Hybrid multi-view Reconstruction by Jump-Diffusion

Florent Lafarge[1,2] Renaud Keriven[1] Mathieu Brédif[3] Hiep Vu[1]

[1] Imagine group, Université Paris Est [2] INRIA Sophia Antipolis [3] French Mapping Agency

florent.lafarge@inria.fr {keriven,vhh}@imagine.enpc.fr mathieu.bredif@ign.fr

Abstract

We propose a multi-view stereo reconstruction algorithm which recovers urban scenes as a combination of meshes and geometric primitives. It 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). A Jump-Diffusion process is designed to sample these two types of elements simultaneously. 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 sampler is 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 multi-view based meshing algorithms.

1. Introduction

Urban structures within the same scene significantly differ in terms of complexity, diversity, and density. The 3D reconstruction of such environments from multi-view stereo images and laser scans is a well known computer vision problem which has been addressed by various approaches but remains an open issue [27, 48].

1.1. Urban scene modeling

Several types of representations can be distinguished in the literature for urban scene modeling (see Fig. 1). **3D-primitive** arrangements constitute the most common type of 3D-representation in building reconstruction. The scenes are represented as layouts of simple geometric objects such as planes, lines or cubes which describe the urban objects well and are interesting in terms of storage capacity. For example, buildings are fully reconstructed by an urban component collection in [8] or by 3D-planes in [40]. More specific works on roof reconstructions from aerial/satellite data [3, 23, 33, 46, 47], facade modeling from terrestrial data [5, 21, 28, 29, 45] or building interior reconstructions [10] underline the efficiency of the 3D-primitive based approaches. They are also used to introduce semantic information in 3D building representations by detecting and inserting various urban objects such as windows, doors or roof superstructures [17, 25]. However, these parametric descriptions remain a simplistic representation and fail to model fine details and irregular shapes. **Depth maps** offer a view dependent 2.5D representation as each point of the map is associated with a single depth value. Such pixel based representations generally remain noisy even if they provide interesting details on the observed scenes as mentioned in various comparative studies [6, 35]. They are particularly well adapted to describe urban areas from aerial/satellite data [19]. **Meshes** provide highly detailed descriptions of urban structures featuring ornaments, statues and other irregular shapes. The mesh generation techniques are mainly performed using laser scanning [4, 9], video sequences [31] or multi-view stereo processes [11, 13, 39, 44]. multi-view stereo techniques have progressed significantly during recent years [1, 37, 42]. However, man made objects contain many regular structures and the meshes generated from multi-view stereo images give a large amount of redundant information concerning these elements which could be more relevantly described by parametric objects such as planar or cylindrical shapes.

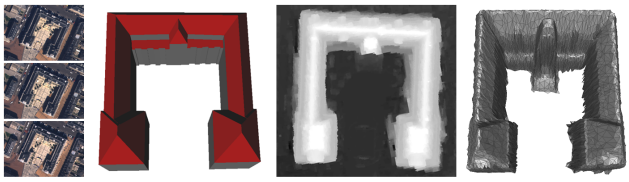

Figure 1. Various 3D representations of an urban object - from left to right: multi-view images, 3D-primitive modeling, depth map and mesh with triangular facets.

1.2. Compaction while preserving details

The representation types mentioned above have complementary advantages : semantic knowledge and model com-

paction for primitive based representations, detailed modeling and non-restricted use for mesh based surfaces. A natural idea but still lightly explored is the combination of 3D-primitives for representing regular elements and meshes for describing irregular structures (see Fig. 2). 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 have to be both compact and detailed. This idea has been partially addressed by several works in specific contexts. In [24], geometrical objects are detected and injected in dense urban meshes according to curvature attributes. However, no photo consistency information is used for detecting primitives and controlling their quality and this drawback limits the accuracy of the models. Roof simplification from urban aerial Digital Surface Models is proposed in [7] by inserting planar constraints but cannot be extended to non-planar shapes and general depth maps easily. Others works [22, 36] provide rough models from point clouds by extracting primitives using Ransac based algorithms but are limited in describing fine details due to outliers contained in point clouds.

Figure 2. Our hybrid reconstruction with associated multi-view images. Irregular elements such as statues are described by mesh based surfaces whereas regular structures such as columns or walls are modeled by 3D-primitives.

In this paper we develop this hybrid concept by reconstructing complex urban scenes from multi-view images. Our method presents significant contributions to the field which are explained below.

Mesh and 3D-primitive joint sampling - As mentioned above, hybrid modeling has been explored lightly in multi-view reconstruction either by generating meshes where primitives are then inserted [24] or by detecting primitives and then meshing the unfitted parts of the scene [22, 36]. Thus 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. We propose a rigorous mathematical formulation to address this problem through an original multi-object energy model based on stochastic sampling techniques especially adapted for exploring complex configuration spaces.

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 favor certain 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 - Most of the multi-view reconstruction algorithms use fast local optimization techniques by assuming good initializations. However they can easily get stuck into local minima and fail to accurately describe the scene. Jump-Diffusion based algorithms [15] are particularly interesting in this case because they can escape from local minima thanks to the stochastic relaxation while gradient descent based dynamics guarantee fast local explorations.

1.3. Overview

Starting from multi-view images and a sparse mesh based initial surface, we aim to obtain a hybrid model with the best accuracy to compaction ratio. We adopt a two-step strategy consisting in first, segmenting the initial mesh based surface and second, 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, a more and more refined hybrid model: the extracted 3D-primitives are accumulated along iterations whereas the mesh patches are subdivided and used as the initialization of the next iteration. An overview of the segmentation method is 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 in Section 5.

2. Mesh based surface segmentation

The sparse initial mesh based surface is partitioned using the multi-label Markov Random Field based algorithm proposed by [24]. We choose this segmentation method because it is especially adapted to non synthetic meshes generated by multi-view stereovision. Such surfaces have meshing irregularities/errors and significant noise corruption that most of the conventional synthetic mesh based methods fail to segment efficiently [2, 38]. [24] proposes a mesh labeling using local curvature properties. The interaction potential, based on label consistency and edge preservation, is de-

signed to segment surfaces corrupted by noise and containing degenerate facets. In addition, this algorithm is adapted to different mesh densities. This constitutes a key point for performing our general iterative refinement procedure. As we can see on Fig. 3, the main structures of urban scenes are detected even from sparse meshes whereas urban details can be located from dense meshes.

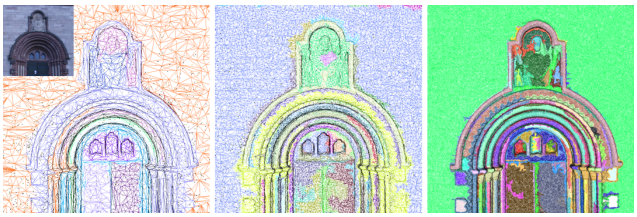

Figure 3. Mesh segmentation using [24]. From left to right: one of the input images and results from approximated mesh based surfaces with various levels of density. Each cluster is drawn with some random color. (NB: on the right column, triangles are so small that the mesh seems plain)

3. Stochastic hybrid reconstruction

3.1. Notation

Let $x^{(0)}$ be the initial rough mesh based surface segmented in N clusters by using the algorithm of [24]. Let x be a hybrid model defined as a set of N_m mesh patches and N_p primitives, each of them associated with an above-mentioned cluster such that $x = (x_i)_{i\in[1,N]} = ((m_i)_{i\in[1,N_m]}, (p_i)_{i\in[N_m+1,N]})$ and $N_m + N_p = N$. m_i represents the mesh patch associated with the cluster i. The primitive p_i is 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. We denote by $\mathcal{H}$, the configuration space of the hybrid models given the segmented initial mesh based surface $x^{(0)}$. $\mathcal{H}$ is 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). In the following, we call *object*, an element x_i of x which can be a primitive or a mesh patch. The quality of a hybrid model is measured through an energy formulation as detailed below.

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 the differentiability in order to perform efficient gradient descent based optimization methods. The energy is expressed as an association of three terms by:

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

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 weighting these three terms.

Photo-consistency U_{pc} This term, based on the work of [32], computes the image reprojection error with respect to the object surface.

$$U_{pc}(x_i) = A(x_i) \sum_{\tau,\tau'} \int_{\Omega^{\mathcal{S}}_{\tau\tau'}} f(I_\tau, I^{\mathcal{S}}_{\tau\tau'})(s)\, ds\ , \tag{2}$$

where $\mathcal{S}$ is the surface of the object x_i, $f(I, J)(s)$ a positive decreasing 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 this reprojection and $A(x_i)$ a function tuning the occurrence of 3D-primitive/mesh based surfaces. $A(x_i) = 1$ if the object x_i is a mesh patch and $A(x_i) = \lambda$ otherwise. λ is a parameter slightly inferior to 1 so that 3D-primitives are favored with respect to the meshes. The lower the value of λ, the higher the primitive to mesh ratio in the hybrid representation.

Mesh smoothness U_s This term allows the regularization of mesh patches by introducing smoothness constraints. We use the thin plate energy E_{TP} proposed in [20] which penalizes strong bending. In particular, this local bending energy is efficient for discouraging degenerate triangles. Our term U_s is then given by:

$$U_s(m_i) = \sum_{v \in V_{m_i}} E_{TP}(v, (\overline{v})) \tag{3}$$

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.

Priors on shape layout U_a This term allows to both improve the visual representation by realistic layouts of 3D-primitives and compensate for the lack of information contained in the images by inserting urban knowledge. It is expressed through a pairwise interaction potential which favors both perpendicular and parallel primitive layouts and object repetitiveness in a scene. For instance, an urban environment composed of perpendicular and parallel planar structures is more probable than multiple structures randomly oriented. By considering two primitives p_i and $p_{i'}$,

we have

$$U_a(p_i, p_{i'}) = w_{ii'}(1 - cos(2.\gamma_{ii'})^{2\alpha}) \quad (4)$$

where $\gamma_{ii'}$ is the angle between the direction of revolution of the two primitives[1]. 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 favors the repetitiveness of primitive types in the scene. It is computed according to the numbers of primitive type in the scene pondered by their areas.

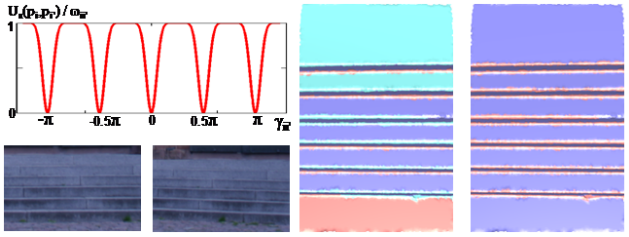

Figure 4. Shape layout prior impact on a simple example. Top left: prior energy behavior in function of $\gamma_{ii'}$; bottom left: some input images of stairs; right section: results (left) without and (right) with the prior seen from above (The following color code will be used in the sequel: *purple=plane, pink=cylinder, blue=cone, yellow=sphere, green=torus, grey=mesh*). See how stairs are correctly recovered thanks to the prior.

Figure 4 shows the impact of this prior on the shape layout. On this simple example, the photo-consistency term does not provide a robust estimation on the vertical parts of the stairs because of the poor visibility in the images. Without this prior, some parts of the stairs are detected as wrongly oriented curved primitives. The prior corrects this problem by favoring an ideal perpendicular/parallel structure arrangement of similar primitive types.

3.3. Jump-Diffusion based sampling

The search for an optimal configuration of objects is performed using a Jump-Diffusion based algorithm [15]. This type of sampler has shown an interesting potential in various applications such as image segmentation [16, 43] or target tracking [41]. This process combines the conventional Markov Chain Monte Carlo (MCMC) algorithms [14, 18] and the Langevin equations [12]. 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 diffusion 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.

[1]In the case of a spherical shape, we have an infinity of axis of revolution and thus the angle is considered as null. In the case of a plane, we take its normal as direction of revolution.

Jump dynamics A jump consists in proposing a new object configuration y from the current one x according to a dynamic $Q(x \to y)$. The proposition is then accepted with the following probability:

$$\min\left(1, \frac{Q(y \to x)}{Q(x \to y)} e^{-\frac{U(y)-U(x)}{T}}\right) \quad (5)$$

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 is proposed randomly. This dynamic, which consists in creating bijections between the parameter sets of the different object types (see [14]), is sufficient to explore the various subspaces of our problem.

Diffusion dynamics Each subspace $\mathcal{H}_n$ is explored using a mesh adaptation dynamic and a primitive competition dynamic.

Mesh adaptation - This dynamic allows the evolution of mesh based objects using variational considerations. The energy gradient restricted to mesh based objects (*i.e.* $\nabla_{m_i} U = \nabla_{m_i}(U_{pc} + \beta_1 U_s)$) is computed by using the discrete formulation proposed by [32]. Brownian motions, which drive the diffusion equations, allow us to ensure the convergence towards the global minimum but make 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 is efficient to escape from local minima. Then we favor computing time with a solution close to the optimal one.

Primitive competition - This dynamic selects relevant parameters θ_i of primitive based objects p_i without changing their types. It is particularly efficient to accelerate the shape layout while keeping the object coherent to the images. The gradient related to this dynamic is given by $\nabla_{\theta_i} U = \nabla_{\theta_i}(U_{pc} + \beta_2 \sum_{i'} U_a)$ where $\nabla_{\theta_i} U_{pc}$ is approximated using [26].

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. Information on relaxation parameter tuning can be found in [34]. Fig. 5 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 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.

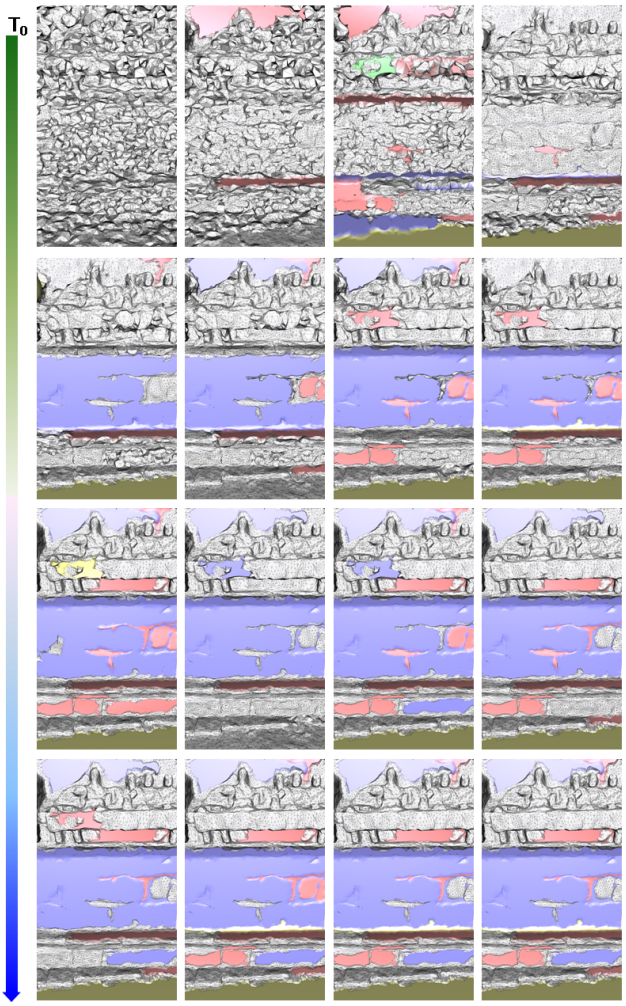

Figure 5. Jump-Diffusion sampling. Evolution of the object configuration as the temperature decreases (from left to right, top to bottom). Note how primitives are finally detected while details are preserved. See Fig. 4 for color code

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. This procedure has several advantages. Firstly, it allows the extraction of the main regular structures of the scene at low resolutions. Thus we save time compared to 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 remeshing. 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 correctly segmented. This iterative refinement procedure is illustrated on Fig. 6. 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 remeshed as scene details (*e.g.* the statue head).

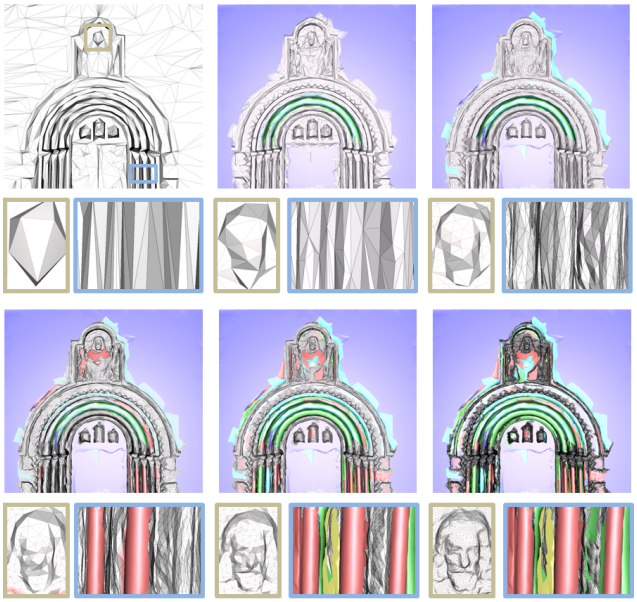

Figure 6. Iterative refinement procedure. From left to right, top to bottom: initial sparse mesh based surface and hybrid models at different iterations. Two details are zoomed. Note how more primitives are detected while details are refined. See Fig. 4 for color code

5. Experiments

Large scene reconstruction Our method has been tested on various datasets commonly used in multi-view stereo. Figure 7 presents some results on different types of urban scenes including facades from terrestrial images and a rock sculpture. The obtained hybrid representations are promis-

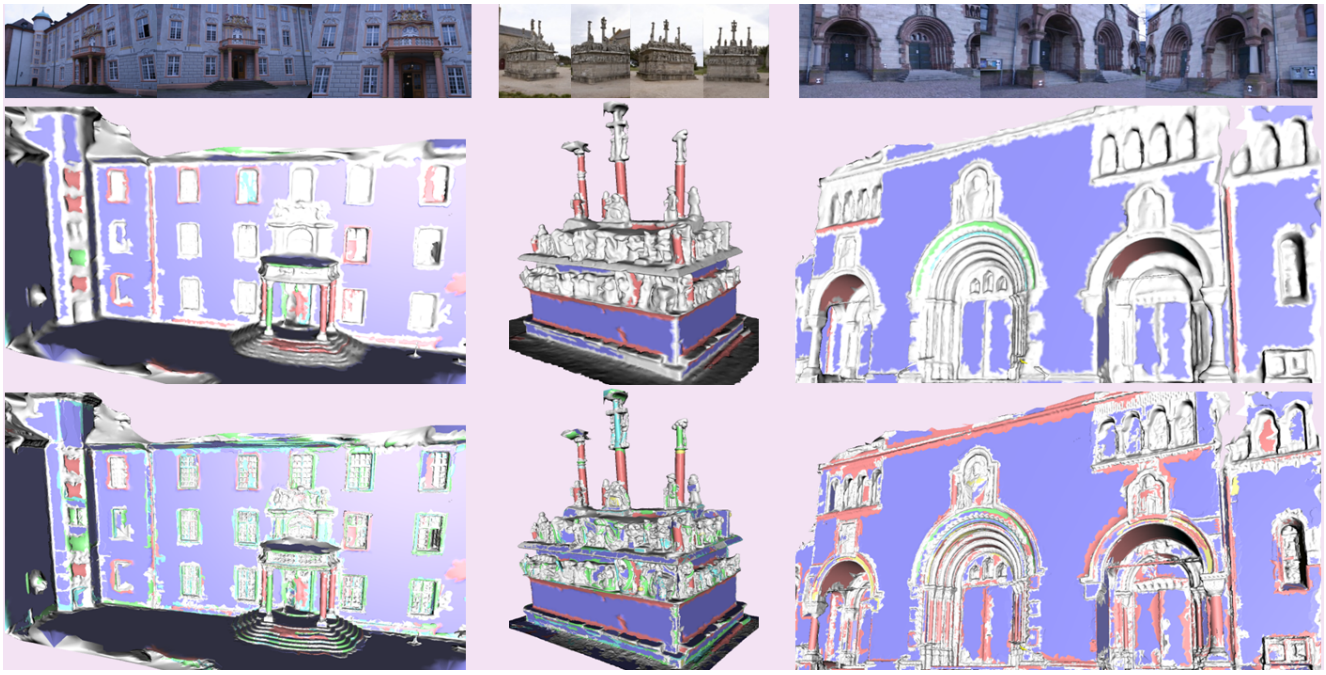

Figure 7. Large scenes. Top row: some input images. Middle and bottom rows: hybrid models at low and high resolutions. From left to right: Entry-P10, Calvary and Herz-Jesu-P25 datasets. Color code: see Fig. 4. Comments: see text.

ing and provide interesting descriptions of the scenes. The main regular components such as walls, columns, vaultings or roofs are mainly reconstructed by 3D-primitives during the first iterations of the refinement procedure (*i.e.* on the models at low resolution). The shape layout prior is especially useful at low resolution in order to obtain coherent structures in spite of the lack of information contained in the images. The hybrid representations at high resolution are more detailed and accurate. The irregular elements are correctly modeled by mesh patches. Certain parts of statues or ornaments are described by small primitives as we can see for example on Calvary where dresses are represented by cones or head backs by spheres. Such object layouts are useful for identifying semantic in the scene by a subsequent basic analysis. Structural components such as walls, roofs, windows and dormer windows can be located easily according to the object type and its parameters. As underlined in

Table 1. Additional information on the hybrid models presented on Fig. 7 (number of primitives, vertex and facets respectively).

	low resolution	high resolution
Entry-P10	$51/20K/37K$	$342/0.33M/0.62M$
Calvary	$37/56K/0.11M$	$426/0.55M/1.04M$
HJ-P25	$41/42K/77K$	$263/0.38M/0.74M$

Table 1, the mesh based area is negligible in the case of facade based environments but remains high for less regular scenes as the rock sculpture. Note that the number of 3D-primitives in the hybrid models at high resolution could be reduced by merging the similar neighboring primitives in post-processing. Fig. 8 shows the impact of the coefficient

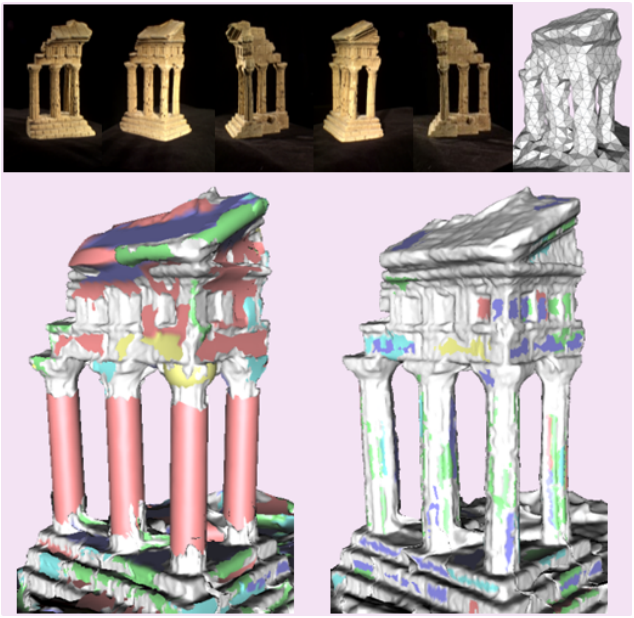

Figure 8. Impact of coefficient λ. Top row: input images and a rough visual hull as initial surface. Bottom row: left, results with a low λ value; right, with a high value. Color code: see Fig. 4. Comments: see text.

λ on the temple model [37]. A low λ value (*e.g.* 0.8) favors the primitive occurrence but tends to generalize the scene as we can see with the four columns. On the contrary, a high λ value (*e.g.* 0.99), provides a representation mainly

composed of mesh patches. On this example, we can also see the impact of the prior: the four columns are extracted by the same type of primitives (*i.e.* cylinders) and with the same orientation.

Accuracy and compaction The method has been compared to the standard mesh based stereo multi-view algorithms evaluated in [42]. The cumulative error histograms[2] presented on Fig. 9 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 on the crops of Fig. 9. First the regular structures which are partially occluded in the images are fully reconstructed by 3D-primitives as we can see on the column and vaulting crops. The standard mesh based stereo multi-view algorithms, which do not take into account semantic considerations, fail to provide a correct description of these elements. 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 local analysis of the scene. Another advantage is the compaction of our hybrid representation. Table 2 provides the storage saving rates with respect to the mesh based representations obtained by [44]. Although the accuracy of both models are similar, the hybrid representation allow us to reduce the storage capacity by a factor close to 5 and 20 for the high and low resolution versions respectively. In particular, these results offer interesting perspectives for integrating both detailed and compact models in public visualization softwares.

Table 2. Storage saving rates with respect to the mesh based representations obtained by [44] for low resolution (LR) and high resolution (HR) hybrid model.

	Entry-P10	Calvary	Herz-Jesu-P25
HR model	5.7	4.3	5.2
LR model	26.4	17.7	21

Limitations First, some details located inside main structures can be lost with a too low λ value due to the primitive accumulation process (see for example the small ornaments on the doors of the Herz-Jesu-P25 model on Fig. 7). One solution could be to use the full hybrid model as initialization of the next refinement iteration instead of just considering the mesh based objects. However it would considerably increase the computing time. Second, the shape layout prior fails to extract repetitive structures through fully identical shapes in some locations (see for example the set of small windows of the background tower of the Entry-P10 which is not represented by similar primitives). It could be improved by taking into account additional primitive attributes [30].

[2]The histograms are measured with respect to the standard deviation Σ of the ground truth accuracy (see [42]).

6. Conclusion

We propose an original multi-view reconstruction method 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 provides high storage savings while having an accuracy similar to the best mesh based algorithms [42]. Then 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 escape from local energy minima 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 repetitiveness. Also of interest would be to extend the current library of 3D-primitives in order to include more complex shapes and even to automatically adapt the library to a given scene by a learning procedure.

Acknowledgments

The authors are grateful to the EADS foundation for partial financial support. We thank C. Strecha and B. Curless for the data and the multiview stereo challenges.

References

[1] N. Agarwal, S.and Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, Kyoto, Japan, 2009. 1

[2] M. Attene, S. Katz, M. Mortara, G. Patane, M. Spagnuolo, and A. Tal. Mesh segmentation - a comparative study. In *Proc. of IEEE International Conference on Shape Modeling and Applications*, Washington, US, 2006. 2

[3] C. Baillard and A. Zisserman. Automatic reconstruction of piecewise planar models from multiple views. In *CVPR*, Los Alamitos, US, 1999. 1

[4] 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*, 78(2-3), 2008. 1

[5] C. Brenner and N. Ripperda. Extraction of facades using RJMCMC and constraint equations. In *Photogrammetric and Computer Vision*, Bonn, Germany, 2006. 1

[6] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in Computational Stereo. *PAMI*, 25(8), 2003. 1

[7] N. Chehata, M. Pierrot-Deseilligny, and G. Stamon. Hybrid DEM generation constrained by 3d-primitives : A global optimization algorithm using graph cuts. In *ICIP*, Genoa, Italy, 2005. 2

[8] A. Dick, P. Torr, and R. Cipolla. Modelling and interpretation of architecture from several images. *IJCV*, 60(2), 2004. 1

[9] C. Fruh and A. Zakhor. An automated method for large-scale, ground-based city model acquisition. *IJCV*, 60(1), 2004. 1

[10] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *ICCV*, Kyoto, Japan, 2009. 1

[11] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *IEEE CVPR*, Minneapolis, US, 2007. 1

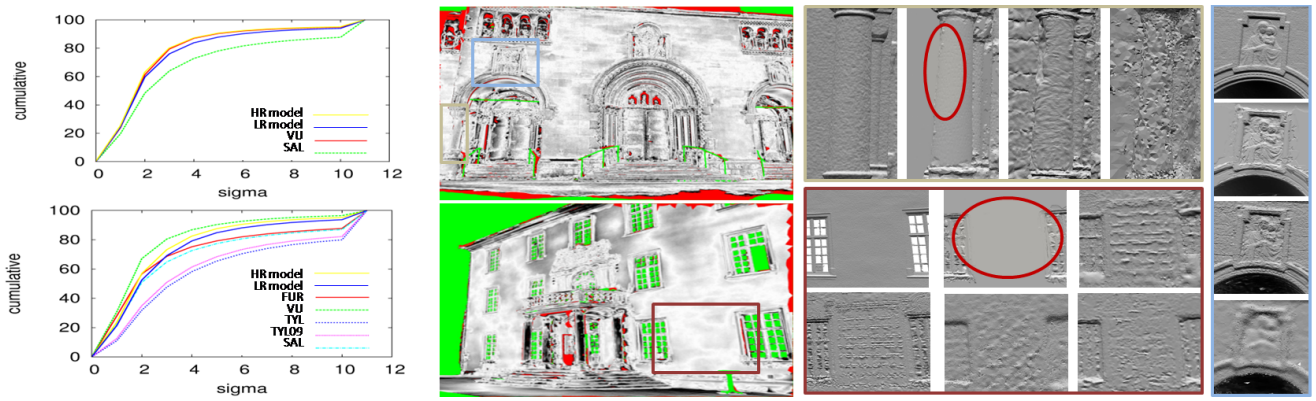

Figure 9. Accuracy evaluation. Left and middle columns: for two different datasets (top and bottom rows), cumulative error histograms (see [42]) and error of our HR model with respect to the ground truth (white=low, black=high). On the right, three sets of details presenting the ground truth, our HR hybrid model and the other mesh based models submitted. Comments: set text.

[12] S. Geman and C. Huang. Diffusion for global optimization. *SIAM Journal on Control and Optimization*, 24(5), 1986. 4

[13] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. In *ICCV*, Rio de Janeiro, Brazil, 2007. 1

[14] P. Green. Reversible Jump Markov Chains Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4), 1995. 4

[15] U. Grenander and M. Miller. Representations of Knowledge in Complex Systems. *J. of Royal Statistical Society*, 56(4), 1994. 2, 4

[16] F. Han, Z. W. Tu, and S. Zhu. Range image segmentation by an effective jump-diffusion method. *PAMI*, 26(9), 2004. 4

[17] F. Han and S. Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *ICCV*, Beijing, China, 2005. 1

[18] W. Hastings. Monte Carlo sampling using Markov chains and their applications. *Biometrika*, 57(1), 1970. 4

[19] H. Hirschmuller. Stereo processing by semi-global matching and mutual information. *PAMI*, 30(2), 2008. 1

[20] L. Kobbelt, S. Campagna, J. Vorsatz, and H.-P. Seidel. Interactive multi-resolution modeling on arbitrary meshes. In *International Conference on Computer Graphics and Interactive Techniques*, 1998. 3

[21] 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. 1

[22] P. Labatut, J.-P. Pons, and R. Keriven. Hierarchical shape-based surface reconstruction for dense multi-view stereo. In *Proc. of 3-D Digital Imaging and Modeling*, Kyoto, Japan, 2009. 2

[23] F. Lafarge, X. Descombes, J. Zerubia, and M. Pierrot-Deseilligny. Building reconstruction from a single DEM. In *CVPR*, Anchorage, US, 2008. 1

[24] F. Lafarge, R. Keriven, and M. Brédif. Combining meshes and geometric primitives for accurate and semantic modeling. In *BMVC*, London, UK, 2009. 2, 3

[25] S. Lee and R. Nevatia. Extraction and integration of window in a 3d building model from ground view images. In *CVPR*, Washington, US, 2004. 1

[26] D. Marshall, G. Lukacs, and R. Martin. Robust segmentation of primitives from range data in the presence of geometric degeneracy. *PAMI*, 23(3), 2001. 4

[27] H. Mayer. Object extraction in photogrammetric computer vision. *Journal of Photogrammetry and Remote Sensing*, 63(2), 2008. 1

[28] B. Micusik and J. Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. In *CVPR*, Miami, US, 2009. 1

[29] P. Muller, G. Zeng, P. Wonka, and L. Van Gool. Image-based procedural modeling of facades. *Trans. on Graphics*, 26(3), 2007. 1

[30] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. Guibas. Discovering structural regularity in 3D geometry. *Trans. on Graphics*, 27(3), 2008. 7

[31] M. Pollefeys et al. Detailed real-time urban 3D reconstruction from video. *IJCV*, 78(2-3), 2008. 1

[32] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *IJCV*, 72(2), 2007. 3, 4

[33] C. Poullis and S. You. Automatic reconstruction of cities from remote sensor data. In *CVPR*, Miami, US, 2009. 1

[34] P. Salamon, P. Sibani, and R. Frost. Facts, conjectures, and improvements for simulated annealing. *SIAM Monographs on Mathematical Modeling and Computation*, 2002. 4

[35] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense 2-frame stereo correspondence algorithms. *IJCV*, 47(1-2-3), 2002. 1

[36] R. Schnabel, R. Wahl, and R. Klein. Efficient ransac for point-cloud shape detection. *Computer Graphics Forum*, 26(2), 2007. 2

[37] 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. 1, 6

[38] A. Shamir. A survey on mesh segmentation techniques. *Computer Graphics Forum*, 27(6), 2008. 2

[39] S. Sinha, P. Mordohai, and M. Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *ICCV*, Rio de Janeiro, Brazil, 2007. 1

[40] S. N. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *ICCV*, Kyoto, Japan, 2009. 1

[41] A. Srivastava, M. Miller, and U. Grenander. Multiple target direction of arrival tracking. *SP*, 43(5), 1995. 4

[42] 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. 1, 7, 8

[43] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *IJCV*, 63(2), 2005. 4

[44] H. Vu, R. Keriven, P. Labatut, and J. Pons. Towards high-resolution large-scale multiview. In *CVPR*, Miami, US, 2009. 1, 7

[45] J. Xiao, T. Fang, P. Tan, P. Zhao, E. Ofek, and L. Quan. Image-based faade modeling. *Trans. on Graphics*, 27(5), 2008. 1

[46] L. Zebedin, J. Bauer, K. Karner, and H. Bischof. Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In *ECCV*, Marseille, France, 2008. 1

[47] Q. Zhou and U. Neumann. A streaming framework for seamless building reconstruction from large-scale aerial lidar data. In *CVPR*, Miami, US, 2009. 1

[48] Z. Zhu and T. Kanade. Special issue on modeling and representations of large-scale 3D scenes. *IJCV*, 78(2-3), 2008. 1