A PROBABILISTIC FRAMEWORK FOR GEOMETRY RECONSTRUCTION USING PRIOR INFORMATION

Wende Zhang[*†] *and Tsuhan Chen*[*]

[*] Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
{wendez, tsuhan}@andrew.cmu.edu

[†] General Motors Corporation
30500 Mound Road, Warren, MI 48090
wende.zhang@gm.com

ABSTRACT

In this paper, we propose a probabilistic framework for reconstructing scene geometry utilizing prior knowledge of a class of scenes, for example, scenes captured by a camera mounted on a vehicle driving through city streets. In this framework, we assume the video camera is calibrated, i.e., the intrinsic and extrinsic parameters are known all the time. While we assume a single camera moving during capturing, the framework can be generalized to multiple cameras as well. Traditional approaches try to match the points, lines or patches in multiple images to reconstruct scene geometry. The proposed framework also takes advantage of each patch's appearance and location to infer its orientation using prior information based on statistical learning from training data. The prior hence enhances the geometry reconstruction performance. We show that prior-based 3D reconstruction outperforms traditional 3D reconstruction with both synthetic data and real data, especially in the textureless areas.

Index Terms— Visual Learning, Geometry, and Stereo

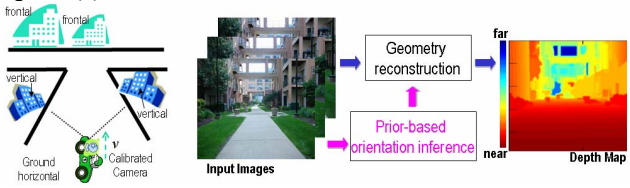
1. INTRODUCTION

Scene reconstruction and rendering have been a popular research topic for decades [1]. Given a set of images captured by one or more cameras, the goal of scene reconstruction and rendering is to reproduce a realistic image of the scene at an arbitrary viewpoint.

Using a multi-camera system, Collins [2] proposed an efficient multi-image matching technique using plane-sweeping for geometry reconstruction. Recently, Akbarzadeh *et al.* [3] extended the plane-sweeping algorithm by sweeping planes in multiple directions for urban geometry reconstruction. Zitnick *et al.* [4] used the modified plane-sweeping algorithm to estimate the current scene's geometry with a smoothness constraint between patches and a spatial consistency constraint between images. Other approaches for geometry estimation include voxel coloring [5] and stereo [6] methods, etc.

Most existing reconstruction approaches match points, lines or patches among multiple images for scene reconstruction. Considering that humans can easily understand the geometry structure of the scene from a single image based on prior knowledge, Hoiem *et al.* [7] proposed prior-based geometry estimation for outdoor scenes using statistical learning. They could reconstruct a coarse 3D model from a single image by classifying each patch into ground, vertical or sky. Saxena *et al.* [8] applied supervised learning to predict the depth map of an outdoor scene also from a single image. Their depth-map estimation model used a Markov Random Field that contained multi-scale local and global image features, and modeled both the depth at each individual point and the relation between depths at neighboring points.

In this paper, we reconstruct scenes from multiple images captured by a single calibrated camera mounted on a moving vehicle as illustrated in Figure 1(a). We assume that the camera is calibrated based on the vehicle's GPS sensor, speed sensor, and gyro/yaw-rate sensor. We represent the scene by small patches with different orientations: horizontal (e.g., ground), vertical (e.g., building facets towards the street and parallel to the camera motion), and frontal (e.g., building facets towards the street and perpendicular to the camera motion). Our prior-based geometry reconstruction algorithm extends Hoiem's approach to reconstruct dense depth maps from a moving calibrated camera as shown in Figure 1(b).



(a) Capturing illustration (b) Prior-based geometry reconstruction

Figure 1. Prior-based geometry capturing and reconstruction. The depth is represented by a color map.

The paper is organized as follows. In the next section, we describe the prior-based geometry reconstruction. In Section 3, we show experimental results and compare the geometry reconstruction quality between the approaches with and without prior information. Conclusions are given in Section 4.

2. PRIOR-BASED GEOMETRY RECONSTRUCTION

In this section, we describe the prior-based geometry reconstruction using a calibrated moving camera. We first provide an overview of the prior learning method and the prior-based geometry reconstruction method, and then introduce each component in detail.

2.1. Overview

As shown in Figure 2, for prior learning, we first segment training images into patches using the efficient graph-based image segmentation technique in [9]. We then train the orientation estimator based on the labeled patches.

For the prior-based geometry reconstruction, input images are first segmented into patches S_j [9]. We then infer each patch's orientation distribution $P_j(o)$ using the orientation estimator. We calculate the color consistency of every patch among multiple

images at the assumed depth d with a given orientation o to estimate the conditional probability $P_j(d \mid o)$. The initial likelihood of patch's geometry $P_j^0(d,o)$ is approximated by the product of the prior probability $P_j(o)$ and the conditional probability $P_j(d \mid o)$. A coarse patch-based smoothing algorithm is then applied to refine the initial geometry likelihood $P_j^0(d,o)$ between its neighboring patches and between its corresponding regions at different viewpoints iteratively. The maximum likelihood estimates of patch's depth $\hat{d}$ and orientation $\hat{o}$, based on the resulting $P_j^t(d,o)$, determine the initial depth map $d^0(x)$ and the orientation map at each pixel position x. The initial depth map $d^0(x)$ is further smoothed iteratively per pixel between images to create the refined depth map $d^t(x)$. Next we will explain each of these steps in more detail.

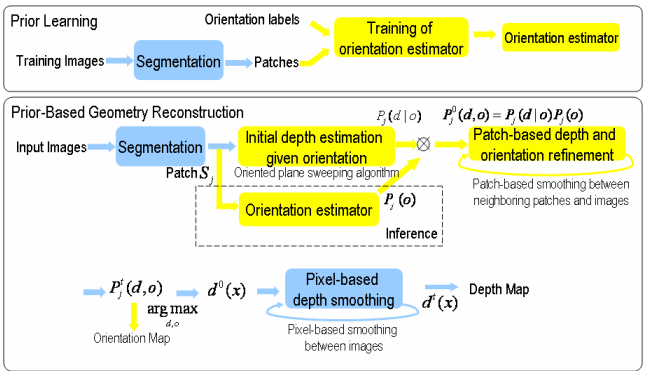


Figure 2. Prior learning and prior-based geometry reconstruction

2.2. Prior-Based Orientation Estimation

Prior-based orientation estimation contains two stages: learning and inference. Similar to human vision system, texture, color and location features are extracted from each patch. The texture feature is the 15 mean values of the absolute responses of the Leung-Malik filter bank [10]. The color feature is the 6 mean values of RGB and HSV. And the location feature is the 2D mean location in the image coordinates.

In the prior learning stage, we first extract the features from the training patches, and then train the orientation estimator using Support Vector Machines (SVM) probability estimation [11] based on the labeled (frontal, vertical, or horizontal) patch features. Compared to [7], we apply a weaker statistic learning approach only using patch's features without further grouping the patches.

In the inference stage, we calculate the prior distribution of patch's orientation. We first extract patch S_j's features, and then determine its orientation distribution $P_j(o)$ using the orientation estimator. The SVM-based estimator provides the probabilities of all possible orientations.

2.3. Initial Geometry Estimation

The initial distribution of the patch's geometry $P_j^0(d,o)$ is evaluated by the product of the orientation probability $P_j(o)$ and the conditional probability $P_j(d \mid o)$ of the patch's depth d given the orientation o.

$$P_j^0(d,o) = P_j(o)P_j(d \mid o) \tag{1}$$

The conditional probability $P_j(d \mid o)$ is determined based on color consistency between images using the oriented plane-sweeping algorithm with the given orientation [3] as shown in Figure 3. Patch S_j's each depth with any given orientation is evaluated by its color consistency $e_{diff}(S_j)$ between multiple images at the current viewpoint using the following robust function.

$$e_{diff}(S_j) = \frac{1}{\text{num}_{S_j}} \sum_k \sum_{pixel \in S_j} \frac{\gamma_k^2}{\gamma_k^2 + th^2}, \tag{2}$$

where $\gamma_k = |r_{cur} - r_{cor,k}| + |g_{cur} - g_{cor,k}| + |b_{cur} - b_{cor,k}|$ is the RGB color difference, parameter th is a constant, and num_{S_j} is the number of the pixels in S_j.

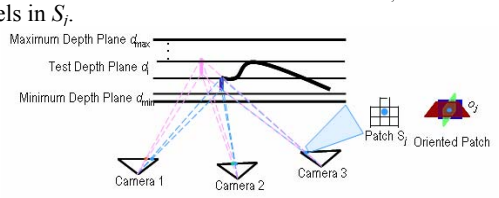


Figure 3. Oriented plane-sweeping algorithm

The conditional probability $P_j(d \mid o)$ is determined by the color consistency measures $e_{diff}(S_j)$:

$$P_j(d \mid o) = \frac{g(d,o)}{\sum_{d'} g(d',o)}, \text{where } g(d,o) = 1 - e_{diff}(S_j). \tag{3}$$

2.4. Patch-Based Smoothing

We refine patch's initial distribution $P_j^t(d,o)$ between its neighboring patches and between its corresponding regions at multiple viewpoints iteratively, which is similar to [4], with the extension of smoothing for additional orientation estimation.

The likelihood distribution of the patch's geometry $P_j^t(d,o)$ is updated iteratively with two constraints.

$$P_j^{t+1}(d,o) = \frac{n_j(d,o)\prod_{k \in N} c_{j,k}(d,o)}{\sum_{d',o'} n_j(d',o')\prod_{k \in N} c_{j,k}(d',o')} \tag{4}$$

where $n_j(d,o)$ enforces patch's smoothness constraint between the neighboring patches, and $c_{j,k}(d,o)$ enforces patch's consistency constraint in each projected region in multiple images.

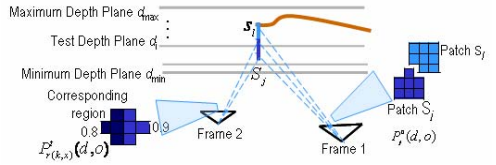


Figure 4. Patch-based smoothing

Let s_l denote one of patch S_j's neighboring patches as shown in Figure 4. The geometry smoothness coefficient $n_j(d,o)$ enforces that the neighboring patches (S_j and s_l in Frame 1) with similar colors should have similar depths ($d \approx \hat{d}_l$) and the same orientations ($o = \hat{o}_l$). $\hat{d}_l$ and $\hat{o}_l$ are the maximum likelihood estimates of its depth and orientation based on $P_l^t(d,o)$.

We assume that if patches S_j and s_l have the same orientation, the depth d of patch S_j is modeled by a contaminated Gaussian distribution with the mean $\hat{d}_l$ and variance σ_l^2 as follows.

$$n_j(d,o) = \begin{cases} \prod_{s_l} N(d;\hat{d}_l,\sigma_l^2) + \varepsilon & o = \hat{o}_l \\ \varepsilon & o \neq \hat{o}_l \end{cases} \tag{5}$$

where $N(x;mean,\sigma^2)$ is Gaussian distribution, and ε and c are small constants. We estimate the variance σ_l^2 using 1) the color similarity of the patches $\Delta_{j,l}$, which measures the color difference between patches S_j and s_l, 2) the neighboring measure $b_{j,l}$, which is the percentage of the patch S_j's border between patches S_j and s_l, 3) and the geometry maximum likelihood for patch s_l: $P_l(\hat{d}_l,\hat{o}_l)$, which represents the accuracy of the maximum likelihood estimates for patch s_l's geometry.

σ_l^2 is defined to be: $\sigma_l^2 = v/[P_l(\hat{d}_l,\hat{o}_l)^2 b_{j,l} N(\Delta_{j,l};0,\sigma_\Delta^2)]$ (6)
where v and σ_Δ^2 are constants. Therefore, if patch S_j and its neighboring patch s_l have similar colors, and patch S_j's depth and orientation are consistent with its neighbor's depth and orientation maximum likelihood estimates, we expect $n_j(d,o)$ to be large.

The spatial consistency coefficient $c_{j,k}(d,o)$ ensures that the patch S_j's depth and orientation estimates are consistent with the depth and orientation estimates at the viewpoint k. We compute $c_{j,k}(d,o)$ based on spatial consistency without occlusion, visibility, and patch S_j's initial geometry likelihood $P_j^0(d,o)$.

1. Spatial consistency without occlusion. We first project patch S_j with the depth d and orientation o onto a neighboring image. We then calculate patch S_j's projecting distribution $b_{j,k}^t(d,o)$ based on the geometry distribution at the projected viewpoint k to estimate the spatial consistency without occlusion.

$$b_{j,k}^t(d,o) = \frac{1}{num_{s_j}} \sum_{x \in S_j} P_{r(k,x)}^t(d,o) \quad (7)$$

where $r(k,x)$ is the patch index at the viewpoint k, on which the corresponding pixel of the pixel position x on patch S_j is. And num_{S_j} is the number of the pixels on patch S_j. If the projected region's depth and orientation maximum likelihood estimates are consistent with patch S_j's estimates, we expect $b_{j,k}^t(d,o)$ to be large when patch S_j is visible in Frame 2 as shown in Figure 4.

2. Visibility. Due to the possible occlusions, a patch might not have the corresponding pixels at another viewpoint. We estimate the overall visibility likelihood $v_{j,k}$, that the patch is visible.

$$v_{j,k} = \min(1.0, \sum_{d',o'} b_{j,k}^t(d',o')) \quad (8)$$

If the patch S_j is visible at the viewpoint k (Frame 2) as shown in Figure 4, we can find its corresponding region when we search the space of depth d and orientation o. The ground-truth solution and its neighboring solutions offer large $b_{j,k}^t(d',o')$ values. Otherwise, we can not find its corresponding region when we search the space of depth d and orientation o. No solution provides large $b_{j,k}^t(d',o')$ value. Therefore, we use $v_{j,k}$ as a robust and computational-efficient measure of patch's visibility.

We also estimate the specific visible likelihood $vc_{j,k}(d,o)$ that patch S_j is visible at the viewpoint k, given the depth d and orientation o. $vc_{j,k}(d,o) = \sum_{x \in S_j} P_{r(k,x)}^t(d,o) h(\hat{d}_{r(k,x)} - d) / num_{S_j}$, where $h(x)$ is the Heaviside step function and $\hat{d}_{r(k,x)}$ is the maximum likelihood depth estimate of patch $s_{r(k,x)}$.

This suggests that if patch S_j is visible at the viewpoint k, its estimated depth should not be under the surface of the estimated depth map at the viewpoint k.

Now, we combine the visible and occluded cases. If the patch is visible, $c_{j,k}(d,o)$ is calculated from the visible consistency likelihood $b_{j,k}^t(d,o)P_j^0(d,o)$. Otherwise, its occluded consistency likelihood is $(1 - vc_{j,k}(d,o))P^0$, where the uniform prior $P^0 = 1/size(d)size(o)$. $size(d)$ and $size(o)$ are the sizes of the depth and orientation hypothesis spaces. Therefore,

$$c_{j,k}(d,o) = v_{j,k} b_{j,k}^t(d,o) P_j^0(d,o) + (1 - v_{j,k})(1 - vc_{j,k}(d,o))P^0. \quad (9)$$

2.5. Pixel-Based Smoothing

The maximum likelihood estimates of each patch's depth $\hat{d}$ and orientation $\hat{o}$ based on $P_j^t(d,o)$ determine the initial depth map $d^0(x)$ and the orientation map for each pixel x.

$$< \hat{d} \quad \hat{o} >= \arg\max_{d,o} P_j^t(d,o) \quad (10)$$

We further refine the depth map $d^t(x)$ iteratively between images [4]. For each pixel x at the current viewpoint, we find its corresponding pixel y at the neighboring viewpoint k. If the corresponding pixel's depth $d_k^t(y)$ is similar to pixel x's depth $d^t(x)$, $d^{t+1}(x)$ is replaced by the average of $d^t(x)$ and $d_k^t(y)$.

3. EXPERIMENTAL RESULTS

We first showed the experimental results of the prior-based orientation estimator based on a single image.

We trained the SVM-based orientation estimator with 6670 labeled patches, extracted from 49 color images at 320x240 pixels. The sample images for training the orientation estimator were shown in Figure 5.

Figure 5. Sample images for training the orientation estimator

We inferred the orientation distribution of each image patch using the orientation estimator. In Figure 6, we showed the classification results of the orientation estimator on a sample image with the maximum likelihood estimates represented by the shaded colors: red (horizontal), green (vertical), and blue (frontal). It achieved the classification accuracy: 85%.

(a) Sample image (b) Classification results

Figure 6. Prior-based orientation estimation results

We also compared our simple orientation estimator with Hoiem's orientation estimator [7] using their online database [12]. We trained and tested our orientation estimator on their training and testing data provided. On a test set of 62 novel images, Hoiem reported that 87% of the pixels were correctly labeled into ground, vertical, or sky. We achieved the accuracy of 85% of the pixels correctly labeled in the same classification task, while our simple algorithm ran more than 3 times faster than Hoiem's algorithm as shown in Table 1.

Table 1. Performance comparison of the orientation estimators

	Our algorithm	Hoiem's algorithm
Classification Accuracy	85%	87%
Time/Frame	1.8 sec	7.6 sec

Next, we showed the experimental results of the prior-based geometry reconstruction of stationary scenes.

We ran experiments on multiple synthetic images of a stationary street simulated by POV-ray [13]. As illustrated in Figure 7, six images were captured by a backward-moving camera at 320x240 pixels with the known intrinsic and extrinsic camera parameters as inputs.

Each image was segmented into small patches, and patch's prior orientation probabilities were inferred based on patch's appearance and location. We applied the prior-based geometry reconstruction algorithm on these input images to reconstruct the depth map at each viewpoint.

We compared the reconstruction results of the proposed prior-based algorithm with the estimated prior distribution $P_j(o)$ and the results without using any prior, which were the oriented plane-sweeping algorithm [3] and our smoothed version of the oriented plane-sweeping algorithm with the patch-based smoothing and the pixel-based smoothing.

Figure 7. Input images of a stationary street scene

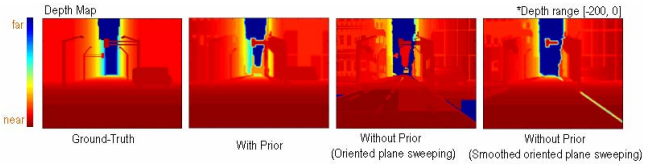


Figure 8. Depth map comparison of a stationary street scene

Table 2. Performance comparison of different algorithms

Depth map reconstruction	% Error/Pixel
With prior	3.5%
Without prior (Oriented plane sweeping)	7.2%
Without prior (Smoothed oriented plane sweeping)	4.1%

We compared the resulting depth maps at the first frame's viewpoint as shown in Figure 8 and Table 2. The dynamic depth range of the scene is assumed to be 200. Compared with the ground-truth depth map, the prior-based method provided the reconstructed depth map with 7/200=3.5% error per pixel on average, which is better than the approaches without any prior knowledge. The oriented plane sweeping algorithm offered 14.4/200=7.2% reconstruction error per pixel. The smoothed oriented plane-sweeping algorithm achieved 8.4/200=4.1% error per pixel. The algorithms without prior knowledge had difficulty in reconstructing the depth of the school bus, buildings, and ground areas as shown in Figure 8.

We also showed the experimental results in a real garden scene. A forward moving camera captured seven input images at 320x240 pixels as shown in Figure 9. We calibrated the camera's intrinsic parameters (camera's focal length and optical center) with checker board patterns offline and the extrinsic parameters (the translation vector and the rotation matrix) with markers on the ground using Zhang's method [14].

We applied the prior-based geometry reconstruction algorithm on these input images to reconstruct the depth map and orientation map at the fourth frame's viewpoint. We compared the results of these three algorithms again. The prior-based method provided better orientation map than the uniform prior approaches, especially in the textureless areas (ground and sky) in Figure 10. Although the smoothed oriented plane-sweeping algorithm had better and smoother results than the oriented plane-sweeping algorithm, it was still difficult to find the correct orientation in the textureless areas without any prior knowledge. Therefore, the prior-based method had better estimated depth maps than the uniform prior approaches in Figure 11.

4. CONCLUSIONS

In this paper, we proposed a probabilistic framework for reconstructing scene geometry utilizing prior knowledge. Traditional approaches try to match the points, lines or patches in multiple images to reconstruct scene geometry. Our framework also takes advantage of each image patch's appearance and location to infer its orientation using statistical learning. We showed that the prior-based reconstruction methods outperformed traditional reconstruction methods with both synthetic data and real data, especially in the textureless areas (a challenge problem for most traditional approaches).

Figure 9. Input images of a garden scene

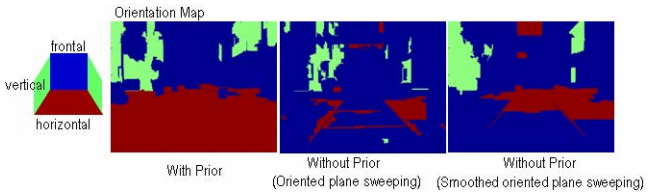


Figure 10. Orientation map comparison of a garden scene

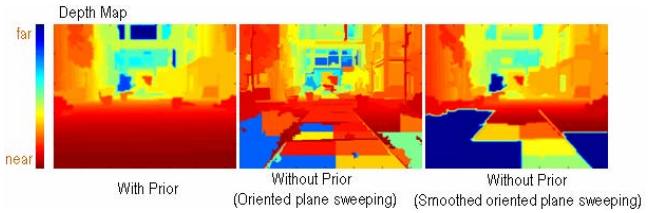


Figure 11. Depth map comparison of a garden scene

5. REFERENCES

[1] H.Y. Shum, S.B. Kang, and S.C. Chan, "Survey of image-based representations and compression techniques," *CirSysVideo(13)*, No. 11, pp. 1020-1037, November 2003.

[2] R. T. Collins, "A space-sweep approach to true multi-image matching," *IEEE CVPR*, pp. 358-363, June 1996.

[3] A. Akbarzadeh, J.-M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, H. Towles, D. Nistér and M. Pollefeys, "Towards urban 3D reconstruction from video," *3DPVT 2006*.

[4] L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *SIGGRAPH'04*, Aug. 2004.

[5] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by Voxel coloring", *CVPR'97*,1997.

[6] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", *IJCV*, pp. 7-42, Vol. 47, No.1-3, April-June 2002.

[7] D. Hoiem, A.A. Efros, and M. Hebert, "Automatic photo pop-up," *SIGGRAPH'05*, August 2005.

[8] A. Saxena, S. H. Chung, A. Y. Ng. "Learning depth from single monocular images," In NIPS 18, 2005.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, 59(2), 2004.

[10] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *IJCV*, 43(1):29-44, June 2001.

[11] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, 5:975–1005, 2004.

[12] http://www.cs.cmu.edu/~dhoiem/projects/data.html

[13] Ray tracking software at http://www.povray.org/.

[14] Z. Zhang, "A flexible new technique for camera calibration," *Microsoft Technical Report*-98-71, Dec 1998.