as image segmentation [47], [48] and target tracking [49]. In particular, it can escape local minima through the use of a stochastic relaxation while gradient descent based dynamics guarantee fast local explorations of the configuration space.

This process combines the conventional Markov Chain Monte Carlo (MCMC) algorithms [50], [51] and the Langevin equations [52]. Both dynamic types play different roles in the Jump-Diffusion process: the former performs jumps between the subspaces of different dimensions, whereas the latter realizes diffusions within each continuous subspace. The global process is controlled by a relaxation temperature T depending on time t and approaching zero as t tends to infinity. The diffusions are interrupted by jumps following a discrete time step Δt.

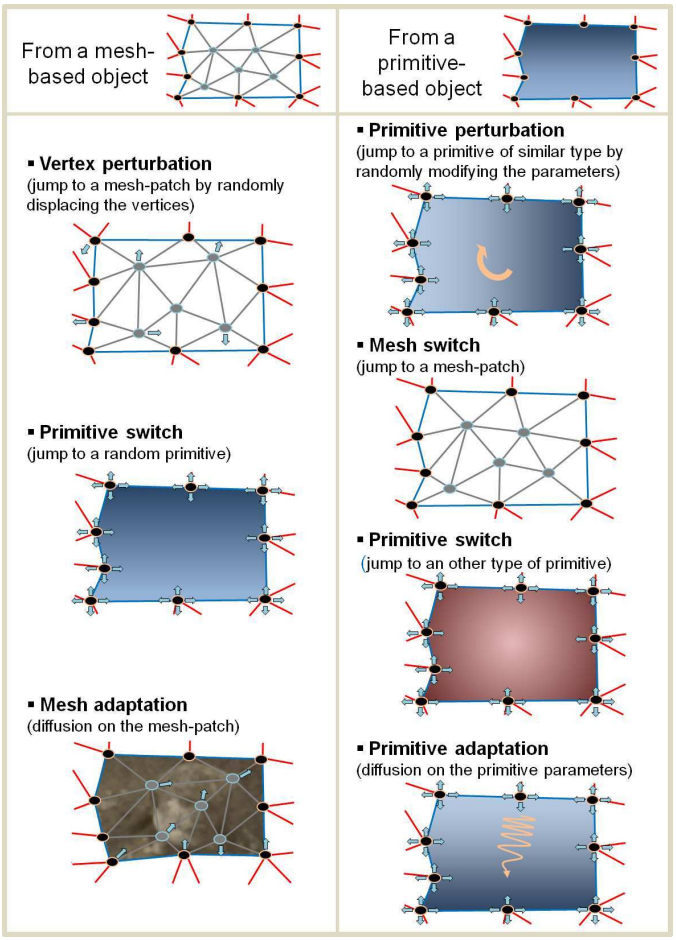

Fig. 6. List of the possible moves for mesh-based and primitive-based objects during the Jump-Diffusion.The bordering vertices (respectively edges) of the object are illustrated by black dots (resp. blue segments) whereas the interior vertices and edges (respectively primitive surfaces) are displayed in grey (respectively color gradation). Note that the bordering vertices of the object are displaced in the different moves without changing the adjacency relation with the neighboring objects illustrated by the red segments.

3.3.1 *Jump dynamics*

A jump consists in proposing a new object configuration $\boldsymbol{y}$ from the current object configuration $\boldsymbol{x}$ according to a dynamic $Q(\boldsymbol{x} \rightarrow \boldsymbol{y})$. The proposition is then accepted with probability

$$\min\left(1, \frac{Q(\boldsymbol{y} \rightarrow \boldsymbol{x})}{Q(\boldsymbol{x} \rightarrow \boldsymbol{y})} e^{-\frac{U(\boldsymbol{y})-U(\boldsymbol{x})}{T}}\right) \quad (7)$$

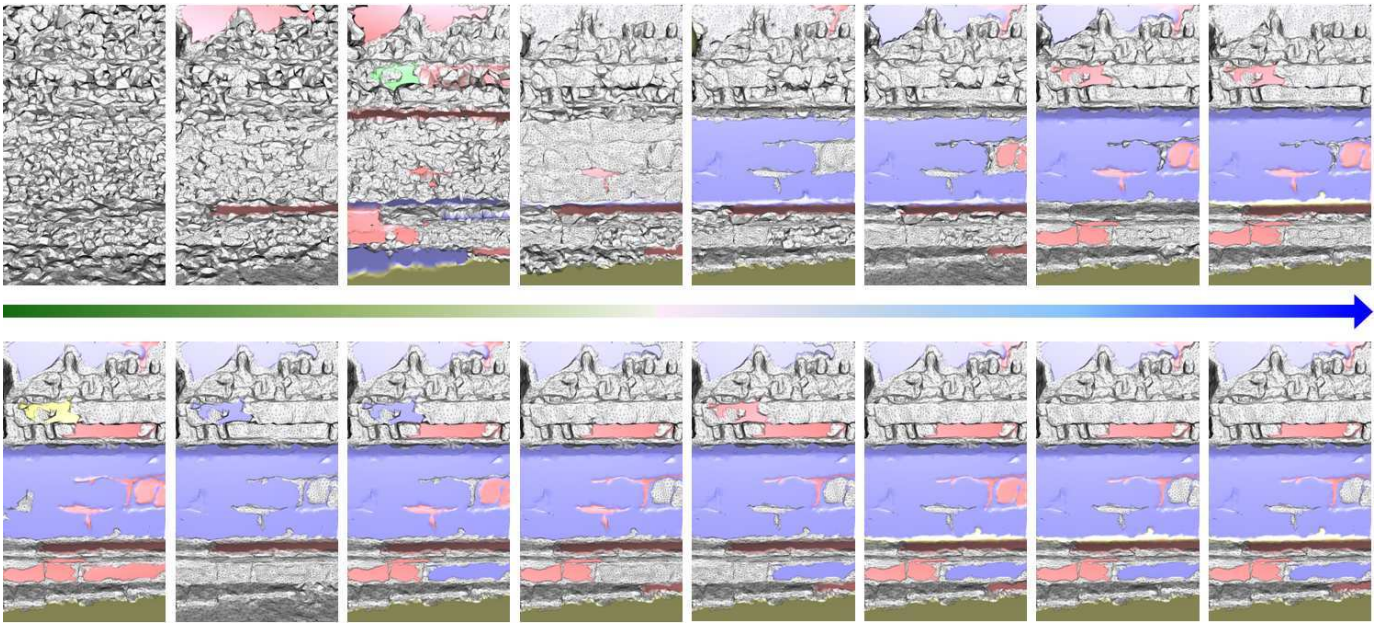

Fig. 7. Jump-Diffusion sampling. Evolution of the object configuration as the temperature decreases (from left to right, top to bottom). The following color code will be used in the sequel: *purple=plane*, *pink=cylinder*, *blue=cone*, *yellow=sphere*, *green=torus*, *gray=mesh*. Note how primitives are finally detected while details are preserved.

A jump must be reversible (*i.e.* the inverse jump $\boldsymbol{y} \rightarrow \boldsymbol{x}$ must be possible): it constitutes a necessary condition to the convergence of the algorithm. One single type of dynamic is used to perform jumps between the subspaces: switching the type of an object in the configuration x (*e.g.* a mesh patch to a cylinder or a torus to a plane). The new object type is proposed randomly. This dynamic consists in creating bijections between the parameter sets of the different object types [50]. Note that the switching kernel also includes jumps toward the same object type but with a different parameter set, as illustrated in Fig. 6 with the vertex perturbation and primitives perturbation moves.

The switching kernel is sufficient to explore the all configuration space of our problem. In particular, any configuration in $\mathcal{H}$ can be reached from any initial configuration in a finite number of iteration, which is a necessary condition to guarantee the irreductibility of the Markov chain. However, diffusion dynamics are also introduced in order to speed-up the optimization procedure during the subspace explorations.

3.3.2 Diffusion dynamics

The diffusion process controls the dynamics of the object configuration in their respective subspaces. Stochastic diffusion equations driven by Brownian motions depending on the relaxation temperature T are used to explore the subspaces $\mathcal{H}_n$. If $\boldsymbol{x}(t)$ denotes the variables at time t, then

$$d\boldsymbol{x}(t) = -\frac{dU(\boldsymbol{x})}{d\boldsymbol{x}}dt + \sqrt{2T(t)}dw_t \tag{8}$$

where $dw_t \sim N(0, dt^2)$. At high temperature ($T \gg 0$), the Brownian motion is effective in avoiding local pits. At low temperature ($T \simeq 0$), the role of the Brownian motion becomes negligible and the diffusion dynamics act as a gradient descent.

Two diffusion dynamics are proposed to explore each subspace $\mathcal{H}_n$: a mesh adaptation dynamic and a primitive parameter adaptation dynamic.

- *Mesh adaptation* - This dynamic allows the evolution of mesh-based objects using variational considerations. The energy gradient restricted to the mesh-based object m_i is given by

$$\nabla_{m_i} U = \nabla_{m_i} U_{pc} + \beta_1 \nabla_{m_i} U_s \tag{9}$$

  Brownian motion, which drives the diffusion equations, allows us to ensure the convergence towards the global minimum but makes the process extremely slow. In practice we found that the Brownian motion is not necessary to explore mesh-based object configurations because the switching dynamic, which proposes random mesh patches, efficiently enables the escape from local minima. Therefore we favor computing time with a solution close to the optimal one.
- *Primitive adaptation* - This dynamic selects relevant parameters θ_i of primitive-based objects p_i without changing their types. It is particularly efficient in accelerating the shape layout while keeping the object coherent to the images. The gradient related to this dynamic is given by

$$\nabla_{\theta_i} U = \nabla_{\theta_i} U_{pc} + \beta_2 \sum_{i'} \nabla_{\theta_i} U_a \tag{10}$$

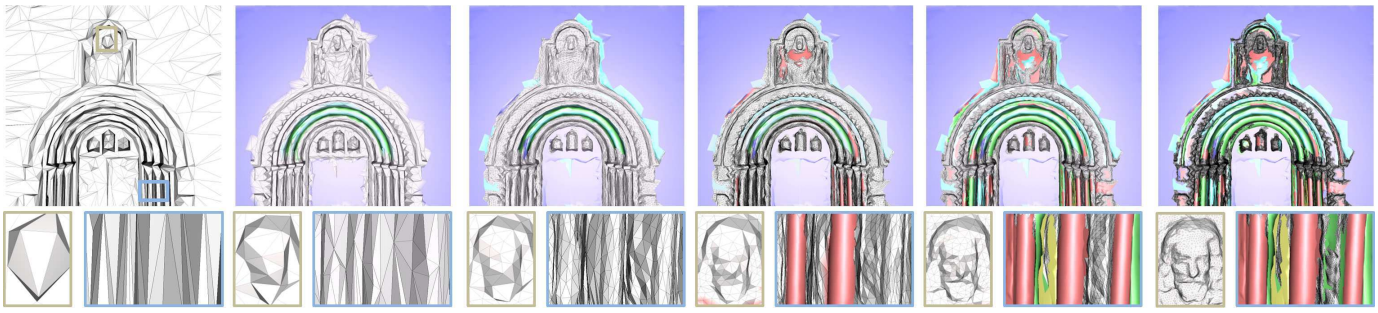

Fig. 8. Iterative refinement procedure. From left to right, top to bottom: initial sparse mesh-based surface and hybrid models at different iterations. Two details are illustrated on the bottom part. Note how more primitives are detected while details are refined during the iterations. See Fig. 7 for color code.

3.3.3 *Stochastic relaxation*

Simulated annealing theoretically ensures convergence to the global optimum from any initial configuration using a logarithmic decrease of the temperature. In practice, we use a faster geometric decrease which gives an approximate solution close to the optimum [53]. Fig. 7 shows the evolution of the object configuration during the jump-diffusion sampling. At the beginning, i.e., when the temperature is high, the process is not especially selective: the density modes are explored by perturbing mesh patches randomly and detecting the many various primitives mainly by using the jump dynamics. The two diffusion dynamics allow the fast exploration of the modes. At low temperatures, the process is stabilized in configurations close to the global optimum solution. Regular structures are extracted while being organized according to the shape prior. The parts of the wall are correctly detected as planar primitives whereas the curved ground is represented by a slightly spherical object. The irregular components are described by mesh patches which evolve towards the optimal meshing representation.

4 Iterative Refinement

The segmentation and the multi-object sampling are embedded into a general iterative refinement procedure in order to provide a more and more accurate hybrid model. At each iteration, the extracted 3D-primitives are accumulated with the primitives detected at the previous iterations whereas mesh patches are subdivided according to image resolution and used as the initialization of the next iteration.

The mesh subdivision is performed using a classical one-to-four triangle scheme, which has the advantage of preserving sharp edges. A triangular facet is subdivided while it exists one camera pair from which the visible facet projection exceeds a user-defined number of pixels in both images. In our experiments this number is fixed to 16 pixels which constitutes a good compromise between photo-consistency computation robustness and mesh refinement. The general algorithm stops when the triangular facets of the remaining mesh-based surfaces cannot be subdivided anymore.

This procedure has several advantages. First, it allows the extraction of the main regular structures of the scene at low resolutions. Thus we save time in comparison with multi-view based meshing algorithms where these regular urban elements are iteratively refined as the rest of the mesh. We then focus on irregular components represented by the remaining mesh-based surface in order to find other smaller regular structures at higher resolution levels. Secondly, the eventual irrelevant clusters generated by the segmentation algorithm are corrected at the next iterations as a result of a more accurate re-meshing. In particular, the first iteration from the initial surface, which usually does not provide relevant clusters, mainly consists in making the mesh patches evolve toward a better mesh-based representation which is correctly segmented.

This iterative refinement procedure is illustrated on Fig. 8. The main regular components of the scene such as the wall, the door and some toroidal ornaments are extracted from the first iterations whereas the ambiguous elements are refined in order to be either extracted as primitives (for example, the vertical columns on each side of the door), or re-meshed as scene details (*e.g.* the statue head).

5 Experiments

Our method has been tested on various datasets commonly used in multi-view stereo. The input surface is generated using a 3D Delaunay triangulation based approach [12] from a cloud of 3D points which is extracted from the multi-view stereo images. The obtained initial meshes are particularly rough, noisy, and spatially heterogeneous, but they are topologically close to the real surfaces.

5.1 Implementation details

The meshing operations have been implemented by using the Computational Geometry Algorithms Library [54]. Graphical processors (GPUs) have been used for the photo-consistency based computations.

The parameter β introduced in Section 2 is constant during the successive segmentations and fixed to 0.2. The parameters β_1 and β_2 weighting the various terms

of the energy in Eq. 3 are more sensitive. β_1 is fixed to 0.5, and is constant during the iterative procedure because the term U_s, which takes into account the variations of the adjacent edge length in the mesh-patches, evolves proportionally to U_{pc} when the mesh patches are subdivided. β_2 has to be updated at each iteration. Indeed, the importance of the shape layout prior U_a is progressively reduced because the number of primitives increases among the iterations whereas their size and semantic meaning decrease. In practice, β_2 is reduced by a factor 4 at each iteration. For Entry-P10 and Herz-Jesu-P25 datasets, the value of β_2 at the first iteration is 0.1. λ has been fixed to 0.9 in the experiments presented in this section. The weights $w_{ii'}$ in Eq. 6 are fixed to $1 - \frac{\mathcal{A}_{t_i} \cdot \mathcal{A}_{t_{i'}}}{4\mathcal{A}^2}$ where $\mathcal{A}_{t_i}$ (respectively $\mathcal{A}_{t_{i'}}$) is the cumulated area of the primitives of type t_i (resp. $t_{i'}$), and $\mathcal{A}$ is the cumulated area of all the primitives.

The simulated annealing is performed using a classical geometric decrease scheme of the form

$$T_t = T_o.\alpha^t \tag{11}$$

where α and T_0 are, respectively, the decrease coefficient and the initial temperature. The decrease coefficient α is fixed according to the number of clusters N in the object configurations (in practice, $\alpha = 0.9995^{\frac{1}{N}}$). T_0 is estimated by using the variation of the energy U on random configurations. More precisely, T_0 is chosen as twice the standard deviation of U at infinite temperature [55]:

$$T_0 = 2.\sigma(U_{T=\infty}) = 2.\sqrt{\langle U^2_{T=\infty} \rangle - \langle U_{T=\infty} \rangle^2} \tag{12}$$

where $\langle U \rangle$ is the means of the energy of the samples. Additional information on relaxation parameter tuning can be found in [53].

Details concerning the switching kernel computation can be found in the works of Green [50]. In our case, the completion parameters are randomly chosen. Thus the computation of the kernel is made easier because the bijection function can be taken as the identity function. The mesh-based object type is chosen with a probability $\frac{1}{3}$ whereas the primitive-based object types are chosen with a probability $\frac{2}{3}$. Note that, in the case of a jump from a primitive to a mesh-patch, the triangulation is created by projecting a regular 2D-grid on the primitive surface, and by then randomly displacing the grid vertices under a tolerance distance which guaranties a non-degenerated mesh-patch. The mesh adaptation dynamic, which consists in computing the energy gradient restricted to mesh-based objects (*i.e.* $\nabla_{m_i} U = \nabla_{m_i}(U_{pc} + \beta_1 U_s)$), is estimated by using the discrete formulation proposed by Pons *et al.* [10]. Finally, the primitive adaptation dynamics is performed by computing the gradient $\nabla_{\theta_i} U = \nabla_{\theta_i}(U_{pc} + \beta_2 \sum_{i'} U_a)$. While $\nabla_{\theta_i} U_a$ is trivial to compute, the photo-consistency term $\nabla_{\theta_i} U_{pc}$ requires advanced estimation methods for non planar primitives. $\nabla_{\theta_i} U_{pc}$ is estimated by using the first order approximation of the Euclidean distance from points to primitives proposed by Marshall *et al.* [56].

In order to improve the visual rendering of an hybrid model, note also that, in practice, the set of edges $E_{\partial C_i, \partial C_j}$ between two adjacent clusters C_i and C_j associated with primitives can be collapsed such that the exact geometric intersection of the two primitives is imposed. However, this operation is only allowed for small edge lengths in order to avoid degenerated configurations.

5.2 Large scene reconstruction

Fig. 9 and 10 present some results on different types of urban scenes including facades from terrestrial images, roofs from aerial data and a rock sculpture.

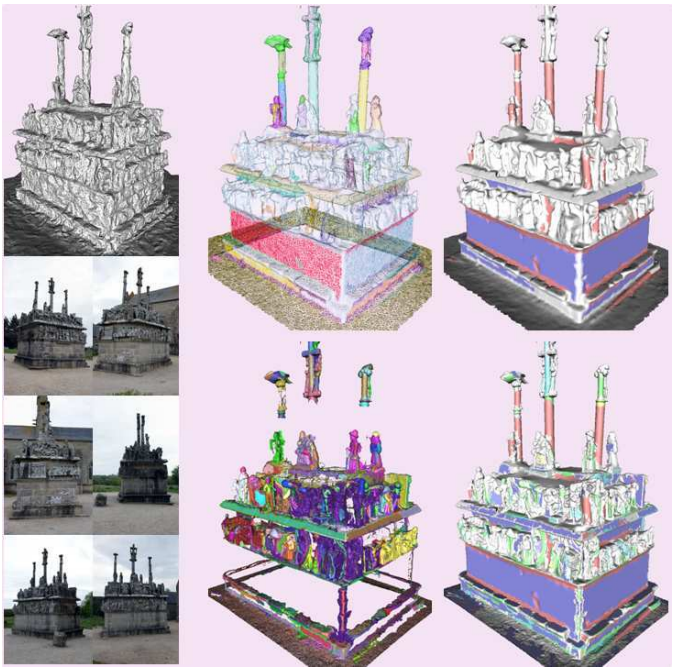

Fig. 9. Rock sculture reconstruction. Left: the initial rough surface and some input images of the Calvary dataset. Middle: mesh segmentation at low (top) and high (bottom) resolutions. Right: associated hybrid models. The low resolution model is composed of both low density mesh-based surfaces mainly representing the statues and large primitives corresponding to rock facets and columns. The refinement procedure allows the reconstruction of many small primitives on the high resolution model, especially some regular parts of statues, bricks and missing small facets whereas remaining mesh patches are very accurate.

The obtained hybrid models provide interesting representations of the scenes. The main regular components such as walls, columns, vaultings or roofs are largely reconstructed with 3D-primitives during the first iterations of the refinement procedure (*i.e.* on the models at low resolution). As shown on the church model, the shape layout prior is especially useful at low resolutions in order to obtain coherent structures in spite of the lack of information contained in the images. The church is

Fig. 10. Facade and roof reconstruction. Top row: some input images. Middle and bottom rows: hybrid models at low and high resolutions. From left to right: Entry-P10, Herz-Jesu-P25 and Church datasets. Color code: see Fig. 7.

then mainly reconstructed by plane arrangements representing roofs and facades. The primitive-based objects are completed by mesh patches which allow us to easily localize the urban details such as superstructures (chimneys, dormer-windows...) and ornaments on the church walls. The hybrid representations at high resolution are more detailed and accurate. The irregular elements are correctly modeled by mesh patches. Some statue or ornament elements are described by small primitives such as those, for example, on Calvary model where dresses are represented by cones, head backs by spheres and legs by tori. Note that the large primitives are useful for identifying semantics in the scene by a subsequent basic analysis. Structural components such as walls, roofs, windows and dormer windows can be located from the hybrid models using primitive attributes such as the type, the positioning, the orientation or the adjacency with neighboring primitives.

The segmentation stage plays an important role in the general algorithm by defining the partitions of the initial surface to sample (see Fig. 9). However, this stage is not crucial because the hybrid sampling automatically corrects the eventual bad clusters by re-meshing the patches which are not relevant enough to be turned into primitive-based objects. For instance, the segmentation of the bricks surrounding the Calvary is not relevant at low resolution because one single cluster is associated to severals bricks. At the next iterations of the refinement procedure, the cluster is subdivided into more relevant clusters, each representing one brick.

The area of the meshes is negligible in the case of facade and roof based environments but remains high for less regular scenes as the rock sculpture (see Table 1). Note that the number of 3D-primitives in the hybrid models at high resolution can be reduced by merging the similar neighboring primitives in post-processing.

TABLE 1
Additional information (number of primitives and mesh size) on the hybrid models presented on Fig. 9 and 10.

	initial surface	LR hybrid model	HR hybrid model
Entry-P10 (10 images)	$9K$ vert. $16K$ fac.	51 prim. $20K$ vert. $37K$ fac.	342 prim. $0.33M$ vert. $0.62M$ fac.
Calvary (27 images)	$23K$vert. $47K$ fac.	37 prim. $56K$ vert. $0.11M$ fac.	426 prim. $0.55M$ vert. $1.04M$ fac.
Herz-Jesu-P25 (25 images)	$14K$ vert. $17K$ fac.	41 prim. $42K$ vert. $77K$ fac.	263 prim. $0.38M$ vert. $0.74M$ fac.
Church (37 images)	$21K$ vert. $34K$ fac.	143prim. $82K$ vert. $0.15M$ fac.	406 prim. $0.13M$ vert. $0.22M$ fac.

Fig. 12 shows the impact of the coefficient λ from the Temple dataset [18]. A low λ value (*e.g.* 0.8) favors the primitive occurrence but tends to generalize the scene as illustrated with the four columns. Note also that the shape layout prior impacts on the regularization of the scene: the four columns are extracted by the same type of primitives (*i.e.* cylinders) and with the same orientation. On the contrary, a high λ value (*e.g.* 0.99), provides a representation mainly composed of mesh patches, with a better accuracy (0.48 mm *vs* 1.03 mm), but with less semantic meaning.

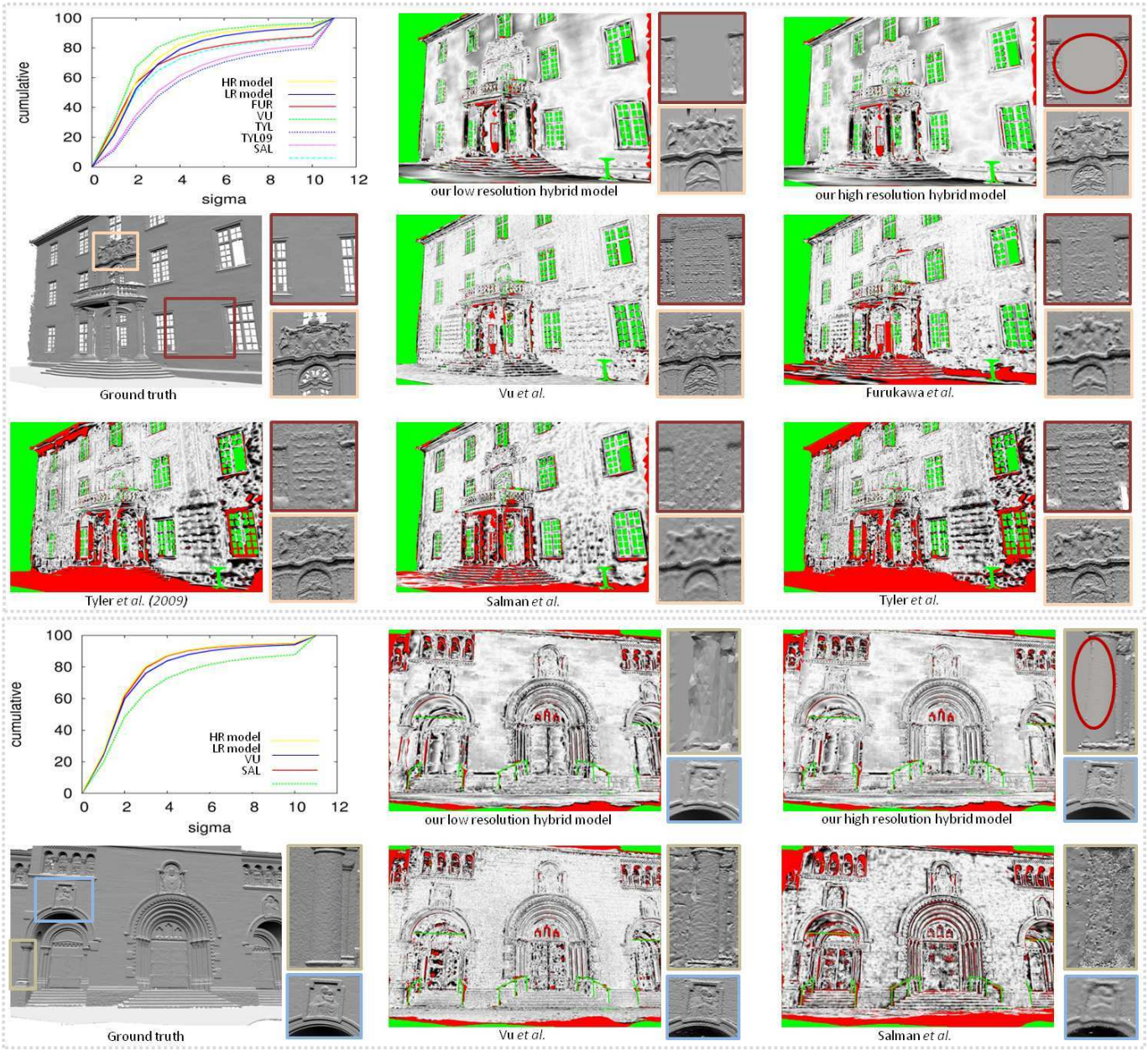

Fig. 11. Accuracy evaluation on Entry-P10 (top) and Herz-Jesu-P25 (bottom). The cumulative error histograms are measured with respect to the standard deviation Σ of the ground truth accuracy [19]. The error maps of the different models are established with respect to the ground truth (white=low, black=high, red=off-the-scale). Our high resolution hybrid models obtain the first and second best accuracies for Herz-Jesu-P25 and Entry-P10 respectively.

5.3 Accuracy and compaction

The method has been compared to the standard mesh-based algorithms which are evaluated in the multi-view stereo benchmark proposed by Strecha *et al.* [19]. The cumulative error histograms presented in Fig. 11 show that we obtain the first and second best accuracies for Herz-Jesu-P25 and Entry-P10 respectively.

Our method has several interesting advantages illustrated in the cropped elements of Fig. 11. First, the regular structures which are partially occluded in the images are more accurately reconstructed by using 3D-primitives than by standard mesh-based multi-view stereo algorithms, as shown on the column crops. Second the *trompe l'oeil* structures (see for example the textures representing fake ornaments on the walls of Entry-P10) are correctly reconstructed contrary to the mesh-only models which are based on a local analysis of the scene. Another advantage is the compaction of our hybrid representation. The level of compaction depends on the type of scenes: the more regular the scene structures, the more compact the hybrid model. On the *Entry P10* and *Herz-Jesu P25* datasets, our models at high resolution allow the reduction of the storage capacity by a factor superior to 5 with respect to the mesh-based models

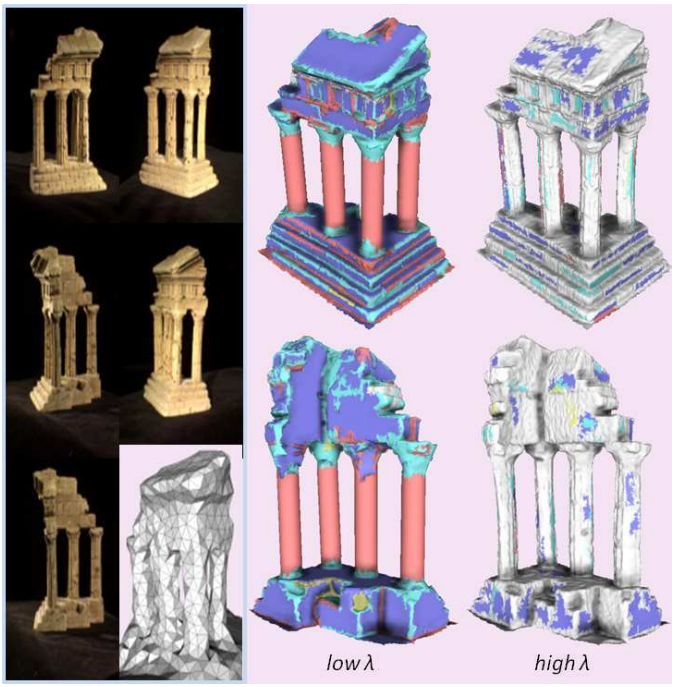

Fig. 12. Impact of the λ coefficient from the Temple dataset. Left: input images and a rough visual hull as initial surface; center: result with a low λ value; right: result with a high value. Color code: see Fig. 7. The lower the value of the λ parameter, the higher the primitives to mesh-patches ratio in the final result. The high λ version has a 0.48 mm accuracy and a 99.7% completeness on the Middlebury benchmark [18] whereas the low λ version has a 1.03 mm accuracy and a 95.7% completeness. Note that the reconstruction of columns and pieces of walls by cylinders and planes engenders a loss of accuracy in such a case, *i.e.* when the structures are not regular.

obtained by [11], for a similar accuracy as shown on Fig. 11. Note that the compaction of our hybrid models is not obtained at the expense of their visual quality as shown in Fig. 14. The presence of large primitives in our models even simplifies the texturing process in comparison with the mesh-based representations. In particular, these results offer interesting perspectives for integrating both detailed and compact models in public visualization softwares.

The approach has also been compared to a hybrid mesh simplification method proposed in former works [38]. Our models have a better accuracy, as shown in Fig. 13. In particular, the hybrid mesh simplification method [38] cannot correct the errors of the input mesh surface (see the red patch in the crops). Globally speaking, the mesh simplification models contain more primitives than our results. However a large amount of these primitives is not relevantly located as illustrated on the crop of the statue head where a spherical primitive corrupts the face. Our approach has a better preservation of the details due to the photo-consistency considerations, and also a better restitution of the urban structures as shown on the crop of the door.

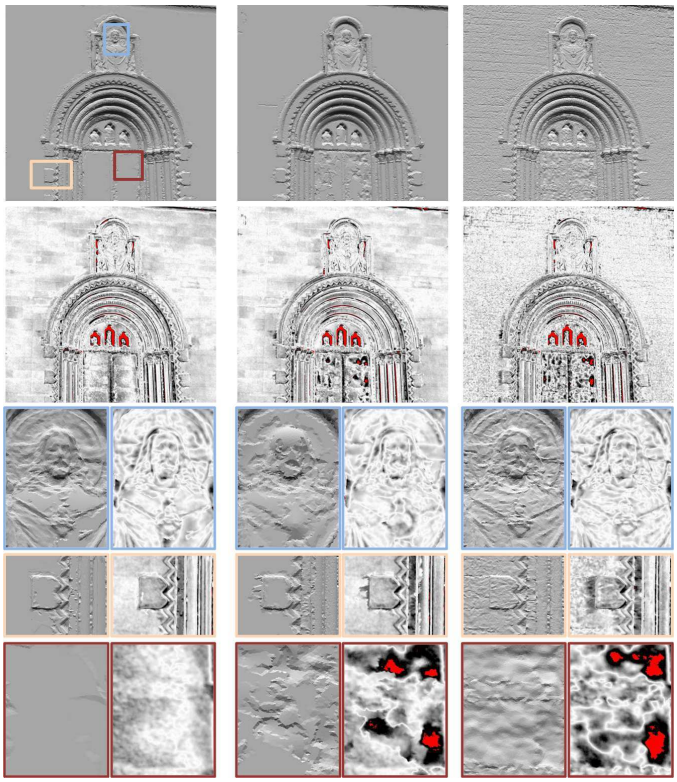

Fig. 13. Comparison with a primitive-driven mesh simplification algorithm [38].Our results (left) have a better accuracy than the mesh simplification models (center) performed from the mesh surface obtained by [11] (right). Both rendering (first row) and accuracy map (second row) are provided by the facade benchmark [19]. The various cropped parts underlines the benefits of our approach which takes into account both photo-consistency and urban shape layout considerations.

5.4 Limitations

First, some details located inside the main structures can be lost with low λ values due to the primitive accumulation process (see, for example, the small ornaments on the doors of the Herz-Jesu-P25 model on Fig. 10 incorrectly modeled by a spherical shape). One solution would be to use the full hybrid model as the initialization of the next refinement iteration instead of simply considering the mesh-based objects. However this would considerably increase the computing time.

Secondly, the shape layout prior fails to extract repetitive structures by fully identical shapes in some locations (see, for example, the set of small windows on the background tower of the Entry-P10 which are not represented by similar primitives). Even if this prior allows us to reduce the uncertainty and to give more coherence to the global primitive layout in the scene, it can be improved by taking into account additional primitive attributes such as those proposed by Pauly *et al.* [57].

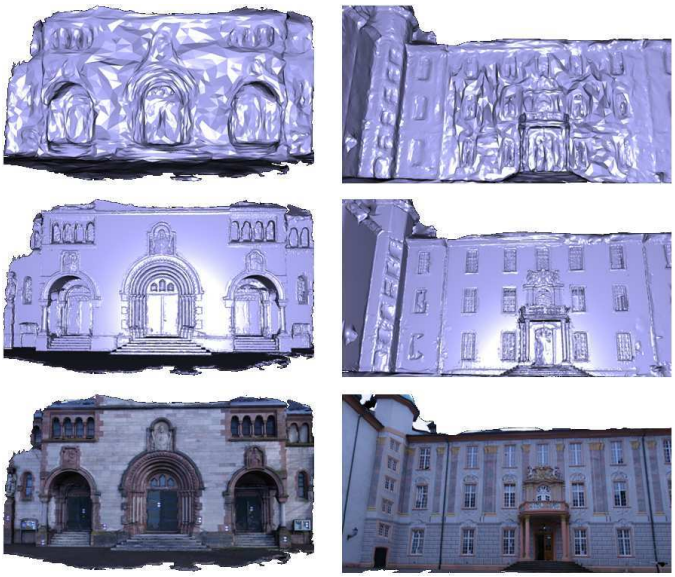

Fig. 14. Visualization of hybrid models. From top to bottom: the initial rough surface, our hybrid model without and with texture. The primitive-based objets inserted in the hybrid models are valuable to reduce the texture distortion problems.

Note also that the rough initial surface has to be topologically close to the real surface to guarantee good results. Finally, a more complex approach would be possible by embedding the segmentation stage into the sampling procedure. New kernels dealing with partition changes, *e.g.* object merging/splitting dynamics, could be introduced in order to unify segmentation and object sampling. However, such an algorithm would be performed at the expense of robustness and computation times.

6 Conclusion

We propose an original multi-view reconstruction algorithm based on the simultaneous sampling of 3D-primitives for describing regular structures of the scene and mesh patches for detailing the irregular components. Our approach offers several interesting characteristics compared to standard mesh-based multi-view algorithms. First, it affords high storage savings while having an accuracy similar to the best mesh-based algorithms [19]. Secondly, our hybrid model takes into account semantic knowledge in the scene which allows the reconstruction of partially occluded regular structures and of surfaces with *trompe l'oeil* textures. Finally, an efficient Jump-Diffusion sampler combining two different types of 3D representation tools has been developed to explore complex configuration spaces while preserving fast computing times.

In future work, it would be interesting to improve the shape layout prior by imposing more constraints on the structure repetition. It could allow the correction of errors at some locations. Also of interest is the extension of the current library of 3D-primitives in order to include more complex shapes and even to automatically adapt the library to a given scene with a learning procedure. Finally, embedding the segmentation step into the sampling procedure is probably the most motivating challenge in our future works.

Acknowledgments

The authors are grateful to the EADS foundation for partial financial support, and to the reviewers for their helpful comments. We thank C. Strecha, B. Curless and D. Scharstein for the data and the multi-view stereo challenges.

References

[1] H. Mayer, "Object extraction in photogrammetric computer vision," *JPRS*, vol. 63, no. 2, 2008.
[2] Z. Zhu and T. Kanade, "Special issue on modeling and representations of large-scale 3D scenes," *IJCV*, vol. 78, no. 2-3, 2008.
[3] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," in *CVPR*, Minneapolis, US, 2007.
[4] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz, "Multi-view stereo for community photo collections," in *ICCV*, Rio, Brazil, 2007.
[5] N. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *ECCV*, Marseille, France, 2008.
[6] K. Kolev and D. Cramers, "Integration of multiview stereo and silhouettes via convex functionals on convex domains," in *ECCV*, Marseille, France, 2008.
[7] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense 2-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-2-3, 2002.
[8] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in Computational Stereo," *PAMI*, vol. 25, no. 8, 2003.
[9] H. Hirschmuller, "Stereo processing by semi-global matching and mutual information," *PAMI*, vol. 30, no. 2, 2008.
[10] J.-P. Pons, R. Keriven, and O. Faugeras, "Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score," *IJCV*, vol. 72, no. 2, 2007.
[11] H. Vu, R. Keriven, P. Labatut, and J. Pons, "Towards high-resolution large-scale multiview stereo," in *CVPR*, Miami, US, 2009.
[12] 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 *ICCV*, Rio, Brazil, 2007.
[13] C. Fruh and A. Zakhor, "An automated method for large-scale, ground-based city model acquisition," *IJCV*, vol. 60, no. 1, 2004.
[14] 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," *IJCV*, vol. 78, no. 2-3, 2008.
[15] Pollefeys, M. et al., "Detailed real-time urban 3D reconstruction from video," *IJCV*, vol. 78, no. 2-3, 2008.
[16] N. Agarwal, S.and Snavely, I. Simon, S. Seitz, and R. Szeliski, "Building rome in a day," in *ICCV*, Kyoto, Japan, 2009.
[17] J.-M. Frahm, P. Fite Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys, "Building rome on a cloudless day," in *ECCV*, Hersonissos, Greece, 2010.
[18] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, New York, US, 2006.
[19] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *CVPR*, Anchorage, US, 2008.
[20] C. Baillard and A. Zisserman, "Automatic reconstruction of piecewise planar models from multiple views," in *CVPR*, Los Alamitos, US, 1999.
[21] A.-L. Chauve, P. Labatut, and J.-P. Pons, "Robust piecewise-planar 3D reconstruction and completion from large-scale unstructured point data," in *CVPR*, San Francisco, US, 2010.

[22] S. N. Sinha, D. Steedly, and R. Szeliski, "Piecewise planar stereo for image-based rendering," in *ICCV*, Kyoto, Japan, 2009.

[23] J. Xiao, T. Fang, P. Tan, P. Zhao, E. Ofek, and L. Quan, "Image-based facade modeling," *Trans. on Graphics*, vol. 27, no. 5, 2008.

[24] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski, "Manhattan-world stereo," in *CVPR*, Miami, US, 2009.

[25] J. M. Coughlan and A. L. Yuille, "The Manhattan world assumption: Regularities in scene statistics which enable Bayesian inference," in *NIPS*, Denver, US, 2000.

[26] A. Dick, P. Torr, and R. Cipolla, "Modelling and interpretation of architecture from several images," *IJCV*, vol. 60, no. 2, 2004.

[27] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny, "Building reconstruction from a single DEM," in *CVPR*, Anchorage, US, 2008.

[28] C. Brenner and N. Ripperda, "Extraction of facades using RJMCMC and constraint equations," in *Photogrammetric and Computer Vision*, Bonn, Germany, 2006.

[29] P. Muller, G. Zeng, P. Wonka, and L. Van Gool, "Image-based procedural modeling of facades," in *SIGGRAPH*, 2007.

[30] P. Koutsourakis, O. Teboul, L. Simon, G. Tziritas, and N. Paragios, "Single view reconstruction using shape grammars for urban environments," in *ICCV*, Kyoto, Japan, 2009.

[31] C. Vanegas, D. Aliaga, and B. Benes, "Building reconstruction using manhattan-world grammars," in *CVPR*, San Francisco, US, 2010.

[32] S. Lee and R. Nevatia, "Extraction and integration of window in a 3d building model from ground view images," in *CVPR*, Washington, US, 2004.

[33] F. Han and S. Zhu, "Bottom-up/top-down image parsing by attribute graph grammar," in *ICCV*, Beijing, China, 2005.

[34] M. Brédif, D. Boldo, M. Pierrot-Deseilligny, and H. Maître, "3d building reconstruction with parametric roof superstructures," in *ICIP*, San Antonio, US, 2007.

[35] C. Shen, J. F. O'Brien, and J. Shewchuk, "Interpolating and approximating implicit surfaces from polygon soup," *Trans. on Graphics*, vol. 23, no. 3, 2004.

[36] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 2003.

[37] R. Keriven and O. Faugeras, "Variational principles, surface evolution, PDEs, level set methods and the stereo problem," *Trans. on Image Processing*, vol. 7, no. 3, 1998.

[38] F. Lafarge, R. Keriven, and M. Brédif, "Insertion of 3d-primitives in mesh-based representations: Towards compact models preserving the details," *Trans. on Image Processing*, vol. 19, no. 7, 2010.

[39] D. Gallup, J.-M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *CVPR*, San Francisco, US, 2010.

[40] P. Labatut, J.-P. Pons, and R. Keriven, "Hierarchical shape-based surface reconstruction for dense multi-view stereo," in *3DIM*, Kyoto, Japan, 2009.

[41] U. Grenander and M. Miller, "Representations of Knowledge in Complex Systems," *J. of Royal Statistical Society*, vol. 56, no. 4, 1994.

[42] F. Lafarge, R. Keriven, M. Brédif, and H. Vu, "Hybrid multi-view reconstruction by jump-diffusion," in *CVPR*, San Francisco, US, 2010.

[43] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *PAMI*, vol. 23, no. 11, 2001.

[44] M. Attene, S. Katz, M. Mortara, G. Patane, M. Spagnuolo, and A. Tal, "Mesh segmentation - a comparative study," in *SMA*, Washington, US, 2006.

[45] A. Shamir, "A survey on mesh segmentation techniques," *Computer Graphics Forum*, vol. 27, no. 6, 2008.

[46] L. Kobbelt, S. Campagna, J. Vorsatz, and H.-P. Seidel, "Interactive multi-resolution modeling on arbitrary meshes," in *SIGGRAPH*, 1998.

[47] F. Han, Z. W. Tu, and S. Zhu, "Range image segmentation by an effective jump-diffusion method," *PAMI*, vol. 26, no. 9, 2004.

[48] Z. Tu, X. Chen, A. Yuille, and S. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *IJCV*, vol. 63, no. 2, 2005.

[49] A. Srivastava, M. Miller, and U. Grenander, "Multiple target direction of arrival tracking," *Trans. on Signal Processing*, vol. 43, no. 5, 1995.

[50] P. Green, "Reversible Jump Markov Chains Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, 1995.

[51] W. Hastings, "Monte Carlo sampling using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, 1970.

[52] S. Geman and C. Huang, "Diffusion for global optimization," *SIAM Journal on Control and Optimization*, vol. 24, no. 5, 1986.

[53] P. Salamon, P. Sibani, and R. Frost, "Facts, conjectures, and improvements for simulated annealing," *SIAM Monographs on Mathematical Modeling and Computation*, 2002.

[54] www.cgal.org.

[55] S. White, "Concepts of scale in simulated annealing," in *Proc. of IEEE International Conference on Computer Design*, 1984.

[56] D. Marshall, G. Lukacs, and R. Martin, "Robust segmentation of primitives from range data in the presence of geometric degeneracy," *PAMI*, vol. 23, no. 3, 2001.

[57] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. Guibas, "Discovering structural regularity in 3D geometry," *Trans. on Graphics*, vol. 27, no. 3, 2008.

Florent Lafarge received the MSc degree in applied mathematics from the University of Toulouse, France, and the PhD degree in applied mathematics from the Ecole des Mines ParisTech, in 2007. He has been a postdoctoral researcher at the University of Auckland, New Zealand and at the IMAGINE group of École des Ponts ParisTech. He is now a permanent research at the INRIA Sophia-Antipolis. His research interests focus on urban scene analysis and probabilistic modeling in vision.

Renaud Keriven is General Manager of the Acute3D company. Previously, he was professor of Computer Science at École des Ponts ParisTech where he headed the IMAGINE group and associate professor at École Polytechnique, France. He received the M.S. degree from Ecole Polytechnique in 1988, obtained his Ph.D. from the École des Ponts ParisTech in 1997, and his HDR ("Habilitation a Diriger les Recherches") from Paris-Est University in 2006. From 2002 to 2007, he was assistant director of the INRIA Odyssée team (leader Pr. O. Faugeras) at École Normale Superieure, Paris. His research interests include 3D photography, Multiview Stereovision, Shapes analysis, Discrete and continuous optimization in Computer Vision and Generic Programming on Graphics Processing Units."

Mathieu Brédif received the BSc degree from Ecole Polytechnique in2003, the MSc degree in 2005 from both Telecom ParisTech and Stanford University, and the Ph.D. degree from Telecom ParisTech in 2010, working at the French Mapping Agency (IGN). He is now a permanent researcher at IGN, leading the computer graphics research group. His research interests include Computational Geometry, Computational Photography and Computer Vision.

Hoang-Hiep Vu is a graduate student in IMAGINE group, at École des Ponts ParisTech, under the supervision of Renaud Keriven. He received the M.S. degree of Computer Science from École Polytechnique, France in 2008. His research interests include Multiview Stereovision, Mesh Processing, and GPGPU.