

Aerial Reconstructions via Probabilistic Data Fusion

Randi Cabezas Oren Freifeld Guy Rosman John W. Fisher III
Massachusetts Institute of Technology
{rcabezas, freifeld, rosman, fisher}@csail.mit.edu

Abstract

We propose an integrated probabilistic model for multi-modal fusion of aerial imagery, LiDAR data, and (optional) GPS measurements. The model allows for analysis and dense reconstruction (in terms of both geometry and appearance) of large 3D scenes. An advantage of the approach is that it explicitly models uncertainty and allows for missing data. As compared with image-based methods, dense reconstructions of complex urban scenes are feasible with fewer observations. Moreover, the proposed model allows one to estimate absolute scale and orientation, and reason about other aspects of the scene, e.g., detection of moving objects. As formulated, the model lends itself to massively-parallel computing. We exploit this in an efficient inference scheme that utilizes both general purpose and domain-specific hardware components. We demonstrate results on large-scale reconstruction of urban terrain from LiDAR and aerial photography data.

1. Introduction

The increasing availability of multi-modal data sets, including aerial imagery and Light Detection and Ranging (LiDAR), provides new opportunities for visualization and analysis of extended geographic areas. A critical challenge for enabling such analysis is to develop methods for fusing the available data within a mathematically-consistent framework. While much progress has been made on image-based scene reconstruction, the existing literature on *multi-modal* scene reconstruction is less extensive.

Here, we propose an integrated probabilistic model for multi-modal fusion of Wide Area Motion Imagery (WAMI) and LiDAR for 3D scene analysis. We formulate the reconstruction as a statistical inference problem within a Bayesian framework that encompasses the following key modeling issues: (1) integration and exploitation of multi-modal measurements within a mathematically consistent model; (2) explicit modeling of uncertainty in both measurements and the resultant reconstruction; and (3) allows for straightforward incorporation of additional data sources.

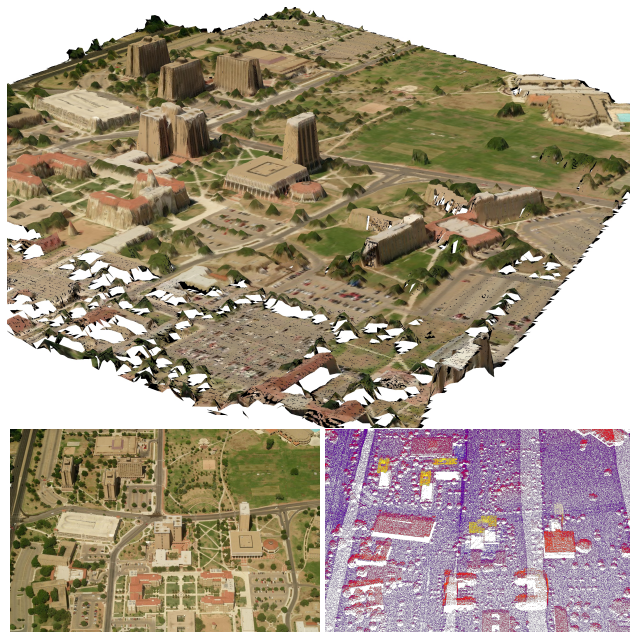


Figure 1: *Top*: Lubbock scene reconstruction (3 images, and 1M LiDAR points). *Bottom*: Sample measurements.

The resulting model lends itself to highly parallelizable inference, a property we exploit using existing graphics hardware. We discuss and empirically demonstrate several advantages of the approach including: (1) the ability to obtain dense reconstructions in both geometry and appearance by utilizing higher order primitives and images to represent the 3D scene; (2) the need for fewer images given the geometric information provided by LiDAR; and (3) the ability to perform higher level reasoning within the model, e.g., detecting moving objects and obtaining absolute scale and orientation.

2. Background and Related Work

Structure from Motion (SfM) is a widely-used technique for estimating 3D scene geometry and camera pose from a collection of images. Traditional reconstruction methods [12, 32, 35] rely solely on images, neglecting avail-

able intrinsic and extrinsic information. Newer methods exploit additional noisy information such as Global Positioning System (GPS) measurements or focal-length estimates (information readily available from most digital cameras). Such additional information is commonly used for initialization [31] and/or reconstruction [9]. Recent advances in SfM [1, 28, 31, 34, 37] achieve remarkable reconstructions of urban scenes from a large collection of street views. The proposed approach differs from these methods in the choice of geometric primitive used (*e.g.*, modeling geometry as higher-order primitives as opposed to single points), the use of additional noisy geometric measurements (*e.g.*, LiDAR), and the replacement of explicit pixel correspondences with dense implicit correspondences.

Previous attempts to recover higher-order primitives include piecewise-planar reconstructions [17, 30] that produce accurate reconstructions when the underlying scene is primarily planar; however, these methods typically utilize SfM as an initial step and are therefore susceptible to errors in the SfM reconstructions.

The utility of geometric information in SfM is generally agreed upon but methods for exploiting it vary widely [2, 3, 4, 5, 8, 11, 33]. These methods typically introduce geometry as constraints that provide regularity and reduce computational complexity. Common constraints include time-consuming manual annotation of points, lines, or planar structures. The proposed method differs in that it utilizes LiDAR as a source of geometric information while eliminating the need for manual interaction.

LiDAR has been exploited extensively in aerial reconstructions [10, 15, 19, 22, 23, 24, 25, 38, 39, 40]. However, in sharp contrast to the proposed method, these methods assume that LiDAR provides a noise-free and accurate geometry, relegating images solely as a source of texture information and neglecting image-based geometric information. Previous image-based aerial reconstructions methods include [20, 21] while probabilistic formulations of SfM include [7, 13, 29]. To our knowledge, no previous method for 3D reconstruction combines geometry and appearance information within a joint probabilistic model.

The work presented herein is closest in spirit to the variational approaches of [18, 37]. The primary differences being the use of LiDAR and probabilistic modeling as well as the use of multi-image color differences as a comparison metric rather than pairwise image correlation.

LiDAR is an optical remote-sensing technology that measures distance and/or material properties of a point of reflection. For airborne collections, the system is mounted on an aircraft along with a position tracking system such as GPS. During collection, the ground is scanned continuously with light pulses. Combining the return delay of a pulse with the platform position and velocity yields accurate 3D point measurements with errors on the order of a few cen-

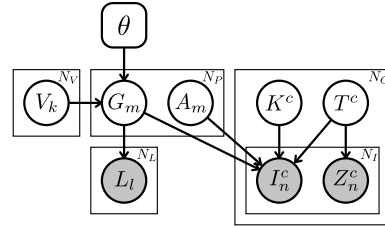


Figure 2: Graphical Model Representation.

timeters. Collecting a dense set of such measurements over a large area facilitates automated methods for scene modeling by providing a rich source of geometric information.

3. The Probabilistic Model

The proposed model consists of the following latent parameters: a collection of 3D primitives described by their geometry and appearance; a camera-trajectory model (described by camera extrinsic and intrinsic parameters), as well as several observation models that couple the latent variables with observed data.

3.1. The Proposed Model

The graphical model in Fig. 2 depicts the generative probabilistic model used herein. The latent 3D model consists of the variables G , V , and A which collectively explain the *geometry* and *appearance* of the underlying scene. Intrinsic and extrinsic camera parameters are denoted by K and T , respectively. Observations are LiDAR points (L), images (I), and (if available) GPS positions (Z). The observations are assumed to be statistically independent conditioned on the latent variables. Conceptually, the model may be divided into two parts. The first part (the leftmost and middle plates in Fig. 2) encodes the 3D scene structure, while the second (the rest of the plates in Fig. 2) encodes the multi-modal observations.

The scene structure consists of N_P geometry primitives $\mathbf{G} = \{G_m\}_{m=1}^{N_P}$ and associated canonical appearances $\mathbf{A} = \{A_m\}_{m=1}^{N_P}$. The appearance variable, A_m , represents a square image of known size which contains the texture information for G_m . For each $m \in \{1, \dots, N_P\}$, θ_m encodes the choice of vertices (for the primitive G_m) among $\mathbf{V} = \{V_k\}_{k=1}^{N_V}$, the set of all vertices in the scene (as a shorthand we also write $\theta = \{\theta_m\}_{m=1}^{N_P}$). This enables different connectivity assumptions and primitive types to be used in the model, and allows us to represent triangulated meshes, quad-meshes, and polygon-soup models. For concreteness we will focus on triangular meshes.

The LiDAR measurements, $\mathbf{L} = \{L_l\}_{l=1}^{N_L}$, are expressed in the bottom plate of Fig. 2 as conditionally independent points. The rightmost plate explains image measurements arising from N_C independent cameras, each with

its own intrinsic parameters K^c and extrinsic camera trajectory T^c . The set of all intrinsic parameters and extrinsic trajectories are denoted by $\mathbf{K} = \{K^c\}_{c=1}^{N_C}$ and $\mathbf{T} = \{T^c\}_{c=1}^{N_C}$ respectively; note that set complements are denoted using subscript \setminus (e.g., $\mathbf{T}_{\setminus n} = \{T^c\}_{c=1, c \neq n}^{N_C}$). Henceforth, we reserve superscript notation for camera indices unless clear by context; all other indices appear in subscript. Each camera c generates N_f^c independent images, denoted by $\mathbf{I}^c = \{I_n^c\}_{n=1}^{N_f^c}$, and GPS measurements, denoted by $\mathbf{Z}^c = \{Z_n^c\}_{n=1}^{N_f^c}$. The ensembles across all cameras are denoted by $\mathbf{I} = \{\mathbf{I}^c\}_{c=1}^{N_C}$ and $\mathbf{Z} = \{\mathbf{Z}^c\}_{c=1}^{N_C}$.

The probability model depicted in Fig. 2 is

$$p(\mathbf{L}, \mathbf{I}, \mathbf{Z}, \mathbf{G}, \mathbf{V}, \mathbf{A}, \mathbf{T}, \mathbf{K}; \theta) = p(\mathbf{V}, \mathbf{A}, \mathbf{T}, \mathbf{K}) \prod_{l=1}^{N_L} p(L_l | \mathbf{G}) \\ \times \prod_{c=1}^{N_C} \prod_{n=1}^{N_f^c} p(I_n^c | \mathbf{G}, \mathbf{A}, K^c, T^c) p(Z_n^c | T^c) \prod_{m=1}^{N_P} p(G_m | V_\theta), \quad (1)$$

where

$$p(\mathbf{V}, \mathbf{A}, \mathbf{T}, \mathbf{K}) = \prod_{m=1}^{N_P} p(A_m) \prod_{c=1}^{N_C} p(T^c) p(K^c) \prod_{k=1}^{N_V} p(V_k).$$

Next, we explain the individual terms of Eq. (1).

3.2. Camera Trajectories

Assuming a rigid attachment, the airborne platform and camera trajectories are the same. Let T^c denote the trajectory of such a camera c . We use a *Gaussian Process* (GP) as a prior over T^c ; see our Supplemental Material for details [6]. This serves as a prior model on smooth trajectories and enables efficient inference. While SfM and/or aerial-imagery-based methods may also incorporate such a prior over camera trajectories (even if they do not exploit LiDAR information), we are not aware of any published works which do so.

3.3. LiDAR Observation Model

LiDAR measurements are modeled as either being generated by some point on a primitive and corrupted by additive Gaussian noise or as outliers (as in the case of spurious measurements). Mathematically, we write the former case as $L_l = L_l^{m(l)} + W$, where $L_l^{m(l)}$ is the point on primitive $G_{m(l)}$ that generated the measurement, and $W \sim \mathcal{N}(w; 0, \sigma^2 \mathbf{I})$. Thus, $L_l | L_l^{m(l)} \sim \mathcal{N}(L_l; L_l^{m(l)}, \sigma^2 \mathbf{I})$.

A-priori, it is unknown which primitive, let alone which point on the primitive, generated each observation. We address this data association problem by sampling the measurement-to-primitive association, and assuming that the measurement is always generated by the closest point on the generating primitive. The LiDAR statistical model then becomes $p(L_l | \mathbf{G}) = p(L_l | G_{m(l)}) = p(L_l | L_l^{m(l)})$ where $m(l) \sim \text{Cat}(\alpha p(L_l | G_1), \dots, \alpha p(L_l | G_{N_P}), 1 - \alpha)$ and $1 - \alpha$ is the probability of an outlier (set to a small value).

3.4. GPS and Image Observation Models

The GPS observation model is an additive Gaussian model, $Z_n^c = T_n^c + W$ where $W \sim \mathcal{N}(w; 0, \sigma_w^2 \mathbf{I})$ and $T_n^c = T^c(t_n)$ is the location of camera c at t_n , the time of the n^{th} measurement. This leads to a Gaussian likelihood, $Z_n^c | T^c \sim \mathcal{N}(z; T_n^c, \sigma_w^2 \mathbf{I})$. We model color intensity of image I_n^c at pixel (u, v) via a Gaussian noise model,

$$I_n^c(u, v) \sim \mathcal{N}(i_n^c(u, v); A_{m^*}(u', v'), r_{m^*}^2) \quad (2)$$

where A_{m^*} is the mean of the appearance of G_{m^*} (the primitive that generated $I_n^c(u, v)$) at some coordinate (u', v') , $r_{m^*}^2$ is the variance of the noise which depends on the angle between G_{m^*} and the camera's viewing direction (see [6] for more details). Importantly, m^* , u' and v' are functions of u, v, K^c, T^c , and \mathbf{G} ; this can be seen by interpreting Eq. (2) as a mapping of color values from an image pixel to the appearance of the primitive m^* . This map depends on camera parameters and the visibility of pixel (u', v') of G_{m^*} . With these choices,

$$p(I_n^c | \mathbf{G}, \mathbf{A}, K^c, T^c) = \prod_{k \in \mathcal{S}_n^c} \mathcal{N}(i_k; a_{m^*(k)}, r_{m^*(k)}^2) \quad (3)$$

where \mathcal{S}_n^c is the set of pixels in image I_n^c .

3.5. Lie Algebraic Representation of Primitives

Here, we describe the representation for the geometric primitives in the model. Since we wish to allow surfaces to be composed of planar primitives, it is intuitive to handle the transformation of each primitive rather than treat displacements of each primitive's vertex separately and then enforce primitive planarity via constraints. Motivated by the differential-geometric tangent-plane approach, we constrain the representation to linear maps between planar primitives that preserve the primitive's orientation. In the case of triangulated meshes (the type used in our implementation), these maps capture the full range of triangle transformations in \mathbb{R}^3 . For triangular meshes, one choice of parameterization of these transformations is the approach suggested in [14], which conveniently decomposes local transformations into their translation, rotation, scale and skew components. This decomposition is exploited in our inference algorithm, by allowing us to make coordinate descent steps along more meaningful directions. Due to space limitations, we provide the mathematical details in [6].

4. Inference

We now describe our inference algorithm. To simplify notation and without loss of generality, the explanation below assumes that only one camera is used, $N_C = 1$, and we thus drop the superscript c . For computational reasons, we focus on Maximum-a-Posteriori (MAP) estimates.

Algorithm 1 General Inference Procedure

- 1: Initialize world primitives (see Sec. 4.4).
 - 2: Initialize camera pose (see Sec. 4.4).
 - 3: **for** $iter = 1 : N_{iter}$ **do**
 - 4: Estimate Appearance using Eq. (5).
 - 5: Optimize over each camera pose using Eqs. (6-7).
 - 6: Estimate Appearance using Eq. (5).
 - 7: Optimize over each world primitive using Eq. (8).
 - 8: **end for**
-

4.1. Appearance

The posterior distribution over the appearance is

$$p(\mathbf{A}|\mathbf{I}, \mathbf{G}, K, T) \propto \prod_{n=1}^{N_I} p(I_n|\mathbf{G}, \mathbf{A}, K, T) \prod_{m=1}^{N_P} p(A_m). \quad (4)$$

By Sec. 3.4, the first term in the RHS of Eq. (4) is normally distributed. Selecting the prior on A_m to be Gaussian, $A_m \sim \mathcal{N}(a; \mu, \sigma^2)$, leads to a closed-form Gaussian posterior; *i.e.*, $p(\mathbf{A}|\mathbf{I}, \mathbf{G}, K, T) = \mathcal{N}(a; \hat{\mu}, \hat{\sigma}^2)$ where

$$\hat{\mu} = \frac{\mu + \sigma^2 \sum_{i=0}^{n-1} \frac{z_i}{r_i^2}}{1 + \sigma^2 \sum_{i=0}^{n-1} \frac{1}{r_i^2}}; \quad \hat{\sigma} = \frac{\sigma}{\sqrt{1 + \sigma^2 \sum_{i=0}^{n-1} \frac{1}{r_i^2}}}, \quad (5)$$

and z_i are observed pixels generated by the same primitive appearance pixels (see [6] for derivation). This is weighted-least-squares estimation with weights that are inversely proportional to the variance of the pixel likelihood.

4.2. Camera Parameters

Inference over camera parameters is decomposed into inference over intrinsic and extrinsic parameters.

Intrinsic Parameters. The posterior distribution is

$$p(K|\mathbf{I}, \mathbf{G}, \mathbf{A}, T) \propto p(K) \prod_{n=1}^{N_I} p(I_n|\mathbf{G}, \mathbf{A}, K, T). \quad (6)$$

Note that Eq. (6) does not have a closed-form solution due to the intricate dependency of $p(I_i|\mathbf{G}, \mathbf{A}, K, T)$ on K . This dependency, implicit in Eq. (2), necessitates the computation of A_{m^*} and (u', v') which requires reasoning over both a projective transformation and occlusions and cannot be achieved in closed form. However, we can optimize numerically to obtain a MAP estimate of K .

Extrinsic Parameters. The posterior distribution related to extrinsic parameters is given by

$$p(T_n|\mathbf{G}, \mathbf{A}, I_n, K, \mathbf{T}_{\setminus n}, Z_n) \propto p(I_n|\mathbf{G}, \mathbf{A}, T_n, K_n) p(Z_n|T_n) p(T_n|\mathbf{T}_{\setminus n}). \quad (7)$$

Note that Eq. (7) contains not only the image and GPS likelihoods, but also the likelihood of the current camera parameters conditioned on all the other latent camera parameters in the trajectory. This is due to the GP prior placed

on T (Sec. 3.2), which can be evaluated as shown in [6]. Intuitively, $p(T_n|\mathbf{T}_{\setminus n})$ favors an extrinsic-parameters configuration that fits well with the rest of the trajectory.

4.3. Geometry

The posterior distribution over geometric parameters is

$$p(\mathbf{V}|\mathbf{I}, \mathbf{L}, \mathbf{G}; \theta) \propto \prod_{n=1}^{N_I} p(I_n|\mathbf{G}, \mathbf{A}, K, T) \prod_{l=1}^{N_L} p(L_l|\mathbf{G}) \prod_{m=1}^{N_P} p(G_m|\mathbf{V}; \theta). \quad (8)$$

The complicated form of $p(I_n|\mathbf{G}, \mathbf{A}, K, T)$ prevents closed-form exact-inference solutions for Eq. (8). Moreover, $\prod_{l=1}^{N_L} p(L_l|\mathbf{G})$ is computationally intensive as the data association problem requires computing distances for each LiDAR observation to every world primitive. We seek MAP solutions for the scene geometry using the representation from Sec. 3.5 which, as mentioned earlier, enables us to optimize over a linear space. Empirically, when compared with a naive 3D vertex representation, the Lie-algebraic representation produced results that are either similar or better – in equivalent run-times. Importantly, the transformation decomposition property is utilized to reduce local optima by first searching over rotations and translations, then optimizing over all deformations, *i.e.*, rotation, translations, scale and skew.

4.4. Practical Considerations

Thus far we have discussed inference for individual latent parameters; this section combines them and discusses a few implementation details (see [6] for more details). The full inference procedure can be seen in Alg. 1. In a fully-Bayesian model the initialization of parameters can be done by sampling from the prior distributions; however, due to the high complexity of the model, these poor initializations are likely to produce equally-poor parameter estimates. As a result, in this work the model parameters are initialized from the input data. World primitives are initialized by triangulating a subsampled input LiDAR data at the ground level. The camera parameters are initialized from GPS information, if available, or by first registering the images and LiDAR [25]. As a final note, the optimization scheme used in this work was downhill simplex optimization [26]. Source code for our implementation can be downloaded from: <http://people.csail.mit.edu/rcabezas/code>

5. Experiments

In our experiments we use the CLIF 2007 dataset [36] and LiDAR from [27]. The CLIF dataset is partitioned into three scenes: *intersection*, *stadium* and *multi-camera*. See [6] for detailed scene descriptions, parameters used in our experiments and more results.

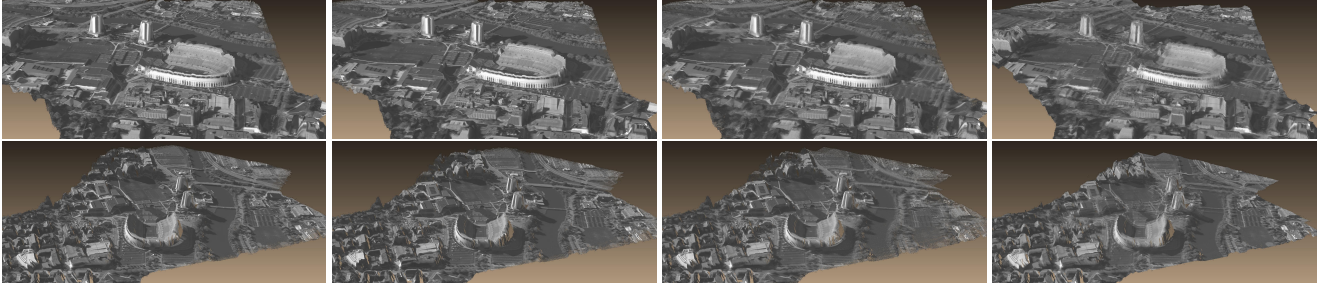


Figure 3: Reconstruction as a function of number of images used (2 views). From left to right: 49, 24, 10 and 5 images (fixed world geometry and LiDAR, same number of optimizations run for all).

5.1. Appearance and Geometry

LiDAR and images complement each other as the former provides geometric information while the latter primarily provides information about the appearance (though images also convey implicit geometric information). We characterize this property by comparing reconstruction quality as a function of images used for a fixed set of LiDAR points.

In this experiment we use a single camera and vary the number of images used to reconstruct the scene between 5, 10, 24, and 49. Images were chosen (equally spaced in time) from the observation sequence always beginning and ending with the same images. This ensured that the baseline was maintained between reconstructions, but also that there was maximal separation between images used. We optimized over the camera pose for each image in each set while maintaining all other model parameters fixed. Each set of camera poses was optimized using three runs of ten batches with image noise standard deviation of 10, 5, and 2 respectively. This was followed by a final run of five batches with a smaller optimization step-size. The term batch refers to updating all the poses in the set once. Empirically, we found that varying image-noise level had no noticeable effect on the reconstruction output, but decreasing the number of batches could have significant impact.

Two views for each of the image sets are shown in Fig. 3. Note that the reconstruction quality degrades very little when reducing the number of images from 49 to 10. This degradation is manifested as black pixels which occur due to missing information in the latent appearance model (*i.e.*, no observation pixel is mapped to that particular appearance location). While mildly distracting, this is easily mitigated by either increasing the number of appearance pixels each observation pixel is allowed to affect or by post-processing the appearance maps to fill in missing data. Despite the aforementioned artifacts, the first three sets of images converged to a similar configuration of camera parameters, and are aligned well with each other; *e.g.*, note the sharp features in Fig. 3. When we further decrease the number of images to 5, the figure shows that this no longer holds. We hy-

pothesize that the limited image evidence, together with the wide-baseline and the high uncertainty in the initial camera pose caused the optimization to converge to an undesired local optimum.

5.2. SfM comparison

In this section the reconstructions obtained with the proposed model are compared with the results of Bundler+PMVS2 [16, 31]. As pointed out earlier, these reconstructions are fundamentally different, beginning with the choice of primitives. Due to such variations and the lack of ground truth of real-world large-scale urban scenes, only qualitative comparisons are considered.

The three scenes of the CLIF dataset were used to compare the reconstruction quality between SfM and the proposed approach. For brevity, only the results of the multi-camera scene are shown here; see [6] for more reconstructions. The results of Bundler+PMVS2 (top row of Figs. 4-5) contain over 151k points and were obtained using 77 images out of the 100 given as their algorithm failed to identify camera pose for the remaining cameras. From a far-off distance the point cloud lets us see the underlying scene structure. Horizontal surfaces are well reconstructed, leading to excellent ground coverage; building sides are fairly dense and vary from being highly vertical to slightly inclined (*e.g.*, the front wall of the stadium or the towers, as seen in Fig. 5).

The results using the proposed method are shown on the bottom row of Figs. 4-5. Note that we can easily identify fine scene details such as pavement markings, roof details and cars. Moreover, horizontal and vertical surfaces are well reconstructed. We note that foliage typically does not follow the locally-planar assumption made in this model; as such, reconstructions of trees are highly irregular. All images are used in this model, yielding more ground coverage of the scene than the SfM counterpart. Furthermore, note that small but crucial scene details can be identified from the reconstructions obtained using the proposed method.

Computation Time. Runtimes of Bundler+PMVS2 and the proposed method are shown in Table 1 (see [6] for detailed timing breakdown). The Bundler+PMVS2 compu-

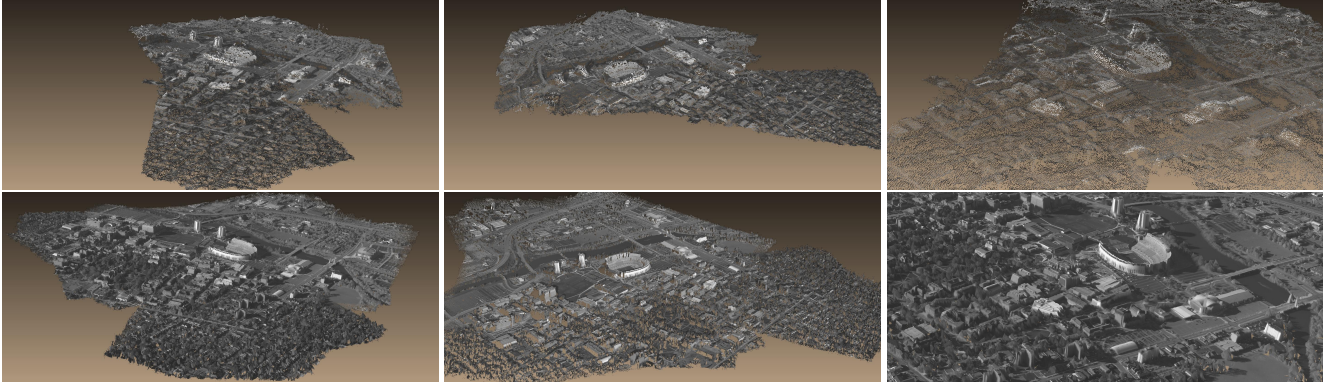


Figure 5: CLIF reconstruction, 3 views. *Top*: Bundler+PMVS2, (151k points using 77 out of 100 images since not all camera pose parameters were found). *Bottom*: proposed method (280k visible primitives).

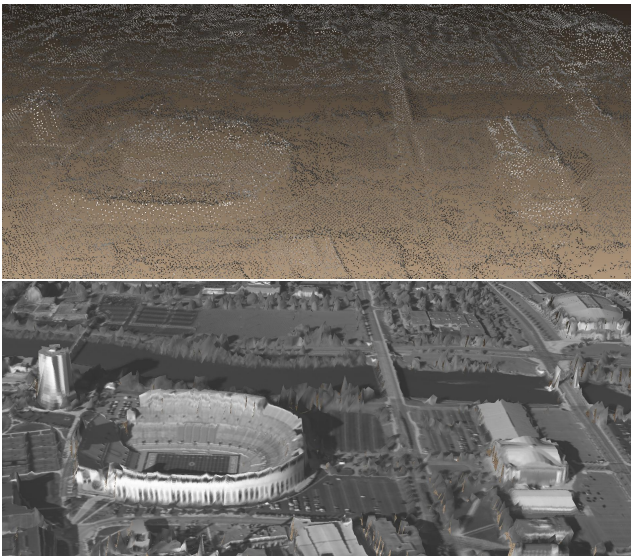


Figure 4: CLIF reconstruction. *Top*: Bundler+PMVS2. *Bottom*: proposed method. Note that small details can be easily seen in the reconstruction obtained using the proposed approach.

tation time for each of the scenes is quite low; this is unsurprising as SfM typically requires many images, on the order of thousands, and hence implementations are highly efficient. The proposed method is considerably slower; the main contributing factor is the large number of visible primitives needed to reason about in order to compute scene geometry. Recall that scene geometry requires reasoning about projective transformations and occlusions. This places significant computation burden on the rendering pipeline. Optimizing the rendering pipeline should significantly reduce computation times of the proposed method. As an additional note, the probabilistic model allows one to reason about uncertainty; this feature can be used to allocate

Scene	Bundler+PMVS2	Proposed Model
<i>Intersection</i>	8.15	37.72
<i>Stadium</i>	51.90	445.08
<i>Multi-Camera</i>	143.80	2,238.97

Table 1: Running Time Comparisons (all times in minutes).

computation resources given a fixed budget (*e.g.*, given such a budget, the model can be used to identify and rank scene parts that could benefit the most from added computation).

5.3. Beyond Reconstructions

This section discusses two aspects of the proposed model that go beyond 3D scene reconstruction. These added features are: (1) having absolute scale and orientation; and (2) being able to identify moving objects in the scenes. In our model, these features are obtained directly as a simple by-product of the inferred model. This is in sharp contrast to traditional SfM.

Absolute Scale and Orientation. LiDAR measurements may be used to identify the absolute scale and orientation of reconstructions. This is possible since the measurements are geocoded, and allows us to recover the scale factor that is unknown in traditional SfM. Of course, additional metric information can also be used for similar purposes in SfM; however, unlike the proposed method, this information must be explicitly incorporated in SfM. Knowing the absolute scale and orientation is important in many applications where measuring distances are crucial (*e.g.* route planning). Furthermore, being able to reason about physical units via remote sensing, as opposed to traditional surveying techniques, can result in both financial and time savings. As an example of how scale and orientation knowledge may be used within the proposed model, we demonstrate the ability to measure distances in the stadium scene; see Fig. 6. The

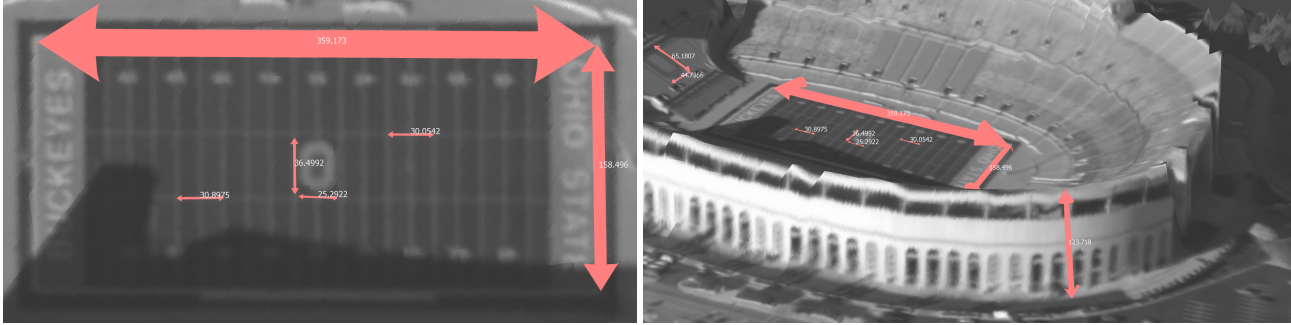


Figure 6: Two distance annotation examples of OSU Stadium. *Left: Field, Right: Stadium*

figure shows the measured dimensions of the football field to be 359.2×158.5 [ft], which corresponds to the actual dimensions of 360×160 [ft] for the field; this estimate is consistent with the uncertainty measurements. We can use the same distance measuring tools to directly measure unknown quantities such as the size of the “O” in the field or the height of the stadium (see Fig. 6).

Identifying Moving Objects. An implicit assumption of traditional reconstruction algorithms is that the scene is static. In practice, however, this assumption is often violated and, as a result, moving objects are often pruned or treated as outliers. In contrast, having a complete model for the measurement process allows us to reason about, and detect, moving objects as part of the inference process.

To demonstrate this ability, we computed image likelihood of the *intersection* scene after inferring the model parameters. Typical results are shown in Fig. 7, where low likelihood pixels have been color-coded in red on the original images. We see that the model captures cars traversing the intersection fairly well. Some low likelihood pixels do not correspond to movers, *e.g.*, building edges; these sharp discontinuities have a low likelihood due to poor geometry reconstruction. Despite these outliers, it is clear that movers are reliably detected.

6. Conclusion

We proposed a Bayesian data-fusion approach for scene reconstruction from multi-modal data, specifically LiDAR and aerial imagery. We defined a novel generative model for combining the data sources and showed that it leads to not only fast approximate inference but also to several important advantages when compared with traditional image-based approaches. First, our method requires fewer images to achieve reconstruction results that are qualitatively similar (and sometimes superior). Second, the use of higher-order geometry primitives and dense appearance allows small but crucial scene details to be detected and visualized. Third, as presented, the model allows scene reasoning to extend beyond reconstruction, *e.g.*, the model allows

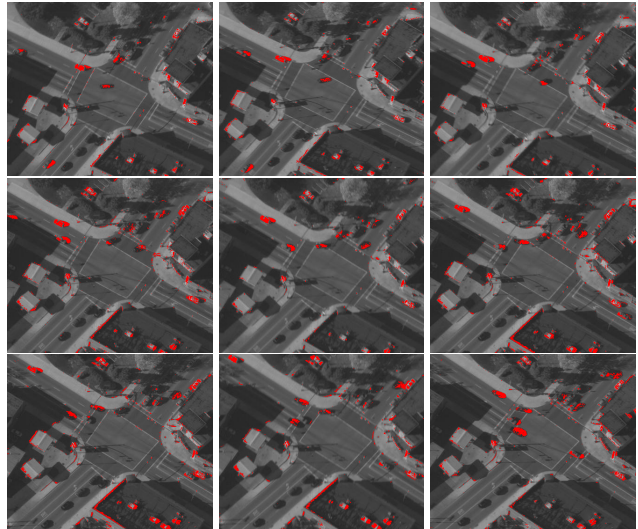


Figure 7: Low likelihood regions colored in red. Regions correspond to moving objects and sharp edges (sequence left to right CLIF Intersection images 19-27)

effortless detection of moving objects and reasoning about scene dimensions in absolute scale and orientation.

An important extension to the work presented here is to perform quantitative model evaluation and comparisons. These evaluations were not carried out due the lack of ground truth data or standard performance metrics for large-scene reconstruction algorithms. Future work should either obtain ground truth data to validate the reconstructions or outline a comprehensive set of criteria in which to quantify reconstructions. An additional avenue to consider is the incorporation of other modalities. The two observation types currently in the model complement each other, but more realistic reconstructions could be achieved if material properties and/or lighting conditions were added. Additional prior beliefs should also be investigated; particularly, smoothness priors on scene primitives might help regularize the optimization and reduce sensitivity to initialization.

Acknowledgments. The authors thank Sue Zheng, Jason Chang and Julian Straub for general and helpful discussions. R.C., O.F., and J.F. were partially supported by the Office of Naval Research Multidisciplinary Research Initiative program, N00014-11-1-0688 and by the Defense Advanced Research Projects Agency, FA8650-11-1-7154. G.R. was partially funded by the MIT-Technion Postdoctoral Fellowship Program.

References

- [1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a day. *ICCV*, 2009.
- [2] J. Alon and S. Sclaroff. Recursive Estimation of Motion and Planar Structure. *CVPR*, 2000.
- [3] C. Baillard and A. Zisserman. Automatic reconstruction of piecewise planar models from multiple views. *CVPR*, 1999.
- [4] A. Bartoli and P. Sturm. Constrained Structure and Motion From Multiple Uncalibrated Views of a Piecewise Planar Scene. *IJCV*, 2003.
- [5] A. Bartoli, P. Sturm, and R. Haraud. Projective structure and motion from two views of a piecewise planar scene. *ICCV*, 2001.
- [6] R. Cabezas, O. Freifeld, G. Rosman, and J. W. Fisher III. Aerial reconstructions via probabilistic data fusion. Supplemental material, *CVPR*, 2014. <http://people.csail.mit.edu/rcabezas/publications/>.
- [7] F. Dellaert, S. M. Seitz, C. E. Thorpe, and S. Thrun. Structure from motion without correspondence. *TMI*, 2000.
- [8] A. R. Dick, P. H. S. Torr, S. J. Ruffle, and R. Cipolla. Combining Single View Recognition and Multiple View Stereo For Architectural Scenes. *ICCV*, 2001.
- [9] C. Ellum. Integration of raw GPS measurements into a bundle adjustment. *ISPRS Congress*, 2004.
- [10] F. Tsai, T. A. Teo, L. C. Chen, S. J. Chen. Construction and visualization of photo-realistic three-dimensional digital city. *Joint Urban Remote Sensing Event*, 2009.
- [11] O. Faugeras and F. Lustman. Motion and Structure from Motion in a Piecewise Planar Environment. *IJPRAI*, 1988.
- [12] A. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. *ECCV*, 1998.
- [13] D. A. Forsyth, S. Ioffe, and J. Haddon. Bayesian structure from motion. *ICCV*, 1999.
- [14] O. Freifeld and M. Black. Lie bodies: a manifold representation of 3D human shape. *ECCV*, 2012.
- [15] K. Fujii and T. Arikawa. Reconstruction of 3D urban model using range image and aerial image. *IGARSS*, 2001.
- [16] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2010.
- [17] D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. *CVPR*, 2010.
- [18] B. Goldluecke and D. Cremers. Superresolution texture maps for multiview reconstruction. *ICCV*, 2009.
- [19] P. Gurram, H. Rhody, J. Kerekes, S. Lach, and E. Saber. 3D Scene Reconstruction through a Fusion of Passive Video and Lidar Imagery. *AIPR Workshop*, 2007.
- [20] H. H. Liao, Y. Lin, G. Medioni. Aerial 3D reconstruction with line-constrained dynamic programming. *ICCV*, 2011.
- [21] K. Hammoudi and F. Dornaika. A Featureless Approach to 3D Polyhedral Building Modeling from Aerial Images. *Sensors*, 2010.
- [22] J. Hu, S. You, and U. Neumann. Integrating LiDAR, Aerial Image and Ground Images for Complete Urban Building Modeling. *3DPVT*, 2006.
- [23] M. Huber, W. Schickler, S. Hinz, and A. Baumgartner. Fusion of LIDAR data and aerial imagery for automatic reconstruction of building surfaces. *GRSS Workshop*, 2003.
- [24] L.C. Chen, T.A. Teo, J.Y. Rau, J.K. Liu, W.C. Hsu. Building reconstruction from LIDAR data and aerial imagery. *IGARSS*, 2005.
- [25] A. Mastin, J. Kepner, and J. Fisher. Automatic registration of LIDAR and optical images of urban scenes. *CVPR*, 2009.
- [26] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 1965.
- [27] Ohio Statewide Imagery Program and Ohio Geographically Referenced Information Program. Ohio LiDAR Data. <http://gis3.oit.ohio.gov/geodata/>.
- [28] M. Pollefeys, D. Nistér, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénus, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3d reconstruction from video. *IJCV*, 2007.
- [29] G. Qian and R. Chellappa. Structure from Motion Using Sequential Monte Carlo Methods. *IJCV*, 2004.
- [30] S. N. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. *ICCV*, 2009.
- [31] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 2007.
- [32] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. *ECCV*, 1996.
- [33] R. Szeliski and P. Torr. Geometrically Constrained Structure from Motion: Points on Planes. *SMILE*, 1998.
- [34] E. Tola, C. Strecha, and P. Fua. Efficient Large Scale Multi-View Stereo for Ultra High Resolution Image Sets. *Machine Vision and Applications*, 2012.
- [35] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 1992.
- [36] US Air Force. Columbus large image format dataset 2007. <https://www.sdms.af.mil/index.php?collection=clif2007>.
- [37] H. Vu, P. Labatut, J. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *PAMI*, 2012.
- [38] S. W. S. Weng, G. Z. G. Zhao, and B. H. B. He. Rapid reconstruction of 3D building models from aerial images and LiDAR data. *ICACTE*, 2010.
- [39] M. Xie, K. Fu, and Y. Wu. Building Recognition and Reconstruction from Aerial Imagery and LIDAR Data. *International Conference on Radar*, 2006.
- [40] Y. Zhang, Z. Zhang, J. Zhang, and J. Wu. 3D Building Modelling with Digital Map, Lidar Data and Video Image Sequences. *The Photogrammetric Record*, 2005.