Robust Probabilistic Occupancy Grid Estimation from Positive and Negative Distance Fields

Xiaoyan Hu
Stevens Institute of Technology
Hoboken, New Jersey, USA
xhu2@stevens.edu

Philippos Mordohai
Stevens Institute of Technology
Hoboken, New Jersey, USA
mordohai@cs.stevens.edu

Abstract

We present an approach for estimating occupancy grids with an emphasis on robotics applications, where collision avoidance and robustness to severe noise are of more importance than high resolution. We build upon probabilistic techniques, typically used in robotics, and techniques based on signed distance fields, typically used in computer vision, to obtain an approach that is robust and also allows probabilistic reasoning on free and occupied space. The uniqueness of our method lies in the use of separate accumulators for positive and negative evidence for the occupancy of each voxel. This enables our representation to capture the uncertainty due to potential conflicts among the measurements instead of allowing contradictory evidence to cancel each other out. We show occupancy grids computed from multi-view stereo inputs on precisely and imprecisely calibrated image sequences. The ground truth that is available with the former dataset allows quantitative evaluation of the performance of our algorithm.

1. Introduction

Probabilistic occupancy grids have arguably been the dominant paradigm for map building in robotics [25, 4, 32] partly because they are a world-based representation that enables the fusion of measurements from mobile sensors, even in the presence of large uncertainties in motion estimation. The space in which the robot operates is divided into voxels, which should be classified as occupied or empty. For ground robots, the grid can be 2D, but we will focus our attention on 3D grids in this paper. While several methods for computing such occupancy grids have been proposed in the literature (see Sec. 2), most of them share one characteristic: range measurements are assumed to follow a normal distribution centered around the true range, possibly contaminated by another distribution representing false positives, missed obstacles or both. These probabilities are fused in voxels, rays or cones depending on the sensor model to produce posterior occupancy probabilities. Final decisions are made by thresholding the posteriors to minimize risk.

The resolution of occupancy grids in robotics is typically low since this, on the one hand, acts as low-pass filtering on the measurements and reduces the effects of noise, and, on the other hand, it is sufficient for tasks such as obstacle avoidance and path planning without putting undue strain on the limited memory and processing resources of robots. Not bound by these constraints, researchers in computer graphics and vision also adopted volumetric representations in order to merge multiple range scans into complete surface models. In one of the most significant papers in this line of research, along with that of Hoppe et al. [12], Curless and Levoy [3] criticized methods based on occupancy probability computations arguing that "a difficulty with this technique is the fact that the best description of the surface lies at the peak or ridge of the probability function" and such peaks and ridges are hard to localize robustly. To overcome this difficulty and extract high-resolution, detailed surfaces from the voxel grid, Curless and Levoy used truncated signed distance functions (TSDF) which are generated from the range scans and aggregated on the voxel grid. Surfaces can be extracted as the zero-crossings of the cumulative signed distance function. This is a more stable problem and its solution can lead to 3D surface models of outstanding quality. See Fig. 1 for a simple 1D illustration and Sec. 2 for more details on related work.

Both representations - occupancy probabilities or signed distance fields - share the following property: two contradictory measurements with the same weight cancel each other out either after they are added or after Bayesian updates of the posterior. This makes a voxel that has received, for example, two votes for being occupied and none for being empty equivalent to a voxel that has received six votes for being occupied and four for being empty. We argue that the representation of these voxels should not be oblivious to the difference in uncertainty between them. The latter

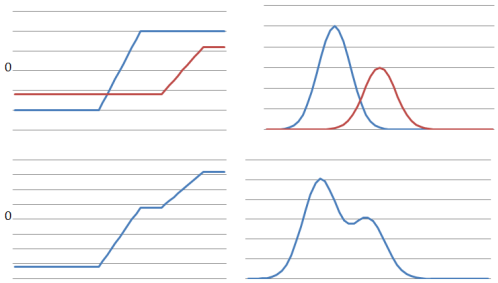

Figure 1. Left: the addition of truncated signed distance functions leads to zero-crossings that are easy to detect. Right: the same is not true for the addition of normal densities, which may result in multiple local maxima. (The x-axis is defined on a ray of the sensor.)

voxel may be in a part of space where the input range maps are particularly noisy, where motion may have occurred or some other source of uncertainty is present. The former voxel should be more straightforward to classify as occupied. We would like to provide a risk-averse autonomous vehicle with information that would allow it to stay away from uncertain regions.

We accomplish this by maintaining two separate accumulators: one for positive distance values and one for negative ones. At the end, the ratio of the accumulated evidence is compared to a threshold determined by a user-specified loss function, or simply the zero-one loss function, to make final decisions about occupancy. *This approach inherits the robustness of the TSDF representations, but allows the specification of prior probabilities for occupancy and free-space as well as generalized loss functions during the decision making process.* To the best of our knowledge, no previously published work combines these properties.

In this paper, we work solely with depth maps computed via multi-view stereo, but other inputs can be used. Viewpoint-based depth estimation can be carried out at higher resolution than volumetric processing due to its quadratic, instead of cubic memory, requirements. Thus, the input depth maps are of higher resolution than the voxel grid and each voxel projects on multiple pixels. We take this explicitly into account, similar to [1, 27], to increase information utilization from the depth maps to the voxel grid. We also use the confidence of each depth estimate [13] as a weight during evidence aggregation.

We show the effectiveness of our method on two diverse datasets. First, we show quantitative results on the accurately calibrated data provided by Strecha et al. [30] and, then, we show robustness on image sequences we collected for which calibration is less accurate. We acknowledge that one could try to improve the structure from motion estimation using techniques such as those proposed by Tylecek and Sara [33] and Furukawa and Ponce [6], but we opt for an alternative route and attempt to suppress the adverse effects of these errors via the robustness of our approach.

2. Related Work

In this section, we review related work on estimating occupancy grids from range maps. Methods that operate on image inputs such as voxel coloring [29] or space carving [21] are considered out of scope in this discussion, as are silhouette-based methods [31, 22, 5]. We distinguish between methods that model surfaces as local maxima of some probability distribution and those that model surfaces as zero-crossings of an appropriate function.

Occupancy grids were introduced in the robotics literature as representations to facilitate navigation using sensors such as sonar [25] or stereo [4]. Due to severe sensor noise, emphasis was placed on robustness and probabilistic approaches able to fuse uncertain evidence. Critical for the estimation is the sensor model, which is used to describe either the probability of observing a measurement conditioned on the state of the environment (forward model) or the probability of the state conditioned on the measurements (inverse model). The state here comprises variables representing the occupancy of each cell of the grid. Inverse models were preferred by early researchers because they allow the processing of each voxel independently of other voxels resulting in simpler and faster algorithms. Elfes and Matthies [4] proposed a Bayesian approach that employs an inverse sensor model and integrates evidence for occupancy. Konolige [19] proposed several improvements to this approach by using better designed sensor models and accounting for specularities and redundant measurements. A forward sensor model was adopted by Thrun [32] who solved the problem taking into account the dependencies between voxels using the EM algorithm. A comparison of these algorithms was performed by Collins et al. [2] and showed that forward sensor models are more accurate but very demanding in computational resources. Around the same time, Pathak et al. [26] presented a formulation using a forward sensor model that does not require EM and is thus more efficient.

An approach designed specifically for stereo depth maps inputs was presented by Andert [1]. It uses an inverse model and, unlike previous work, explicitly takes into account the fact that pixel-voxel correspondences are not one to one. Since depth maps are of higher resolution than the slices of the occupancy grid, each voxel covers multiple pixels. The minimum depth among all pixels covered by a voxel is used to update its occupancy probability in order to be conservative. Pyramids are computed from the depth maps to improve efficiency. Pirker et al. [27] adapted this approach to process inputs from a Kinect and to interpolate between pyramid levels for increased accuracy. We use all pixels, instead, to obtain multiple positive and negative measurement for the voxel.

Occupancy grids are widely used in robotics because they provide robustness to sensor pose estimation errors

and noise in range estimates. The voxels are typically large since the size of the robot determines the scale of the representation. In the computer vision and graphics literature, on the other hand, the typical goal is to obtain very accurate surface models preserving fine details. As mentioned in the previous section, the use of truncated distance fields has proven to be more robust than the estimation of occupancy probabilities.

Hoppe et al. [12] presented a seminal method for inferring an implicit surface representation from an unorganized point cloud, by estimating local tangent planes to the points and from those estimating a volumetric distance function to the surface. In separate papers in 1996, Curless and Levoy [3] as well as Hilton et al. [11] proposed algorithms for merging range maps by combining the signed distance functions induced by them in volumetric grids to generate a cumulative distance function, the zero level-set of which was the surface. Wheeler et al. [35] adapted the method of [3] to increase its robustness to outliers by extracting surfaces only in voxels that are supported by a minimum level of consensus, measured by accumulating the confidence of range estimates incident to each voxel. More recently, Kazhdan et al. [17] formulated implicit surface estimation from unorganized oriented points as a spatial Poisson problem whose solution is an indicator function that labels voxels as interior or exterior to the surface. Several authors [9, 23, 7] use the Poisson Surface Reconstruction software [17] to extract watertight manifold surfaces from partial reconstructions. Zach et al. [37, 36] proposed a global optimization framework based on minimization of total variation and an L_1 data term defined using truncated signed distance fields computed from the input range images.

Techniques for directly estimating occupancy have also been published in the computer vision literature. Koch et al. [18] presented a voting-based approach for depth map fusion under which depth estimates contained in voxel are used as evidence of its occupancy. Surfaces can be extracted by thresholding the accumulated votes. Sato et al. [28] also advocated a volumetric method based on voting. Each depth estimate votes not only for the most likely surface but also for free space between the camera and the surface. The ratio of votes for surface over free-space is thresholded to decide on the occupancy of a voxel. Hernández et al. [10] adopted a Gaussian distribution for the depth on rays, conditioned on the true depth, contaminated by an outlier process. The final surface is extracted as a graph cut separating interior form exterior nodes using evidence of visibility to compute the unary potentials. Under certain conditions, they argue that this computation of visibility is equivalent to the use of signed distance functions as in [3]. Recently, impressive results have been shown by volumetric methods [34, 16] based on 3D Delaunay triangulation to tessellate the volume into tetrahedra followed by interior/exterior segmentation computed via a graph cut. Reconstructed sparse features are clustered in 3D and the cluster centroid are the vertices of the Delaunay tetrahedra. The cost for labeling a tetrahedron as occupied is determined by evidence of visibility accumulated from the cameras in which the vertices are visible.

Our method adopts an inverse model and ignores dependencies between voxels. Signed distance functions are chosen for their robustness, but positive and negative distances are accumulated separately and used as evidence for occupancy and emptiness. This allows a probabilistic interpretation and the use of decision rules that take risk into account, as shown in (7). This has not been pursued in the computer vision literature where the aim is minimization of errors regardless of their type. The method of Hernández et al. [10] is, in principle, capable of doing that with minor modifications to the unary terms of the objective function, but it has not been shown. It should also be noted that all the methods mentioned above, except for [36], use a singe accumulator of probability or distance per voxel. Ours is the only method that uses separate accumulators for the two types of evidence, while Zach [36] modified the data term of [37] by dividing the interval $[-1, 1]$ into bins and histogramming the distance field values at each voxel. The key difference between our method and that of Zach is that our focus is on robustness, at low resolution, as opposed to obtaining high geometric accuracy on cleaner inputs.

3. Input Depth and Confidence Maps

In this section, we briefly describe how the input depth and confidence maps are estimated from the images. Camera poses are assumed to be externally provided.

Raw depth maps are computed from a set of calibrated images using plane-sweeping stereo implemented on the GPU similar to [8] with normalized cross-correlation (NCC) as the similarity function. NCC is computed in small windows defined in the reference image and mapped to several target images via a homography through the hypothesized plane. For all the experiments in this paper, we use four sets of planes: one horizontal and three vertical. One of the vertical sets is fronto-parallel and the other two are rotated 45^o clockwise and counter-clockwise around the gravity direction. Horizontal planes are very effective for reconstructing surfaces such as the ground, which are heavily distorted when projected via homographies defined on vertical planes. To compute the photoconsistency of a candidate depth along a sweeping orientation, all pairwise NCC scores are averaged. The depth for every pixel is selected as the one with maximum score among all sweeping directions independently of its neighbors.

Concurrently with the depth maps, *confidence maps* are also computed. Based on a recent comparison of confidence measures [15], we selected a form of confidence introduced

by Merrell et al. [24] to obtain a probability mass function for depth given NCC scores, assuming a uniform depth prior. The confidence of a depth candidate d_i is defined as follows:

$$C(d_i) = \frac{e^{-\frac{NCC(d_0)-NCC(d_i)}{2\sigma^2}}}{\sum_j e^{-\frac{NCC(d_0)-NCC(d_j)}{2\sigma^2}}}, \quad (1)$$

where d_0 is the depth with the maximum NCC score for that sweeping direction. Taking into account $NCC(d_0)$ is crucial for allowing confidence values to transfer across depth maps. Here, we only consider the candidate with the highest NCC value, d_0.

The raw depth maps are subsequently fused to generate *fused depth maps* which are the inputs to the occupancy grid computation. We follow our least commitment depth map fusion approach [14], published concurrently with this paper. Fusion begins by rendering the input depth and confidence maps onto the reference view. Support for each depth candidate is accumulated from depth candidates that are likely to have been generated by the same 3D point. Consistent depth candidates are then merged by taking their average weighted by confidence. These merged depth estimates are penalized according to violations of visibility constraints, namely occlusions and free space violations, which indicate inconsistencies among depth maps. Both support and penalties are weighed according to the confidence of the participating depth estimates. Unlike the greedy approach of [24], we adopt a least commitment strategy and fully evaluate a number of depth candidates for each pixel before making hard decisions. *Fused confidence maps* are also generated during this process by aggregating the support and penalties of the fused depth values. See [14] for details. It should be noted that any depth maps can be used as inputs. If confidence maps are not available, all pixels can be treated equally in the subsequent stages resulting in some loss of accuracy.

4. Occupancy and Free Space Estimation

In this section, we describe how the accumulators of positive and negative evidence are populated from the depth maps and how decisions are made for the occupancy of each voxel. We adopt an inverse sensor model and consider the occupancy of each voxel as independent of the occupancy of all other voxels. The grid is defined so that it covers the area of interest. Including all visible surfaces, however, is not necessary. Depths beyond the far end of the grid provide evidence in support of the free space hypothesis for all affected voxels.

Given n input depth maps, each with an associated confidence map, every voxel of the grid is projected to each depth map. We handle the notorious voxel-pixel correspondence problem similar to [1, 27] by determining the area a voxel covers on the depth map. Specifically, we project the center of the voxel onto the depth map and denote its pixel coordinates by (u, v). If the length of the voxel's edge is S and its depth with respect to the camera center is Z, then the projection of the voxel is approximately $s = Sf/Z$ pixels wide. (This would be exact if the voxels were spherical.) Then, we aggregate data from all pixels in a $s \times s$ window centered around (u, v). Unlike [1, 27], we do not use image pyramids during this operation. This process is repeated for all depth maps in which the voxel is visible.

A pixel contributes to a voxel positive or negative evidence for its occupancy according to the truncated signed distance function model. Our implementation follows that of Zach et al. [37]. Let D_i denote the distance from the camera center to voxel m_i and $d(u, v)$ denote the depth of pixel (u, v), then the signed distance function is:

$$f_i = \frac{D_i - d(u, v)}{\delta}, \quad (2)$$

where δ defines the width of the "near surface" region. *Positive distances correspond to occupied voxels and negative distances to free space.* The signed distance function is truncated to be within $[-1\ 1]$, as shown below. If the confidence of the depth estimate is $C(u, v)$, then the accumulators are updated as follows. We use p_i and n_i for the positive and negative accumulator and p_i^- and p_i^+ for the values of p_i before and after the update respectively. Only one accumulator is affected by each measurement.

$$n_i^+ = n_i^- - C(u, v) \quad \text{if } f_i < -1 \quad (3)$$

$$n_i^+ = n_i^- - C(u, v) f_i \quad \text{if } -1 \le f_i < 0 \quad (4)$$

$$p_i^+ = p_i^- + C(u, v) f_i \quad \text{if } 0 < f_i \le 1 \quad (5)$$

$$p_i^+ = p_i^- + C(u, v) \quad \text{if } 1 < f_i \le \eta\delta \quad (6)$$

The influence of each depth estimate stops at some distance behind the estimated surface determined by η, which can be viewed as the minimum thickness of objects allowed. While a measurement with high confidence and $f_i = 0$ does not affect either p_i or n_i, it affects the voxels in front and behind the current voxel and prevents large errors.

Optionally, for noisy data, we perform a few iterations of diffusion to impose smoothness to the voxel space. Diffusion is applied separately in the positive and negative accumulators with larger weights for connections between vertically neighboring voxels to favor vertical surfaces.

While in 3D reconstruction for visualization purposes over and under-estimation of depth have the same significance, this is not true for navigation and other applications. If depth is over-estimated, the vehicle may collide with an obstacle and, if it is under-estimated, a traversable path may be rejected. We name these two types of errors from the perspective of occupancy and use e_{fp} to represent

false positives (obstacles detected where they do not exist) and e_{md} for missed detections (true obstacles labeled as free space). We cast our decision making process as one of risk minimization and assign potentially different costs, λ_{fp} and λ_{md}, to the two types of errors. This leads to the following decision rule, where z denotes all the measurements:

$$\begin{aligned} \text{if } \frac{P(z|OCC)}{P(z|EMP)} > \frac{\lambda_{md} P(EMP)}{\lambda_{fp} P(OCC)} = \theta \qquad & m_i = OCC \\ \text{else} \qquad & m_i = EMP \end{aligned} \tag{7}$$

θ is the threshold of the likelihood ratio test and depends on the priors and the relative costs for each type of error. In the remainder, we will use θ as a single parameter encompassing all these factors, but if priors were available, they could have been used explicitly. So far, we have not estimated the likelihoods on the left-hand side of (7), but we are able to approximate their ratio by the ratio of the evidence we have accumulated, which leads to the final decision rule:

$$\text{if } \frac{p_i}{n_i} > \theta,\ m_i = OCC,\ \text{else } m_i = EMP \tag{8}$$

This formulation is similar to the TSDF paradigm where signed distances are aggregated in one accumulator, but it captures the presence of conflicting information better. Revisiting the example of Sec. 1, a voxel that received six positive and four negative votes has a likelihood ratio of 1.5 which is equivalent to a 60% probability of occupancy, but not equivalent to the 100% probability of a voxel that received only two positive votes. A mobile robot should avoid the former voxel to reduce the risk of collisions.

5. Experiments on Controlled Data

In this section, we present an evaluation of our algorithm on two real outdoor datasets with ground truth [30]. We used the *fountain-P11* and *Herz-Jesu-P8* datasets which contain 11 and 8 images respectively, which we downsampled to 1536×1024. All experiments were performed with fixed parameters: three adjacent images were used for the computation of each depth map, the NCC window was 7×7, σ in (1) was set to 0.2 and the sweeping planes where distributed in space so that the disparity step between them was no more than 0.2, with disparity defined between the reference view and the farthest target view. Finally, median filtering in 13×13 windows was applied to fill the holes in the fused depth maps. See [14] for details.

Starting from the fused depth and confidence maps as inputs (Fig. 2), we computed occupancy grids according to the previous section. The grids were configured according to the provided bounding box information. The grid dimensions for *fountain-P11* were $342 \times 228 \times 200$ with 4.25cm

Figure 2. One of the input images for each of the *fountain-P11* and *Herz-Jesu-P8* datasets. Input fused depth and confidence maps for the former.

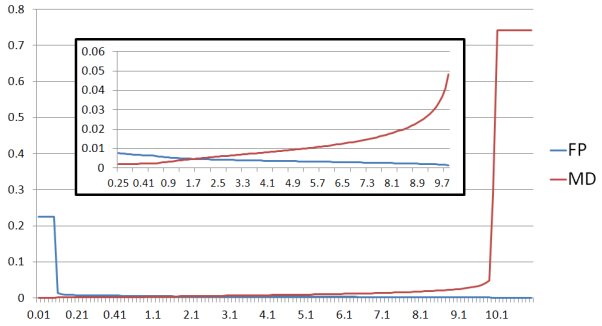

Figure 3. ROC curves for false positives (FP) and missed detections (MD) as a function of θ for the *Herz-Jesu-P8* dataset. At each end, all voxels have been assigned the same label producing the maximum possible error of the relevant type. The abrupt transitions occur when θ causes the label assigned to voxels that received only positive or only negative evidence to change. The inset is a zoom in on the middle part of the curves. The equal error rate is 0.47% for $\theta \approx 1.8$.

voxels, while the grid for *Herz-Jesu-P8* was $328 \times 218 \times 200$ and the voxel size was 7.5cm. We set δ equal to one half of the depth range of the grid and η to a large value. (Smaller values could have been more effective for objects with thin parts, but no such data with ground truth was available to us.) No diffusion was performed to isolate the core of the algorithm for the evaluation and also because independent computation for each voxel generated very accurate results.

In the first experiment, we verified that using fused depth maps as inputs is superior to using raw depth maps. With $\theta = 1$, the occupancy grid computed for the *fountain-P11* from raw depths had e_{fp} (false obstacles) of 0.596% and e_{md} (missed obstacles) of 0.020%. The same figures using fused depth maps for $\theta = 1$ were 0.546% and 0.016% respectively. Only fused depth maps were used in all subsequent experiments.

Varying the threshold θ in (8), we obtained ROC curves for the two types of error. The ROC curves for *Herz-Jesu-P8* are shown in Fig. 3. The curves for *fountain-P11* look similar and have been omitted. As expected, for very small

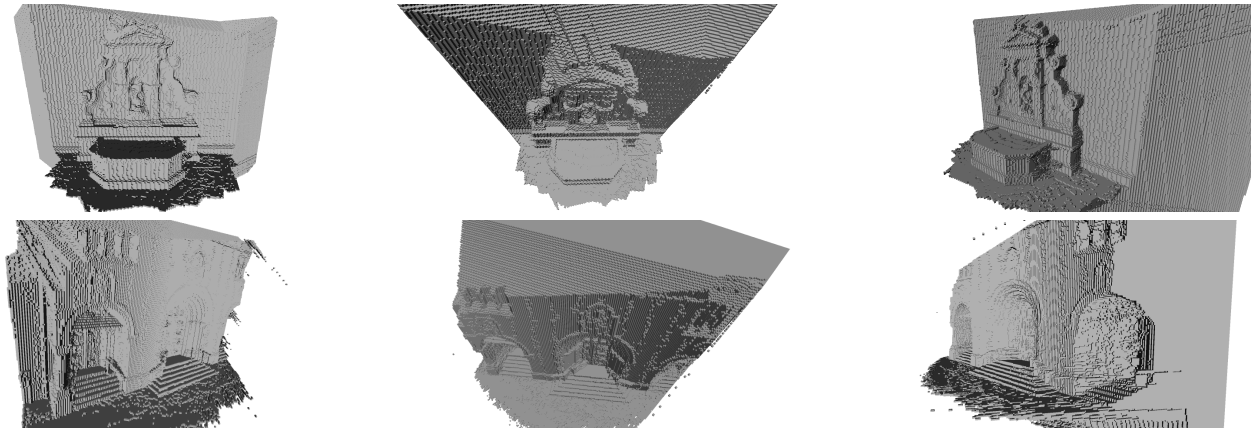

Figure 4. Screenshots of occupancy grids for both datasets.

or very large values of the threshold all voxels are assigned the same label. The important observation from this figure is that for a broad range of θ values, both types of errors are very small. In fact they are in the order of 0.5% which corresponds to one voxel per ray on average. Considering that a large error can cause several voxels to be wrong, this is an encouraging result. Screenshots of occupancy grids for both datasets can be seen in Fig. 4.

Besides the stability mentioned above, our algorithm also allows the user to direct the solution toward over and under-estimation of occupancy. Figure 5 shows difference grids in which voxels that agree with the ground truth are blank, voxels that are labeled as occupied, but are actually empty, are colored red and voxels that are labeled as empty, but are occupied in the ground truth, are colored green. As shown in the figure, results can be adjusted to be more or less conservative without producing gross outliers that would be dangerous for a mobile robot.

6. Experiments on Imprecise Data

The previous section contains results that show good performance on imagery acquired very carefully and for which calibration was aided by fiducial markers in the scene. Obtaining similar image quality and calibration accuracy is hard for a number of reasons, especially outdoors. In this section, we show how our method can still be effective under less than ideal conditions. We collected video sequences using two video cameras rigidly mounted on a cart, synchronized and pre-calibrated in the lab so that their intrinsic parameters and the relative transformation from one camera to the other were known with high accuracy. Then, we used a slight modification of the binocular stereo visual SLAM package of the Robotics Operating System (ROS)[1] [20] to estimate their motion.

While the estimated trajectory is reasonably accurate and reprojection errors are small in nearby frames, it is not perfect. Instead of trying to improve the quality of camera motion estimation, in this paper we apply our occupancy grid method to reduce the effects of noise and produce low resolution 3D models of acceptable quality for navigation. We kept the same parameters as in Section 5 with the following differences. Due to the different spacing between camera poses and the inability to use wide baselines due to drift in pose estimation, we compute raw depth maps using seven images and then fuse seven raw depth maps to obtain each fused one. We divide the space using multiple voxel grids with 10.5cm voxels and apply the algorithm of Section 4 using five fused depth and confidence maps as input for each grid. Finally, we apply 50 iterations of diffusion in the voxel grid weighing the central voxel by 8, its two vertical neighbors by 2 and its four horizontal neighbors by 0.5. (These weights were chosen arbitrarily and no tuning was attempted.) We use $\theta = 0.4$ to facilitate hole filling.

Due to the much lower quality of our images and calibration data, there are significant errors in the inputs to our algorithm, as can be seen in Fig. 6. Accumulating evidence independently for each voxel results in a 3D model with large holes between the first and second floor, in windows and other surfaces for which depth was estimated incorrectly. At the same time, large protrusions further diminish the visual quality and usefulness of the model. These errors are systematic, in the sense that they appear in all depth maps and thus cannot be corrected by fusion. The confidence, however, of the noisy parts of the depth maps is much lower than that of well-imaged, textured parts that produce accurate matches. After diffusion, reliable, occupied and empty, voxels dominate unreliable ones resulting in large reductions of both holes and protrusions. See Fig. 7 for images of the input point cloud and of our final results. In the absence of ground truth, we overlayed the input sparse point cloud consisting of the most confident reconstructed points with the final result after diffusion and

[1] http://www.ros.org/wiki/vslam

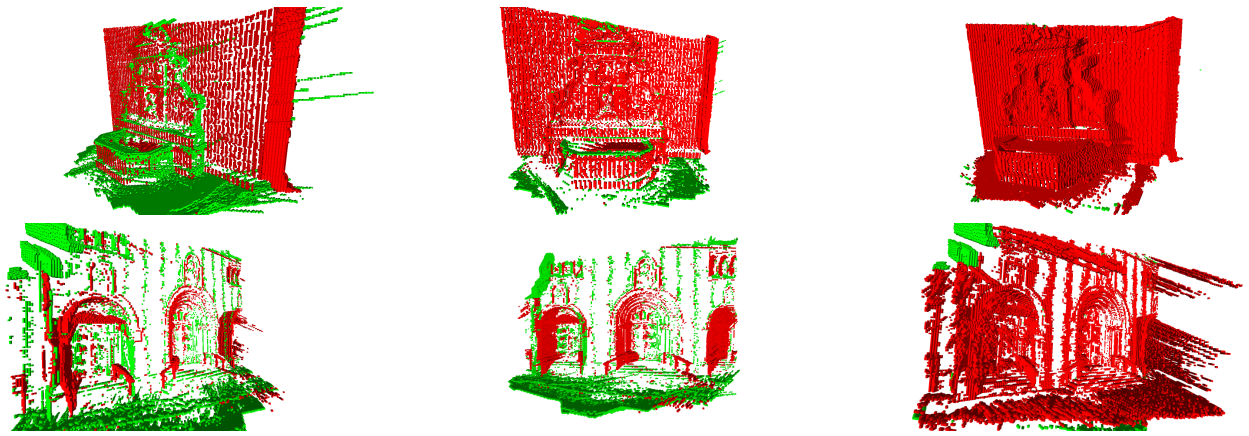

Figure 5. Screenshots of difference grids for both datasets. False positives of occupancy are colored red, missed detections are colored green and voxels in agreement with the ground truth are left blank. Each screenshot is from a different model using different values for θ, resulting in bias for one type of error over the other.

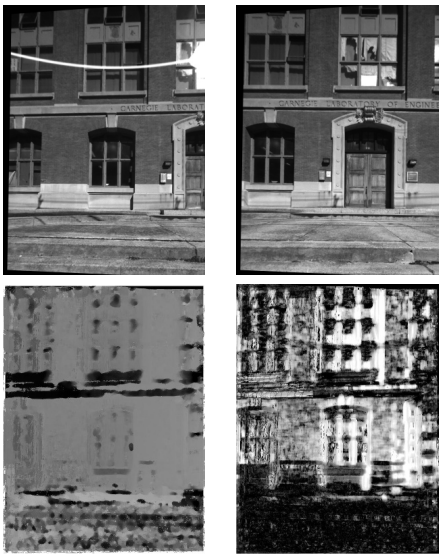

Figure 6. Two of the input images from our sequence and the fused depth and confidence maps for the image on the left. Notice the CCD blooming in the top left image, which appears curved after radial undistortion. Depth maps contain errors due to lack of texture, reflections on the windows and miscalibration. Fortunately, most of the depth errors are associated with low confidence values.

observed that the surfaces do not "move" further or closer to the viewer. Since the geometry of this building is less characteristic than *fountain-P11* and *Herz-Jesu-P8*, we generated colored point clouds by taking boundary occupied voxels and coloring them using the central image of the relevant grid.

7. Conclusions

We have presented a novel approach for volumetric occupancy estimation that combines the advantages of probabilistic methods with those based on signed distance functions. Using this approach, we have shown very accurate, low-resolution results on real data with ground truth, as well as acceptable performance on data captured "in the wild".

Figure 7. Screenshots of results on our sequence. Top point cloud reconstructed from input fused depth maps. Bottom: point cloud of surface voxels after diffusion.

The initial success of this line of research opens several directions for future work. On the one hand, we plan to improve the core of the approach by investigating the benefits of a forward sensor model and of more sophisticated diffusion guided by the input images. On the other hand, we plan to develop a real-time version of our software by leveraging octrees, depth map pyramids, parallel processing on the GPU and by determining the degree to which depth map resolution can be reduced without significant adverse effects on accuracy.

Acknowledgements

This research has been supported in part by the Domestic Nuclear Detection Office of the U.S. Department of Homeland Security, by the National Science Foundation (grant CNS-0855218) and by Google Inc. via a Google Research Award. We are grateful to Konstantinos Batsos for his help in data collection and structure from motion estimation.

References

[1] F. Andert. Drawing stereo disparity images into occupancy grids: Measurement model and fast implementation. In *IROS*, pages 5191–5197, 2009.

[2] T. Collins, J. Collins, and C. Ryan. Occupancy grid mapping: An empirical evaluation. In *Mediterranean Conference on Control Automation*, 2007.

[3] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *ACM Trans. on Graphics*, 30:303–312, 1996.

[4] A. E. Elfes and L. Matthies. Sensor integration for robot navigation: combining sonar and range date in a grid-based representation. In *IEEE Conference on Decision and Control*, volume 3, pages 1802–1807, 1987.

[5] J. S. Franco and E. Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. In *ICCV*, pages II: 1747–1753, 2005.

[6] Y. Furukawa and J. Ponce. Accurate camera calibration from multi-view stereo and bundle adjustment. *IJCV*, 84(3):257–268, 2009.

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010.

[8] D. Gallup, J. M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *CVPR*, 2007.

[9] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007.

[10] C. Hernández Esteban, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007.

[11] A. Hilton, A. Stoddart, J. Illingworth, and T. Windeatt. Reliable surface reconstruction from multiple range images. In *ECCV*, pages I:117–126, 1996.

[12] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. *ACM SIGGRAPH*, 26(2):71–78, 1992.

[13] X. Hu and P. Mordohai. Evaluation of stereo confidence indoors and outdoors. In *CVPR*, 2010.

[14] X. Hu and P. Mordohai. Least commitment, viewpoint-based, multi-view stereo. In *3DIMPVT*, 2012.

[15] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, 2012.

[16] M. Jancosek and T. Pajdla. Robust, accurate and weakly-supported-surfaces preserving multi-view reconstruction. In *CVPR*, 2011.

[17] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, pages 61–70, 2006.

[18] R. Koch, M. Pollefeys, and L. Van Gool. Robust calibration and 3d geometric modeling from large collections of uncalibrated images. In *DAGM*, 1999.

[19] K. Konolige. Improved occupancy grids for map buildingf. *Autonomous Robots*, 4(4):351–367, 1997.

[20] K. Konolige. Sparse sparse bundle adjustment. In *BMVC*, 2010.

[21] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000.

[22] A. Laurentini. The visual hull concept for silhouette-based image understanding. *PAMI*, 16(2):150–162, 1994.

[23] Y. Liu, X. Cao, Q. Dai, and W. Xu. Continuous depth estimation for multi-view stereo. In *CVPR*, pages 2121 –2128, 2009.

[24] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. M. Frahm, R. Yang, D. Nistér, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *ICCV*, 2007.

[25] H. Moravec and A. E. Elfes. High resolution maps from wide angle sonar. In *ICRA*, pages 116 – 121, 1985.

[26] K. Pathak, A. Birk, J. Poppinga, and S. Schwertfeger. 3d forward sensor modeling and application to occupancy grid based sensor fusion. In *IROS*, pages 2059–2064, 2007.

[27] K. Pirker, M. Ruther, H. Bischof, and G. Schweighofer. Fast and accurate environment modeling using three-dimensional occupancy grids. In *IEEE Workshop on Consumer Depth Camera for Computer Vision*, 2011.

[28] T. Sato, M. Kanbara, N. Yokoya, and H. Takemura. Dense 3-D reconstruction of an outdoor scene by hundreds-baseline stereo using a hand-held video camera. *IJCV*, 47(1-3):119–129, 2002.

[29] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *IJCV*, 35(2):151–173, 1999.

[30] C. Strecha, W. von Hansen, L. J. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

[31] R. Szeliski. Rapid octree construction from image sequences. *CVGIP*, 58(1):23–32, 1993.

[32] S. Thrun. Learning occupancy grid maps with forward sensor models. *Autonomous Robots*, 15(2):111–127, 2003.

[33] R. Tylecek and R.Sara. Depth map fusion with camera position refinement. In *Computer Vision Winter Workshop*, pages 59–66, 2009.

[34] H. Vu, P. Labatut, J. P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multi-view stereo. *PAMI*, 2011.

[35] M. D. Wheeler, Y. Sato, and K. Ikeuchi. Consensus surfaces for modeling 3D objects from multiple range images. In *ICCV*, pages 917–924, 1998.

[36] C. Zach. Fast and high quality fusion of depth maps. In *3DPVT*, 2008.

[37] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *ICCV*, 2007.