Fast Scene Analysis using Image and Range Data

Camillo J. Taylor and Anthony Cowley

Abstract— **This paper presents a scheme which takes as input a 3D point cloud and an associated color image and parses the scene into a collection of salient planar surfaces. The scheme makes use of a fast color segmentation scheme to divide the color image into coherent regions and the groupings suggested by this procedure are used to inform and accelerate a RANSAC based interpretation process. Results on real data sets are presented.**

I. INTRODUCTION

Over the past decade the advent of smaller, cheaper range sensors has made it more attractive to field robots that can acquire 3D range maps of their environment. Early systems made use of single scan range finders such as the SICK or Hokuyo sensors which were mounted on moving platforms or pan-tilt heads and scanned across the scene to produce a 3D point cloud. More recently, range sensors such at the SR 4000 'Swiss Ranger' from Mesa Imaging and the Velodyne scanning range sensor have been used to produce two dimensional range images at high frame rates.

The recently announced 2D range camera systems from Canesta and Primesense promise to further accelerate this trend by providing real time range imagery at a very compelling price point. The Primesense sensor, which will be employed in the Xbox Kinect system, is a particularly interesting example since it acquires a color video stream along with the range imagery which makes it easier to deploy schemes that exploit both sources of information simultaneously.

These developments prompt us to consider the following research question; how do we go about programming our robots to make use of the volumes of raw data that these sensors can produce? Ideally we would like to endow our robots with the ability to extract relevant high level percepts from the stream of sensor data. For instance, in an indoor environment it would be useful for the robot to be able to quickly detect relevant objects such as walls, doors, windows, tables and chairs.

As a step towards this goal this paper proposes a scheme which can be used to rapidly parse a scene into a collection of planar surfaces as illustrated in Figure 1. The algorithm takes as input a 3D point cloud and an associated color image. The system then makes use of a recently developed real time segmentation scheme to divide the color image into coherent regions. These regions are then used to suggest groupings of the 3D points to a RANSAC based interpretation process which extracts the relevant planar surfaces. Figure 1a shows an example of an input image that has been passed through the segmentation process. The associated point cloud has been projected into the frame and overlaid on the image. Figure1 b shows the planar surfaces that have been extracted by the procedure, each of the recovered planes is assigned a unique color.

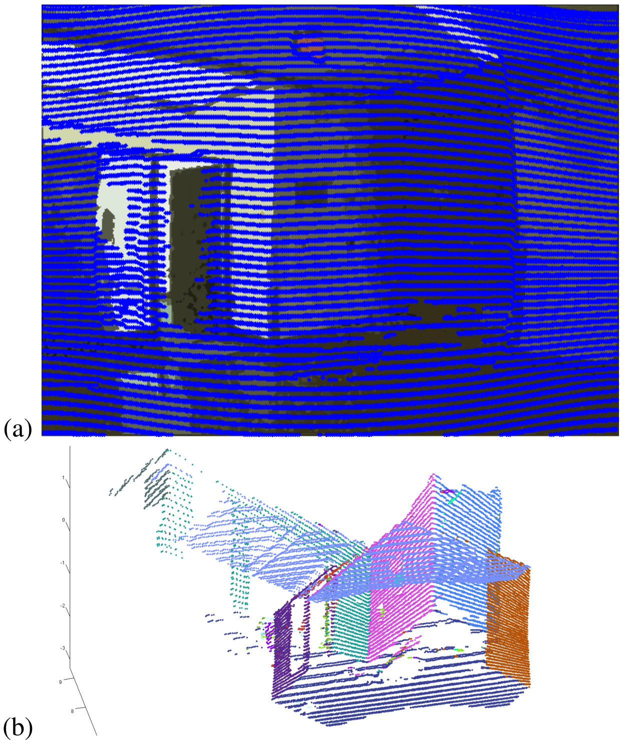

Fig. 1. This figure shows the input to and output from the scene interpretation process. (a) shows the input image after color based segmentation with the 3D points projected into the frame (b) shows the 3D planes recovered by the procedure. (This figure is best viewed in color.)

Image segmentation is a well studied problem in the Computer Vision literature and a number of effective algorithms have been developed including the Mean Shift [3] method proposed by Comaniciu and Meer, the Normalized Cut method proposed by Shi and Malik[9], and the superpixel method developed by Felzenswalb and Huttenlocher [5]. Recently, segmentation schemes based on the pB edge detector developed by Arbelaez et al. [2] have been shown to produce state of the art results on standard datasets.

One issue with these segmentation methods is that they all require a considerable amount of computational effort and, as such, they are not ideally suited to a robotic context where

C.J. Taylor is on the faculty of Computer and Information Science, University of Pennsylvania, Philadelphia PA, USA `cjtaylor@cis.upenn.edu`

A. Cowley is with the GRASP Laboratory, University of Pennsylvania, Philadelphia PA, USA `acowley@cis.upenn.edu`

we are typically interested in processing large volumes of data quickly. The segmentation scheme that will be leveraged in this paper makes use of an efficient randomized hashing scheme which produces competitive segmentation results in real time with modest computational effort. This makes it particularly attractive for the types of applications we envision.

Several schemes have been proposed to address the problem of interpreting point cloud data. Rusu et al. [8] describe an impressive system for parsing range scans acquired from indoor kitchen scenes. Toshev et al. [11] describe a scheme that has been used to automatically parse range scans to produce building models. Anguelov et al. [1] and Lalonde et al. [6] describe schemes for classifying regions in point cloud data sets to identify, buildings, trees and other structures.

Most of these schemes were designed to work offline in a batch fashion. In this context one can afford to make several passes over the data to identify nearest neighbors, or to fuse neighboring regions. The goal in this work is to develop a scheme that can ultimately be run in an online fashion so that it can be used to parse the data as it is being acquired. Another salient difference is the fact that this approach seeks to exploit the relationship between the 2D range image and the associated imagery to accelerate the interpretation process.

The idea that one can make use of the mutual information between color and range imagery has also been explored by Dolson et al. [4] and by Yang et al.[12]. These papers describe innovative techniques that can be used to upsample a relatively low resolution range image to the resolution of an associated color image. In contrast, the goal in this work is not to upsample the range imagery but rather to provide an interpretation in terms of salient planar surfaces.

The idea of using image segmentation information to constrain or inform a RANSAC based interpretation process has also been suggested by Ni, Jin and Dellaert [7]. In this work, these researchers show how image based groupings can be used to improve the search for the fundamental matrix that relates two views. Here we make use of image segmentations to search for structurally salient regions in the scene. This interpretation process will typically return multiple groupings in the scene as opposed to a single interpretation of the data such as a fundamental matrix.

Section 2 of this paper presents the approach that has been developed to interpret the range and image data. Section 3 describes some of the results that have been obtained by applying this approach to actual data sets. Finally Section 3 discusses conclusions and future work.

II. TECHNICAL APPROACH

A. Data Acquisition

In our experiments range and image data were acquired using the PR2 humanoid robot from Willow Garage shown in Figure 2. This platform is equipped with a Hokuyo range finder mounted on a tilting platform and an array of cameras mounted on a pan-tilt head. A calibration procedure was performed to recover the intrinsic parameters of the cameras and the geometric relationship between the camera platform and the range finder. This allows us to project the 3D points acquired with the range scanner onto the image as shown in Figure 1. This projection procedure takes into account the radial and tangential distortion in the camera projection model and the displacement between the range scanner and the camera.

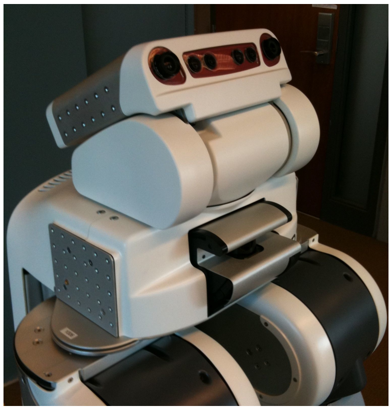

Fig. 2. The PR2 humanoid robot was used to capture the input imagery and range scans. This platform is equipped with an array of cameras and a Hokuyou range scanner mounted on a tilting platform.

While the PR2 platform was a convenient choice for development and debugging, the algorithms that have been developed do not rely on any special capabilities of this robot. The only assumption is that the 3D point cloud can be accurately registered to the associated color imagery. As such, the scheme could equally be applied to the data produced with a Swiss Ranger and a color camera or to the imagery gathered with a Primesense 'Kinect' sensor.

B. Image Segmentation

As previously mentioned, the parsing procedure makes use of a novel segmentation scheme based on randomized hashing which is described in more detail in [10]. The segmentation scheme employs a feature based approach. Each pixel in the image is represented by a feature vector which encodes a set of properties used to describe that pixel. In all of the experiments described in this paper, we employ a simple HSV color descriptor vector but one could equally easily use more sophisticated feature vectors such as a histogram of color values or a vector of texture coefficients.

Given this set of feature vectors, the goal of the segmentation procedure is to divide them into a set of clusters which capture the most salient groupings in the distribution. To do this, the scheme employs a randomized hashing procedure where the feature vectors are hashed onto binary codes using a series of randomly chosen splitting planes.

For each of the hash codes the clustering procedure records how many feature vectors are mapped to that code. We expect that clusters in feature space will induce population maxima in the code space. That is, if we interpret the set of hash codes as the nodes of a hypercube graph we would expect to observe that some of the hash codes have a greater population than their neighbors. This motivates us to replace the original problem of clustering vectors in the feature space with the simpler problem of looking for population maxima in the code space graph.

The scheme is similar in spirit to the Mean Shift segmentation algorithm which also seeks to identify modes in the distribution of feature vectors. Where the mean shift algorithm uses a Parzen Window based scheme to estimate density in feature space, this scheme uses randomized hashing to identify relevant groupings of feature vectors.

A significant advantage of the proposed segmentation scheme is that the computational effort required scales linearly in the number of pixels and the operations required are simple and regular. In order to demonstrate this fact, a real time version of the scheme was implemented on a Macbook Pro laptop computer. This implementation was used to segment 640 by 480 video frames at a rate of 10 frames per second using a single core of an Intel Core 2 Duo processor running at 2.33 GHz. This rate includes the time taken for all phases of the algorithm, image acquisition, randomized hashing, local maxima detection and connected components processing. Since almost all of the steps in the procedure are embarrassingly parallel, the algorithm is a well suited to implementation on modern multi-core processors and GPUs and should be amenable to further acceleration.

C. Scene Interpretation

Once the color image has been segmented, all of the 3D points that have been projected into the frame can be tagged with the label of the image segment that they project to. This provides us with a vey useful partitioning of the 3D point data sets into subsets that are quite often semantically meaningful.

We use this partitioning to suggest groupings to a RANSAC based interpretation process. More specifically the scheme considers each of the image regions in turn and pulls out all of the 3D points that project to that segment. Using this subset, the system runs a RANSAC loop wherein it randomly selects groups of 3 points and constructs the plane passing through those selections. Each candidate plane is scored based on the number of inliers it attracts in the point cloud and the best candidate is retained. An iteratively reweighted least squares procedure is then invoked to refine the parameters of the plane. Finally all of the inliers of the resulting refined plane in the point data set are removed from further consideration and the next image segment is considered.

At the beginning of the interpretation process the image segments are sorted by population so that the larger regions are considered first. This heuristic tends to speed up the interpretation process since, in indoor environments, large coherent regions in the image often correspond to planar regions. Detecting these regions early in the interpretation process removes those points from further consideration and, thus, speeds the overall interpretation procedure.

It is important to note that the number of RANSAC iterations required to find an inlier set is typically quite low since 3D points that project to the same image segment are very often on the same surface. This means that the chances of picking an acceptable set of inliers from each group is relatively high and the computational effort required to find such a grouping is concomitantly low.

Since the procedure considers every image segment it is quite effective at finding relatively small planar regions that may represent a small fraction of the overall data set but that are grouped together by the image segmentation procedure. Such groupings would be particularly difficult to find if we were to search for them by picking triples of points from the data set at random.

Algorithm 1 Fast Scene Interpretation

```
Segment the color image using randomized hashing
Project the 3D points onto the image
Sort the image segments by population
for $i = 0$ to npasses do
    for all image segments do
        Find all of the 3D points that project to this segment
        for $j = 0$ to ransac-its do
            Select 3 points from the subset
            Construct the plane through those points
            Score the plane by counting inliers in the point cloud
            Retain the best plane
        end for
        Refine the best plane using iteratively reweighted least squares fitting
        Find all of the inliers to this plane in the point cloud and remove them from further consideration
    end for
end for
```

III. EXPERIMENTAL RESULTS

The proposed scene interpretation procedure was applied to data sets that were collected in and around our office complex. Like most indoor scenes these examples contained a number of planar surfaces along with a number of distracting point measurements caused by clutter and spurious range readings. The results obtained on a few of these scenes are shown in Figure 3. The first column shows the color image of the scene, the second column shows the segmented image along with the projection of the point cloud onto that frame. The third column shows various views of the processed 3D point cloud where each of the extracted planes is given a unique color.

Table I summarizes some of the relevant numbers associated with each of the scan data sets shown in Figure 3. These results show that the total number of candidate triples considered by the interpretation procedure is fairly low in comparison to the the total number of 3D points in each of the data sets. This indicates that the segmentation procedure is doing a good job of focusing the efforts of the interpretation procedure since even with this small number of candidates the procedure does a creditable job of parsing

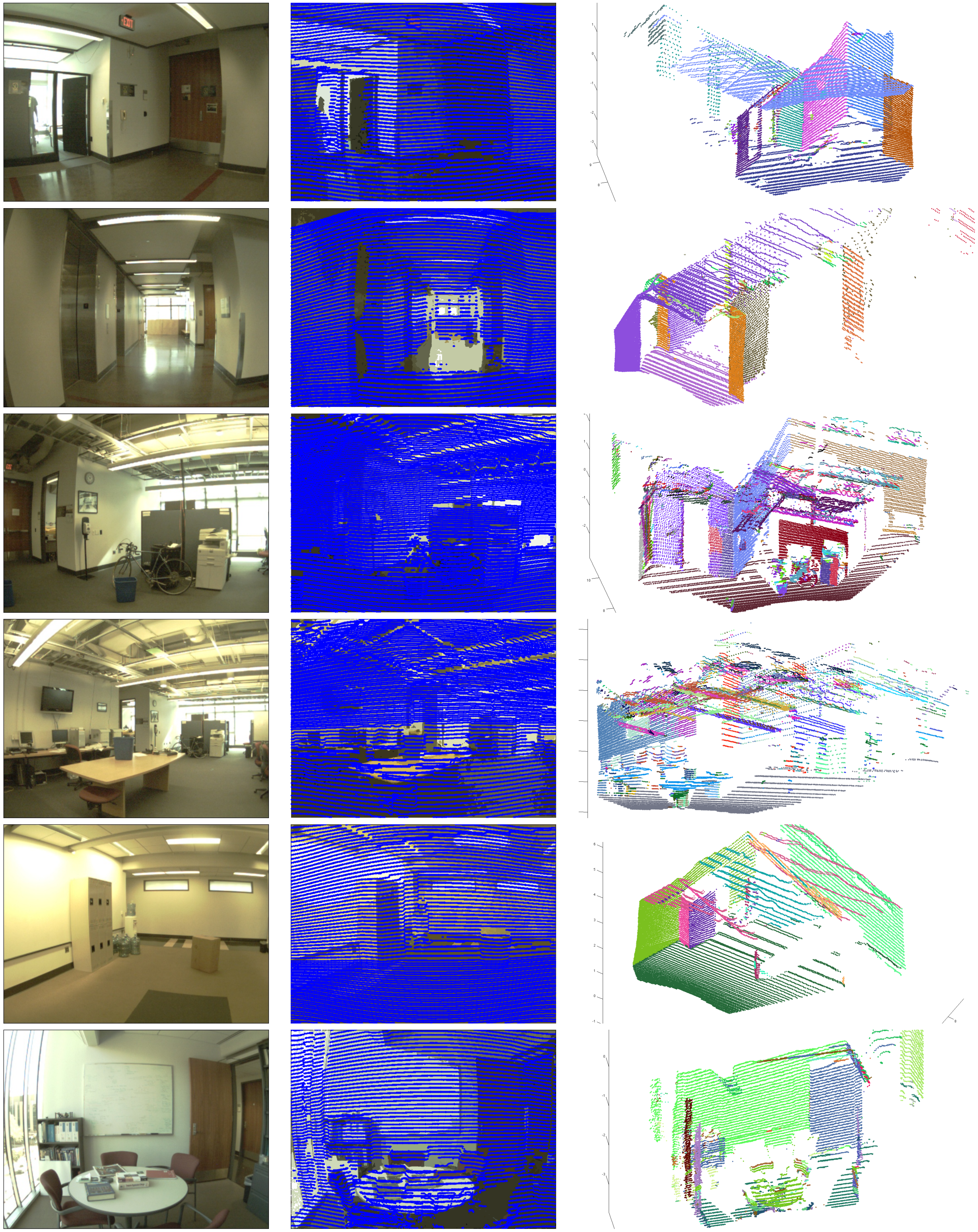

Fig. 3. This figure shows the input and output to the scene interpretation process for a variety of different scenarios. The first column shows the original color image. The second column shows the segmented color image with the 3D points projected into the frame. The third column is a 3D rendering of the point cloud where each recovered plane is assigned a unique color. (This figure is best viewed in color.)

TABLE I

PROCESS MEASUREMENTS

Data Set no.	1	2	3	4	5	6
Number of 3D points	59137	57368	60432	60107	60458	58324
Number of Image Segments	2534	2859	3854	4271	2847	2781
Number of Recovered planes	14	15	33	40	11	20
Total number of RANSAC its	560	720	1320	1600	440	800
Estimated RANSAC its *without* segmentation	2391	300859	97499	268110	476	6614
Running time of interpretation procedure in seconds	0.301	0.335	0.676	0.813	0.252	0.433

the scene into a relatively small number of salient planar surfaces. Note that the total number of RANSAC iterations is even less than the number of extracted image segments since earlier stages of interpretation typically absorb the points associated with several image regions as inliers so those segments are not considered in subsequent stages.

For comparison we compute an estimate for the number of RANSAC iterations that would have been required to find the same set of planes without the benefit of image segmentation information. Here the number of iterations required to find a set of 3 inliers on a plane containing k points in a point cloud of N points is approximated as $(N/k)^3$. The estimate for the number of RANSAC iterations required is the sum of the iterations to required to recover each of the extracted planes starting from the most populous and proceeding to the smallest removing the detected inliers at each stage. This estimate is typically several times higher than the number of RANSAC iterations carried out by the proposed procedure.

Table I also records the total time required to perform the scene interpretation procedure which was implemented in MATLAB without any significant optimization and run on a MacBook Pro laptop. The average running time across all six trials was 0.47 seconds. This suggests that the entire interpretation procedure, including the image segmentation step, could be performed several times a second on a laptop class machine.

As was noted previously, the scheme is quite effective at pulling out small surfaces with relatively few inliers. This illustrated in Figure III which shows a closeup view of the interpretation produced for the first data set in Figure 3. This closeup highlights the fact that the EXIT sign on the ceiling and the card reader on the wall are both pulled out as distinct surfaces even though they constitute a relatively small proportion of the point cloud. Similarly in the fourth data set the box on the table is successfully distinguished from the surface of the table and the background. Most of the apparent clutter in this particular data set is the result of the interpretation procedure detecting and delineating small surfaces on the structures in the ceiling. In the sixth scan data set the table in the foreground is distinguished from the books lying on that surface, the chair backs are also recovered as separate regions. This is due to the fact that these subsets are identified as coherent groups in the segmented image and as such are eventually discovered and reported.

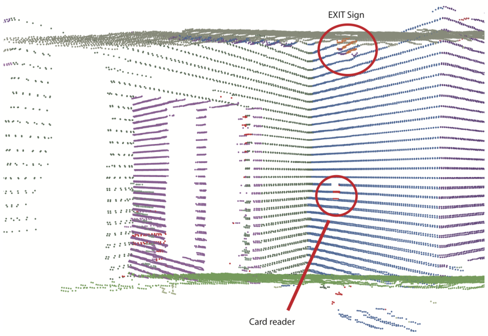

Fig. 4. Figure showing a closeup view of the interpretation of the first scan data set. This view shows that the EXIT sign hanging from the roof and the card reader on the wall are detected as separate planes even though they are relatively small sets. (This figure is best viewed in color.)

IV. CONCLUSION

The research in this paper was motivated, in part, by the advent of relatively inexpensive sensors that can be used to acquire range imagery at video rates. These range video cameras can be paired with standard image sensors to provide rich descriptions of the scene in front of the robot. The goal of this work has been to develop methods that could be used to parse these kinds of data streams in an online manner to provide the robot with a higher level understanding about its environment.

The proposed scheme leverages a novel real-time segmentation scheme based on randomized hashing which is used to group the 3D points in the point cloud into subsets which are likely to lie on the same surface in the scene. This grouping is used to inform a RANSAC based interpretation scheme which uses this grouping as a prior to bias its sampling procedure.

Fusing the image and range data in this manner proves to be a very effective method for guiding the interpretation scheme towards fruitful interpretations. The scheme is able to identify salient surfaces in most scenes with relatively little random exploration. Hence, the scheme can be run relatively quickly on realistic data sets. Furthermore, using the image segmentation data as a prior helps the system to identify small but salient surfaces that would be difficult to detect through purely random sampling. Interestingly the results seem to indicate that even a fairly rough segmentation of the image can provide very useful information since no particular effort was made to tune the segmentation procedure to produce an optimal decomposition and one can easily notice several artifacts in the segmentation image caused by under or over segmentation. This is not terribly problematic since the segments are being used as advice rather than as 'ground truth' and in that role the segmentation is correct much more often than not.

The ancillary cues provided by the segmentation procedure can also be used to focus the interpretation process. As an example the size and position of the image segments in the frame can provide useful cues about the extent or importance of the associated regions. One may want to focus ones attention on large regions in search of major structural features such as walls or floors, or one may be interested in searching for smaller objects like books on tables which would typically correspond to smaller segments.

A. Future Work

We believe that this work opens up several avenues for further exploration. Firstly, for indoor scenes at least, the scheme can be viewed as a black box which reduces the reams of point data to a smaller set of planar primitives that can be matched between frames as the robot moves through the scene. These correspondences could be used to gauge the trajectory of the robot and to fuse 3D information gathered from a range of vantage points.

It is also attractive to consider using the output of this interpretation system as the input to a higher level interpretation process which would seek to explain the observed surfaces in terms of doors, floors, walls, tables, etc. Here again the ability to rapidly reduce tens of thousands of range points to tens of plane candidates simplifies the higher level parsing process and makes it much more tractable.

In a similar vein one could consider using the image segmentation information to delineate the extents of surfaces in the scene. Typically the range images that we can obtain are at a significantly lower resolution than the associated imagery. The boundaries of the regions in the image could be used to interpolate and extrapolate the surfaces in order to provide a clearer picture of the layout of the scene.

Another place where one can leverage the image data is in distinguishing between salient surface regions that may be geometrically coplanar. For instance in a hallway closed doorways may lie flush with the walls but in the image they differ in appearance. By leveraging the image segmentation data one may be able to discriminate between surfaces or regions in the scene based on geometry *and* appearance.

Lastly while planar surfaces are prevalent in indoor environments they are not the only structures that we are interested in extracting. One could equally consider using the grouping information provided by the image to extract other relevant structures or more general shape models.

V. ACKNOWLEDGMENTS

Research was partially sponsored by the Army Research Laboratory Cooperative Agreement Number W911NF-10-2-0016. The authors also gratefully acknowledge the support of Willow Garage.

REFERENCES

[1] D. Anguelov, B. Taskarf, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng. Discriminative learning of markov random fields for segmentation of 3d scan data. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 169 – 176, jun. 2005.

[2] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. From contours to regions: An empirical evaluation. In *CVPR*, pages 2294–2301, 2009.

[3] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

[4] J. Dolson, J.M. Baek, C. Plagemann, and S. Thrun. Upsampling range data in dynamic environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1141–1148, 2010.

[5] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, 2004.

[6] J. F. Lalonde, N. Vandapel, D. F. Huber, and M. Hebert. Natural terrain classification using three-dimensional ladar data for ground robot mobility. *Journal of Field Robotics*, 23(10):839–861, 2006.

[7] K. Ni, H.L. Jin, and F. Dellaert. Groupsac: Efficient consensus in the presence of groupings. In *International Conference on Computer Vision*, pages 2193–2200, 2009.

[8] Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, Mihai Dolha, and Michael Beetz. Towards 3D Point Cloud Based Object Maps for Household Environments. *Robotics and Autonomous Systems Journal (Special Issue on Semantic Knowledge)*, 2008.

[9] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 1997.

[10] Camillo J. Taylor and Anthony Cowley. Fast segmentation via randomized hashing. In *British Machine Vision Conference*, 2009.

[11] A. Toshev, P. Mordohai, and B. Taskar. Detecting and parsing architecture at city scale from range data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 398–405, 2010.

[12] Qingxiong Yang, Ruigang Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, jun. 2007.