Segmentation and Analysis of RGB-D data

Camillo J. Taylor and Anthony Cowley
GRASP Laboratory, University of Pennsylvania

I. Abstract

One particularly interesting aspect of the Kinect sensor is that it makes it relatively easy to acquire color and depth images that are registered to a common viewpoint.This encourages us to pursue the development of algorithms that seek to exploit the advantages of both sensing modalities. Consider the range and image scans shown in Figure 1. The RGB image contains a number of cues that can be used to break the scene into regions. Changes in color, contrast and texture draw our attention to distinct areas, which often correspond to semantically salient objects. In contrast, while the depth image provides us with a sense of the geometry of the scene it is relatively featureless which can make interpretation more difficult. In this work we explore the development of algorithms that allow us to segment the scene geometry into salient planar regions based upon the cues in the RGB imagery (see also [2]) . This segmentation serves as a useful first step towards a semantic interpretation of indoor scenes.

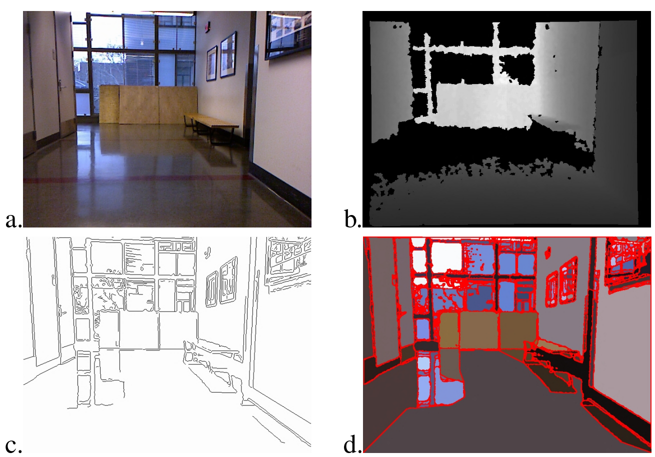

Fig. 1. a. RGB image b. Depth image c. Extracted Intensity Edges d. Image Segmentation

The first step in our analysis is an edge accurate segmentation scheme that breaks the RGB imagery into coherent, disjoint regions. This segmentation is then used as a prior to subdivide the associated depth image. The image segmentation procedure begins with a standard Canny edge extraction step which finds significant discontinuities in the intensity image as shown in Figure 1.

The results of the edge detection scheme are passed to a Delaunay Triangulation algorithm which computes a triangular tessalation of the image as is done in [3]. The Delaunay Triangulation derived in this manner has a number of attractive properties. Firstly the triangle boundaries conform to the extracted edges by construction. Secondly the tessellation naturally adapts to the content in the image, that is, it produces large triangles in boring regions of the image and small triangles in regions where there are a lot of edges. The triangular tessellation of the image induces a planar graph where the nodes are the triangles and the edges indicate adjacency relations between the triangles.

The Delaunay Triangulation procedure guarantees that the circumcircle associated with each of the triangles does not contain any other edgels which implies that these circumcircles must correspond to coherent regions in the image. We can exploit this property to associate attributes with each of the triangles in the graph. More specifically, the procedure interrogates each of the triangles in turn and computes a mean color for each node by considering all of the pixels that lie within its circumcircle as opposed to considering the interior of each triangle. This is an important distinction since the tessellation scheme will typically produce a large number of thin triangles with extreme aspect ratios.

The next stage in the segmentation procedure applies a variant of the Normalized Cut algorithm [1] to the graph derived from the triangulation. This procedure is appropriately modified to reflect the geometry latent in the triangulation. Importantly, we can run the eigenvector procedure associated with the normalized cuts procedure in less than 2 seconds on the triangular graph as opposed to the hundreds of seconds required when the same procedure is run on the pixel grid. The resulting segmentation procedure is particularly well suited to the analysis of indoor scenes which often contain large coherent areas bounded by edges.

The segments derived from the image segmentation are used to suggest groupings of the depth samples from the range imagery. More specifically, the depth samples associated with each of the image regions are passed to a RANSAC routine which is used to recursively divide the point set into planar regions. A key advantage of the proposed approach is that the image segmentation procedure is very effective at suggesting useful groupings so very few RANSAC iterations are needed to discover the structures of interest. Effectively, the image segmentation serves to focus the computational effort of the procedure on groupings that are likely to yield fruitful interpretations so the procedure is able to discover relevant groupings quickly even in complex environments with several surfaces.

It is important to keep in mind that the depth measurements produced by the Kinect sensor are derived from structured light via triangulation as opposed to time of flight. As such

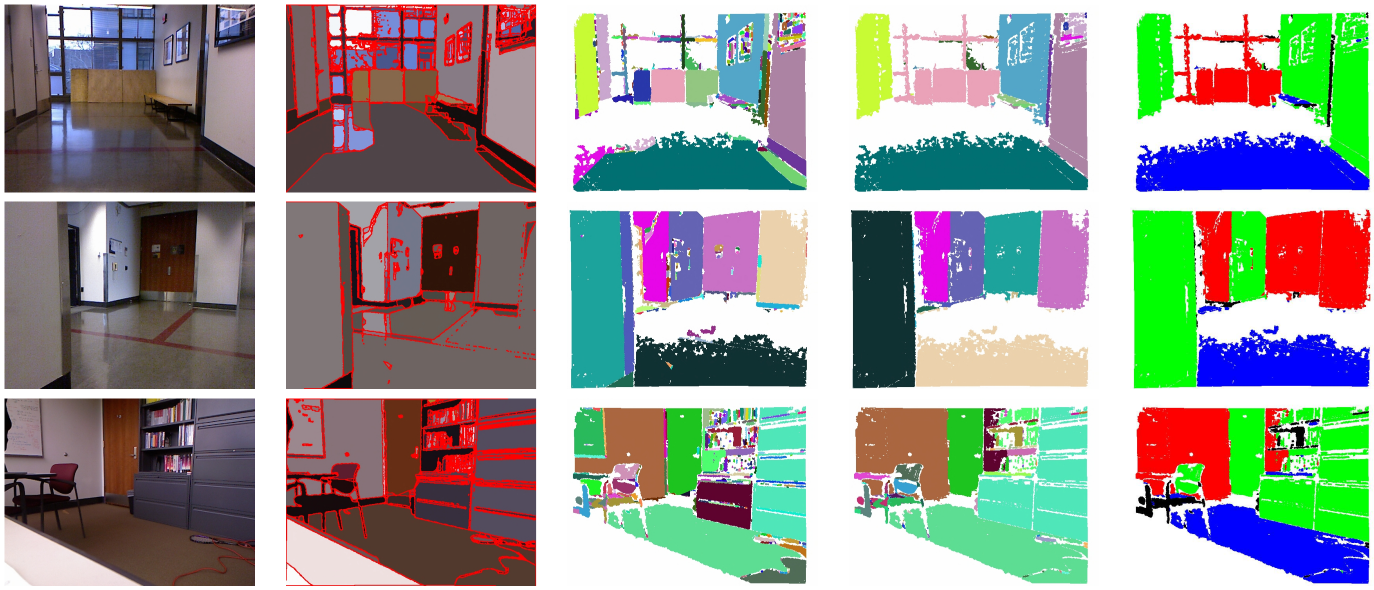

Fig. 2. The second column shows the image segmentations, the third shows the planar segments extracted from the depth imagery and the fourth the results after merging coplanar segments. Different colors are used to indicate different extracted regions. The final column shows the results of the analysis that finds the dominant rectilinear structure, Horizontal surfaces are blue and vertical surfaces are red or green depending on their orientation.

the device is best thought of as measuring disparity which is inversely related to depth. One practical consequence is that the error in the depth estimates increases rapidly as one gets further away from the sensor which argues against standard approaches to fitting planes to points based on the residual error in 3D.

In the proposed scheme the planar surfaces are fit in the image coordinates by exploiting the observation that planar surfaces in the scene will project to planar regions in the disparity image. This can be seen by taking the standard equation for a plane in the coordinate frame of the sensor:

$$n_x X + n_y Y + n_z Z = c$$

dividing through by scene depth, Z, to obtain:

$$n_x \frac{X}{Z} + n_y \frac{Y}{Z} + n_z = c \frac{1}{Z}$$

and noting that $u = \frac{X}{Z}$ and $v = \frac{Y}{Z}$ correspond to the normalized image coordinates while $w = \frac{1}{Z}$ denotes the measured disparity at that coordinate. This means that planar regions in the scene can be extracted by fitting affine models to the disparity in each image region.

Figure 2 shows the results of the planar interpretation procedure on a variety of examples. Here the different colors correspond to different planar segments that were recovered. These planar segments are then passed to a greedy merging procedure which seeks to group coplanar segments into extended regions as shown in the fourth column of Figure 2.

The last column of Figure 2 shows the results of an analysis that extracts the predominant rectilinear structure in each scene. Horizontal surfaces like the floor are colored blue. Vertical surfaces are colored red or green depending on their axis of alignment. This step is carried out by examining the most salient structural planes, those with the largest populations, and using them to infer the direction of gravity and the predominant rectilinear orientation

This analysis provides a useful second step in a semantic analysis since it allows us to clearly identify horizontal surfaces like the floor and tabletops as well as candidate wall segments. In this context it is very useful that the segmentation is edge accurate since it helps to accurately delimit the spatial extents of the extracted surfaces.

Note also that since the scheme relies on an underlying image based segmentation it can identify relevant regions like closed doors and windows that are not apparent in the range scan but are clearly delineated in the RGB imagery and, hence, can be used in subsequent analysis stages.

The entire segmentation and analysis procedure is implemented in Matlab and it takes approximately 5 seconds to run the complete analysis on a typical RGB-D image on a Macbook Pro laptop. Of these 5 seconds 1 second is spent on edge extraction and 2 seconds are spent on the image segmentation procedure. We expect that a more efficient implementation would run significantly faster.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2011.

[2] Camillo J. Taylor and Anthony Cowley. Fast scene analysis using image and range data. In *IEEE International Conference on Robotics and Automation*, 2011.

[3] Qing Wu and Yizhou Yu. Two-level image segmentation based on region and edge integration. In *Proceedings DICTA*, pages 957–966, 2003.