Toward Generating Labeled Maps from Color and Range Data for Robot Navigation

Caroline Pantofaru, Ranjith Unnikrishnan, Martial Hebert
The Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213, USA
{crp,ranjith,hebert}@ri.cmu.edu

***Abstract*— This paper addresses the problem of extracting information from range and color data acquired by a mobile robot in urban environments. Our approach extracts geometric structures from clouds of 3-D points and regions from the corresponding color images, labels them based on prior models of the objects expected in the environment - buildings in the current experiments - and combines the two sources of information into a composite labeled map. Ultimately, our goal is to generate maps that are segmented into objects of interest, each of which is labeled by its type, e.g., buildings, vegetation, etc. Such a map provides a higher-level representation of the environment than the geometric maps normally used for mobile robot navigation. The techniques presented here are a step toward the automatic construction of such labeled maps.**

I. Introduction

The problem of building an internal model of a mobile robot's environment from sensor data has been investigated extensively, in particular in the context of indoor environments. The majority of the prior work focuses on the problem of reconstructing accurate geometric 2- or 3-D maps from sets of images or range scans [21]. Most of the work also concentrates on highly structured environments, such as indoor environments.

The next natural step is to extract a symbolic representation of the environment in which objects and structures of interest are labeled. In this paper, we consider the problem of extracting environment models from sensor data in the context of a robot driving through a typical urban environment.

Such a labeled map may be used by a planner for object-referenced navigation, using commands such as "Go to building X" rather than "Go to location (x,y)", or for providing a compact representation for a user interface.

The problem of scene understanding in its most general form is very challenging and largely unsolved. To make the problem manageable, we limit ourselves to a particular class of environments, outdoor urban environments. As a first step, we focus on the problem of extracting large structures, such as buildings, and segmenting the remaining clutter into individual objects.

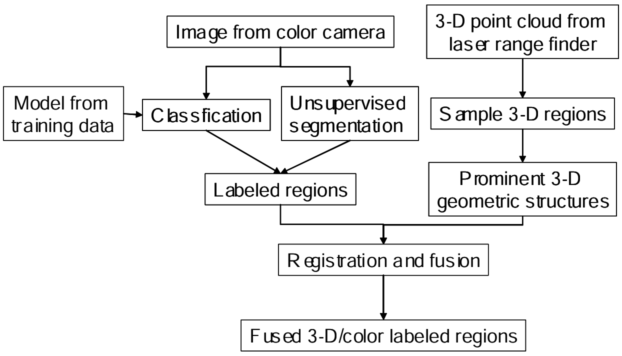


Image from color camera
3-D point cloud from laser range finder
Model from training data
Classfication
Unsupervised segmentation
Sample 3-D regions
Labeled regions
Prominent 3-D geometric structures
Registration and fusion
Fused 3-D/color labeled regions

Fig. 1. Overall approach to scene segmentation and classification from video and range data.

The concept is summarized in Fig. 1. Data from a laser range finder, providing geometric information, and data from cameras, providing appearance information, is combined to generate a map augmented with object labels. The image data is processed in two ways. First, a segmentation of the scene into homogeneous regions is generated. Second, each pixel is classified based on a statistical model of the distribution of features in different classes - in the examples presented here, the two classes are "building" and "non-building". The models are built from prior training data. Segmentation and classification are combined to form the final segmentation of the scene. As a final step, the 3-D structures extracted from the range data and the segmentation extracted from the image are combined into a single representation.

It is important to combine the two sources of information since they lead to different types of segmentations. For example, video imagery contains rich information on texture and color distributions but it is not suitable for interpreting scene geometry. This paper is organized as follows. We describe the extraction of geometric structures from 3-D data and the interpretation of images in Sec-

[1] This work was supported in part by DARPA MARS2020 program under Grant NBCH1020014.

tions II and III, respectively. We describe our initial results from combining the two sources of data in Section IV. The relevant prior work is discussed throughout these three sections. Additional details and results from the 3-D segmentation work (Section II) are described in a companion paper [23].

II. Extracting Geometric Structures from 3-D Data

In this section, we consider the part of the system that takes as input a cloud of 3-D points from, for example, a scanning laser range finder. From this point cloud we extract prominent geometric features, such as large planar structures corresponding to walls, and segment the remaining clutter into clusters corresponding to different objects. The planar structures are combined with the regions extracted from images (as in Section III) to form hypotheses of building locations. Details of the estimation procedure are described in a companion paper [23]. Here, we summarize the components of the system used for fusion with the image data.

For our experiments, we used a SICK laser range finder actuated so that it scans in two directions, producing a point cloud in a 100° field of view with 0.25° angular resolution sub-sampled by 20. The maximum range is 80m. We consider here the general problem of extracting structure from point clouds, rather than assuming a specific spatial arrangement of the points, so that the technique is applicable to point clouds accumulated over several scans. Fig. 2 shows an example data set, sub-sampled for display. Note that the density of points on the structures - walls of buildings - varies substantially with distance, and that the amount of clutter - bushes and trees - is large and cannot be modeled by a simple noise model.

The problem of extracting structure from 3-D point clouds is challenging for two main reasons. First, the density of data varies drastically over the entire scanning volume, due to the large variations in scanning angle and in range. This rules out any technique that relies on fixed prediction of expected data density. Second, a large amount of clutter is always present and partially obscures the 3-D structures. The clutter does not follow any simple predictive noise model.

Popular approaches towards recovering planar structure from outdoor range data have addressed the problem in either the original input domain (3-D points) or in the sampling domain (as range images). Work in [1] identifies planar regions from raw 3-D points to build photorealistic models of buildings. The limitation of this approach lies in its dependence on high point density and low clutter. Recent work [10] tackles a more challenging scenario, segmenting buildings from foreground clutter for building textured meshes of building facades. Our work attempts to tackle the more general scenario where buildings are observed at varied distances, the manner of data acquisition is not controlled, and a global prior model of noise and clutter distribution is not defensible.

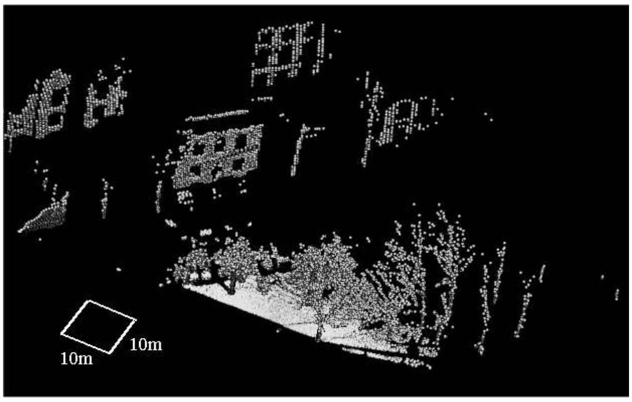


Fig. 2. Oblique view of points sampled by a laser rangefinder. The triangular noise region near the front of the image corresponds to laser readings off the ground plane.

Work in [15] uses a variant of the EM algorithm to fit planar surfaces to indoor laser range data. The number of planar regions is controlled by using the AIC model selection criterion. Extensions to the standard model selection criterion that are robust to more realistic noise models have been proposed. However, such scoring criteria are typically biased toward explaining a larger fraction of data points, even though the models themselves may be of low complexity. Hence the approaches mentioned earlier are limited in practice to controlled or low-clutter environments. Recently, kernel-based approaches [5] have been proposed to deal with the scale variation issue, and robust estimators for extracting structures in that framework have also been introduced [6]. We use this approach as the basis for our system and we summarize below the main extensions to the basic estimation techniques.

In this approach, the point cloud is divided into voxels which are grouped into 3-D regions grown from randomly selected seed voxels. The selection of the seed voxels and the growing of the regions are guided by voxel-specific weights that are proportional to the density of data within that voxel.

For each region, the parameters of a possible planar structure are estimated. In order to take into account the potentially high percentage of clutter points in the region, the planar structure is estimated by using a robust estimator. Following the notations of [6], if the measurements in a sample region are $\gamma_i \in \Re^3, i = 1,..,n$, the parameter vector is defined for a planar structure as $\Theta = (\alpha, \theta)$ such that:

$$\gamma_i^T \theta - \alpha = 0,\ i = 1,\ldots,n,\ \|\theta\| = 1$$

Θ is estimated from the data points by using a robust estimator ρ:

$$[\hat{\alpha},\hat{\theta}] = \arg\min_{\alpha,\theta} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{1}{s}\|\gamma_i^T\theta - \alpha\|\right)$$

Equivalence between this robust estimation formulation on the one hand, and kernel estimation and projection pursuit estimation in statistics on the other hand, are derived in [6] and used in [23].

Once the vector of structure parameters is estimated for every sample group, the distribution of these vectors is analyzed in order to determine its modes. Informally, a parameter vector at a mode of this distribution is one that is consistent with a large number of sampled 3-D regions and is therefore a dominant structure of the point cloud. Once the structured regions are extracted, the remainder of the point cloud is segmented into groups of connected objects.

This general approach does not immediately scale to problems with large variations of data density and large amounts of clutter. In [23], we describe extensions to this approach that make it practical for real-world application. We summarize below the key extensions.

A crucial extension of the algorithm is to modify the weights associated with each voxel so that they characterize local planarity as determined by the distribution of the eigenvalues of the scatter matrix, instead of the local density of data only. When used in the selection of sample groups, these weights take into account the variations in data density and the presence of clutter.

A second extension of the algorithm post-processes the regions found as dominant structures. An additional test is applied to all the points within the regions and only those points that are consistent with the structure parameters are retained. This test is based on a robust estimate of the variance of the structure parameters. Finally, to address the fact that large structures may be split because, for example, they are partially obscured by clutter, 3-D regions with similar parameters are merged based on a similar statistical test.

Fig. 3 shows an example output in which the boxes indicate the location of the detected structures and different colors indicate the objects in the rest of the data. In this example, the extracted structure corresponds to the walls of buildings, and the clutter regions are primarily bushes and trees.

III. Segmenting and Interpreting Images

The scene segmentation process has three key ingredients. First, we need to extract features that will facilitate segmentation of the scene into regions and the classification of those regions into known object classes. Second, we need to actually segment the image into regions corresponding to coherent distributions of those features. This is an unsupervised step in that it does not rely on prior models of the feature distributions. Third, we need to classify the regions based on learned feature distributions for each object class. This is a supervised step in that it uses models built from prior data. In principle, it would be possible to use only supervised classification for scene analysis. However, in practice, it is not possible to build sufficiently accurate models and the classification step must be combined with the result of the unsupervised segmentation.

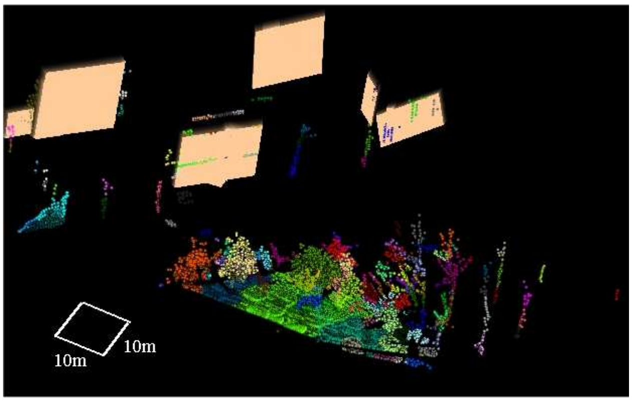


Fig. 3. Output of point cloud processing: Structure regions (walls) and clutter regions (trees and bushes). The beige walls indicate bounding boxes of detected planar points. Groups of non-planar points (clutter) are shown in different colors.

The problem of scene understanding from images has been explored extensively in prior work. Of particular interest here is prior work in the interpretation of natural images and the identification of building structures in natural images. Techniques based on perceptual organization have been proposed for the detection of buildings in a scene for image retrieval [11]. In [18] a technique was proposed to learn the parameters of a large perceptual organization using graph spectral partitioning. However, these techniques require the low-level image primitives to be computed explicitly, and to be relatively noise-free. Recent work addresses the classification of the whole image as a landscape or an urban scene [16][24]. Oliva and Torralba [16] obtain a low-dimensional holistic representation of the scene using principal components of the power spectra. Vailaya et. al. [24] use the edge coherence histograms over the whole image for scene classification, using edge pixels at different orientations. Olmos and Trucco [17][9] have recently proposed a system to detect the presence of man-made objects in underwater images using properties of the contours in the image. The techniques which classify the whole image in a certain class implicitly assume the image to be exclusively containing either man-made or natural objects, which is not true for many real-world images.

The techniques described in [9][13] perform classification in outdoor images using color and texture features, but employ different classification schemes. These papers report poor performance on the classes containing man-made structures, since color and texture features are not very informative for these classes [24].

A. Features

Our definition of features follows the one described in [12]. The input image is first divided into non-overlapping 16×16 pixel blocks. A feature vector is generated for each block independently. The features are computed at each block instead of at each pixel because integration over image blocks is necessary to compute the more complex features required for region classification.

The basic features used for region segmentation are the average color within each block after mapping to L*u*v space (3 components) and the spatial position of each block (2 components). This is a standard configuration for image segmentation; the color features ensure appearance consistency within each region, while the position features ensure spatial connectivity of the regions.

For the application considered here, however, it is necessary to also use features that are better tuned to the segmentation of regions corresponding to man-made structures, such as buildings. Approaches have been developed in prior work for identifying man-made structure in scenes containing both man-made and natural elements using the distribution of a multi-scale feature vector as a classifier. Here, we use these multi-scale features for segmentation as well. These features attempt to capture the linearity of the lines and edges in man-made structures, as opposed to the less structured lines and edges in natural objects. This is accomplished using a set of 14 features derived from histograms of gradient orientations in a region weighted by gradient magnitudes - termed “orientograms” - generated at three different scales; 1×1, 2×2, and 4×4 blocks. Each orientogram is smoothed using kernel smoothing in order to minimize the aliasing effects due to binning.

Fig. 4 shows examples of orientograms computed at three different points (“building”, “tree”, and “car”) in a typical image. Note that the orientograms for the building point show clearly two extrema separated by approximately 90°, while the orientograms at the tree point are more uniformly distributed.

Intrascale features are then computed from the orientograms at each scale separately. The first six features are the first and third heaved central-shift moments at each of the three scales. The next three features are the absolute locations of the highest bin at each scale. These features capture the dominant direction within each block at each scale. The next three features examine the relationship between the two most dominant orientations at each scale. Let δ_1 and δ_2 be the two dominant orientations at a given scale, then the offset feature is $\beta = | \sin(\delta_1 - \delta_2) |$.

It is also desirable to model the dependence of the features in neighboring blocks, since structures often take up multiple adjacent blocks. Hence, a set of interscale features are computed between the first and second, and second and third, scales. Let δ_1^i and δ_2^i be the highest and second highest peaks at scale i, respectively, then $\beta_p = | \cos 2(\delta_p^i - \delta_p^{i+1}) |$. This choice of features was shown in [12] to be well-suited for discrimination between urban structures and natural environments. Intuitively, for buildings, the orientograms tend to be uni- or bi-modal, and the angular difference between peaks tends to be distributed near $\frac{\pi}{2}$.

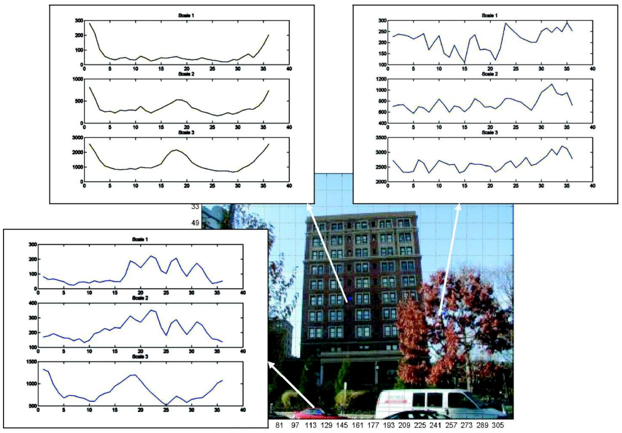


Fig. 4. Orientograms computed at three sample points. Three orientograms corresponding to the three different scales are displayed for each point. The orientograms show the total magnitude (vertical axis) of the gradient with respect to orientation (horizontal axis) integrated over three different neighborhood sizes.

In summary, we use a 19-component feature vector. All of the components are used in the segmentation stage and only those 14 components computed from the orientograms are used for classification in Section III-C.

B. Segmentation

Given the features described above, the next step is to generate a segmentation of the image into regions that are uniform with respect to these features. Several classical approaches can be used for segmentation including spectral techniques [19] and parametric clustering techniques [2]. We prefer to use the non-parametric mean shift segmentation technique proposed in [7] as the basis for our segmentation algorithm. This is because the spectral method is sensitive to the exact definition of a distance or a similarity metric in feature space, which is difficult in our case. Also, the parametric method does not perform well when the clusters do not follow the parametric model, and it requires the explicit estimation of the cluster parameters, which we do not need.

The mean shift technique of [7] performs segmentation by finding the modes of the distribution features in feature space, and by computing which mode corresponds to each

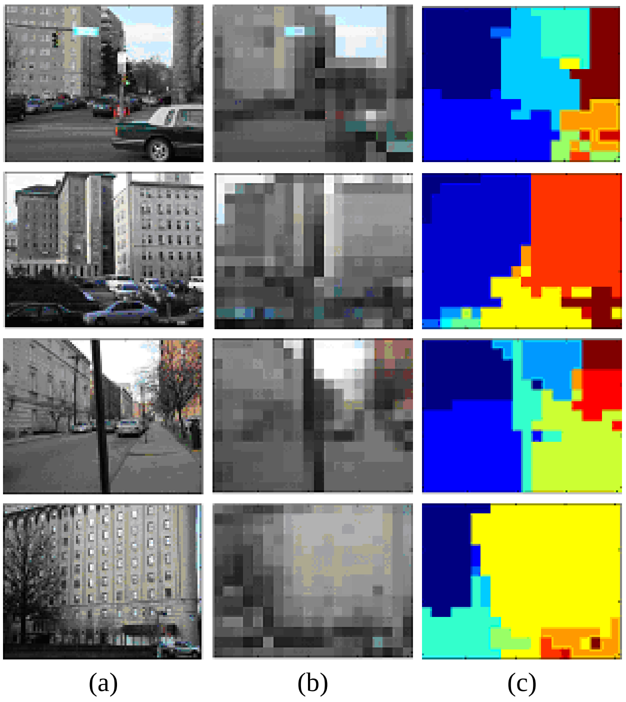


Fig. 5. Results from unsupervised segmentation using the features of Section III-A. (a) Input image; (b) Blocks used for computing the features; (c) Initial segmentation.

pixel in the input image. Regions are extracted from the input image by grouping those pixels that are attracted to a common mode into a single image region.

We adapt this approach to our problems as follows. The initial mean-shift filtering is applied to pixel blocks rather than to individual pixels. This reflects the fact that the multi-scale features are computed at the block-level. We use a uniform kernel in the mode-finding step. The final regions are defined by grouping the modes extracted by mean-shift filtering into clusters. The blocks corresponding to each individual cluster are grouped into a separate region. We use Kruskal's algorithm for this final clustering step.

We tested the image-based interpretation approach on sequences of images taken in an urban environment. Fig. 5 shows results of the unsupervised segmentation on typical real-world images.

C. *Classification*

The 14 orientogram-based features, i.e. all the features described above except for color and block position, have been shown to be effective in separating man-made structures from unstructured elements in natural images by using a Bayesian classification approach [12]. Specifically, parametric models of the feature distributions of each of man-made structures and natural scene elements are constructed from training data. The models of the feature distributions are Gaussian Mixture Models (GMM) with three mixture components. The GMM allows us to compute explicitly, for any block with a feature vector f, the probability that the block was generated by a structured object, $P(S \mid f)$. For classification, every block i in the input image is classified based on the likelihood ratio

$$l_i = \frac{P(S \mid f_i)}{(1 - P(S \mid f_i))}.$$

It should be noted that one difficulty is that this approach assumes that each block is classified independently, which assumes implicitly that the feature vectors f_i and f_j at neighboring blocks i and j are computed independently. In fact, that is not the case since the neighborhoods used for computing f_i and f_j do overlap. This has been noted in prior work in which random fields models [3][26] or Bayesian networks [9] are used in order to take into account the dependency between features computed at neighboring locations. Such an approach that uses the features described here is introduced in [12]. The approach is based on a Multi-Scale Random Field (MSRF) model and it has shown good results in detecting structure in images containing both structured and non-structured classes. This model was originally proposed by Bouman and Shapiro [3] and further used in [9] for semantic image segmentation.

We use this approach to train a GMM model on images of building and non-building regions taken in urban environments. In the current experiments, we use 55 images for training for a total of 4814 blocks on building regions and 11686 blocks on non-building regions.

There are several issues that the block-level classification procedure of [12] cannot address. First, an urban scene is filled with man-made structures which do not belong to the class we are interested in. This causes the models for the building and non-building classes to have a large overlap, which leads to much higher false positive and false negative rates. Second, the classifier misses parts of buildings that do not have strong features, such as bare walls. Finally, the classifier does not deal with the issue of spatial consistency in a region larger than a 4×4 block, which a building may occupy in a ground-level urban scene. For these reasons, block-based classification is not sufficient and needs to be integrated with the unsupervised segmentation described in Section III-B.

D. *Integration*

The last step is the integration of the mean shift segmentation with the classification from the multi-scale features. Consider an image region as extracted by the mean shift segmentation. We wish to classify this region as either building or non-building. Given the classification for each block in the region as defined by the multi-scale feature set distribution method described above, we can compute

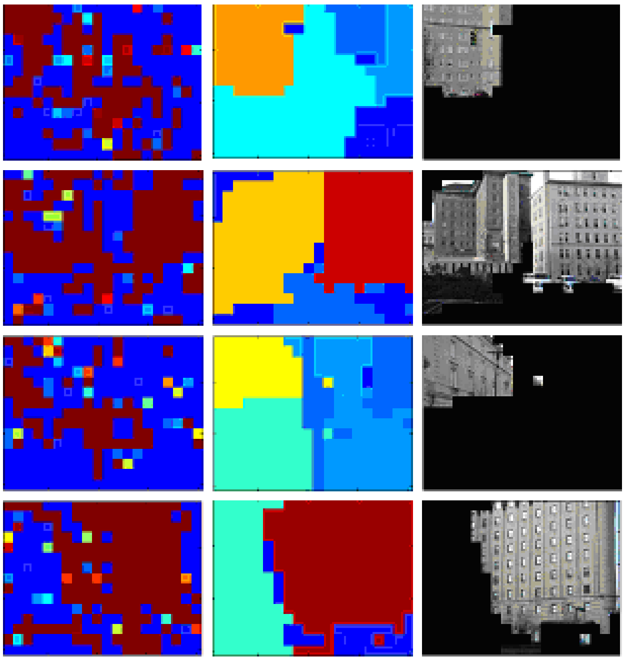


Fig. 6. Combination of block-wise classification and unsupervised image segmentation. (a) Confidence of classification of blocks as type "building" from red = high confidence to blue = low confidence; (b) Final segmentation; (c) Buildings extracted from the image based on the final segmentation.

a confidence measure for the region being classified as a building using the mean of the likelihood ratios l_i of the component blocks $i = 1,\ldots,n$ of the region, passed through a sigmoid function:

$$c = \frac{1}{n}\sum_i \omega_i \frac{1}{1+e^{-\alpha(l_i-1)}}$$

where c is the overall region confidence, α is a constant. A final binary building/non-building classification is obtained by thresholding the confidence. Since the multi-scale features are evaluated at 1×1, 2×2, and 4×4 block scales, there are some border effects which we need to eliminate. Namely, the borders of non-building regions may be classified as building blocks since their features overlap with the buildings at some scales. To lessen this problem, we have eroded each segmented region twice, and weighted the blocks found to be on each border less than the blocks in the central region. This is reflected by the ω_i weight in the formula above.

Fig. 6 illustrates the integration process on the examples of Fig 5. These examples illustrate the importance of combining unsupervised segmentation with supervised classification: by evaluating each region for class membership instead of each block, we have taken into account the spatial consistency of objects in an image. Although, as anticipated, the block-wise classification does make a substantial number of mistakes, blocks which were mistakenly or weakly classified with the multi-scale feature method alone are now combined with other blocks in their region which were correctly classified. Similarly, regions that are artificially separated by the unsupervised segmentation are grouped after integration with the classification result. Finally, clutter in the image is segmented away from the building regions by the mean shift segmentation in the preprocessing step. Since it has, on average, lower confidence than the building regions, the multi-scale feature classifier now has better accuracy on this clutter.

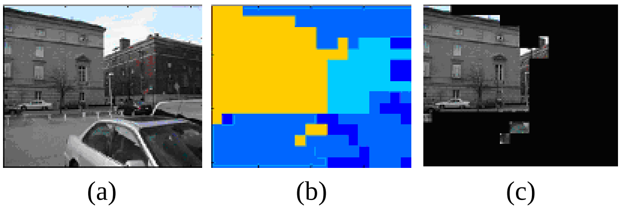


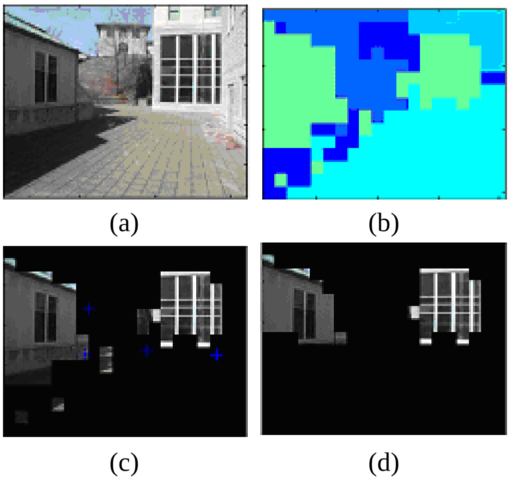
Fig. 7. Example of incorrect classification. (a) Input image; (b) Combined segmentation/classification output; (c) Extracted "building" region.

The classification tends to perform poorly on images in which surfaces do not have enough texture or in which surfaces are too slanted with respect to the viewing direction. In the first case, not enough information can be extracted to correctly classify the surface; in the second case, the distribution of the features is too different from the one learned from images of surfaces at more typical orientations. Fig. 7 shows one example of a failure in which the far away building is correctly segmented but incorrectly classified.

IV. Putting it all together

The LADAR and video sensors are calibrated with respect to each other by using calibration targets visible from both sensors. After calibration, it is a simple matter to merge the two data sets into a combined colored 3-D point cloud. However, a more difficult question is how to combine the structures extracted from the 3-D data and the region obtained from the segmentation and classification algorithms in the color images. In particular, segmentation from LADAR and color data leads to different types of errors. Therefore, it is important to combine the segmentations in a way that best compensates for the errors in the two sources of scene interpretation. The approach we follow is to first process separately the range and color data sets. Then, those points from the 3-D data that have been identified as belonging a common geometric structure, e.g., a single plane, are mapped to the color image. Finally, the formula used in Section III-C to integrate the classification likelihood over each region found by the image segmenter is modified as:

$$c = \frac{1}{n}\sum_i k_i\omega_i \frac{1}{1+e^{-\alpha(l_i-1)}}$$

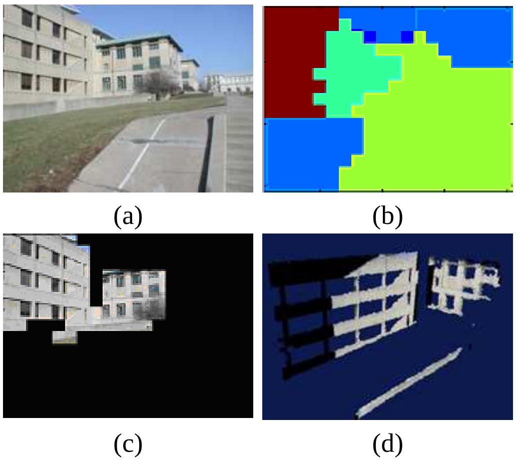


(a) (b) (c) (d)

Fig. 8. Example of incorrect classification. (a) Input image; (b) Segmentation from image data; (c) Regions classified as buildings from the image only; (d) Classification after fusion of the range data.

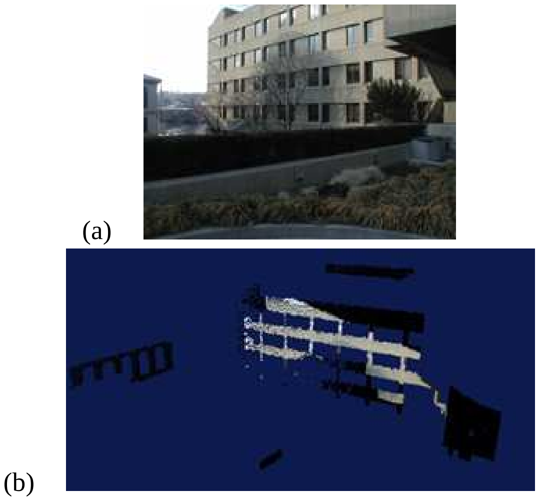
The additional weight k_i is proportional to the likelihood that point i belongs to a plane. This fusion has the following effects: It increases the weights of points that are correctly classified in the 3-D data so that regions that are mis-classified in the image can be corrected. Conversely, those regions that have a very low classification likelihood in the image, i.e., we know that they do not belong to a building, will remain correctly labeled even if the corresponding points are labeled as being on planes in the 3-D data, so that errors in the 3-D data segmentation are corrected.

Fig. 8 shows an example in which the classification from the image - (b) and (c) - contains several small mis-classified regions near and on the ground plane. After fusion with the 3-D segmentation, the two correct building regions are extracted (d). Fig. 9 shows another example of fused segmentation. Fig. 9(d) shows an oblique view of the 3-D data points texture-mapped using the values from the corresponding pixels in the color image. Only the points that fall on regions classified as structure regions are displayed here, showing that the correct scene interpretation is generated from the fusion of range and color data. Because of the different fields of view of the range and color sensors, those 3-D points that do not have corresponding points in the color image are shown in black in Fig. 9(d). Fig. 10 shows a similar result on another example. These examples show that this approach produces the type of output that we defined initially: The locations of labeled regions (buildings) in the environment with their precise location and shape, as well as a set of other objects segmented in the rest of the environment. This type of output is a first step toward a fully labeled map of the environment.



(a) (b) (c) (d)

Fig. 9. Example of fused segmentation. (a) Input image; (b) Segmentation from the color image; (c) Regions classified as building structures from the image; (d) Oblique view of the 3-D points on the building structures after fusion.



(a) (b)

Fig. 10. Another example of fused segmentation.

V. Conclusion

We have presented an approach for the combined segmentation of video and 3-D data as a first step toward generating labeled environment maps. The current approach extracts 3-D structures and regions from the data and identifies the regions corresponding to building structures. The salient features of the approach are the use of multi-scale features adapted to structure detection, the fusion of unsupervised image segmentation with image classification from models of feature distributions, and the fusion of the outputs of the 3-D and image segmentation. Many issues remain to be addressed.

First, the choice of image features can be improved. In particular, these features have been shown to be effective in tasks that involve extracting man-made structures in natural environments. In our task, which involves scene interpretation in urban environments, the features are no longer optimal. This can be verified empirically as the classification rate is substantially worse than in the case of natural images.

Second, the statistical techniques used for structure extraction from 3-D point clouds provide a sound basis. However, more work is needed to incorporate sensor-specific information, such as the directionality of the measurements, to incorporate an explicit model of sensor noise, and to extend the approach to the extraction of complex structures. Also, other approaches for combining the segmentation results from 3-D data and image data need to be explored. For example, one alternative approach would be to combine information from the 3-D data and the features from the image into combined features prior to segmentation and classification.

Finally, the current approach does not take advantage of obvious contextual information to guide the scene analysis. For example, we know that a particular class of objects has different probabilities of occuring in different parts of the image. This information could be used to compute priors to bias the classification as was suggested by Torralba [22], for example.

VI. REFERENCES

[1] P. K. Allen, I. Stamos, A. Troccoli, B. Smith, M. Leordeanu and Y. C. Hsu, "3D Modeling of Historic Sites using Range and Image Data", *Proc. of the IEEE Intl. Conference on Robotics and Automation*, 2003.

[2] S. Belongie, C. Carson, H. Greenspan, J. Malik, "Color- and texture-based image segmentation using EM and its application to content-based image retrieval", *Proc. of the Intl. Conf. on Computer Vision*, 1998.

[3] C. A. Bouman, M. Shapiro, "A Multiscale Random Field Model for Bayesian Image Segmentation", *IEEE Trans. On Image Processing*, Vol. 3, No. 2, 1994.

[4] B. Bradshaw, J. C. Platt, B. Scholkopf, "Kernel Methods for Extracting Local Image Semantics", *Tech. Report MSR-TR-2001-99*, Microsoft Research, 2001.

[5] H. Chen, P. Meer, "Robust computer vision through kernel density estimation", *Proc. of the European Conference on Computer Vision*, 2002.

[6] H. Chen, P. Meer, D.E. Tyler, "Robust regression for data with multiple structures", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001, vol. I, 1069-1075.

[7] D. Comaniciu, V. Ramesh, P. Meer, "The variable bandwidth mean shift and data-driven scale selection", *Proc. of the Intl. Conference on Computer Vision*, 2001.

[8] F.Dell'Acqua, F.Stulp, R.B.Fisher, "Reconstruction of surfaces behind occlusions in range images", *Proc. of the Intl. Conf on 3-D Digital Imaging and Modelling*, 2001.

[9] X. Feng, C. K. I.Williams, S. N. Felderhof, "Combining belief networks and neural networks for scene segmentation', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 4, 2002.

[10] C. Fruh, A. Zakhor, "Data processing algorithms for generating textures 3D building faade meshes from laser scans and camera images", *3D Data Processing, Visualization and Transmission (3DPVT)*, 2002.

[11] Q. Iqbal, J. K. Aggarwal, "Applying Perceptual Grouping to Content-Based Image Retrieval: Building Images", *Proc. of the IEEE Intl. Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 42-48, 1999.

[12] S. Kumar, M. Hebert, "Man-Made Structure Detection in Natural Images using a Causal Multiscale Random Field", *Proc. of the IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2003.

[13] S. Konishi, A. L. Yuille, "Statistical Cues for Domain Specific Image Segmentation with Performance Analysis", *Proc. of the IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2000.

[14] S. Kumar, M. Hebert, "An Observation-Constrained Generative Approach for Probabilistic Classification of Image Regions", *Image and Vision Computing*, Vol. 21, No. 1. January 2003.

[15] Y. Liu, R. Emery, D. Chakrabarti, W. Burgard, S. Thrun, "Using EM to learn 3D models with mobile robots", *Proc. of the Intl. Conf. on Machine Learning*, 2001.

[16] A. Oliva, A. Torralba, "The Shape of the Scene: a Holistic Representation of the Spatial Envelope", *Intl. Journal of Computer Vision*, Vol. 42, No. 3, 2001.

[17] A. Olmos, E. Trucco, "Detecting Man-Made Objects in Unconstrained Subsea Videos", *Proc. of the British Machine Vision Conference*, pp. 517-526, 2002.

[18] S. Sarkar, P. Soundararajan. "Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 5, pp. 504-525, 2000.

[19] J. Shi, J. Malik, "Normalized cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8 , 2000.

[20] I. Stamos, P. Allen, "3D Model Construction Using Range and Image Data", *Proc. of the IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2000.

[21] S. Thrun, W. Burgard, D. Fox, "A Real-Time Algorithm for Mobile Robot Mapping With Applications to Multi-Robot and 3D Mapping", *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2000.

[22] A. Torralba, P. Sinha, "Statistical context priming for object detection", *Proc. of the Intl. Conference on Computer Vision*, 2001.

[23] R. Unnikrishnan, M. Hebert, "Robust Extraction of Multiple Structures from Non-uniformly Sampled Data", *To appear in the Proc. of IROS'2003*, 2003.

[24] A. Vailaya, A. K. Jain, H.-J. Zhang, "On Image Classification: City Images vs. Landscapes", *Pattern Recognition*, Vol. 31, pp. 1921-1936, 1998.

[25] R. Wilson, C. T. Li, "A Class of Discrete Multiresolution Random Fields and Its Application to Image Segmentation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 25, No. 1, pp. 42-56, 2003.

[26] C. S. Won, H. Derin, "Unsupervised Segmentation of Noisy and Textured Images using Markov Random Fields", *CVGIP: Graphics Model and Image Processing*, Vol. 54, pp. 308-328, 1992.