2011 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 25-30, 2011. San Francisco, CA, USA

Building Facade Detection, Segmentation, and Parameter Estimation for Mobile Robot Localization and Guidance

Jeffrey A. Delmerico
SUNY at Buffalo
jad12@buffalo.edu

Philip David
Army Research Laboratory
Adelphi, Maryland
philip.j.david4.civ@mail.mil

Jason J. Corso
SUNY at Buffalo
jcorso@buffalo.edu

***Abstract*— Building facade detection is an important problem in computer vision, with applications in mobile robotics and semantic scene understanding. In particular, mobile platform localization and guidance in urban environments can be enabled with an accurate segmentation of the various building facades in a scene. Toward that end, we present a system for segmenting and labeling an input image that for each pixel, seeks to answer the question "Is this pixel part of a building facade, and if so, which one?" The proposed method determines a set of candidate planes by sampling and clustering points from the image with Random Sample Consensus (RANSAC), using local normal estimates derived from Principal Component Analysis (PCA) to inform the planar model. The corresponding disparity map and a discriminative classification provide prior information for a two-layer Markov Random Field model. This MRF problem is solved via Graph Cuts to obtain a labeling of building facade pixels at the mid-level, and a segmentation of those pixels into particular planes at the high-level. The results indicate a strong improvement in the accuracy of the binary building detection problem over the discriminative classifier alone, and the planar surface estimates provide a good approximation to the ground truth planes.**

I. Introduction

Accurate scene labeling can enable applications that rely on the semantic information in an image to make high level decisions. Our goal of labeling building facades is motivated by the problem of mobile robot localization in GPS-denied areas. This problem arises in urban areas, so the approach currently being developed by our group depends on detection of buildings within the field of view of the cameras on a mobile platform. Within this problem, accurate detection and labeling is critical for the high level localization and guidance tasks. We restrict our approach to identifying only planar building facades, and require image input from a stereo source. Since most buildings have planar facades, and many mobile robotic platforms are equipped with stereo cameras, neither of these assumptions is particularly restrictive.

In this paper, we propose a method for building facade labeling in stereo images that further segments the individual facades and estimates the parameters of their 3D models. Our approach proceeds in three main steps: discriminative modeling, candidate plane detection through PCA and RANSAC, and energy minimization of MRF potentials. Our contribution is the use of plane fitting techniques from stereo imagery to the problem of building facade segmentation in the context of mobile platform localization. A diagram of the workflow for candidate plane detection and high-level labeling is provided in Fig. 1.

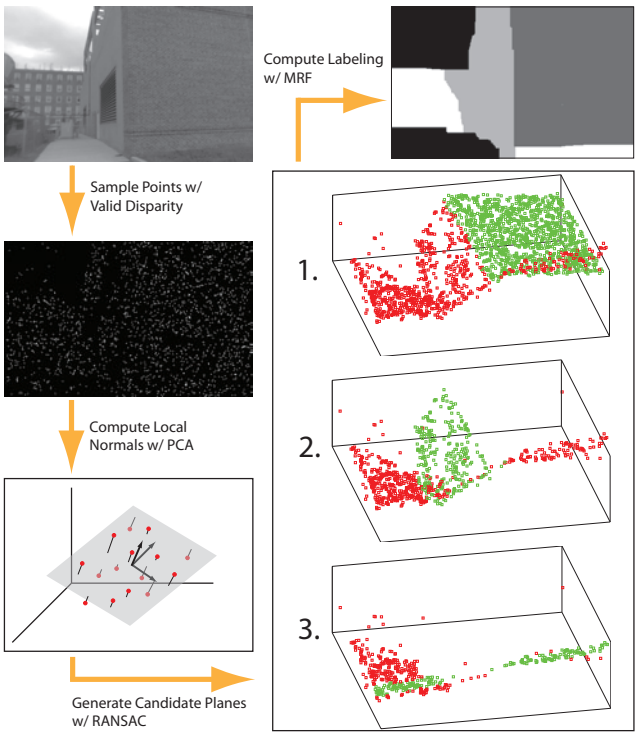

Fig. 1. Workflow of our candidate plane selection and labeling method. We iteratively run RANSAC on a set of sampled points from the image, removing the inliers (green) from the set, to generate a set of candidate planes. Our planar model incorporates the fit of the PCA local normal estimate into the error term. The set of candidate planes provides the label set for the high-level MRF model.

Our work leverages stereo information from the beginning. Our discriminative model is generated from an extension of the Boosting on Multilevel Aggregates (BMA) method [1] that includes stereo features [2]. Boosting on Multilevel Aggregates uses hierarchical aggregate regions coarsened from the image based on pixel affinities, as well as a variety of high-level features that can be computed from them, to learn a model within an AdaBoost [3] two- or multi-class discriminative modeling framework. The multilevel aggregates

The authors are grateful for the financial support provided in part by NSF CAREER IIS-0845282, DARPA W911NF-10-2-0062, and ARO Young Investigator W911NF-11-1-0090.

exploit the propensity of these coarsened regions to adhere to object boundaries, which in addition to the expanded feature set, offer less polluted statistics than patch-based features, which may violate those boundaries. Since many mobile robot platforms are equipped with stereo cameras, and can thus compute a disparity map for their field of view, our approach of using statistical features of the disparity map is a natural extension of the BMA approach given our intended platform. Since buildings tend to have planar surfaces on their exteriors, we use the stereo features to exploit the property that planes can be represented as linear functions in disparity space and thus have constant spatial gradients [4]. We use the discriminative classification probability as a prior for inference of facade labeling.

In order to associate each building pixel with a particular facade, we must have a set of candidate planes from which to infer the best fit. We generate these planes by sampling the image and performing Principal Component Analysis (PCA) on each local neighborhood to approximate the local surface normal at the sampled points. We then cluster those points by iteratively using Random Sample Consensus (RANSAC) [5] to find subsets which fit the same plane model and have similar local normal orientations. From these clusters of points, we are able to estimate the parameters of the primary planes in the image.

We then incorporate both of these sources of information into a Bayesian inference framework using a two-layer Markov Random Field (MRF). We represent the mid-level MRF as an Ising model, a layer of binary hidden variables representing the answer to the question "Is this pixel part of a building facade?" This layer uses the discriminative classification probability as a prior, and effectively smooths the discriminative classification into coherent regions. The high-level representation is a Potts model, where each hidden variable represents the labeling of the associated pixel with one of the candidate planes, or with no plane if it is not part of a building. For each pixel, we consider its image coordinates and disparity value, and evaluate the fitness of each candidate plane to that pixel, and incorporate it into the energy of labeling that pixel as a part of that plane. A more in-depth discussion of these methods can be found in Section II-B.

We use the Graph Cuts energy minimization method [6] to compute minimum energy labelings for both levels of our MRF model.

In principle our approach is modular, in that for each of the three phases (modeling, candidate plane detection, and labeling), a different method that produces the same type of output (probability map, candidate plane set, facade segmentation, respectively) could be substituted. However, the specific techniques we have developed have been motivated by the features of this specific problem.

A. Related Work

Building facade detection and segmentation have been and continue to be well-studied problems. Many recent papers in the literature have focused on segmentation of building facades for use in 3D model reconstruction, especially in the context architectural modeling or geo-spatial mapping applications such as Google Earth. Korah and Rasmussen use texture and other *a priori* knowledge to segment building facades, among other facade-related tasks [7]. Wendel et al. use intensity profiles to find repetitive structures in coherent regions of the image in order to segment and separate different facades [8]. Hernández and Marcotegui employ horizontal and vertical color gradients, again leveraging repetitive structures, to segment individual facades from blocks of contiguous buildings in an urban environment [9].

Several other methods utilize vanishing points for planar surface detection. David identifies vanishing points in a monocular image by grouping line segments with RANSAC and then determining plane support points by the intersection of the segments which point toward orthogonal vanishing point ultimately clustering them to extract the planes of the facade [10]. Bauer et al. implement a system for building facade detection using vanishing point analysis in conjunction with 3D point clouds obtained by corresponding a sweep of images with known orientations [11]. Lee et al. use a line clustering-based approach, which incorporates aerial imagery, vanishing points, and other projective geometry cues to extract building facade textures from ground-level images, again toward 3D architectural models reconstruction [12].

Our work draws on the contributions of Wang et al., whose facade detection method using PCA and RANSAC with LiDAR data inspired our approach with stereo images [13]. Perhaps the approach most similar in spirit to ours is that of Gallup et al. [14], who also use an iterative method for generating candidate plane models using RANSAC, and also solve the labeling problem using graph cuts [6]. However, their approach relies on multiview stereo data and leverages photoconsistency constraints in their MRF model, whereas we perform segmentation with only single stereo images. In addition, on a fundamental level their method involves finding many planes that fit locally, and stitching them together, whereas we aim to extract our planar models from the global data set, without an explicit restriction on locality. We present quantitative results on the accuracy of our planar modeling as well.

Although many of these results are directed toward 3D model reconstruction, some other work using unrelated techniques has been focused toward our intended application of vision-based navigation, namely [10], [15], [16], [17]. Additionally, our work is focused on retrieval of the estimated plane parameters, as implemented in the planar surface model of [4], and not on 3D model reconstruction.

II. Methods

A. BMA+Disparity Classifier

Based on the work in [2], we model building facade features using the Boosting on Multilevel Aggregates (BMA) [1] method, with the extension to stereo features. In principle, any classifier could be used for this step, so long as it could produce a probability map for binary classification in

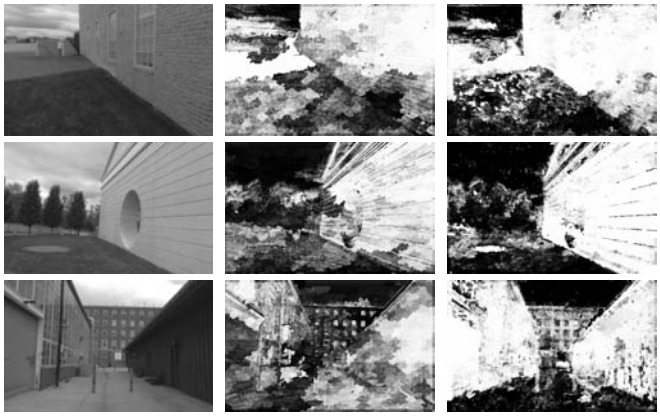

Fig. 2. Several examples of probability maps to be used as priors for our MRF. For each example, they are (L to R) the original image, by standard BMA, and by BMA+Disparity.

identifying building pixels. We choose the BMA+Disparity method because of its superior performance to standard AdaBoost, as well as to BMA without the addition of stereo features (See Fig. 2).

B. Plane Parameters

Throughout this discussion, we assume that we have stereo images which are rectified, but since we do not aim for full 3D reconstruction, for the purposes of the following derivation the camera's calibration parameters are left as unknown constants. We can determine the surface normal parameters up to a constant that describes the camera parameters, and since that constant will be the same across all candidate planes, we can use the computed surface normals to differentiate between planes. Known camera parameters would enable recovery of the surface normal parameters in world coordinates.

A plane in 3D space can be represented by the equation:

$$ax + by + cz = d \tag{1}$$

and for non-zero depth, z, this can be rewritten as:

$$a\frac{x}{z} + b\frac{y}{z} + c = \frac{d}{z} \tag{2}$$

We can map this expression to image coordinates by the identities $u = f \cdot \frac{x}{z}$ and $v = f \cdot \frac{y}{z}$, where f is the focal length of the camera. We can also incorporate the relationship of the stereo disparity value at camera coordinate (u, v) to the depth, z, using the identity $D(u, v) = \frac{fB}{z}$, where D is the disparity and B is the baseline of the stereo camera. Our plane equation becomes:

$$a\frac{u}{f} + b\frac{v}{f} + c = \frac{d \cdot D(u, v)}{fB} \tag{3}$$

which reduces to:

$$\left(\frac{aB}{d}\right)u + \left(\frac{bB}{d}\right)v + \left(\frac{cfB}{d}\right) = D(u, v) \tag{4}$$

Although $\mathbf{n} = (a, b, c)^T$ is the surface normal in world coordinates, for our purposes we can seek to determine the following modified plane parameters $\mathbf{n}' = (a', b', c')$, where:

$$a' = \frac{aB}{d}, b' = \frac{bB}{d}, c' = \frac{cfB}{d} \tag{5}$$

such that

$$\mathbf{n}' \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = a'u + b'v + c' = D(u, v) \tag{6}$$

This new set of plane parameters relates the image coordinates and their corresponding disparity values by incorporating the constant but unknown camera parameters.

C. Candidate Plane Detection

We perform the second phase of our approach by iteratively using RANSAC to extract a set of points which fit a plane model in addition to having a local normal estimate which is consistent with the model. The extracted plane models become the set of candidate planes for our high-level MRF labeling.

1) Local Normal Estimation: Based on our assumption of rectilinear building facades, we can use Principal Component Analysis to determine a local normal to a point in disparity space as in [18]. We first construct the covariance matrix of the neighborhood around the point in question. To do this, we consider any points, in a 5×5 window centered on our point $p = (v, u, -D(u, v))$, that have a valid disparity value. Here, u and v represent row and column indices, respectively. Note that stereo cameras that compute the disparity map with onboard processing in real-time often do not produce dense disparity maps. The pixels that are not matched in the left and right images or are at a distance beyond the usable range of the camera will be labeled with the maximum value for that disparity image, representing that the camera failed to compute the disparity at that pixel. Consequently, the neighborhood we use for PCA may be sparse. Other neighborhood sizes could be used, but we found that a 5×5 window provided good estimates while remaining local. We compute the centroid, $\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$, of the points $\{p_i\}_{i=1\ldots N}$ in the neighborhood with valid disparity, and calculate the 3×3 covariance matrix with:

$$W = \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p}) \otimes (p_i - \bar{p}) \tag{7}$$

where $\otimes$ is the outer product. We then compute the eigenvalues of W, and the eigenvectors corresponding to the largest two eigenvalues indicate the directions of the primary directions on the local planar estimate. The eigenvector corresponding to the smallest eigenvalue thus indicates the direction of the local surface normal, $\mathbf{n}_{(u,v)}$.

2) RANSAC Plane Fitting: We take a sample, S, of points from the image, which all have valid disparity values, and compute the local planar surface normal estimates by the aforementioned method. We then seek to fit a model to some subset of S of the form:

$$\alpha v + \beta u + \epsilon(-D(u, v)) + \theta = 0 \tag{8}$$

where $\tilde{\mathbf{n}} = \frac{1}{\epsilon}(\alpha, \beta, \theta)$ is the surface normal from Eq. (6). Since RANSAC finds the largest consensus set, P_{in}, that it can among S, we will fit the most well-supported plane first [5]. We then remove the inliers, $S' = S \setminus P_{in}$, and repeat this process iteratively, finding progressively less well-supported planes, until a fixed percentage of the original S has been clustered into one of the extracted planes. In our experiments, we used a sample of 2000 points from the image, and concluded the plane extraction once 80% of the points had been clustered, or when RANSAC failed to find a consensus set among the remaining points. We also use a RANSAC noise standard deviation of $\sigma_\eta = 5$, representing the amount of Gaussian noise that is assumed on the positions of the inlier points.

Although we use RANSAC to fit a standard plane model, we use a modified error term in order to incorporate the information in the local normal estimates. Here, since our local normal estimate required the use of a three dimensional coordinate system $(u, v, -D(u, v))$, and produces a normal of that form, we must use a slightly different normal formulation of $\mathbf{n}_m = (\alpha, \beta, \epsilon)$. The standard measure of error for a plane model is the distance of a point from the plane: $E_m = \mid \alpha v + \beta u + \epsilon(-D(u, v)) + \theta \mid$, assuming $\mathbf{n}_m = (\alpha, \beta, \epsilon)$ is a unit vector. We compute another measure of error, E_{norm}, the dot product of the model plane normal $\mathbf{n}_m$ and the local normal estimate $\mathbf{n}_{(u,v)}$, which is the cosine of the dihedral angle between the two planes defined by those normals. If we take its magnitude, this metric varies from 0 to 1, with 1 representing normals which are perfectly aligned, and 0 representing a dihedral angle of $90°$. Since the range of E depends on the properties of the image (resolution, disparity range), we combine these two metrics as follows:

$$E = E_m(2 - E_{norm}) = E_m(2 - \mid \langle \mathbf{n}_m, \mathbf{n}_{(u,v)} \rangle \mid) \quad (9)$$

such that the dihedral angle scales the error term from E_m to $2E_m$, depending on the consistency of the model and local normals.

D. MRF Model

We model our problem in an energy minimization framework as a pair of coupled Markov Random Fields. Our mid-level representation seeks to infer the correct configuration of labels for the question "Is this pixel part of a building facade?" Based on this labeling, the high-level representation seeks to associate those pixels which have been positively assigned as building facade pixels to a particular candidate plane. Figure 3 shows a graphical representation of this MRF model. Our motivation for this design stems from the fact that these are related but distinct questions, and they are informed by different approaches to modeling buildings. The mid-level MRF represents an appearance-based model, while the high-level MRF represents a generative model for the planar facades.

1) Mid-level Representation: We want our energy function for the mid-level model to capture the confidence (probability) of our discriminative classification, and we want there to be a penalty whenever a pixel with a high confidence

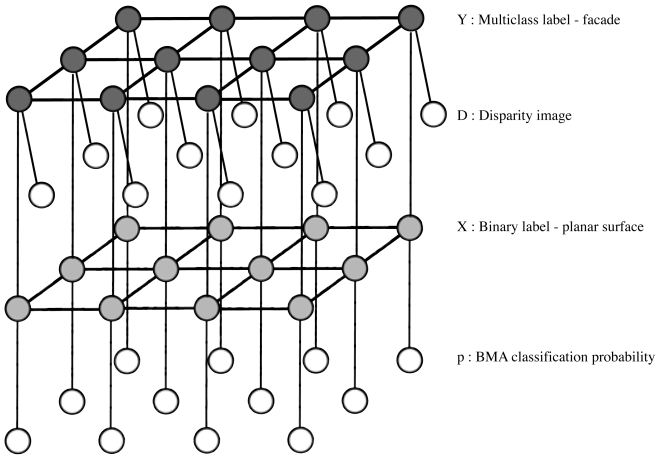

Fig. 3. Our two-layer MRF model.

is mislabeled, but a smaller penalty for pixels with lower confidence in their a priori classification. We will use an Ising model to represent our mid-level MRF, where our labels x_s for $s \in \lambda$, our image lattice, come from the set $\{-1, 1\}$. We define a new variable b_s to represent a mapping of the $X_s \in \{-1, 1\}$ label to the set $\{0, 1\}$ by the transformation $b_s = \frac{X_s+1}{2}$. For a particular configuration of labels l, we define our mid-level energy function as:

$$E(l) = \sum_{s \in \lambda} [(1 - b_s)p(s) + b_s(1 - p(s))] - \gamma \sum_{s \sim t} x_s x_t \quad (10)$$

where $p(s)$ is the discriminative classification probability at s and γ is a constant weighting the unary and binary terms. The b_s quantity in the unary term essentially switches between a penalty of $p(s)$ if the label at s is set to -1, and a penalty of $1 - p(s)$ if the label at s is set to 1. Thus for $p(s) = 1$, labeling $x_s = -1$ will incur an energy penalty of 1, but labeling $x_s = 1$ will incur no penalty. Similarly for $p(s) = 0$, labeling $x_s = -1$ will incur no penalty, but labeling it 1 will incur a penalty of 1. A probability of 0.5 will incur an equal penalty with either labeling. Our smoothness term is from the standard Ising model. In our experiments, we used a γ value of 10.

2) High-level Representation: In designing our energy function for the high-level MRF, we want to penalize points which are labeled as being on a plane, but which do not fit the corresponding plane equation well. Our label set for labels y_s, $s \in \lambda$, is $\{0, \ldots, m\}$, with m equal to the number of candidate planes identified in the plane detection step. It corresponds to the set of candidate planes indexed from 1 to m, as well as the label 0, which corresponds to "not on a plane". We define a set of equations $E_p(s)$ for $p \in \{0, \ldots, m\}$ such that

$$E_p(s) = \mid a'_p u + b'_p v + c'_p - D(s) \mid \quad (11)$$

where the surface normal $\mathbf{n}'_\mathbf{p} = (a'_p, b'_p, c'_p)$ corresponds to the plane with label p, and $D(s)$ is the disparity value at s. We normalize this energy function by dividing by the maximum disparity value, in order to scale the maximum energy penalty down to be on the order of 1. For consistency

in our notation, we define $E_0(s)$ to be the energy penalty for a label of 0 at s, corresponding to the "not on a plane" classification. We set $E_0(s) = b_s$, such that a labeling of -1 in the mid-level representation results in $b_s = 0$, so there is no penalty for labeling s as "not on a plane". Similarly, when $x_s = 1$, $b_s = 1$, so there is a penalty of 1 to label any of the non-planar pixels as a plane.

To construct our overall energy function for the high-level MRF, we incorporate the exponential of the set of planar energy functions E_p with a delta function, so the energy cost is only for the plane corresponding to the label y_s. Since we cannot compute E_p without a valid disparity value, we use an indicator variable $\chi_D \in \{0, 1\}$ to switch to a constant energy penalty for all planes and the no-plane option, in order to rely strictly on the smoothness term for that pixel's label. For the smoothness term, we use a Potts model, weighted like the mid-level representation, with a constant $\gamma' = 1$. Thus the high-level energy function we are seeking to minimize is:

$$E(l) = \sum_{s \in \lambda} \sum_{p=0}^{m} \delta_{y_s = p} \cdot \exp\left(\chi_D E_p(s)\right) + \gamma' \sum_{s \sim t} \delta_{y_s = y_t} \quad (12)$$

E. Energy Minimization

To perform the energy minimization, we use the graph cuts expansion algorithm, specifically the implementation presented in [6]. We perform the minimization in two stages. Although the two labeling problems are coupled, the results, at least in terms of their appearance, were more accurate when performed in this way. We first minimize the energy of the mid-level MRF to obtain an approximation to the optimal labeling of planar surface pixels. This step uses prior knowledge from the discriminative classification. Next, we use the mid-level labeling as well as the detected candidate planes as a prior for the high-level MRF, and we use graph cuts again to compute an approximation to that optimal labeling.

III. Experimental Results

We have performed two experimental studies using our method on a new benchmark dataset[1]. We are not aware of another publicly available, human-annotated, stereo building dataset. Our full data set consists of 142 grayscale images from the left camera of a stereo imager[2], each with a corresponding 16-bit disparity map. All images have 500×312 resolution and human-annotated ground truth. We used 100 randomly selected images for training our discriminative classifier. Of the remaining 42 images, 21 were used for testing; 21 images were excluded from the test set because they are negative examples or the facades were too distant for the stereo camera to generate useful disparity values.

A. Single Plane Segmentation

Our testing data set contains 13 images which feature only one plane in the scene. In order to validate our methods, we first tested the accuracy of our approach for both labeling and parameter estimation on this restricted data set. For each image, we hand-labeled a ground truth segmentation, and used only that region of the image to determine a ground truth plane with RANSAC. Next, we applied our method for inferring the segmentation labels and plane parameters and compared the results with our ground truth by computing the F-score for each mid- and high-level labeling, and the dihedral angle between the ground truth and primary estimated plane from the high-level labeling. In the case of the mid-level labeling, the f-score represents the accuracy of labeling building pixels from background, and for the high-level it represents the accuracy of labeling the correct plane within the foreground. It should be noted that in most cases, several candidate planes were generated and then labeled by the MRF, but for our planar comparisons, we must select only one to compare with the ground-truth. This portion of our method is not yet automated, and we manually selected the best candidate plane from the set of applied labels. In most cases, this choice coincided with the label with the largest segmentation area, and was also generally the most vertical of the candidate planes. But in a few cases, notably image sets 12 and 13, the choice was somewhat ambiguous based on area or the magnitude of the $\mathbf{u}$ component of the normal. Future work will be devoted to refining this portion of our method for disambiguating labels in the absence of ground-truth. Our results are summarized in Fig. 4 and Table I.

TABLE I
Single Plane Modeling Accuracy

Image #	Mid F-score	High F-score	Angle (°)
1	0.8905	0.9738	0.390
2	0.9242	0.9129	2.505
3	0.8859	0.8749	0.472
4	0.9117	0.9094	1.209
5	0.9607	0.7770	3.133
6	0.9168	0.9339	1.172
7	0.9371	0.7933	2.522
8	0.8180	0.9711	0.272
9	0.9836	0.9833	0.866
10	0.9074	0.8200	1.362
11	0.8410	0.9169	1.064
12	0.6298	0.4707	9.076
13	0.7440	0.3135	7.681
Avg:	0.8731	0.8193	2.440

Our results on the single-plane images indicate that our proposed method is robust and highly accurate. The average f-scores for both the mid-level and high-level labeling tasks are both above 80%, and the average dihedral angle is $2.44°$. Additionally, all of the planes were accurate to within $10°$ of their ground truth orientations. It should be noted that although this image set was restricted to those with single facades, all of the images are of natural scenes which include occlusions and differences in illumination, the buildings were captured at different angles, and the buildings themselves have different sizes, textures, and architectural features.

[1] Available at http://www.cse.buffalo.edu/~jcorso/r/gbs

[2] Tyzx DeepSea V2 camera with 14 cm baseline and 62° horizontal field of view.

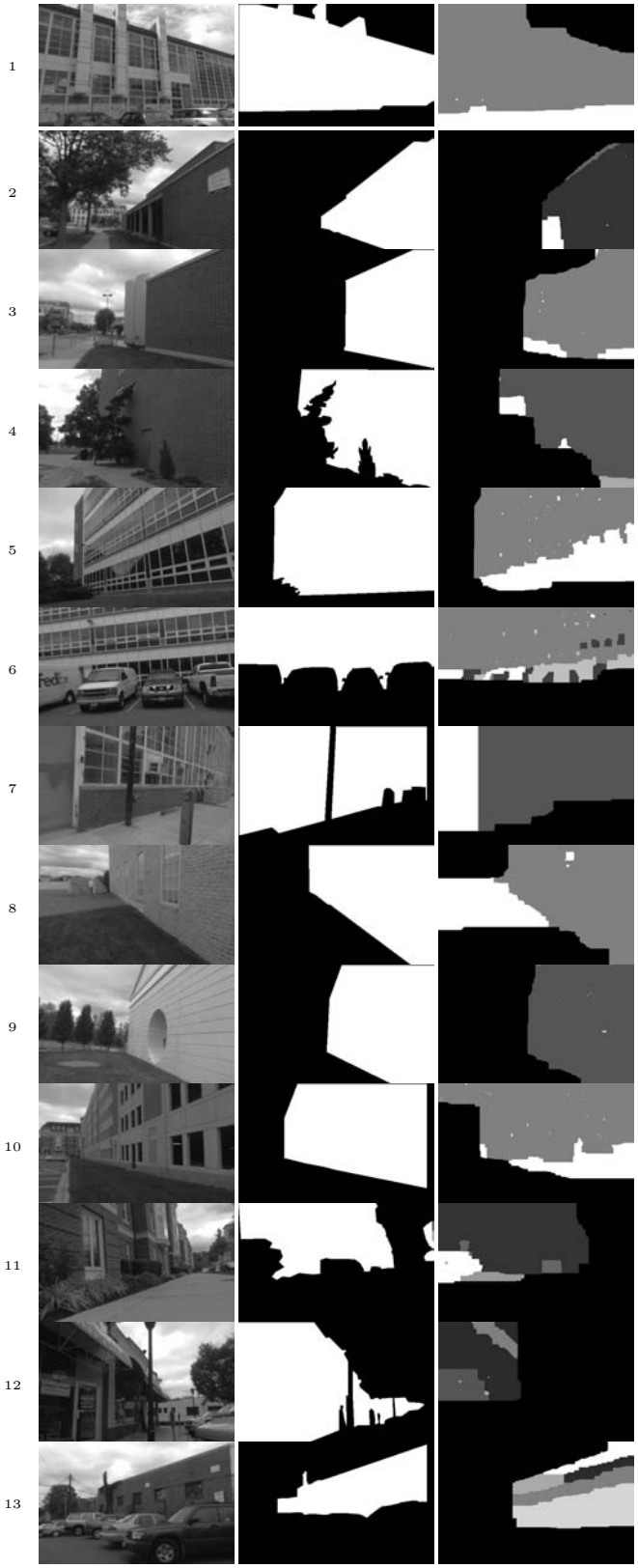

Fig. 4. Segmentations and planar facade estimates on single-facade images. For each example, they are (L to R) the original image, ground truth segmentation, high-level MRF labeling.

TABLE II

MULTIPLE PLANE MODELING ACCURACY

Image #	Mid F-score	High F-score	Angle (°)
1	0.9886	0.9264	15.299 2.113
2	0.9621	0.8751	0.513 27.611
3	0.9452	0.6876	56.510 1.309
4	0.9238	0.9249	0.340 2.035 83.228
5	0.9131	0.8953	0.686 1.173
6	0.9492	0.9118	16.115 0.860 76.594
7	0.9735	0.9595	1.436 0.383
8	0.8864	0.9302	3.011 1.975
Avg:	0.9427	0.8889	16.177

B. Multiple Plane Segmentation

We now proceed to analysis of more sophisticated scenes that include multiple facades at different orientations. We selected the 8 images from our testing set that contained at least 2 facades in the scene. Similarly to the single-plane experiments, we hand-labeled a ground truth segmentation, and for each ground-truth facade, used only that region of the image to determine a ground truth plane with RANSAC. When the MRF segmentation produced more than one plane label over a ground-truth plane region, we again manually chose the best of those candidates for our comparison of angular accuracy to the ground truth. Once again, the best candidate plane often coincided with the largest label area and most vertical plane, but not consistently. For example, in image set 1 of Fig. 5, the segmentation of the left facade is divided among white, light gray, and dark gray labels, but we we compute the dihedral angle only for the dark gray candidate plane because it was the most accurate to the ground-truth for that region. Our results are summarized in Fig. 5 and Table II.

As with the single-plane experiments, the accuracy of labeling the primary facades in the image is generally very good, and in most cases does not require any manual disambiguation for the angular accuracy. However, for some of the minor facades in the image (either small, distant, or obliquely angled) the error in either the labeling or the plane estimation, or both, was large. In particular, image sets 4 and 6 in Fig. 5 show the difficulty in achieving good estimates for distant planes, as they have large errors in the plane estimation, and both plane estimation and labeling, respectively. It should be noted, however, that the plane estimation errors in these cases are primarily in the

rotation about the $\mathbf{v}$-axis, leading to plane estimates which are reasonably accurate in their rotation about the $\mathbf{u}$-axis, but are far from being vertical. Since we are working under the assumption that most buildings have upright facades, it may be possible in the future to apply a constraint to correct these estimates.

C. Analysis

Averaged over all 21 images in our testing set, the mid-level labeling achieved an accuracy of 0.8996, and the high-level labeling achieved an accuracy of 0.8458. The average dihedral angle among all of the 31 ground-truth planes in the 21 images was $10.416°$.

The accuracy of our methods on natural scenes is correlated with the quality of the disparity maps, as well as the location of the facades within the usable range of the stereo camera. For example, the camera used to capture our dataset can only resolve features up to $45\,cm$ at a distance of $15\,m$. Thus, even moderately distant facades are likely to be significantly more prone to large errors in their estimates; they will be both small in the frame and less likely to find an accurate consensus set in RANSAC due to the uncertainty in their disparity values. Similarly, for a facade with many invalid disparity values, it may not be sampled adequately, and the points it does have may erroneously be included as part of an inlier set that does not actually lie on the facade. Limitations of this nature will depend on the specific stereo imager used, but in general, a more accurate and dense disparity map will enable the discovery of a more accurate candidate plane set.

One of the drawbacks of our method is that for validation, we manually disambiguate inconsistent labels for our accuracy measures. For localization and navigation purposes, this may not be an issue, as a mobile platform may have have odometry information that can track the planes in the scene and help distinguish between these labels. In the absence of ground-truth information, the mobile platform could also consider the power set of candidate plane labels to find if one set corresponds well with its other semantic information about the surrounding area. However, we intend to investigate this area further in order to develop a more automated way of determining the best plane estimate for each facade in the absence of ground-truth information. Incorporating other semantic information such as vanishing points may help to improve the quality of both the candidate planes and their segmentations.

IV. Conclusion

We present a system for building facade segmentation from stereo images with parameter estimation of the identified planar surfaces. Our results show high accuracy in both detection of buildings ($\sim 90\%$) and their facades ($\sim 85\%$), and in estimation of their plane parameters ($\sim 10°$). The performance we have demonstrated indicates a promising step toward mobile robot localization and guidance via semantic scene understanding. Our intended future work in that direction includes refinement and improvement of our methods and eventual deployment on a mobile platform as part of a semantic guidance system. A first step will be to perform greater performance anaysis through cross validation with our full data set and the application of our method to further data sets that we intend to gather. We also plan to incorporate more semantic information (e.g. through vanishing point analysis or detection of other objects in the scene) into our framework to improve accuracy.

References

[1] J. J. Corso, "Discriminative modeling by boosting on multilevel aggregates," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[2] J. A. Delmerico, J. J. Corso, and P. David, "Boosting with Stereo Features for Building Facade Detection on Mobile Platforms," in *e-Proceedings of Western New York Image Processing Workshop*, 2010.

[3] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[4] J. J. Corso, D. Burschka, and G. Hager, "Direct plane tracking in stereo images for mobile navigation," in *IEEE International Conference on Robotics and Automation*, 2003.

[5] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[6] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2002.

[7] T. Korah and C. Rasmussen, "Analysis of building textures for reconstructing partially occluded facades," *Computer Vision–ECCV 2008*, pp. 359–372, 2008.

[8] A. Wendel, M. Donoser, and H. Bischof, "Unsupervised Facade Segmentation using Repetitive Patterns," *Pattern Recognition*, pp. 51–60, 2010.

[9] J. Hernández and B. Marcotegui, "Morphological segmentation of building façade images," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2010, pp. 4029–4032.

[10] P. David, "Detecting Planar Surfaces in Outdoor Urban Environments," ARMY Research Lab, Adelphi, MD. Computational and Information Sciences Directorate, Tech. Rep., 2008.

[11] J. Bauer, K. Karner, K. Schindler, A. Klaus, and C. Zach, "Segmentation of building models from dense 3D point-clouds," in *Proc. 27th Workshop of the Austrian Association for Pattern Recognition*. Citeseer, 2003, pp. 253–258.

[12] S. Lee, S. Jung, and R. Nevatia, "Automatic integration of facade textures into 3D building models with a projective geometry based line clustering," in *Computer Graphics Forum*, vol. 21, no. 3. Wiley Online Library, 2002, pp. 511–519.

[13] R. Wang, J. Bach, and F. Ferrie, "Window detection from mobile LiDAR data," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE, 2011, pp. 58–65.

[14] D. Gallup, J. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1418–1425.

[15] J. Kosecka and W. Zhang, "Extraction, matching, and pose recovery based on dominant rectangular structures," *Computer Vision and Image Understanding*, vol. 100, no. 3, pp. 274–293, 2005.

[16] W. Zhang and J. Kosecka, "Image Based Localization in Urban Environments," in *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*. IEEE Computer Society, 2006, pp. 33–40.

[17] D. Robertson and R. Cipolla, "An image-based system for urban navigation," in *Proc. of British Machine Vision Conf.*, vol. 1. Citeseer, 2004, pp. 260–272.

[18] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface reconstruction from unorganized points." *Computer Graphics(ACM)*, vol. 26, no. 2, pp. 71–78, 1992.

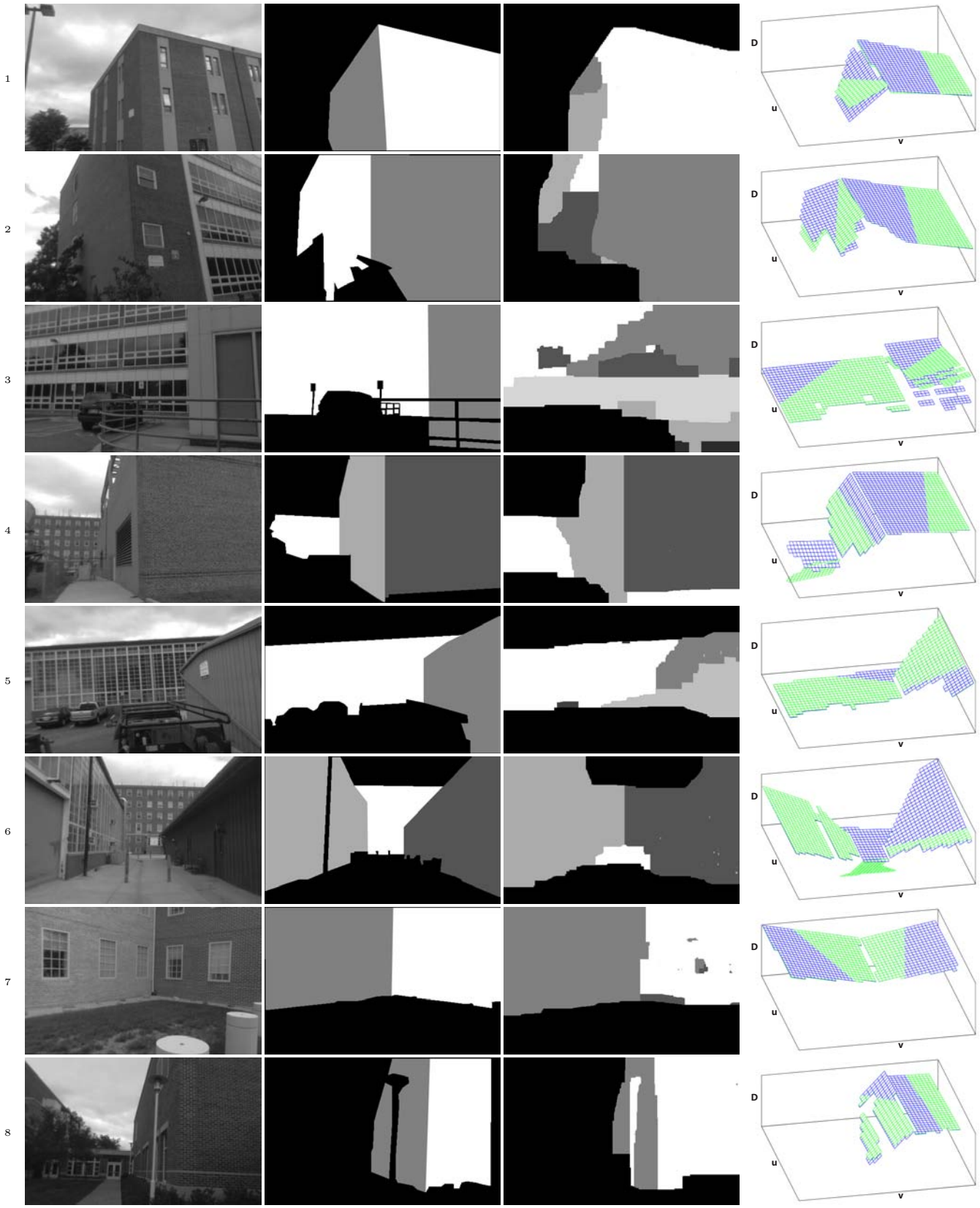

Fig. 5. Segmentations and planar facade estimates on multi-facade images. For each example, they are (L to R) the original image, ground truth segmentation, high-level MRF labeling, and 3D plane projection. In the plane projection plots, the perspective of the original image is looking down on the 3D volume from the positive D-axis. The ground-truth planes are in blue, and the estimated planes are in green (view in color).