A RAO-BLACKWELLIZED MCMC ALGORITHM FOR RECOVERING PIECEWISE PLANAR 3D MODELS FROM MULTIPLE VIEW RGBD IMAGES

Natesh Srinivasan and Frank Dellaert

Georgia Institute of Technology

ABSTRACT

In this paper, we propose a reconstruction technique that uses 2D regions/superpixels rather than point features. We use pre-segmented RGBD data as input and obtain piecewise planar 3D models of the world. We solve the problem of superpixel labeling within single and multiple views simultaneously by using a Rao-Blackwellized Markov Chain Monte Carlo (MCMC) algorithm. We present our output as a labeled 3D model of the world by integrating out over all possible 3D planes in a fully Bayesian fashion. We present our results on the new SUN3D dataset [**?**].

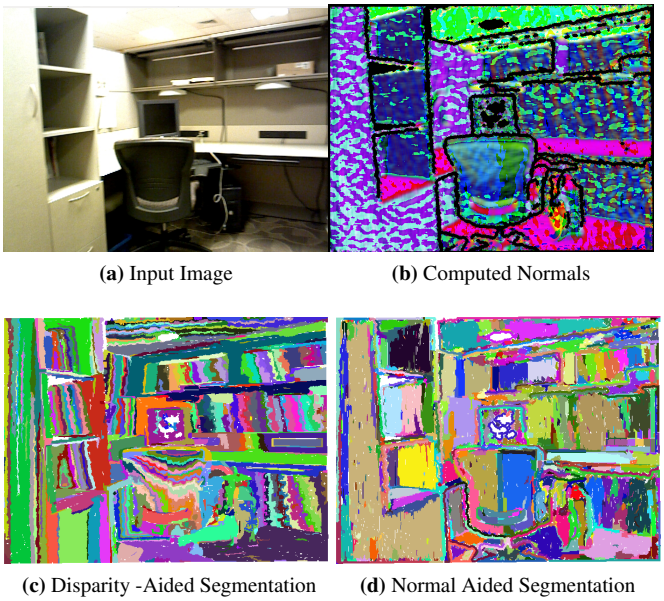
Index Terms— Segmentation, Point Clouds, Reconstruction, Piecewise Planar

1. INTRODUCTION

Mixed input, RGBD data allows dense 3D reconstruction [2] even in featureless environments. We are no longer encumbered by the need to obtain sparse 3D point clouds using costly structure-from-motion pipelines such as in [3]. RGBD data allows us to use model-based methods for reconstruction. Model-based methods assume the world to be comprised of geometric primitives like planes [4, 5], or voxels [6] provide a dense, photorealistic, 3D reconstruction as compared to sparse point clouds.

We can achieve dense 3D reconstruction by using model-based techniques. Inference using model-based methods requires (1) a geometric low-level segmentation to reduce search space and (2) a technique to match presegmented regions (a.k.a superpixels) across multiple views. This is a challenging problem because the superpixels are of radically different dimensions and suffer from occlusions.

There is a need for low level segmentation methods that jointly clusters pixels based on both depth and RGB data. [7, 8] are some of the seminal papers on low level segmentation methods that consider RGB data alone. Depth induces more challenges due to noise and methods like [9] use learning approach to co-segment RGB and depth data. Other methods like [10] focus on high level scene labeling. [1] uses disparity in conjunction with RGB data and is perhaps closest to our work. However this also tends to break physical boundaries as the distance of the scene distance increases.



(a) Input Image **(b)** Computed Normals

(c) Disparity -Aided Segmentation **(d)** Normal Aided Segmentation

Fig. 1: Normal Aided Oversegmentation. The superpixels $\mathcal{V}$ in the normal aided segmentation obey planar boundaries more rigorously than disparity alone as [1]. Helps form larger superpixels.

Matching superpixels across multiple views is a largely unsolved problem. Appearance based methods such as [11, 12] are developed for image de-noising applications and are not accurate enough for superpixel matching across multiple views. [13] assume a fixed affine transformation between the regions under consideration and hence fail in indoor regions. [14, 15] matches superpixels by locally fitting planes to each superpixels and then achieves pair-wise matching using a set of plane heuristics. The quality of the match is affected by the quality of the locally fit plane.

With the emergence of region-based reconstruction methods, correspondence-less SFM algorithms as discussed in [16] are of much greater practical value. This is because incorrect region (superpixel) correspondences affect reconstruction more profoundly than incorrect pixel correspondences and we cannot afford to have incorrect correspondences at the

This project is supported by the office of Naval Research, grant 00014-11-1-0593, for which the authors are grateful

superpixel level. We tackle this problem using a fully generative model that allows us to jointly model the superpixel association and the 3D model. As a result ,we can choose to obtain either the 3D model of the superpixel association by marginalizing out the other parameter from the joint density.

The first contribution of this paper is to present a novel low-level segmentation algorithm. Developing on the earlier work done in [1], we improve the RGBD "over-segmentation" using more discriminative features than disparity alone. The superpixels that we obtain obey planar boundaries Our fundamental framework uses a graph based formulation [8].

The second contribution of this paper is a generative model for doing multi-view RGBD segmentation. By treating the superpixel association in a completely Bayesian way, we model a joint distribution in a discrete-continuous space of superpixel association and the continuous plane parameters that govern the world model. We then obtain the marginal corresponding to the superpixel association by analytically integrating out the world model in a fully analytical way. Our technique allows us to associate superpixels within single as well as multiple views simultaneously.

The third contribution of this paper is a Rao Blackwellized MCMC algorithm to sample the discrete space of superpixel associations. Kaess et al.[17] uses this idea for multi-view reconstruction from RGB images. Erdogan et al uses this idea for single image segmentation [1]. Dellaert et al uses this idea for the sparse structure-from-motion problem without correspondences [16]. We extend this idea to multi-view RGBD data. Such an approach is far more useful/potent in the context of planar reconstruction because it is more resilient to noise in the measured depth data. We obtain the 3D model of the world by probabilistically weighting the individual 3D models that correspond to different superpixel associations. MCMC as an inference technique allows us to get this expectation rather than a MAP estimate.

2. LOW-LEVEL SEGMENTATION

Given a set of RGBD images as evidence $\mathcal{E}$, our goal is to obtain a 3D model of the world $\mathcal{X}$ in terms of piecewise planar patches. We reduce the search space by clustering similar pixels into superpixels by first over-segmenting the RGB image. Below we present a new heuristic for over-segmentation and obtain a set of superpixels $\mathcal{V}$ spanning multiple views.

2.1. Low-Level Geometric Segmentation

We present a graph-based approach for over-segmenting the image. This is is a preprocessing step and is used as an input to the model described in section **??**.We use a modified version of [8] because of the benefit it offers in terms of speed. The edge weight of the underlying graph connecting each pixel is a linear combination of 4 factors - (1) Spatial distance (2) Color difference (3) Disparity difference and (4) Normal Difference between the adjacent pixels\nodes.

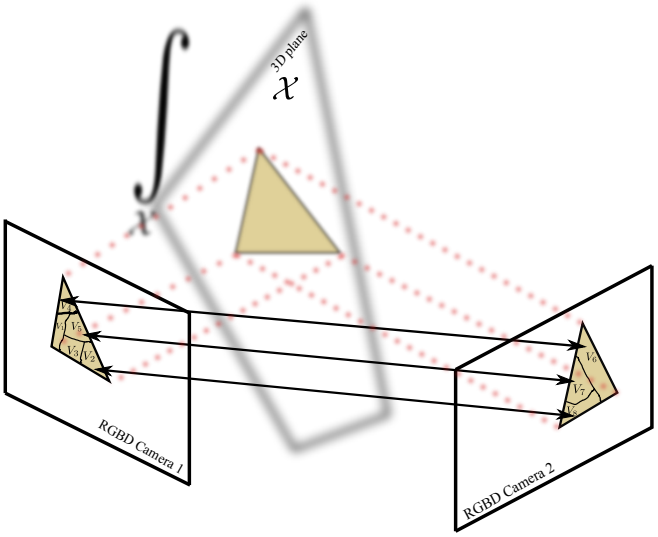


Fig. 2: Multi-view labeling problem using model selection. There are 8 superpixels in the superpixel set $\mathcal{V} = \{V_k\}_{k=1}^{8}$. The initial over-segmentation ensures that the superpixels do not straddle across depth discontinuities. The model selection has to correctly identify that all the superpixels here belong to the same plane and hence have the same label (plane). Bayesian Model Selection and Rao-Blackwellization: We directly evaluate the probability of superpixel correspondences across multiple views by integrating out the hidden world model in Bayesian model selection scheme. In this particular case, we are evaluating the marginal probability distribution corresponding to the label assignment $\mathcal{L}$. We integrate out the plane parameters θ corresponding to the 3D plane.

Before the creation of superpixels, the RGBD data is subject to the following preprocessing steps. (1) The normals are precomputed by fitting a local plane to every 3D point in every view within a predefined radius r. (2) A cross bilateral filter is used to smooth the RGB images. The co-registered disparity data is used to define the kernel weights in the bilateral filter. As a result, we obtain a set of smoothed RGB images that preserve edge boundaries. We call our superpixel set $\mathcal{V} = \{V_k\}_{k=1}^{M}$ where M is the total number of superpixels across all images.Figure 1 shows the improved superpixels obtained by using normals as a factor in the edge weights and compares our approach to [1].

3. GENERATIVE MODEL

To evaluate the likelihood the true measurement, i.e the disparity δ, we build a generative model that can generate the disparities across multiple views given the we have the model of the world $\mathcal{X}$.

The world model $\mathcal{X}$ spans a continuous space and models the plane normals. the label assignment $\mathcal{L}$ is a discrete variable that spans the space of positive integers $\mathcal{L} \in \{1 \dots B_M\}$

where B_M is the bell number of M where $M = |\mathcal{V}|$ and $\mathcal{V}$ is the set of all superpixels from all views. $\mathcal{L}$ models the association of superpixels to a 3D plane. An assignment associates one of the possible 2^M labels to each superpixel V in the superpixel set $\mathcal{V}$. Each label in $\mathcal{L}$ is associated with a continuous vector $\mathcal{X}$ that corresponds to a set of 3D plane segments.

The posterior that jointly models the superpixel labels $\mathcal{L}$ and world model $\mathcal{X}$ can be written using Bayes' law as

$$p\left(\mathcal{L}, \mathcal{X} | \mathcal{E}; \mathcal{V}\right) \propto p\left(\mathcal{E} | \mathcal{L}, \mathcal{X}; \mathcal{V}\right) p\left(\mathcal{L}, \mathcal{X}\right) \tag{1}$$

where $p\left(\mathcal{E} | \mathcal{L}, \mathcal{X}; \mathcal{V}\right)$ is the likelihood and $p\left(\mathcal{L}, \mathcal{X}\right)$ is a joint prior on the label assignment $\mathcal{L}$ and the world model $\mathcal{X}$.

Assuming that the 3D planes are independent of each other, we cluster the evidence $\mathcal{E}$ corresponding to each segment $\mathcal{S}$ as $\mathcal{E}_{\mathcal{S}}$ (A segment is a set of superpixels spanning multiple views). We further assume that every measured data follows an i.i.d Gaussian distribution. This factorizes Eq. 1 as

$$p\left(\mathcal{L}, \mathcal{X} | \mathcal{E}; \mathcal{V}\right) \propto p\left(\mathcal{L}, \mathcal{X}\right) \prod_{\mathcal{S}} \prod_{i} p\left(\mathcal{E}_{\mathcal{S}i} | \mathcal{L}, \mathcal{X}; \mathcal{V}\right) \tag{2}$$

where i is used to index all the measured data in a segment $\mathcal{S}$.

To evaluate Eq. 2 , we need a model that will allow us to predict the disparity δ_i given the plane parameters θ.

$$\delta_{\mathcal{S}i} = h\left(\theta_{\mathcal{S}}; u_{\mathcal{S}i}, v_{\mathcal{S}i}\right) + \mathcal{N}\left(0, \sigma_{\delta}\right) \tag{3}$$

We use geometry to obtain the function $h\left(\cdot\right)$. We model a plane using the parameters $\theta \triangleq \left(\hat{n}, d\right) \in \mathbb{S}^2 \times \mathbb{R}$, i.e., a normal vector $\hat{n} \in \mathbb{S}^2$ and the distance $d \in \mathbb{R}$ to the origin. The sphere $\mathbb{S}^2$ is the space of all unit vectors in $\mathbb{R}^3$. The distance d is signed, and hence we consider planes that are oriented towards/away from origin to be different .

To derive a relationship between disparity δ and the plane parameters θ we note that the plane equation in the global *world* coordinate frame W can be expressed as

$$\left[\begin{array}{cc} \hat{n}^T & d \end{array}\right] p^w = 0 \tag{4}$$

where $p^w = \left[\begin{array}{cccc} x^w & y^w & z^w & 1 \end{array}\right]^T$ is the 3D homogeneous coordinate of a point. We use superscripts to indicate the frame of reference.

Given that we have the camera pose $x_c^w = \left(R_c^w, t_c^w\right)$, we can transform the 3D point p^c in the camera frame to the global frame as

$$p^w = \left[\begin{array}{cc} R_c^w & t_c^w \\ 0 & 1 \end{array}\right] p^c$$

Substituting this in (4), the plane equation in terms of the homogeneous camera coordinates p^c becomes

$$\left[\begin{array}{cc} \hat{n}^T & d \end{array}\right] \left[\begin{array}{cc} R_c^w & t_c^w \\ 0 & 1 \end{array}\right] p^c = 0$$

Multiplying on both sides by $\frac{f\beta}{z^c}$ and collecting the terms corresponding to the disparity δ, where f is the camera focal length, β is the camera baseline, and z^c is the depth of the point from the camera, we obtain

$$\delta_p = \frac{-1}{\left(\hat{n}^T t_c^w + d\right)} \hat{n}^T R_c^w \left[\begin{array}{cccc} u & v & f\beta \end{array}\right]^T \tag{5}$$

where $h\left(\cdot\right)$ is the difference between the predicted and actual disparity.

4. FAST, RAO-BLACKWELLIZED MCMC

We can use jump diffusion sampling [18] to sample the joint posterior given in Eq. 2. However this is slow and we are not interested in obtaining the full posterior, but only the marginal corresponding to the label assignment $\mathcal{L}$.

We use the Rao-Blackwell theorem which predicts that, by sampling over a marginal distribution we need fewer samples to approximate the posterior $p\left(\mathcal{L} | \mathcal{E}; \mathcal{V}\right)$ rather than sampling from the entire joint posterior $p\left(\mathcal{L}, \mathcal{X} | \mathcal{E}; \mathcal{V}\right)$ and then integrating out the continuous variable $\mathcal{X}$.

To evaluate the marginal corresponding to a particular label assignment $\mathcal{L}$, we use a Bayesian model selection scheme [19]

$$p\left(\mathcal{L} | \mathcal{E}; \mathcal{V}\right) = \int_{\mathcal{X}} \left(p\left(\mathcal{L}, \mathcal{X}\right) \prod_{\mathcal{S}} \prod_{i} p\left(\mathcal{E}_{\mathcal{S}i} | \mathcal{L}, \mathcal{X}; \mathcal{V}\right) \right) \tag{6}$$

where per-pixel likelihood can be evaluated using the generative model given in is given in Eq. 3.

The normalization constants of the Gaussian distribution are constant across all models and hence do not play a role in model selection. We can marginalize over the world models $\mathcal{X}$ by integrating out the parameter $\theta_{\mathcal{S}}$. The marginalization step requires us to compute the integral corresponding to an unnormalized Gaussian

$$\mathcal{B}\left(\theta_v, \sigma_{\theta_v}\right) \mathcal{B}\left(\theta_h, \sigma_{\theta_h}\right) \prod_{\mathcal{S}} \prod_{i} \mathcal{B}\left(\left(\delta_{\mathcal{S}i} - h\left(\theta_{\mathcal{S}}; u_{\mathcal{S}i}, v_{\mathcal{S}i}\right)\right), \sigma_{\delta}\right) \tag{7}$$

where $\mathcal{B}\left(\cdot\right)$ represents the fact that we have dropped the normalization constant corresponding to the Gaussian $\mathcal{N}(\cdot)$. We have expanded the prior $p\left(\mathcal{L}, \mathcal{X}\right)$ as a product of two terms, a vertical and horizontal prior $\mathcal{B}\left(\theta_v, \sigma_{\theta_v}\right) \mathcal{B}\left(\theta_h, \sigma_{\theta_h}\right)$. We do not need to explicitly model the discrete variable corresponding to the label assignment $\mathcal{L}$. The segment $\mathcal{S}$, implicitly models this in Eq. 7. The integral of an unnormalized Gaussian is given by

$$\int_{\mathcal{X}} \mathcal{B}\left(\theta, \mathcal{L}; \sigma_{\delta}, \sigma_{\theta_v}, \sigma_{\theta_h}\right) = \mathcal{B}\left(\theta^{\star}, \Sigma_{\theta}\right) \sqrt{\left|2\pi\Sigma_{\theta}\right|} \tag{8}$$

where $\mathcal{B}\left(\theta^{\star}, \Sigma_{\theta}\right)$ is the function evaluated at the mean of the distribution and $|\cdot|$ is the determinant. Eq 8 evaluates the probability of a particular label assignment $\mathcal{L}$.

We use a Markov Chain Monte Carlo algorithm to sample from this marginal distribution.However, this can be slow;

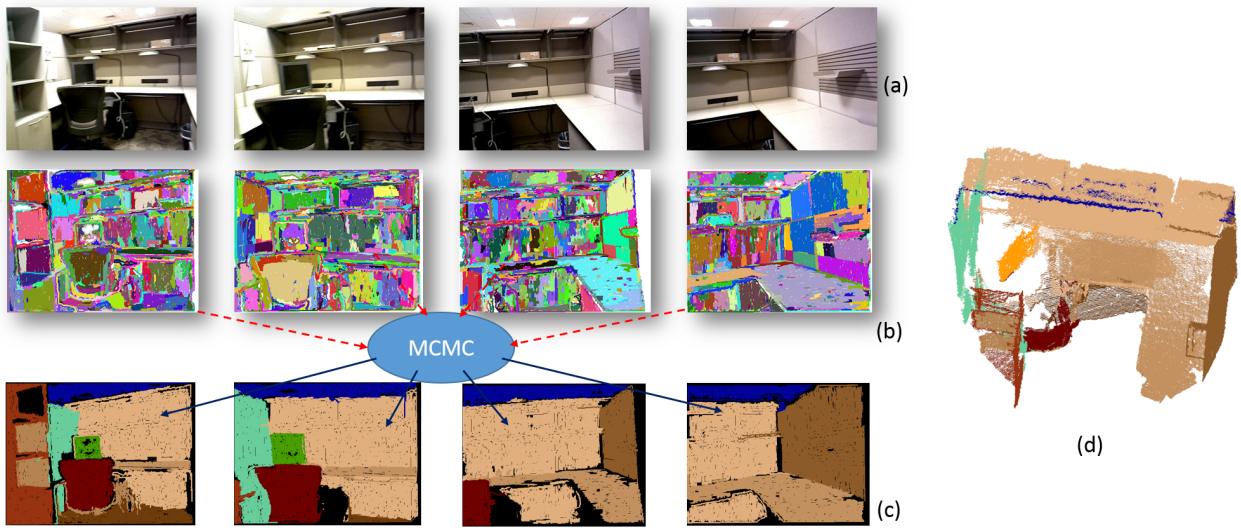


Fig. 3: Multiple View Semantic Segmentation and Reconstruction using multiple images in the brown_bm sequence in [**?**]; [Top Row] RGB images, the depth components are not shown. [Second row] Output of the low level segmentation algorithm. [Bottom row] The output of the multiple-view superpixel association. [Right] 3D reconstruction givne the semantic labels.

this is because we are required to evaluate Eq. 8 for every proposal. However, we can exploit the factorized form obtained in Eq. 2 to quickly make proposals that needs to only re-evaluate local changes.

Evaluation of Eq. 8 requires us to find out the MAP estimate $(\theta^\star, \Sigma_\theta)$ that minimize the negative log probability given in Eq. 7. This can be formulated as

$$\theta^\star = \arg\min_{\theta} \left(\log\left(\mathcal{B}\left(\theta, \mathcal{L}; \sigma_\delta\right)\right)\right)$$

The term inside the logarithm converts the products into summations as

$$\log\left(\mathcal{B}\left(\theta, \mathcal{L}; \sigma_\delta\right)\right) = \|\theta - \theta_v\|_{\sigma_{\theta_v}} + \|\theta - \theta_h\|_{\sigma_{\theta_h}}$$

$$+ \sum_{\mathcal{S}} \sum_{i} \|(\delta_{\mathcal{S}i} - h\left(\theta_{\mathcal{S}}; u_{\mathcal{S}i}, v_{\mathcal{S}i}\right))\|_{\sigma_\delta}$$

which can be regrouped as

$$\log\left(\mathcal{B}\left(\theta, \mathcal{L}; \sigma_\delta\right)\right) = -\sum_{\mathcal{S}} \sum_{i} \Big(\|\theta_{\mathcal{S}} - \theta_{v\mathcal{S}}\|_{\sigma_{\theta_v}} + \dots$$

$$\dots + \|\theta_{\mathcal{S}} - \theta_{h\mathcal{S}}\|_{\sigma_{\theta_h}} + \|(\delta_{\mathcal{S}i} - h\left(\theta_{\mathcal{S}}; u_{\mathcal{S}i}, v_{\mathcal{S}i}\right))\|_{\sigma_\delta} \Big) \quad (9)$$

Eq. 9 has a partial structure that only requires the re-evaluation of segments $\mathcal{S}$ that were locally affected in the MCMC state transition and hence *leaves a large portion of Eq. 9 unchanged.* This allows us to make *fast proposals* and hence use a Rao-Blackwellized MCMC algorithm as an inference tool.

5. RESULTS AND DISCUSSION

Figure 3 shows the results of the multi-view segmentation and reconstruction. Since we use the model selection idea, our algorithm automatically infers the number of 3D planes. In this case, the number of planes are 10. The 3D reconstruction is simply a transformation of the local point clouds in each frames using the camera pose and we do not use a bundle-adjustment like frame work here. Our future work seeks to model the camera transformation within the joint distribution to form a full Semantic SLAM system.

6. REFERENCES

[1] C. Erdogan, M. Paluri, and F. Dellaert, "Planar segmentation of rgbd images using fast linear fitting and markov chain monte carlo," in *Computer and Robot Vision (CRV), 2012 Ninth Conference on*, pp. 32–39, IEEE, 2012.

[2] R. Newcombe, A. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR)*, (Basel, Switzerland), pp. 127–136, Oct 2011.

[3] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, 2011.

[4] D. Gallup, J.-M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruc-

tion," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1418–1425, 2010.

[5] A. Kowdle, Y.-J. Chang, A. Gallagher, and T. Chen, "Active learning for piecewise planar 3D reconstruction," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 929–936, 2011.

[6] B.-S. Kim, P. Kohli, and S. Savarese, "3D scene understanding by Voxel-CRF," in *Intl. Conf. on Computer Vision (ICCV)*, 2013.

[7] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[8] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Intl. J. of Computer Vision*, vol. 59, pp. 167–181, 2004.

[9] S. N. S. Adarsh Kowdle and R. Szeliski, "Multiple view object cosegmentation using appearance and stereo cues," in *European Conference on Computer Vision (ECCV 2012)*, October 2012.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conf. on Computer Vision (ECCV)*, 2012.

[11] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized PatchMatch correspondence algorithm," in *European Conference on Computer Vision ? ECCV 2010*, no. 6313, pp. 29–43, Springer Berlin Heidelberg, 2010.

[12] Y. Eshet, S. Korman, E. Ofek, and S. Avidan, "Dcsh-matching patches in rgbd images," in *International Conference on Computer Vision ICCV*, IEEE, 2013.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *Intl. J. of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.

[14] K. Pathak, A. Birk, N. Vas?kevic?ius, and J. Poppinga, "Fast registration based on noisy planes with unknown correspondences for 3-d mapping," *Robotics, IEEE Transactions on*, vol. 26, no. 3, pp. 424–441, 2010.

[15] A. Trevor, J. Rogers, and H. Christensen, "Planar surface SLAM with 3D and 2D sensors," in *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3041–3048, 2012.

[16] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun, "Structure from motion without correspondence," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2000.

[17] M. Kaess, R. Zboinski, and F. Dellaert, "MCMC-based multiview reconstruction of piecewise smooth subdivision curves with a variable number of control points," in *European Conf. on Computer Vision (ECCV)*, vol. 3023 of *Lecture Notes in Computer Science*, (Prague, Czech Republic), pp. 329–341, Springer, 2004.

[18] Z. Tu and S. C. Zhu, "Image segmentation by data-driven markov chain monte carlo.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 657–673, 2002.

[19] C. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, Secaucus, NJ, USA: Springer-Verlag, 2006.