Towards Probabilistic Volumetric Reconstruction using Ray Potentials

Ali Osman Ulusoy Andreas Geiger Michael J. Black
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{osman.ulusoy,andreas.geiger,black}@tuebingen.mpg.de

Abstract

This paper presents a novel probabilistic foundation for volumetric 3D reconstruction. We formulate the problem as inference in a Markov random field, which accurately captures the dependencies between the occupancy and appearance of each voxel, given all input images. Our main contribution is an approximate highly parallelized discrete-continuous inference algorithm to compute the marginal *distributions of each voxel's occupancy and appearance. In contrast to the MAP solution, marginals encode the underlying uncertainty and ambiguity in the reconstruction. Moreover, the proposed algorithm allows for a Bayes optimal prediction with respect to a natural reconstruction loss. We compare our method to two state-of-the-art volumetric reconstruction algorithms on three challenging aerial datasets with LIDAR ground truth. Our experiments demonstrate that the proposed algorithm compares favorably in terms of reconstruction accuracy and the ability to expose reconstruction uncertainty.*

1. Introduction

Over the last decades, multi-view stereo algorithms have steadily improved in terms of accuracy and completeness [29]. More recently, researchers have shifted their focus from reconstructing isolated single objects to more challenging general scenes that involve significant occlusions, textureless or reflective surfaces, transient objects, varying illumination conditions, and camera mis-registration errors [7, 16, 17, 24, 32]. Such factors cause fundamental ambiguities for 3D reconstruction from images [2]. Consider the grass region in Fig. 1a. The surface contains little texture and therefore, multiple reconstructions satisfy the input images equally well. Fig. 1b depicts the area of ambiguity for two views, where many surfaces residing inside the green quadrangle are valid solutions. Similar ambiguities arise due to occlusions, reflective surfaces, etc., making it critical to explicitly model, or expose, this uncertainty.

Most previous work on multi-view stereo does not address such reconstruction ambiguities. While some methods assign confidence scores for each reconstructed 3D point [11, 15] or model the ambiguity in view-centric depth maps [9, 25, 31], we focus on *volumetric* reconstruction in this paper. A volumetric representation allows for encoding ambiguities *everywhere* in the scene and yields dense reconstructions. Probabilistic and volumetric reconstruction methods associate an occupancy variable with each voxel and infer the probability of occupancy from the input images [3, 1, 4, 2, 23, 35]. However, the resulting inference procedure either requires strong visibility approximations [4] or slow stochastic search [2]. Other, more efficient algorithms lack a global objective function, making it unclear what is being optimized and what the probabilistic interpretation should be [3, 1, 35, 23].

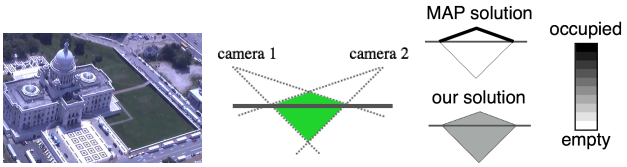


Figure 1: (a) The grass region contains very little texture. (b) Side view of the grass region. The solid line is the true ground plane. The limited viewpoints and lack of texture cause reconstruction ambiguity: all voxels inside the green quadrangle are equally photo-consistent, leading to multiple valid reconstructions. (c) The maximum-a-posteriori (MAP) solution is the closest surface to the camera since these surface voxels are not occluded by any photo-consistent voxel. Our solution assigns uniform and lower probability of occupancy throughout the region, thereby encoding the reconstruction uncertainty.

In this paper, we propose a novel and principled probabilistic approach to volumetric reconstruction. We formulate the problem as inference in a Markov random field (MRF) that models the joint distribution over discrete occupancy and continuous appearance (color) variables at each voxel, given all input images. High-order ray potentials capture dependencies between occupancy and appearance

variables along each input camera ray, accurately modeling visibility constraints. While previous works on ray potentials aim at the maximum-a-posteriori (MAP) solution of the occupancy variables [20, 28], our goal is to infer the *marginal* distributions of occupancy *and* appearance at each voxel. In contrast to the MAP solution, the marginals directly expose the uncertainty in the reconstruction.

Unfortunately, computing per-voxel marginal distributions in the proposed MRF is very challenging. The model comprises both discrete and continuous variables. Further, each input image contains millions of pixels, leading to a huge number of ray potentials, each connecting to hundreds of variables, resulting in highly loopy graphs. We tackle these challenges by deriving an approximate, highly parallelized, inference algorithm based on sum-product belief propagation. As a by-product, our algorithm yields a Bayes optimal prediction for a natural 3D reconstruction metric.

We evaluate our algorithm on three challenging aerial datasets with LIDAR ground truth. Experiments demonstrate that the algorithm is able to produce accurate 3D models while exposing the uncertainty inherent in the reconstruction. Our method also compares favorably to two state-of-the-art volumetric reconstruction methods [20, 23] both in terms of reconstruction accuracy, as well as in terms of its ability to encode reconstruction uncertainty.

2. Related Work

This section reviews the most relevant work on probabilistic approaches to volumetric reconstruction. Please refer to [29, 8] for a more complete overview of multi-view stereo methods.

In their early work, Bonet and Viola proposed an algorithm that iterates between estimating the occupancies and the colors of each voxel [3]. To accommodate video streams, their ideas have been extended to the online setting [1], where voxel occupancy and color are updated one image at a time. Similar online algorithms with a Bayesian interpretation have been proposed in [23, 35]. More recently, Pollard and Mundy's framework [23] has been adapted to use efficient octree representations [6] and implemented on a GPU [21], leading to some of the most accurate and efficient volumetric 3D reconstruction pipelines to date [5, 33]. However, Pollard and Mundy's method, as well as [3, 1, 35], lack a global formulation that relates voxel occupancy and color to all the input images. Therefore, it is not clear how the resulting probabilities should be interpreted. Moreover, Pollard and Mundy's update algorithm is typically iterated many times over the same input images [26], effectively treating each image as a new independent observation at each iteration. Our experiments show that this approach leads to a self-reinforcing behavior and results in overly confident occupancy probabilities.

A more principled approach is to directly integrate all image observations using ray potentials into a MRF model, which can be optimized using message passing [10, 20] or graph cuts [28]. Our approach follows this line of work but differs in two main aspects: First, we estimate voxel occupancy and appearance jointly via a global inference algorithm. In contrast, [10, 20] decouple voxel appearance estimation from the inference of occupancies and [28] does not model voxel appearance but instead relies on pre-estimated depth maps. Second, we compute marginal distributions of occupancy and appearance, rather than the most likely occupancy assignment. This allows our method to expose the uncertainty in the reconstruction, which can be utilized by subsequent processing stages. Our experiments confirm that the proposed approach correctly assigns uniform and low occupancy probabilities to ambiguous regions, whereas the MAP solution favors the first, i.e., most visible, photo-consistent voxel along the ray, therefore leading to reconstructions that "bulge out" as illustrated in Fig. 1c. Interestingly, Space-Carving exhibits similar behavior in featureless regions [19].

3. Probabilistic and Volumetric 3D Model

Let us assume a decomposition of the 3D space into a grid of voxels, which we identify with a unique index from the index set $\mathbb{X}$. Let us further assume that the scene contains only solid objects and empty space. We associate each voxel $i \in \mathbb{X}$ with two random variables: a binary occupancy variable $o_i \in \{0, 1\}$ indicating whether the voxel is occupied ($o_i = 1$) or free ($o_i = 0$) and a real-valued appearance variable $a_i \in \mathbb{R}$ describing the intensity or color of the 3D surface at voxel i. Note that appearance variables are defined for every voxel since the surface locations are unknown a-priori.

In the following, we first describe the image formation process for a single viewing ray, followed by the specification of our full probabilistic model.

3.1. Image Formation Process

Let $\mathcal{R}$ denote the set of viewing rays originating from one or multiple calibrated cameras. For a single ray $r \in \mathcal{R}$, let $\mathbf{o}_r = \{o_1^r, \dots, o_{N_r}^r\}$ and $\mathbf{a}_r = \{a_1^r, \dots, a_{N_r}^r\}$ denote the ordered sets of occupancy and appearance variables associated with voxels intersecting ray r as illustrated in Fig. 2. The ordering is defined by the distance to the respective camera, i.e., we have $i < j$ if the voxel associated with (o_i^r, a_i^r) is closer to the camera from which ray r originates than the voxel associated with (o_j^r, a_j^r).

We model the image formation process by assigning the appearance of the *first occupied voxel* along ray r to the corresponding pixel. This process can be expressed as

$$I_r = \sum_{i=1}^{N_r} o_i^r \prod_{j<i} (1 - o_j^r)\, a_i^r \tag{1}$$

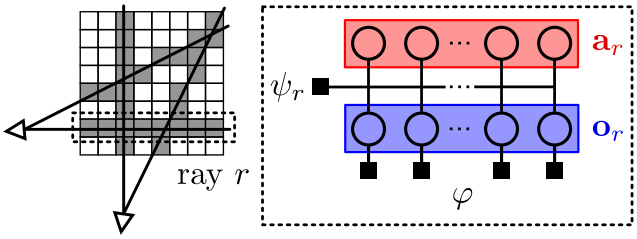


Figure 2: Probabilistic Graphical Model. Left: 2D slice through 3D voxel grid with four rays and the associated voxels marked in gray. Right: Factor graph for a single ray with ray potential ψ_r and unary potentials φ.

where I_r is the intensity/color at the pixel corresponding to ray r. The term $o_i^r \prod_{j<i}(1-o_j^r)$ evaluates to 1 for the first occupied voxel along the ray and 0 for all other voxels. Note that the term $\prod_{j<i}(1-o_j^r)$ is an indicator for visibility; it is 1 if there exist no occupied voxel before the ith voxel, and 0 otherwise. Hence, the summation amounts to the color of the first occupied voxel. In the next section, we use this image formation model to estimate the marginal probabilities of voxel occupancy and color from the pixel observations.

3.2. Probabilistic Model

We phrase the problem of volumetric 3D reconstruction as inference in a Markov random field. Let $\mathbf{o} = \{o_i | i \in \mathbb{X}\}$ and $\mathbf{a} = \{a_i | i \in \mathbb{X}\}$ denote the sets of all occupancy and appearance variables. We specify the joint distribution over $\mathbf{o}$ and $\mathbf{a}$ in terms of its factorization into unary and ray potentials

$$p(\mathbf{o}, \mathbf{a}) = \frac{1}{Z} \prod_{i \in \mathbb{X}} \underbrace{\varphi_i(o_i)}_{\text{unary}} \prod_{r \in \mathcal{R}} \underbrace{\psi_r(\mathbf{o}_r, \mathbf{a}_r)}_{\text{ray}} \tag{2}$$

where Z denotes the partition function, $\mathbb{X}$ is the set of all voxels and $\mathcal{R}$ denotes the set of viewing rays from all cameras observing the scene. Fig. 2 (right) illustrates the corresponding graphical model for a single ray.

The **unary potentials** encode our prior belief that, in many natural scenes, most voxels are empty. Thus, we model $\varphi_i(o_i)$ using a simple Bernoulli distribution

$$\varphi_i(o_i) = \gamma^{o_i} (1-\gamma)^{1-o_i} \tag{3}$$

where γ is the prior probability that voxel i is occupied.

The **ray potentials** model the image generation process as specified by Eq. 1:

$$\psi_r(\mathbf{o}_r, \mathbf{a}_r) = \sum_{i=1}^{N_r} o_i^r \prod_{j<i} (1-o_j^r)\, \nu_r(a_i^r). \tag{4}$$

Here, $\nu_r(a)$ denotes the probability of observing intensity/color a at ray r. Assuming Gaussian noise, we model $\nu_r(a) = \mathcal{N}(a | I_r, \sigma)$.

4. Inference

In this section, we present our inference algorithm for approximately computing the marginal distribution of occupancy and appearance of each voxel. We also propose a natural objective function to evaluate our 3D model against ground truth and show that the Bayes optimal solution for this loss can be computed as a by-product of the proposed inference algorithm.

4.1. Approximate Marginal Inference

For inference, we exploit the well-known sum-product algorithm extended to mixed discrete-continuous distributions. Originally, the sum-product algorithm has been proposed for computing marginals in tree-structured graphs. However, it often also finds high quality solutions when the graph has loops as in our case [22]. In this section, we first briefly review the general sum-product algorithm and then derive the necessary message equations for our model.

The sum-product algorithm for factor graphs works by passing messages between factor and variable nodes [18]. In loopy graphs, messages are initialized to some prior distribution and then updated iteratively until convergence or until a maximum number of iterations has been reached. The factor-to-variable and variable-to-factor messages are defined as

$$\mu_{f \to x}(x) = \sum_{\mathcal{X}_f \setminus x} \phi_f(\mathcal{X}_f) \prod_{y \in \mathcal{X}_f \setminus x} \mu_{y \to f}(y) \tag{5}$$

$$\mu_{x \to f}(x) = \prod_{g \in \mathcal{F}_x \setminus f} \mu_{g \to x}(x) \tag{6}$$

where $\mathcal{X}_f$ denotes all variables associated with factor f and $\mathcal{F}_x$ is the set of factors to which variable x connects. Upon termination, the approximate marginal distribution of each variable can be computed as the product of messages from all neighboring factors:

$$p(x) \propto \prod_{g \in \mathcal{F}_x} \mu_{g \to x}(x). \tag{7}$$

Unfortunately, the application of these message equations to our graphical model (Eq. 2) is not straightforward. First, the ray potentials involve both discrete and continuous variables. While the sum-product equations can be adapted to the continuous domain by replacing sums with integrals, tractable continuous message representations have to be found and the arising integrals need to be calculated efficiently. A second challenge is due to the large number of variables connecting to each ray potential. A naïve application of the sum-product algorithm would require summing over 2^N states for each ray factor-to-variable message in Eq. 5, which is intractable for typical values of N, which are on the order of hundreds. We show that the special algebraic form of the ray potentials in Eq. 4 can be exploited

to reduce this complexity to linear time.

The messages from and to the unary factors, i.e. $\mu_{\varphi_i \to o_i}$ and $\mu_{o_i \to \varphi_i}$, as well as the messages from occupancy variables to ray factors, $\mu_{o_i \to \psi_r}$, are simple and can be calculated using Eq. 5+6. In the following, we will thus focus on the remaining messages: $\mu_{\psi_r \to o_i}$, $\mu_{\psi_r \to a_i}$ and $\mu_{a_i \to \psi_r}$. We consider a single ray r, dropping the subscript for clarity.

Occupancy messages: Before presenting the general form of the message equations, we provide some intuition by analyzing the equations for the first voxel along the ray, o_1. For $o_1 = 1$, the factor-to-variable message reads as

$$\mu_{\psi \to o_1}(o_1 = 1) = \sum_{o_2} \cdots \sum_{o_N} \int_{a_1} \cdots \int_{a_N} \psi(\mathbf{o}, \mathbf{a}) \times \prod_{i=2}^{N} \mu(o_i) \prod_{i=1}^{N} \mu(a_i) \tag{8}$$

where we have abbreviated the incoming occupancy and appearance messages by $\mu(o_i) = \mu_{o_i \to \psi}(o_i)$ and $\mu(a_i) = \mu_{a_i \to \psi}(a_i)$, respectively. Given that the first voxel is occupied, the ray potential ψ in Eq. 4 simplifies to $\psi(o_1 = 1, o_2, ...o_N, \mathbf{a}) = \nu(a_1)$, resulting in

$$\mu_{\psi \to o_1}(o_1 = 1) = \left[\int_{a_1} \nu(a_1) \mu(a_1) \, \mathrm{d}a_1 \right] \times \sum_{o_2} \cdots \sum_{o_N} \int_{a_2} \cdots \int_{a_N} \prod_{i=2}^{N} \mu(o_i) \prod_{i=2}^{N} \mu(a_i). \tag{9}$$

Provided that the incoming messages are normalized such that they integrate (or sum) to 1, all terms in the second line in Eq. 9 evaluate to 1 and the message simplifies to:

$$\mu_{\psi \to o_1}(o_1 = 1) = \int_{a_1} \nu(a_1) \mu(a_1) \, \mathrm{d}a_1. \tag{10}$$

This integral has an intuitive interpretation: it measures the correlation between the observed color, $\nu(a_1)$, and the belief about the voxel's appearance according to the other image measurements, $\mu(a_1)$. If projections of a voxel yield similar colors in multiple images (indicating photo-consistency), the value of the integral will be high and will increase the probability of occupancy for this voxel. We address how to compute integrals of this form in Section 5.

Similarly, for the case where the first voxel is empty, i.e. $o_1 = 0$, we obtain

$$\mu_{\psi \to o_1}(o_1 = 0) = \sum_{j=2}^{N} \mu(o_j = 1) \prod_{k=2}^{j-1} \mu(o_k = 0) \, \rho_j \tag{11}$$

where we use the shorthand notation:

$$\rho_i = \int_{a_i} \nu(a_i) \mu(a_i) \, \mathrm{d}a_i. \tag{12}$$

Note that, owing to the special form of the ray potentials, this message involves only a single sum over the voxel indices and is therefore tractable as well. We provide the full derivation in the supplementary document. Intuitively, this equation measures how well *the voxels after the first voxel* explain the observed pixel intensity assuming that the first voxel is empty. If a voxel $j > 1$ is likely to be occupied ($\mu(o_j = 1)$ is high), visible ($\prod_{k=2}^{j-1} \mu(o_k = 0)$ is high), and matches the appearance of the pixel (ρ_i is high), then this voxel is likely to be the first visible surface voxel and therefore all voxels in front (including the first voxel) should be empty. In this case, the value of Eq. 11 is high, thereby lowering the probability of occupancy for the first voxel.

By following an inductive argument (see supplementary document for details), the general occupancy message equations for voxel i can be written as:

$$\mu_{\psi \to o_i}(o_i = 1) = \sum_{j=1}^{i-1} \mu(o_j = 1) \prod_{k=1}^{j-1} \mu(o_k = 0) \, \rho_j + \prod_{k=1}^{i-1} \mu(o_k = 0) \, \rho_i \tag{13}$$

and

$$\mu_{\psi \to o_i}(o_i = 0) = \sum_{j=1}^{i-1} \mu(o_j = 1) \prod_{k=1}^{j-1} \mu(o_k = 0) \, \rho_j + \frac{1}{\mu(o_i = 0)} \sum_{j=i+1}^{N} \mu(o_j = 1) \prod_{k=1}^{j-1} \mu(o_k = 0) \, \rho_j. \tag{14}$$

The message equations have an intuitive interpretation: The messages cause the probability of occupancy for photo-consistent voxels to increase. The occupancy probability of voxels between the camera and the likely surface voxel are decreased. For voxels that are likely to be occluded, the messages turn out to be uninformative; i.e. $\mu_{\psi \to o_i}(o_i = 1) = \mu_{\psi \to o_i}(o_i = 0)$. Importantly, the ray potential messages can be computed efficiently, in *linear* time, as opposed to the exponential complexity of general high-order cliques.

Appearance messages: The factor-to-variable appearance messages can be written as (see supplementary document for derivation):

$$\mu_{\psi \to a_i}(a_i) = \underbrace{\sum_{j \neq i} \mu(o_j = 1) \prod_{k<j} \mu(o_k = 0) \, \rho_j}_{\text{constant in } a_i} + \underbrace{\mu(o_i = 1) \prod_{k<i} \mu(o_k = 0)}_{\text{weight}} \times \underbrace{\nu(a_i)}_{\text{Gaussian}}. \tag{15}$$

First of all, note that the message computation is again linear in the number of occupancy variables. Second, this message has a special form: it can be written as a constant plus a weighted Gaussian distribution. The constant measures

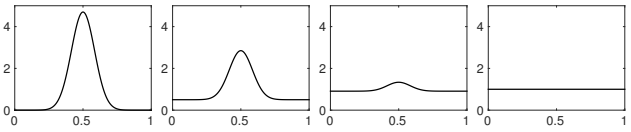


Figure 3: Appearance message for different constants.

how well all voxels except i explain the image observation. The weight measures how likely voxel i is to be the first occupied voxel along the ray. Finally, the term $\nu(a_i)$ measures the agreement between a_i and the observed pixel color I. This special form is highly advantageous since it admits a very compact representation. In practice, we exploit the scaling invariance of the message and store only the weight divided by the constant.

We analyze the message for two cases. First, consider the case when voxel i is the first visible voxel; i.e. $o_i = 1$ and $o_j = 0, \forall j < i$. In this case, it can be verified that the constant term vanishes and the weight evaluates to 1. Thus, the message becomes the normal distribution centered at the observed pixel color I (Fig. 3 left). This is intuitive since the color of the pixel should match the observed voxel i. Second, consider the case when the voxel is either empty or occluded. In either case, it can be verified that the weight is zero and the message becomes a flat distribution (Fig. 3 right). In other words, the pixel observation is not informative for empty or occluded voxels, which is intuitive.

The variable-to-factor appearance messages, $\mu_{a_i \to \psi}(a_i)$, and the appearance marginals, $p(a_i)$, are of continuous form and thus more challenging to represent and compute than their occupancy counterparts. Even for a 1D appearance space, discretization is not an option as the fine level of granularity required would quickly exceed the memory limits for large volumetric reconstructions. Instead, we explicitly represent the current belief about the appearance marginal, $p(a_i)$, at each iteration with a Mixture-of-Gaussians (MoG). The MoG is a suitable approximation for the appearance of many surfaces in natural scenes as it can represent uni-modal (Lambertian surfaces), multi-modal (reflective surfaces), as well as flat distributions (empty or occluded voxels). Moreover, its multi-modal nature helps to cope with wrong observations during the initial iterations when visibility/occlusion relationships are not yet resolved. Finally, the MoG can be stored compactly using the mean, variance and weight of its modes. After each message update, we update the MoG distribution that approximates $p(a_i)$ as

$$p(a_i)^{\text{new}} \propto p(a_i)^{\text{old}} \times \frac{\mu^{\text{new}}_{\psi \to a_i}(a_i)}{\mu^{\text{old}}_{\psi \to a_i}(a_i)} \tag{16}$$

using a Monte Carlo approach as detailed in Section 5.

Finally, the variable-to-factor message, $\mu_{a_i \to \psi}(a_i)$, can be obtained from $p(a_i)$ via:

$$\mu_{a_i \to \psi}(a_i) = \prod_{g \in \mathcal{F}_{a_i} \setminus \psi} \mu_{g \to a_i}(a_i) \propto \frac{p(a_i)}{\mu_{\psi \to a_i}(a_i)}. \tag{17}$$

4.2. Bayes Optimal Depth Prediction

Given a ground truth model or ground truth depth maps, a natural measure of reconstruction performance is the sum of depth errors at each pixel of the input images. While other performance measures (e.g., based on meshes) are applicable as well, we consider this simple metric as it is able to directly measure the performance of the volumetric representation without need for additional meshing steps.

Let $\Delta(\cdot, \cdot)$ denote the loss function that measures the pixel-wise absolute depth error summed over all images. Since $\Delta(\cdot, \cdot)$ decomposes over images and pixels, we will consider a single ray and depth D in the following. According to Bayes decision theory, the optimal depth D^* is given by the depth D that minimizes the expected loss:

$$D^* = \arg\min_D \mathbb{E}_{p(D')}[\Delta(D, D')]. \tag{18}$$

For the ℓ_1-loss, $\Delta(D, D') = |D - D'|$, considered in this paper, the minimizer to Eq. 18 is given by D^* where $p(D < D^*) = p(D \geq D^*) = 0.5$, i.e. the median of $p(D)$ [22].

Similar to the image formation equation (Eq. 1), the depth forward process (i.e. rendering) can be specified as

$$D = \sum_{i=1}^{N} o_i \prod_{j<i} (1 - o_j)\, d_i \tag{19}$$

where d_i denotes the depth of voxel i along an arbitrary ray. According to our graphical model, the depth distribution can be written as (see supplementary document for the derivation)

$$p(D = d_i) \propto \mu(o_i = 1) \prod_{j=1}^{i-1} \mu(o_j = 0)\, \rho_i. \tag{20}$$

Intuitively, voxel i is at the observed depth if the voxel is likely occupied, visible, and explains the observed pixel intensity. Note that this equation highly resembles the message equations (Eq. 13, Eq. 14) and it can be easily computed as a by-product of the inference algorithm.

5. Implementation

This section provides the details of our implementation. The ray-factor messages are initialized to uniform distributions. We estimate an initial belief for each appearance variable by fitting a MoG distribution (using the EM algorithm) to all the pixel colors that the voxel projects to. This appearance initialization helps bootstrap the inference process and leads to faster convergence. After the initialization, the

sum-product algorithm is iterated until convergence.

The presented sum-product belief propagation algorithm is implemented using an asynchronous message passing schedule that is suitable for parallel execution on a GPU. Images are processed one at a time, where each pixel/ray in the image is assigned a thread. All threads simultaneously compute the incoming messages to their ray-factor based on Eq. 6, and then compute the outgoing messages to each occupancy and appearance variable along their ray (Eq. 13, Eq. 14, Eq. 15). Our current implementation can process a 1 MP image in about 7 seconds for a scene with roughly 30 million voxels. This allows us to process hundreds of images with a fine discretization of the voxel grid.

We implemented our algorithm using an octree. The octree allows allocating high resolution cells only near surfaces and therefore saves significant amounts of processing in comparison to a regular voxel grid. In particular, we use shallow octrees amenable to GPU processing proposed by Miller et al. [21]. The inference procedure and octree refinement are carried out in an alternating fashion. Further details can be found in the supplementary document.

The occupancy and appearance beliefs of each voxel are updated based on the new messages computed at each iteration. For the occupancy variables, this belief update is straightforward based on Eq. 7. For the appearance variables, we follow a sampling approach. The belief update equation (Eq. 16) suggests the new MoG distribution will have modes similar to that of the old MoG distribution and/or similar to the mode of $\mu^{\text{new}}_{\psi \to a_i}(a_i)$. Note that both $\mu^{\text{new}}_{\psi \to a_i}(a_i)$ and $\mu^{\text{old}}_{\psi \to a_i}(a_i)$ have the same mode. Therefore, a mixture of $p(a_i)^{\text{old}}$ and $\mu^{\text{new}}_{\psi \to a_i}(a_i)$ constitutes a reasonable proposal distribution. We draw a total of 128 samples from $w\, p(a_i)^{\text{old}} + (1-w)\, \mu^{\text{new}}_{\psi \to a_i}(a_i)$ where $w = 0.5$. The samples are weighted according to the right hand side of Eq. 16 and EM is used to fit the new MoG as an estimate of $p(a_i)^{\text{new}}$. The EM algorithm is initialized with the parameters of $p(a_i)^{\text{old}}$ and iterated until the MoG parameters do not change or a maximum number of iterations (250 for our experiments) are reached. For all experiments, we use MoG with three modes, which we found sufficient for representing the intensities of our gray scale input images.

The integrals that arise during message computation, i.e. Eq. 12, cannot be computed in closed form. We compute an approximation using Monte Carlo integration. The details can be found in the supplementary document.

6. Experimental Evaluation

We evaluate our algorithm on three challenging aerial datasets with LIDAR ground truth and compare the results to the state-of-the-art in terms of reconstruction accuracy as well as its ability to expose uncertainty in the reconstruction. As baselines, we use the algorithms of Liu and Cooper ("LC") [20] and Pollard and Mundy ("PM") [23], both of which have achieved some of the best volumetric reconstruction results for general 3D scenes [5, 20, 30].

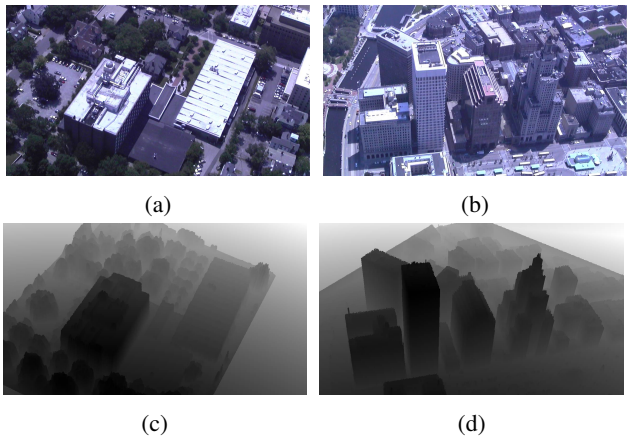


Figure 4: (a,b) Example images from the BARUS&HOLLEY and DOWNTOWN datasets taken from [27]. (c,d) Depth map renderings of the LIDAR ground truth.

We use the publicly available code of the PM algorithm in the VXL project[1] and we have reimplemented the LC algorithm as described in [20]. To enable a direct comparison, we have omitted the pairwise smoothness factors in the original LC formulation as neither PM nor the proposed algorithm contains any spatial regularization. However, note that pairwise smoothness terms could be easily integrated into our approach.

The original LC algorithm estimates the color of each voxel as the *mean* pixel color from which it is visible. We found this procedure to be sensitive to outliers during the first iterations, when visibility/occlusion relationships are not yet resolved. We also implemented a more robust, improved version where the appearance is modeled using a MoG distribution and estimated via EM. The mean of the dominant mode in the MoG is selected as the appearance estimate. We refer to this algorithm as "LC with MoG".

We present a quantitative evaluation of all four algorithms in terms of reconstruction accuracy by comparing the results to LIDAR ground truth. We also provide a qualitative analysis of the occupancy probabilities computed by our algorithm. For the purposes of this paper, we omit evaluation on the relatively easy Middlebury multi-view benchmark [30] as it contains single isolated objects against a uniform background with little ambiguity in the 3D reconstruction. All of the methods investigated by us are able to produce reasonable results on this dataset [5, 20, 30]. Instead, our evaluation focuses on three complex real-world scenes that contain significant ambiguities due to large textureless regions, shadows and highly reflective surfaces.

[1] http://vxl.sourceforge.net/

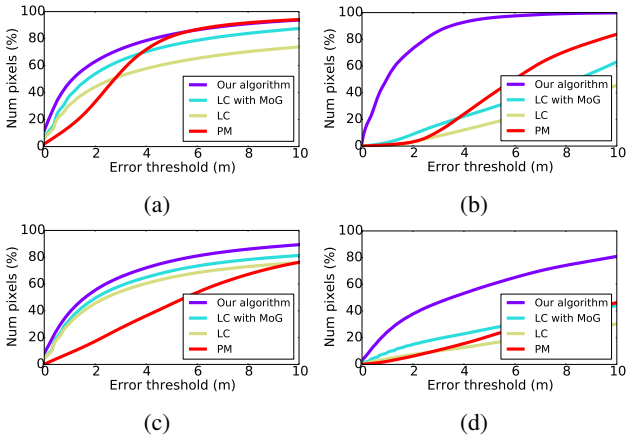


Figure 5: Percentage of correctly estimated pixels in the BARUS&HOLLEY (a,b) and CAPITOL (c,d) datasets. The legend for all figures is the same as in (a). Figures (a) and (c) quantify performance for the entire scenes whereas (b,d) focus on the textureless roof region for BARUS&HOLLEY and on the grass lawn for CAPITOL.

Datasets: We use three aerial datasets from Restrepo et al. [27] and adopt their naming convention: we refer to the sequences as DOWNTOWN, BARUS&HOLLEY and CAPITOL. Sample images from these datasets are shown in Fig. 1a, 4a, and 4b. The images are 1280x720 pixels in size and have an approximate resolution of 30 cm/pixel. 180 views are available for DOWNTOWN, 240 views for CAPITOL and 226 views for BARUS&HOLLEY.

The datasets provide ground truth point clouds obtained from airborne LIDAR [27]. While the LIDAR points are quite dense and precise, they lack coverage on the sides of the buildings. We extrude the points to the ground plane in order to create a more complete ground truth, assuming mostly non-concave surfaces. Note that this assumption mostly holds for these datasets. Trees are the main exception but, since they occupy a negligible fraction of the scene, this does not significantly affect evaluation. The resulting point clouds are then triangulated to obtain a dense surface ground truth, see Fig. 4c and 4d for examples.

We empirically set the occupancy prior to a low value since most voxels are empty in the datasets. Ideally this parameter should be inferred from training data. We use $\gamma = 0.001$ for BARUS&HOLLEY and CAPITOL, and $\gamma = 0.05$ for DOWNTOWN since it has more occupied surface voxels. The effect of γ for the quantitative experiments will be studied in future work.

Evaluation Protocol: We evaluate the reconstruction accuracy as the absolute depth error over all pixels and images with respect to the ground truth depth maps. For each input camera ray, each algorithm's result is used to compute a depth estimate and compared to the ground truth depth. For the LC algorithm's MAP solution, the depth estimate is the depth of the first occupied voxel along the ray. For our algorithm, we compute the estimate according to Section 4.2. As Pollard and Mundy do not provide a method to compute depth estimates, we follow Crispell [6] and use the expected depth $D^* = \mathbb{E}_{p(D)}[D]$ as the prediction.

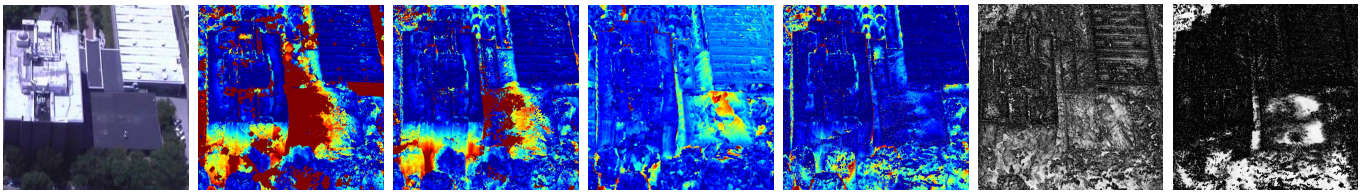
Experimental Results: Fig. 5a, 5c and 8a show the percentage of correctly estimated pixels w.r.t. the error threshold. Our algorithm achieves lower error compared to the other algorithms in both CAPITOL and BARUS&HOLLEY, and produces comparable results for DOWNTOWN.

We visualize the errors in Fig. 6 and Fig. 7. For all three competing algorithms, the dominant locations of error are featureless surfaces: the rooftop in BARUS&HOLLEY and the grass lawn in CAPITOL. All three algorithms yield similar results for these regions: the reconstructions tend to "bulge out" as can be seen in the error maps. Since these featureless surfaces are flat in reality, PM and LC's reconstructions contain gross errors. We explain the reasons for this behavior below.

The LC algorithm computes a MAP estimate which prefers the surface to lie on the outermost layer of the ambiguous region (see Fig. 1c) because any other solution would be occluded by this photo-consistent layer of voxels, thus resulting in a higher energy. PM's Bayes update equation increases the occupancy probability of voxels that are both visible and photo-consistent. Initially, all voxels in the ambiguous region meet this criteria equally well. However, after repeated updates, the interior voxels become occluded and their occupancy probability is no longer increased. Voxels at the outermost layer of the ambiguous region remain visible, thus their occupancy probability continuously increases until it reaches 1. We visualize the occupancy probabilities inferred by the PM algorithm in Fig. 7f by color coding each voxel according to its probability. The color scale is shown in Fig. 1c. It can be seen that the outer visible layer of the ambiguous featureless region has high probability values similar to that of the textured building surface.

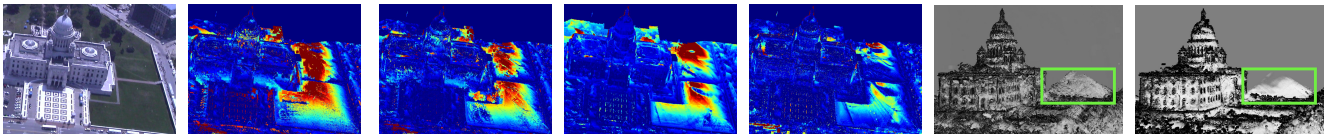
In contrast, the occupancy probabilities inferred by our algorithm (see Fig. 7g) results in a clear distinction. The grass region and textureless patches on the building walls are assigned low probability whereas textured surfaces receive high probability of occupancy. For ambiguous regions, the image evidence is weak, i.e. the ray-potential messages are close to uniform, and therefore, the beliefs are dictated by the prior, which favors empty voxels. For highly textured regions, the image evidence dominates the prior and leads to high probability of occupancy. Similar effects can be observed for the BARUS&HOLLEY dataset in Fig. 6f and Fig. 6g.

Since our algorithm assigns low probability of occupancy *uniformly* over the ambiguous region, the predicted



(a) Image (b) LC (c) LC with MoG (d) PM (e) Our (f) PM Occ. (g) Our Occ.

Figure 6: Analysis of errors for the BARUS&HOLLEY dataset. (a) Reference image. Heat-maps of depth error for LC (b), LC with MoG (c), PM (d) and our algorithm (e). Cooler colors correspond to lower error. Errors for (b-d) are concentrated on the featureless black rooftop, whereas our algorithm's errors are mostly around the tree regions where the LIDAR ground truth is possibly not accurate. Visualization of the occupancy probabilities computed by PM (f) and our algorithm (g).



(a) Image (b) LC (c) LC with MoG (d) PM (e) Our (f) PM Occ. (g) Our Occ.

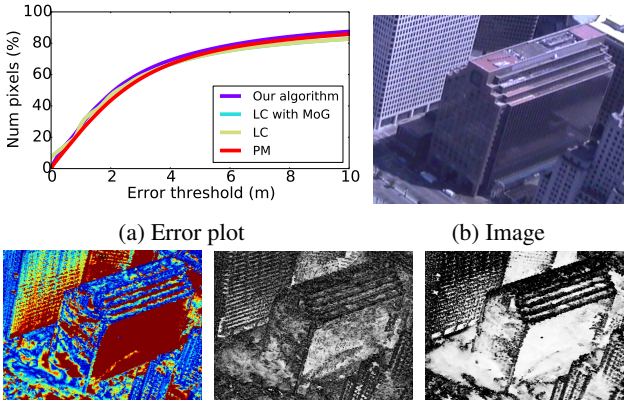
Figure 7: Analysis of errors for the CAPITOL dataset. (a) Reference image. Heat-maps of depth error for LC (b), LC with MoG (c), PM (d) and our algorithm (e). Cooler colors correspond to lower error. Visualization of the occupancy probabilities for PM (f) and proposed algorithm's reconstructions (g).

(Bayes optimal) depth, that is the median of the depth distribution (Section 4.2), is roughly in the middle of the region. This results in lower error than the PM and LC reconstructions, which, by construction, prefer surfaces closest to the camera. We confirm our hypothesis by evaluating all four algorithms in the featureless regions only. Fig. 5b and Fig. 5d show that the proposed algorithm is indeed significantly more accurate in those regions.

Due to the absence of large textureless regions, the performance of all algorithms is roughly the same for the DOWNTOWN sequence as seen in Fig. 8a. The highly reflective building surfaces cause the largest errors. In particular, the building shown in Fig. 8b has a mirror like surface that strongly violates the Lambertian surface assumption. All algorithms yield similar large errors in this region. The error map for our algorithm is shown in Fig. 8c. Occupancy probabilities of our algorithm are visualized in Fig. 8e. The building roof, which contains strong edge features, is localized with high accuracy and the algorithm yields high occupancy probability, indicating strong image evidence. In contrast, the building sides have large errors but also low occupancy probability. The shadow regions are also assigned low probability since their behavior is similar to a featureless surface. In contrast, Fig. 8d shows that the PM algorithm assigns high occupancy to all estimated surface voxels.

We encourage the reader to look at our project page for additional resources [2].

[2] http://ps.is.tue.mpg.de/project/Volumetric_Reconstruction



(a) Error plot (b) Image

(c) Error map (d) PM Occ. (e) Our Occ.

Figure 8: (a) Error plots for the DOWNTOWN dataset. (b) Reference image. (c) Heat-map of errors for our algorithm. (d,e) Visualization of the voxel occupancy probabilities computed by PM (d) and our algorithm (e).

7. Conclusions

In this paper, we have presented a novel probabilistic method for volumetric reconstruction that faithfully encodes reconstruction ambiguities while achieving high quality 3D models. Our results show that the algorithm compares favorably to the state-of-the-art both in terms of reconstruction accuracy as well as in the ability to reveal reconstruction uncertainty. This is a key strength of our approach relative to prior work. While we have utilized weak prior information in this work, the proposed

probabilistic approach provides a foundation with which to combine multi-view image data with a wide range of priors. Priors expressing spatial smoothness, temporal consistency [12, 34] as well as semantic information [13, 14] can be principally integrated into our probabilistic formulation.

References

[1] M. Agrawal and L. S. Davis. A Probabilistic Framework for Surface Reconstruction from Multiple Images. *CVPR*, 2001.

[2] R. Bhotika, D. J. Fleet, and K. N. Kutulakos. A Probabilistic Theory of Occupancy and Emptiness. *ECCV*, 2002.

[3] J. D. Bonet and P. Viola. Poxels: Probabilistic Voxelized Volume Reconstruction. *ICCV*, 1999.

[4] A. Broadhurst, T. W. Drummond, and R. Cipolla. A Probabilistic Framework for Space Carving. *ICCV*, 2001.

[5] F. Calakli, A. O. Ulusoy, M. I. Restrepo, G. Taubin, and J. L. Mundy. High Resolution Surface Reconstruction from Multi-view Aerial Imagery. *3D Imaging Modeling Processing Visualization Transmission (3DIMPVT)*. IEEE, 2012.

[6] D. Crispell. *A Continuous Probabilisitic Scene Model for Aerial Imagery*. PhD thesis, Brown University, 2012.

[7] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing Building Interiors from Images. *ICCV*, 2009.

[8] Y. Furukawa and C. Hernandez. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2013.

[9] P. Gargallo and P. Sturm. Bayesian 3D Modeling from Images Using Multiple Depth Maps. *CVPR*, 2005.

[10] P. Gargallo, P. Sturm, and S. Pujades. An Occupancy-Depth Generative Model of Multi-view Images. *ACCV*, 2007.

[11] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-View Stereo for Community Photo Collections. *ICCV*, 2007.

[12] L. Guan and M. Pollefeys. Probabilistic 3D Occupancy Flow with Latent Silhouette Cues. *CVPR*, 2010.

[13] F. Guney and A. Geiger. Displets : Resolving Stereo Ambiguities using Object Knowledge. *CVPR*, 2015.

[14] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D Scene Reconstruction and Class Segmentation. *CVPR*, 2013.

[15] X. Hu and P. Mordohai. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *PAMI*, 2012.

[16] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes. Large Scale Multi-view Stereopsis Evaluation. *CVPR*, 2014.

[17] J. Kopf, F. Langguth, D. Scharstein, R. Szeliski, and M. Goesele. Image-Based Rendering in the Gradient Domain. *SIGGRAPH Asia*, 2013.

[18] F. Kschischang. Factor Graphs and the Sum-Product Algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001.

[19] K. N. Kutulakos and S. M. Seitz. A Theory of Shape by Space Carving. *IJCV*, 2000.

[20] S. Liu and D. Cooper. Statistical Inverse Ray Tracing for Image-Based 3D Modeling. *PAMI*, 2014.

[21] A. Miller, V. Jain, and J. L. Mundy. Real-time Rendering and Dynamic Updating of 3-d Volumetric Data. *Proceedings of the Fourth Workshop on General Purpose Processing on Graphics Processing Units*, 2011.

[22] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[23] T. Pollard and J. L. Mundy. Change Detection in a 3-d World. *CVPR*, 2007.

[24] J.-P. Pons, P. Labatut, H.-H. Vu, and R. Keriven. High Accuracy and Visibility-Consistent Dense Multiview Stereo. *PAMI*, 2012.

[25] S. Pujades, F. Devernay, and B. Goldluecke. Bayesian View Synthesis and Image-Based Rendering Principles. *CVPR*, 2014.

[26] M. I. Restrepo. *Characterization of Probabilistic Volumetric Models for 3-d Computer Vision*. PhD thesis, Brown University, 2013.

[27] M. I. Restrepo, A. O. Ulusoy, and J. L. Mundy. Evaluation of Feature-based 3-d Registration of Probabilistic Volumetric Scenes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98(0):1–18, 2014.

[28] N. Savinov, L. Ladicky, C. Hane, and M. Pollefeys. Discrete Optimization of Ray Potentials for Semantic 3D Reconstruction. *CVPR*, 2015.

[29] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. *CVPR*, 2006.

[30] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. Multi-View Stereo Evaluation. http://vision.middlebury.edu/mview/eval/, 2011.

[31] C. Strecha, R. Fransens, and L. Van Gool. Wide-baseline Stereo from Multiple Views: a Probabilistic Account. *CVPR*, 2004.

[32] 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. *CVPR*. Ieee, 2008.

[33] A. O. Ulusoy, O. Biris, and J. L. Mundy. Dynamic Probabilistic Volumetric Models. *ICCV*, 2013.

[34] A. O. Ulusoy and J. L. Mundy. Image-based 4-d Reconstruction Using 3-d Change Detection. *ECCV*, 2014.

[35] A. Yao and A. Calway. Dense 3-D Structure from Image Sequences using Probabilistic Depth Carving. *BMVC*, 2003.