High Resolution Surface Reconstruction from Multi-view Aerial Imagery

Fatih Calakli*, Ali O. Ulusoy*, Maria I. Restrepo*, Gabriel Taubin and Joseph L. Mundy
School of Engineering, Brown University, Providence, RI (USA)
{fatih_calakli, ali_ulusoy, maria_restrepo, gabriel_taubin, joseph_mundy}@brown.edu

***Abstract*—This paper presents a novel framework for surface reconstruction from multi-view aerial imagery of large scale urban scenes, which combines probabilistic volumetric modeling with smooth signed distance surface estimation, to produce very detailed and accurate surfaces. Using a continuous probabilistic volumetric model which allows for explicit representation of ambiguities caused by moving objects, reflective surfaces, areas of constant appearance, and self-occlusions, the algorithm learns the geometry and appearance of a scene from a calibrated image sequence. An online implementation of Bayesian learning precess in GPUs significantly reduces the time required to process a large number of images. The probabilistic volumetric model of occupancy is subsequently used to estimate a smooth approximation of the signed distance function to the surface. This step, which reduces to the solution of a sparse linear system, is very efficient and scalable to large data sets. The proposed algorithm is shown to produce high quality surfaces in challenging aerial scenes where previous methods make large errors in surface localization. The general applicability of the algorithm beyond aerial imagery is confirmed against the Middlebury benchmark.**

***Keywords*-computational geometry, object modeling; stereo vision; online Bayesian learning; octrees; optimization;**

I. Introduction

Automated estimation of the geometry and appearance of a scene from multiple images is an important research problem with a wide range of applications, including realistic modeling for the feature film production, mapping and gaming industries, quantitative measures for urban planning, autonomous navigation, as well as various surveillance tasks. The problem of image-based 3-d modeling or multi-view stereo has been widely studied in the field of computer vision, computer graphics and computational photography. While many multi-vew stereo methods resolve with high accuracy the surface geometry for isolated and unoccluded objects, only a few methods are scalable to realistic, cluttered urban scenes where accurate modeling is difficult due to severe occlusions, highly reflective surfaces, varying illumination conditions, misregistration errors and sensor noise. On the other hand, most scalable 3-d reconstruction techniques have demonstrated results that are sufficient for visualization purposes, but do not offer a clear solution to the problem of accurate surface reconstruction.

This paper presents a framework targeted to solve the surface estimation problem from high resolution aerial imagery of challenging large scale urban scenes. The ambiguities caused by moving objects, reflective surfaces, areas of constant appearance, and self-occlusions, among others, are modeled explicitly using probabilistic volumetric scene modeling. The approach allows for a representation that is dense and *general*, where no assumptions are made about the underlying geometry of the scene, such as piecewise planarity. A continuous formulation is used, allowing for various adaptive discretizations of space. The space is finely discretized near the estimated surfaces, and coarsely in regions of empty space, resulting in significant reduction in terms of storage and processing time costs. Once all images have been used to learn the scene in an unconstrained manner, the proposed algorithm uses a novel variational approach to reconstruct a surface that best represents the image data summarized in the probabilistic volumetric model. An overview of the proposed approach is shown in Fig. 1.

The effectiveness of the proposed framework is validated through experiments on various large scale urban scenes. Results indicate extracted surfaces contain high resolution detail while staying smooth in areas of ambiguity. Comparisons show that prior methods can not resolve as much detail and tend to over-smooth important features. Also, the general applicability of the proposed framework is tested in the Middlebury benchmark [26]. Results are highly competitive in terms of accuracy and completeness scores.

II. Related Work

In recent years, multi-view stereo (MVS) has seen much attention and a plethora of algorithms have been introduced. Please refer to [27] for a review and taxonomy of MVS algorithms. In particular, the Middlebury benchmark [26] has fueled the development of algorithms tailored for small objects under controlled environments, where the top performing methods are capable of challenging the accuracy of laser scanners. However, most of these approaches are not suitable for dealing with high resolution aerial imagery of large scale urban scenes. In particular, algorithms that partially rely on shape-from-silhouette techniques [13], [31], [16], under the assumption that a single object is visible and can be segmented from the background, are not applicable to aerial imagery due to unavailability of fully visible silhouettes.

The Patch-based Multi-view Stereo (PMVS) algorithm proposed by Furukawa and Ponce [10] is considered state

*equal contribution

Figure 1. An overview of the proposed system.

of the art amongst methods based on feature extraction and matching. It produces colored oriented point clouds, and performs well for small objects as well as for large urban scenes, and can recover significant geometry in such scenes (see Fig. 4 top right image). However, methods based on feature extraction and matching often show significant errors caused by severe occlusions, highly reflective surfaces and areas of constant appearance. Such phenomena are commonplace in urban scenes, resulting in the generated point clouds often containing gross errors in localization and significant regions where feature points are not detected at all (see Fig. 4 middle right and bottom right images). Surface reconstruction algorithms, such as those reviewed in [15], [8], [2] are typically applied to the resulting point clouds. Although these methods produce excellent quality surfaces when applied to clean and accurate data, they often produce inaccurate surfaces when the input point cloud contains too many inaccuracies and outliers (see Fig. 5 first column).

Probabilistic models have been proposed as an alternative approach to handle complicated scene geometry. These models do not make hard decisions about surface geometry and/or appearance; instead they explicitly represent uncertainties by assigning probabilities to multiple hypothesis within the volume. Early works along this line [4], [3] can be regarded as extensions of Space Carving [17], and more recently, algorithms based on generative models for the reverse image formation process have been introduced [20], [11]. Using Bayesian inference, these algorithms infer the maximum a-posteriori probabilities in the volume from the joint probability of all the images. Pollard and Mundy [23] propose an online alternative, which in theory, can handle an infinite number of images. However, none of these volumetric methods implemented with regular grid discretizations gracefully scale up to large scenes because of the cubic space and time complexity. Crispell *et al.* [6], [7] addressed these limitations with a continuous formulation of Pollard and Mundy's method implemented in an octree. Moreover, a GPU implementation presented in [21] is capable of learning high resolution probabilistic volumetric models of large urban scenes efficiently where one pass of the online update takes approximately one second.

It is not clear how to estimate surfaces from probabilistic volumetric model. A simple approach is to compute the isosurface associated with a certain probability threshold [9], [20], [7]. Since a different threshold may be necessary in different regions of space, this method often fails to produce satisfactory results, and the lack of surface orientation information may result in "double-sided" surfaces. Yezzi *et al.* [32] propose a surface evolution approach to obtain smooth surfaces. Also related are global optimization methods such as graph cuts to extract surfaces from a volumetric photo-consistency function [12], [19], [18]. While graph cuts allows for flexible energy functions and exact solutions, the method is not applicable to large scenes due to its very high memory requirements [18].

The proposed surface reconstruction method extends the Smooth Signed Distance (SSD) approach introduced by Calakli and Taubin [5]. Because of its continuous volumetric formulation, the SSD approach is particularly complementary to the probabilistic volumetric methods described above. SSD estimates a smooth approximation to the signed distance function to output a surface from an oriented point cloud, using adaptive octree-based discretizations which reduce the problem to the solution of a sparse system of linear equations. The resulting algorithm is efficient and scalable to large data sets.

This paper uses a continuous probabilistic volumetric model (CPVM) to explicitly represent ambiguity in both surface geometry and appearance [6] and to learn large scale urban scenes from high resolution aerial imagery in an efficient manner [21]. Once the model is learned, surface reconstruction is performed by fully utilizing uncertainties in geometry in the *entire volume* using a novel extension of the SSD approach, hereby referred to as GSSD. Finally, appearance (already present) in the CPVM is transferred to the surface estimate. This transfer is not only computationally cheap but also avoids the difficulties faced by texture mapping algorithms such as photometric discontinuities, *i.e.* seams between face boundaries, ghosting effects and blurring [1].

III. Proposed Framework

A. Continuous Probabilistic Volumetric Modeling

This framework uses a scalable volumetric approach to model uncertainties in geometry and appearance, as proposed by [6]. Crispell *et al.* propose a variable-resolution model based on a continuous density that removes the dependency on regular-sized grids of previous volumetric models [20], [23]. Although no constraints are imposed about the type of subdivision, an octree is used in practice to approximate the underlying continuous quantities as piecewise constant, achieving several orders of magnitude of storage savings.

Surface probabilities are closely related to a scalar function termed the *occlusion density* $\alpha(x)$. The occlusion density at a point is a measure of the likelihood that the point occludes points behind it along any line of sight, assuming that the point itself is not occluded. Points along a ray (*e.g.* the line of sight) may be parametrized by a distance s from q (*e.g.* camera center) as $x(s) = q + sr \quad s \geq 0$. The visibility probability (1) of a point $x(s)$ is related to the integration of occlusion density from q to $q + sr$ (see [6] for derivation). Intuitively, the visibility along a ray will drop significantly when it hits a point of high occlusion, *i.e.* a surface. An example of this relationship is shown in Fig 2a.

$$vis(s) = e^{-\int_0^s \alpha(t)dt} \tag{1}$$

Learning the occlusion density from images follows an online Bayesian learning algorithm similar to that proposed by Pollard and Mundy [23]. In [23], the probability of a voxel being part of a surface, $P(X \in S)$, is updated with the intensity, I, observed in the pixel associated with a corresponding projection ray. Following Pollard's reasoning, for a discrete set of voxels, the surface update equation is expressed as:

$$P(X \in S|I_{N+1}) = P^N(X \in S)\frac{P^N(I_{N+1}|X \in S)}{P^N(I_{N+1})}, \tag{2}$$

where the conditional $P^N(I_{N+1}|X \in S)$ and marginal $P^N(I_{N+1})$ are computed using the surface probabilities and appearance models stored in the voxels along the corresponding projection ray.

The definitions proposed by Crispell *et al.* allow for a generalization of (2). In [23], information along the projection ray is approximated by a series of voxels that intersect the ray, with no regard to the ray/voxel intersection geometry. However, in Crispell's model, the occlusion density is defined continuously and therefore, the model can account for the exact geometry of ray/voxel intersection, *i.e.* the length of intersection segment. A ray is partitioned into a series of M intervals, where the ith interval is the result of the intersection of the ray and ith cell along the ray. The starting location and length of each interval are denoted by s_i and l_i respectively. Fig. 2b depicts an illustration of this ray reasoning for the octree discretization.

The *segment length occlusion probability* is defined as the probability that a segment starting at s_i of length l_i is occluding and can be expressed as:

$$P(Q_{s_i}^{l_i}) = 1 - \frac{vis(s_i + l_i)}{vis(s_i)} = 1 - e^{-\alpha_i l_i} \tag{3}$$

Equation (2) can now be written in terms of $P(Q_{s_i}^{l_i})$ *i.e.*

$$\begin{aligned} P(Q_{s_i}^{l_i}|I_{N+1}) &= P^N(Q_{s_i}^{l_i})\frac{p^N(I_{N+1}|Q_{s_i}^{l_i})}{p^N(I_{N+1})} \\ &= P^N(Q_{s_i}^{l_i})\frac{pre_i + vis(s_i)p_i(I_{N+1})}{pre_\infty + vis_\infty p_\infty(I_{N+1})} \end{aligned} \tag{4}$$

$$pre_i \equiv \sum_{j=0}^{i-1} P^N(Q_{s_i}^{l_i})vis(s_j)p_i(I_{N+1}) \tag{5}$$

$$vis_\infty \equiv \prod_{i=0}^{M-1} [1 - P^N(Q_{s_i}^{l_i})] \tag{6}$$

The term pre_i accounts for the probability of observing the given intensity I_{N+1} taking into account all segments between the camera center and segment $i-1$. The term vis_∞ measures the probability of a ray passing unoccluded through the model; in such cases the observed appearance can be thought of as the appearance of "background", which is modeled by the density p_∞. The new update equations can be used to update $\alpha(\mathbf{x})$ directly using (3).

Equation (4) has a simple interpretation, the occlusion density of a cell increases if the appearance model at the cell, explains the intensity observed in the image better than any other cell along the ray or by the background. The appearance at each cell is modeled with a Gaussian mixture distribution that is updated as in [23] using an online approach based on Stauffer and Grimson's background modeling algorithm [29].

B. Generalized Smooth Signed Distance Surface Reconstruction (GSSD)

In this section, the original formulation of SSD [5], which reconstructs surfaces from oriented point cloud data, is first summarized. Then, this formulation is generalized, targeted to reconstruct surfaces from probabilistic volumes.

Given an oriented point cloud $\mathcal{D} = \{(x_1, n_1), \ldots, (x_N, n_N)\}$, where x_i is a surface location sample, and n_i is the corresponding surface normal sample oriented towards outside of the object, SSD reconstructs a watertight surface S defined by an implicit equation $S = \{x : f(x) = 0\}$, where $f : V \rightarrow \mathbb{R}$ is a signed distance field defined on a bounded volumetric domain V contained in Euclidean three dimensional space, so that

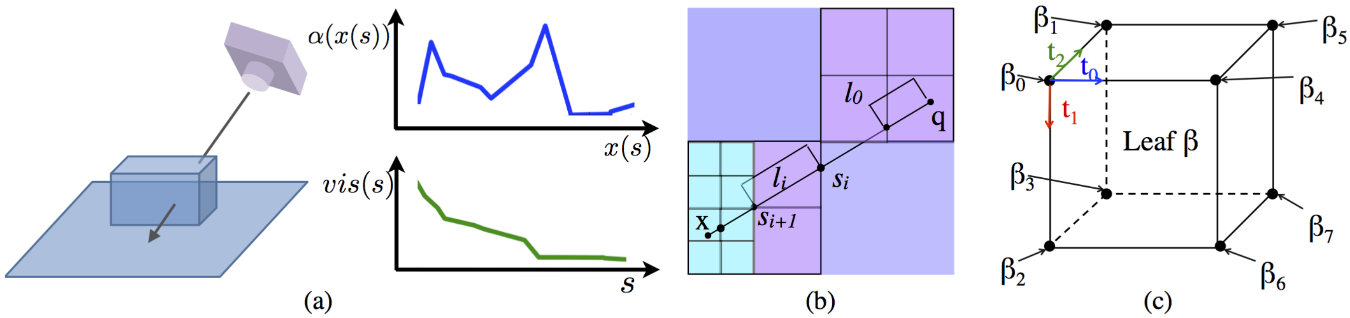

Figure 2. (a) Plots of occlusion density α and vis as a ray travels through volume. α peaks (and visibility drops) as the ray pierces the two walls of the (almost) empty cube. (b) A parametrization of the ray into intervals according to octree cell intersections. (c) Primal vertex indices associated with an octree leaf.

$f(x_i) \approx 0$, $\nabla f(x_i) \approx n_i$, for $i = 1, \ldots, N$. The resulting distance field is $f(x) < 0$ inside and $f(x) > 0$ outside of the object. Since the points x_i are regarded as samples of the surface S, and the normal vectors n_i as samples of the surface normal at the corresponding points, the distance field should ideally satisfy $f(x_i) = 0$ and $\nabla f(x_i) = n_i$ for all the points $i = 1, \ldots, N$ in the data set. These conditions are satisfied in the least squares sense by minimizing the following energy:

$$\mathcal{Q}(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i)^2 + \frac{\lambda_1}{N}\sum_{i=1}^{N} \|\nabla f(x_i) - n_i\|^2 + \frac{\lambda_2}{|V|}\int_V \|Hf(x)\|^2\, dx, \tag{7}$$

where $\{\lambda_1, \lambda_2\}$ are positive constants used to control the weights of the different terms, $Hf(x)$ is the Hessian matrix of f, and the norm of the matrix is the Frobenius matrix norm. The integral is over the volume V and $|V| = \int_V dx$ is the measure of this volume. The first two data terms of the energy function force the implicit function to approximate the signed distance function to the underlying surface. The third regularization term forces the gradient of the function to be close to constant away from the data points.

To generalize SSD, the energy function $\mathcal{Q}(f)$ is modified by replacing the finite oriented point cloud with a continuous distribution of oriented points. The proposed energy function is

$$\mathcal{E}(f) = \int_V f^2 d\mu + \lambda_1 \int_V \|\nabla f - n\|^2 d\mu + \lambda_2 \int_V \|Hf\|^2\, d\sigma, \tag{8}$$

where $d\mu(x)$ and $d\sigma(x)$ are finite measures (in the sense of measure theory), and $n(x)$ is a vector field defined in the volume. When these measures are defined by non-negative continuous densities $\mu(x)$ and $\sigma(x)$ (*i.e.* when $d\mu(x) = \mu(x)dx$ and $d\sigma(x) = \sigma(x)dx$), the values of $\sigma(x)$ and $\mu(x)$ should be chosen to make sure that the first two data terms dominate near high probability areas (*i.e.* $\mu(x)$ should be large and $\sigma(x)$ small), and the regularization term dominates near low probability areas ($\mu(x)$ small and $\sigma(x)$ large). Note that this formulation also allows for singular measures, or generalized functions as densities. In particular, the original SSD formulation (7) can be regarded as a special case of the continuous formulation (8), where

$$d\mu(x) = \sum_{i=1}^{N} \frac{1}{N}\delta(x - x_i)dx \quad \text{and} \quad d\sigma(x) = \frac{1}{|V|}dx \tag{9}$$

or more generally with a weight $\mu_i = \mu(x_i)$ assigned to each point x_i.

In this analysis, f is restricted to belong to a finite dimensional vector space of functions:

$$f(x) = \sum_{\omega\in\Omega} f_\omega\, \phi_\omega(x) = \Phi(x)^T F, \tag{10}$$

where ω denotes an index which belongs to a finite set Ω, say with K elements, $\phi_\omega(x)$ is a basis function, and f_ω is the corresponding coefficient. Then, the energy function $\mathcal{E}(f)$ results in a non-homogeneous quadratic function $F^t A\, F - 2\, b^t F + c$ in the K-dimensional parameter vector $F = (f_\omega)_{\omega\in\Omega}$. The matrix A is symmetric and positive definite, and the resulting minimization problem has a unique minimum. The global minimum is determined by solving the system of linear equations $A\, F = b$. The coefficients of matrix A and vector b requires computing inner products of every pair of basis functions. Depending on how large the support of chosen basis functions, large number of basis functions may overlap at any given point in the volume V. In effect, accumulating the coefficients of the matrix A and the vector b may require significant computation. An octree based finite-element/finite differences scheme is presented in [5] for the minimization of $\mathcal{Q}(f)$ of (7), where the problem still reduces to the solution of linear equations $A\, F = b$, but the matrix A is much sparser, resulting in a fast and space-efficient algorithm.

This discretization is particularly attractive as CPVM estimates surface density $\mu(x)$ and normal vector field $n(x)$ on an octree representation, *i.e.* $\mu(x)$ is a piecewise constant function, one value per octree leaf, and similarly $n(x)$ is a piecewise constant function, one vector per octree leaf. The parameter vector F has K elements, one value per vertex of the primal graph of the octree.

The primal graph represents the signed distance function f, and the dual graph of the octree (which has the octree leaf centroids as vertices) represents the surface density μ. $\{\beta_0, \beta_1, \ldots, \beta_7\} \subseteq \Omega$ denotes the primal vertex indices associated with dual vertex (*i.e.* an octree leaf) β, as depicted in Fig. 2c.

Note that $f(x)$ is a piecewise trilinear function, *i.e.* its value at a point x in the volume V is determined through $f(x) = \sum_{h=0}^{7} w_h \, f_{\beta_h}$ where $w_0, \ldots, w_7$ are the trilinear coordinates of x in leaf β. By taking the piecewise definitions of f and μ into account, the integrals of energy function $\mathcal{E}(f)$ of (8) conveniently become finite sums over the octree leaves:

$$\mathcal{E}(f) \approx \frac{1}{W_1} \sum_{\beta} f(x_\beta)^2 \mu_\beta + \frac{\lambda_1}{W_1} \sum_{\beta} \|\nabla f(x_\beta) - n_\beta\|^2 \mu_\beta + \frac{\lambda_2}{W_2} \sum_{(\beta,\gamma)} \|\nabla f(x_\beta) - \nabla f(x_\gamma)\|^2 \sigma_{\beta\gamma}, \tag{11}$$

where x_β is the centroid location of leaf β, μ_β is the surface density at leaf β, n_β is the oriented normal vector associated with leaf β. Note that μ_β, and n_β are estimated by CPVM, and kept fixed during surface estimation. The signed distance function f and its gradient ∇f are written using the elements of the parameter vector F :

$$f(x_\beta) = \frac{1}{8}(f_{\beta_0} + f_{\beta_1} + f_{\beta_2} + f_{\beta_3} + f_{\beta_4} + f_{\beta_5} + f_{\beta_6} + f_{\beta_7}),$$
$$\nabla f(x_\beta) = \frac{1}{4\Delta_\beta} \begin{pmatrix} f_{\beta_4} - f_{\beta_0} + f_{\beta_5} - f_{\beta_1} + f_{\beta_6} - f_{\beta_2} + f_{\beta_7} - f_{\beta_3} \\ f_{\beta_2} - f_{\beta_0} + f_{\beta_3} - f_{\beta_1} + f_{\beta_6} - f_{\beta_4} + f_{\beta_7} - f_{\beta_5} \\ f_{\beta_1} - f_{\beta_0} + f_{\beta_3} - f_{\beta_2} + f_{\beta_5} - f_{\beta_4} + f_{\beta_7} - f_{\beta_6} \end{pmatrix}, \tag{12}$$

where Δ_β the side length of leaf β. The third energy term is a sum over the pair of octree leaves, β and γ that share a common face. The associated smoothing density $\sigma_{\beta\gamma}$ is determined naturally from the octree subdivision, *i.e.* $\sigma_{\beta\gamma} = \frac{A_{\beta\gamma}}{\Delta_{\beta\gamma}}$, where $A_{\beta\gamma}$ is the area of the common face and $\Delta_{\beta\gamma}$ is the Euclidean distance between the centroids of these leaves. $W_1 = \sum_\beta \mu_\beta$ and $W_2 = \sum_{(\beta,\gamma)} \sigma_{\beta\gamma}$ are normalization factors so that μ_β's sum up to one and $\sigma_{\beta\gamma}$'s sum up to one.

Iterative Multigrid Solver: The discretization described above reduces to the solution of a sparse linear system $AF = b$, which is solved using a cascading multi-grid method, along with a Jacobi preconditioned conjugate gradient solver. The problem is first solved on a much lower depth of the octree than desired, then the solution obtained at a given depth is interpolated to the next depth; and used to initialize the iterative solver.

Polygonization: Once the signed distance function $f(x)$ is estimated, a polygonal approximation (mesh) of the isolevel zero is constructed using the Dual Marching Cubes (DMC) algorithm [25].

Mesh painting: Since CPVM produces color information represented as a volume texture at high resolution, estimating surface colors reduces to volume texture evaluation, and they are represented as polygon mesh vertex attributes. This approach avoids most of the pitfalls of texture mapping from images where occlusion and extreme warping produce unsuitable textures [1].

IV. Implementation

Scene Learning: The probabilistic geometry and appearance of all scenes is learned using the GPU implementation of CPVM as proposed in [21].

Visibility information: The visibility information plays an important role during online Bayesian learning. Surfaces are resolved with increasing accuracy as new views become available. The appearance distributions in empty cells outside of objects fail to explain the intensity of background object, causing α to converge to very low (ideal) values on empty (but visible) space. On the other hand, α converges to infinity (ideal) at surface locations. As the value of the occlusion density increases near surfaces, cells located inside of objects are updated with decreasing weight. Hence, after convergence of the model, the information inside objects could be meaningless due to ambiguous regions that were learned during early stages of the on-line training process. The erroneous information inside objects can potentially hinder the accuracy of surface extraction. These cells can be detected using the visibility information already present in the model. For each cell, a measure of its visibility, $vis\_score(x)$, is computed using a number of viewing directions [7]. Cells with low $vis\_score(x)$ are detected and the corresponding α is set to 0 to eliminate online learning artifacts. In practice, the viewing directions can be selected from the cameras used during learning or simply by defining a canonical set of directions, *e.g.* samples from a unit sphere. For scenes with poorly distributed cameras, the former method might be preferred to avoid possibly unreliable visibility computation due to poorly resolved surfaces.

Normal estimation: Surface normals are computed using the gradient information of the occlusion density $\nabla\alpha$. The gradient direction is computed via convolving first order derivatives of the three-dimensional gaussian kernel with the volume. For all experiments, six oriented kernels were used and their responses interpolated into an estimate of $\nabla\alpha$. Closed objects in the scenes contain mostly empty space, therefore when computing surface normals from the gradient direction, there is an orientation ambiguity. In order to have surface normals oriented consistently towards the outside of objects, the normal directions are oriented to the hemisphere that yields the maximum visibility.

Surface density: The proposed surface density is the following

$$\mu(x) = \alpha(x) \times vis\_score(x) \times \|\nabla\alpha(x)\|. \tag{13}$$

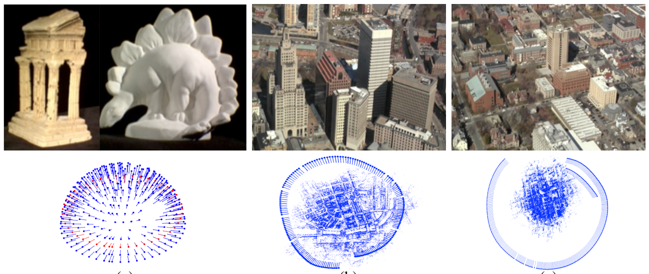

Figure 3. Sample frames and camera locations: (a) Middlebury dataset [26], “Full” set of cameras in blue, “Ring” in red; (b) “Downtown” site; (c) “SciLi” site.

The desired surface should pass through volume of high occlusion density $\alpha(x)$ and high visibility $vis\_score(x)$. Moreover, $\|\nabla\alpha(x)\|$ is included to increase robustness to outliers such as isolated cells (in the air) with high occlusion density.

V. Experiments and Evaluation Results

A series of experiments are conducted to evaluate the proposed framework in three aspects. First, it is examined how well the probabilistic learning estimates the underlying geometry. Then, the quality of the reconstructed surfaces using GSSD is assessed. Finally, comparisons to other MVS reconstruction methods are presented. PMVS [10] with Poisson Reconstruction [15] (PMVS+Poisson) is used as a baseline method because it achieves high ranking in the Middlebury benchmark [26] and is applicable to large scale scenes. The reader is referred to the supplementary material for further comparisons. The results focus on the reconstruction of 3D models from aerial imagery where both the underlying geometry and the reconstructed surfaces are shown to be superior to PMVS+Poisson. In addition, numerical accuracy of the proposed algorithm is demonstrated on the Middlebury benchmark where highly competitive results are obtained.

The dataset consists of images from two challenging urban sites and two benchmark models. Fig. 3a presents a sample image from the Temple and Dino objects that are part of the Middlebury dataset [26]. Both objects were reconstructed using both the “Full” set of cameras, (312 views for Temple, 363 views for Dino), and the “Ring” set (47 views for Temple, 48 view for Dino). Fig 3b and 3c correspond to images from two urban sites in Providence, RI, USA (publicly available in [24]). The urban sites, here referred to as “Downtown” and “SciLi”, cover an area of approximately (500m)$\times$(500m) and have an approximate resolution of 30 cm/pixel. The camera paths follow a circle around each site; the size of all images is 1280x720 pixels, 174 views are available for “Downtown” and 337 for “SciLi”. The camera matrices for all aerial image sequences were obtained using the Bundler software [28]. The resulting models were trained in approximately 2 hours for “Downtown” and 3 hours for “SciLi” and both contain roughly 30 million leaf cells.

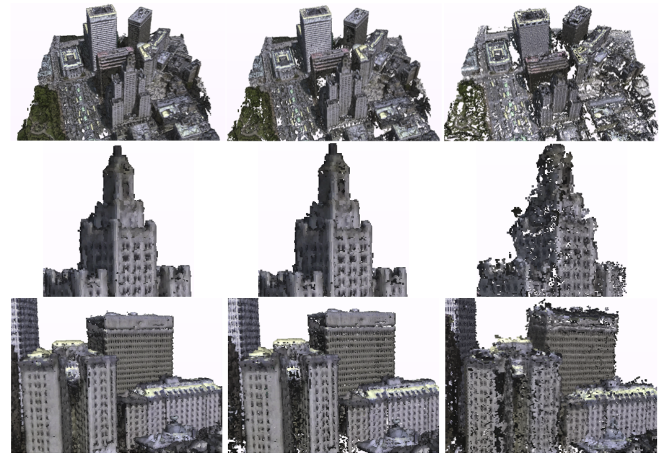

Figure 4. Point cloud visualization of “Downtown” (from left to right column wise): point cloud using the top (13) 70% of CPVM samples; point cloud using the top (13) 10% of CPVM samples; PMVS point cloud.

Urban Scenes: This section begins with an examination of the quality of the surface density learned in the CPVM. In order to make comparisons to PMVS easy to visualize and fair, locations (octree leaf cell centroids) are sampled from CPVM to generate a point cloud. The cell locations are filtered using the surface density (13) as a threshold criteria. The point clouds are visualized in Fig. 4, where the full “Downtown” scene as well as details are shown for the top 70% and 10% of the sampled cells, together with the PMVS point cloud. Comparisons in Fig. 4 demonstrate three clear advantages of CPVM over PMVS. Namely, (a) both systems do well on planar textured surfaces, however in regions of high curvatures, CPVM produces much more accurate results than PMVS, *e.g.* edges of buildings in bottom row; (b) CPVM is able to resolve details at higher resolution than PMVS, *e.g.* pillars and tip of the building in middle row; (c) information is very dense in the probabilistic model, *e.g.* the point cloud of CVPM even with 10% of the surface voxels is denser than PMVS.

Fig. 5 presents the comparisons of surfaces reconstructed for both the “SciLi” and the “Downtown” sites using PMVS+Poisson and the proposed framework. The figure demonstrates that the proposed method produces pleasing surfaces while staying faithful to the high detail information available in the data, notice the sharpness of building walls and corners, and ability to capture small details such as rooftop pipes and tiny windows, as shown in bottom row. The produced surfaces are well defined and crisp, even in regions of high surface curvature, as seen in second row. All these qualities are a direct result of the variational formulation of equation 11. The first two data terms make sure that the surface pass through the highly surface-like regions while the third term forces certain smoothness constraints. It has

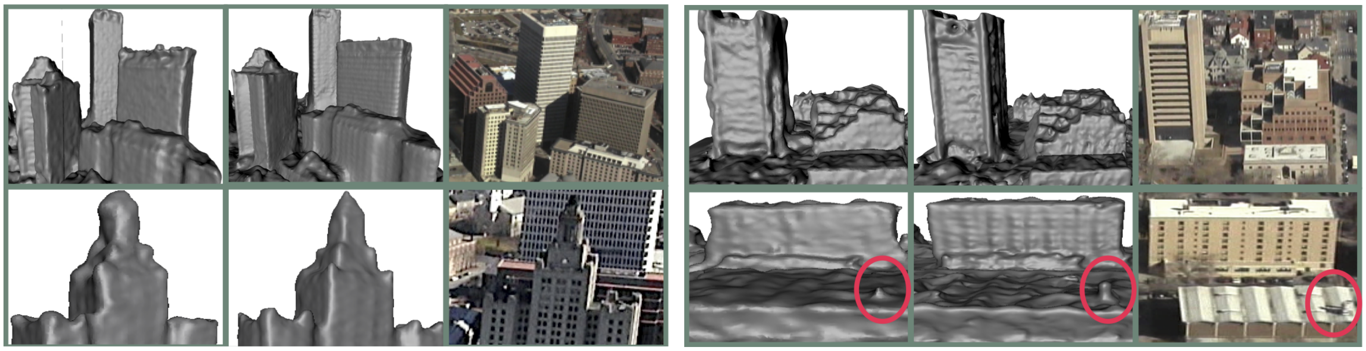

Figure 5. Reconstructed surfaces from the “Downtown” and “SciLi” sites, using PMVS+Poisson (leftmost column) and the proposed framework (mid column). Sample images for the corresponding scenes (rightmost column).

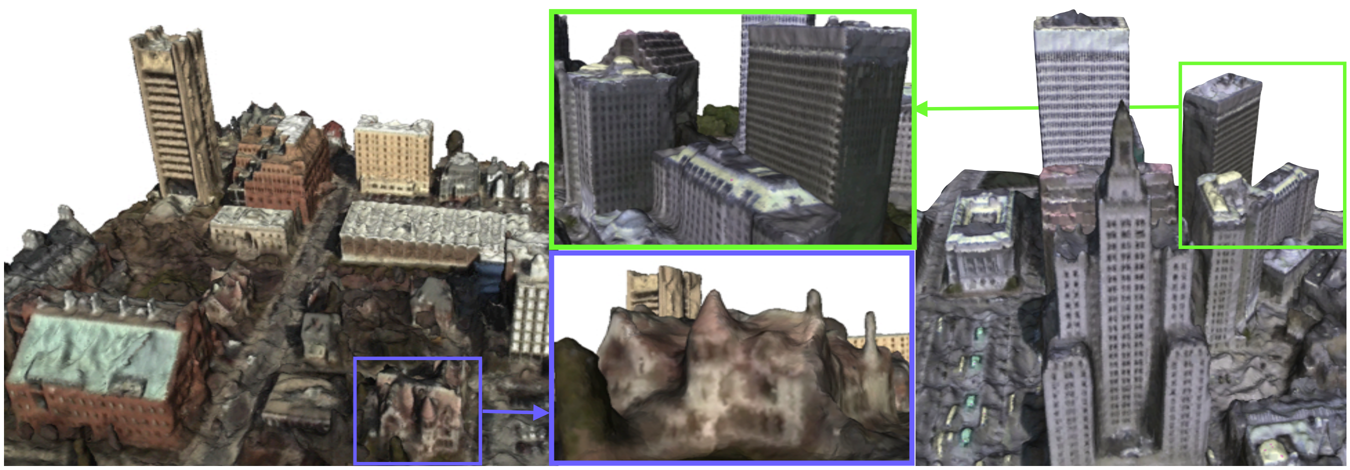

Figure 6. Renderings of 3D models (colored mesh) of the “SciLi” (left) and the “Downtown” (right) sites reconstructed by the proposed framework. Note the high resolution texture on buildings.

been observed that when PMVS and CPVM have difficulties modeling the appearance and underlying geometry of highly specular surfaces, the reconstruction algorithms can not recover accurate surfaces (see supplementary video[1] for an example).

Fig. 6 shows renderings of the colored meshes. Notice the high resolution and completeness of the appearance information such as chimney and cone of the highlighted house (in purple), as well as texture detail on buildings (in green).

All meshes were obtained using GSSD parameters $\lambda_1 = 1$ and $\lambda_2 = 4$ (see (11)). The running time for GSSD was roughly 12 minutes using an Intel Xeon @2.9 Ghz. Both urban scenes contained approximately a million cells with significantly high μ.

Middlebury Evaluation Performance: Quantitative evaluations reported[2] in [26] show that the proposed method is highly competitive both in terms of accuracy(mm) and completeness(%). Table I presents results for a selection of methods obtained from [26]. In [7], Crispell *et al.* reconstruct a surface by extracting the isosurface of the implicit function vis_score using Marching Cubes. The superiority of GSSD compared to this method is clear in both Temple and Dino (Ring) datasets. Also included are the results for PMVS+Poisson obtained in [10]. At the time of writing, this method achieved some of the best results at [26] including top scores in Full Dino. The proposed framework performs competitively, especially in the Full Temple dataset. It should be noted that most of the best performing algorithms reported at [26], including [10], take advantage of the fact that reasonable silhouettes can be extracted. This fact in not exploited in this paper, as the main focus is on applicability to realistic urban scenes.

VI. Conclusion and Future Work

This paper presented a framework targeted to solve surface reconstruction for aerial imagery of large scale urban scenes. The main contribution has been the introduction of Generalized-SSD, capable of extracting highly detailed

[1] http://vimeo.com/45316105

[2] submission under “Generalized-SSD”.

Table I
QUANTITATIVE EVALUATIONS TAKEN FROM [26]. THE FIRST VALUE CORRESPONDS TO ACCURACY (MM) AND THE SECOND TO COMPLETENESS (%). ALL MESHES WERE OBTAINED USING GSSD PARAMETERS $\lambda_1 = 1$ AND $\lambda_2 = 1$ (SEE (11))

	Temple				Dino			
	Full		Ring		Full		Ring	
PMVS+Poisson [10]	0.54	99.3	0.55	99.1	0.32	99.9	0.33	99.6
Proposed Framework	0.53	99.4	0.81	95.8	0.55	98.1	0.6	96.0
CPVM + Marching Cubes[7]	-	-	1.89	92.1	-	-	2.61	91.4

surfaces from probabilistic volumes. The advantages of the proposed framework were shown via comparisons to state of the art methods, namely PMVS [10] and Poisson surface reconstruction [15], chosen due to their popularity and software availability. Comparisons to other related work such as [14], [30], [22] will be performed in the near future. In addition to experiments in aerial scenes, the framework was also tested in the Middlebury evaluation [26] where it performed competitively in terms of accuracy and completeness. Most remarkably, it achieved results on par with algorithms that exploit high resolution silhouettes.

An important future direction is to consider more sophisticated appearance models that can explain specular reflection, which would result in more robust surface estimates. Currently, a difficulty inherited from the CPVM is dealing with voxels inside of surfaces which contain inaccurate information from early stages of the update process. Such voxels are removed as a post-processing step but ideally, they will be handled during the online Bayesian learning. Finally, application of GSSD to other probabilistic volumetric models, such as those found in medical imaging, will be investigated.

The software implementation of GSSD used to create figures shown in this paper is available for download from http://mesh.brown.edu/gssd/.

ACKNOWLEDGEMENT

This material describes work supported by the National Science Foundation under Grants No. IIS-0808718, and CCF-0915661.

REFERENCES

[1] A. Baumberg. Blending images for texturing 3d models. In *BMVC*, 2002. 3, 5
[2] M. Berger, J. Levine, L. Nonato, G. Taubin, and C. Silva. An End-to-End Framework for Evaluating Surface Reconstruction. Sci technical report, University of Utah, 2011. 2
[3] R. Bhotika, D. Fleet, and K. Kutulakos. A probabilistic theory of occupancy and emptiness. In *ECCV*, 2002. 2
[4] A. Broadhurst, T. W. Drummond, and R. Cipolla. A Probabilistic Framework for Space Carving. In *ICCV*, 2001. 2
[5] F. Calakli and G. Taubin. SSD: Smooth Signed Distance Surface Reconstruction. *Computer Graphics Forum*, 30(7), 2011. http://mesh.brown.edu/ssd/. 2, 3, 4
[6] D. Crispell, J. Mundy, and G. Taubin. A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection. In *IEEE Transactions on Geoscience and Remote Sensing*, 2012. 2, 3
[7] D. E. Crispell. *A Continuous Probabilistic Scene Model for Aerial Imagery*. PhD thesis, School of Engineering, Brown University, 2010. 2, 5, 7, 8
[8] T. Dey. *Curve and Surface Reconstruction: Algorithms with Mathematical Analysis*. Cambridge Univ Pr, 2007. 2
[9] J. S. Franco and E. Boyer. Fusion of multiview silhouette cues using a space occupancy grid. In *ICCV*, 2005. 2
[10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *CVPR*, 2007. 1, 6, 7, 8
[11] P. Gargallo, P. Sturm, and S. Pujades. An Occupancy-Depth Generative Model of Multi-View Images. In *ACCV*, 2007. 2
[12] C. Hernández, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007. 2
[13] C. Hernandez Esteban and F. Schmitt. Silhouette and stereo fusion for 3D object modeling. In *CVIU*, pages 367–392, 2004. 1
[14] V. Hiep, R. Keriven, P. Labatut, and J.-P. Pons. Towards high-resolution large-scale multi-view stereo. In *CVPR*, 2009. 8
[15] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Eurographics SGP*, 2006. 2, 6, 8
[16] K. Kolev and D. Cremers. Integration of Multiview Stereo and Silhouettes Via Convex Functionals on Convex Domains. In *ECCV*, Oct. 2008. 1
[17] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. In *ICCV*, 1999. 2
[18] V. Lempitsky and Y. Boykov. Global Optimization for Shape Fitting. In *CVPR*, June 2007. 2
[19] V. Lempitsky, Y. Boykov, and D. Ivanov. Oriented Visibility for Multiview Reconstruction. LNCS. Springer, 2006. 2
[20] S. Liu and D. B. Cooper. A complete statistical inverse ray tracing approach to multi-view stereo. In *CVPR*, 2011. 2, 3
[21] A. Miller, V. Jain, and J. L. Mundy. Real-time rendering and dynamic updating of 3-d volumetric data. In *Workshop on GPGPU*, 2011. 2, 5
[22] P. Mucke, R. Klowsky, and M. Goesele. Surface reconstruction from multi-resolution sample points. In *VMV*, 2011. 8
[23] T. Pollard and J. L. Mundy. Change Detection in a 3-d World. In *CVPR*, 2007. 2, 3
[24] M. I. Restrepo, B. A. Mayer, and J. L. Mundy. Object recognition in probabilistic 3d volumetric scenes. In *ICPRAM*, 2012. 6
[25] S. Schaefer and J. Warren. Dual marching cubes: Primal contouring of dual grids. *Computer Graphics Forum*, 24, 2005. 5
[26] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. Multi-view stereo evaluation. http://vision.middlebury.edu/mview/eval/, 2011. 1, 6, 7, 8
[27] S. M. Seitz, J. Diebel, D. Scharstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *CVPR*, 2006. 1
[28] N. Snavely and S. M. Seitz. Photo Tourism: Exploring Photo Collections in 3D. *ACM Transactions on Graphics*, 2006. 6
[29] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1998. 3
[30] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 27, 2011. 8
[31] G. Vogiatzis, C. Hernández, P. H. S. Torr, and R. Cipolla. Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-consistency. In *PAMI*, Dec. 2007. 1
[32] A. Yezzi, G. Slabaugh, R. Cipolla, and R. Schafer. A surface evolution approach of probabilistic space carving. In *3DPVT*, 2002. 2