Growing Depth Image Superpixels for Foliage Modeling

Daniel Morris, Saif Imran  
Dept. of Electrical and Computer Engineering,  
Michigan State University,  
East Lansing, MI 48824, USA

Jin Chen, David M. Kramer  
MSU-DOE Plant Research Lab,  
Michigan State University,  
East Lansing, MI 48824, USA

Abstract—This paper presents a method for segmenting depth images into superpixels without requiring color images. Typically superpixel methods cluster pixels based on proximity in a multidimensional color space. However, building superpixels from time-of-flight depth images poses a number of new challenges including: pixels do not have color channels for similarity comparisons, the resolution of depth cameras is low compared to color, and there is significant depth noise. To address these we propose a superpixel method that approximates a depth image with set of planar facets. Facets are grown from seed points to cover the scene. Facet boundaries tend to coincide with high curvature regions and depth discontinuities, typically giving an over-segmentation of the scene. This work is motivated by automated foliage modeling, and the data we consider are of dense 3D foliage. Superpixel results are shown on foliage and are quantified using labeled data.

Keywords—Superpixels; depth image; segmentation; energy minimization; RGB-D; time-of-flight camera

I. INTRODUCTION

Superpixels have significant utility in image analysis. With greater information content than individual pixels, they allow for richer low-level descriptors, and at the same time there are far fewer of them than raw pixels, which reduces the dimensionality of graphs and optimization problems built on them. As a result they have been used for numerous computer vision tasks including segmentation [1], [2], [3], object detection [4], object tracking [5], and stereo depth estimation [6]. The key property that makes them useful for these tasks is that they over-segment the image, and boundaries of objects are included in the boundaries between superpixels. This property means superpixels do not need to be split and can be treated as low-level building blocks for scene and object analysis.

The recent availability of consumer depth cameras, such as time-of-flight sensors [7], has provided an important new modality for sensing the world. These depth images can be used for many of the same computer vision tasks as color imagery including segmentation, object detection, object tracking, etc., and similar problems are faced in achieving these tasks. In the application that motivates this paper, a time-of-flight sensor is placed over plants in a growth chamber and leaf growth over a number of days is recorded. A key need is to automatically segment the leaves and leaf fragments within dense foliage. While there are methods for 3D modeling of leaves including [8], [9], [10], [11], automatic segmentation of them from depth imagery has not been solved, and these methods rely on color or intensity images for segmentation, in some cases leveraging superpixels of these [10]. Single-image color segmentation of leaves works well [12] with plane backgrounds, but can fail to segment uniformly-colored, overlapping leaves. While ideally a RGB-D superpixel method such as [13] could use both color and depth, in practice the parallax for close-in plants can be a large fraction of leaf features making this impractical. The alternative is to do superpixel in both color and depth and combine the results afterwards. The missing component is a depth-based superpixel method, and achieving that is the goal of this paper. Future work will address combining superpixels from multiple modalities.

In an ideal superpixel segmentation, the true object boundaries are included in the boundaries of the superpixels. Poor performance occurs when superpixels span the boundaries of
A close-up of SLIC superpixels [15] with boundaries shown in orange. The manually identified outline of the leaves is shown in magenta. The SLIC pixel boundaries align well with the boundary between the leaves and the background. However the superpixel boundaries do not overlap the boundaries between overlapping leaves well. (b) The graph cut superpixel method in [14] is used with a very low threshold to obtain this set of superpixels. More of the leaf boundaries are obtained, but nevertheless a number of true overlapping leaf boundaries (shown in magenta) are not obtained and these are merged within single superpixels.

Figure 2. An illustration of the failures in two of the best color-based superpixel methods when used on overlapping foliage in Fig. 1(c). (a) A close-up of SLIC superpixels [15] with boundaries shown in orange. The manually identified outline of the leaves is shown in magenta. The SLIC pixel boundaries align well with the boundary between the leaves and the background. However the superpixel boundaries do not overlap the boundaries between overlapping leaves well. (b) The graph cut superpixel method in [14] is used with a very low threshold to obtain this set of superpixels. More of the leaf boundaries are obtained, but nevertheless a number of true overlapping leaf boundaries (shown in magenta) are not obtained and these are merged within single superpixels.

Superpixel methods can be broadly divided into graph-based methods [19], [14], [20], [21], [22] and hill-climbing methods [23], [24], [25], [15], [16]. Graph methods optimize an energy function over edge cuts that includes on pixel color similarity and shape and size priors. The more recent hill climbing methods such as SLIC [15] and SEEDS [16] start with a regular grid and refine these leading to fast performance with fairly evenly sized superpixels. Our method follows a similar approach of starting with a regular grid and refining it. However since depth pixels have no color or texture, our underlying local similarity model is quite different from all the color-based methods.

With the availability of RGB-D datasets with a depth pixel for each color pixel, it is natural to do joint color and depth-based superpixel as in [13]. Both color and depth cues can be leveraged potentially achieving better object segmentation. With current sensors the baseline separating the depth and color cameras causes significant parallax and occlusion differences between color and depth cameras, particularly for complicated structures such as foliage. Obtaining a one-to-one mapping of color and depth pixels entails significant feature matching and object modeling often from multiple viewpoints as in [26]. In our case, constrained to viewing the scene from a single viewpoint, it is preferable to use the color and depth images separately, as in [27] where the color image is used to upsample the depth image with an enhanced bilateral filter. Unfortunately it is in regions of detailed structure and occlusions, as exhibited by overlapping leaves, that significant depth errors are reported [27]. In our work we focus on using superpixels to resolve depth discontinuities in the depth image alone, and leave it to future work to improve this with an adjacent color image.

Superpixel can be used as a first step in 3D model building. An alternative is to collect data from a moving sensor, align and fuse it as done with Kinect Fusion using RGB-D images [28], [29], [26]. These methods involve building a voxel model of the scene and extracting an isosurface. However, in our application the camera remains stationary making these approaches not suitable, and that motivates this work to focus on the single-viewpoint segmentation problem.

III. SENSOR DATA

Our data was collected using a Creative Senz3D sensor. Like the Kinect 2, this is a time-of-flight sensor [7] but with a lower resolution of 320 × 240 pixels. Despite the lower resolution, its smaller angular field of view along with closer minimum range of roughly 20cm, compared to 50cm for the Kinect 2, means that it can collect higher resolution depth images. This is important for modeling plants and foliage with fine structure. We used the raw depth values rather than the default smoothed values as these contain more details of the target and return values at greater depth.

Sample data from this sensor is shown in Fig. 1. The depth camera returns both a dense pixel depth image and reflected intensity for each pixel although only the depth image was used here. In addition there is a color camera offset roughly 2.5cm from the depth camera. This offset creates in large parallax for close-in objects and significant occlusion differences.
IV. Method

Unlike color images, depth-image pixels have no color or texture characteristic with which to group them. Instead we propose that depth-image superpixels should model the local geometry as a planar facets. Pixels on flat regions should be grouped together into segments. On the other hand, high curvature edges and depth discontinuities demark the boundaries of objects and so should lie on the boundaries of superpixels. A depth-image superpixel segmentation will thus be a faceted surface model.

In building a superpixel representation we seek the following four characteristics. First, superpixel boundaries should conform to boundaries of 3D objects in the scene. That is, while the superpixels will subdivide objects, they should not span multiple objects. Second, superpixels should be contiguous in the image plane. Third, as argued in [15], it is desirable that superpixels be roughly uniform in size as they will contain similar information content for use in higher level algorithms for scene analysis. This constant size could be in 3D scene space as in [13], although the utility of this depends on the subsequent processing, and here we seek a constant image-space dimension for superpixels. Fourth, since depth-image superpixels provide a planar faceted approximation of the scene, the difference between these planar facets and the measured depths should be minimized. This section describes our approach to achieving these goals.

A. Facet Estimation

Our method starts using a regular grid to divide the depth image into equal-sized square cells. The cell size should be on the order of the smallest features that we wish to approximate, which in our application is the size of the smallest leaves.

The significant challenge we face in building our faceted model is how to avoid crossing depth discontinuities when the depth change across them may be on the order of the sensor noise. Our approach is to let our facet model the largest contiguous planar region in each grid cell. If there are two surfaces spanned by a cell, then the facet should approximate just one. It is important that the facet is contiguous, because simply finding the best planar fit to the points in a cell has a high chance of finding a plane that crosses the discontinuity, as illustrated in Fig. 3. Hence we fit planes to pixel depths in a grid cell using a robust method that grows contiguous regions explained as follows.

First we choose a set of seed points to start the plane fitting. For each seed point we want a depth and two slopes to define a plane. The slope will be noisy hence we sample these seed points from a smoothed depth image. Not all points on the smoothed depth image are likely to a plane. In particular the slopes at depth discontinuities are unlikely to result in a good planar fit. To avoid these cases we exclude pixels corresponding to peaks in the gradient images and also where the smoothed depth differs more than a threshold from the raw depth image.

Next we need to estimate and evaluate the planar region defined by each seed point. There are two main considerations in evaluating this planar region. The first, $H(s)$, for a set of pixels $s$, is a likelihood function of the plane given the data, which we seek to maximize in estimating the planar parameters. The second consideration is a prior on the shape or distribution of pixels belonging a planar region, which we refer to as a penalty term $S(s)$. For instance this can capture a prior that planar regions should be compact. Finding a planar region involves hill climbing an energy function made from these two components:

$$E(s) = H(s) - \alpha S(s).$$  \hspace{1cm} \text{(1)}

In our implementation we chose the following instantiation for this equation. Our likelihood measure for a planar region was simply, $N(s)$, the number of contiguous pixels whose depth points fall within a tight threshold around the plane. Our shape-based penalty term, $S(s)$, was set equal to the number of boundary edges of the pixels on the plane. For a given number of pixels, a compact region will have fewer boundary edges than a region with narrow arms and so will be penalized less. Both measures are easily calculated. The weight, $\alpha$, selects how compact the regions will be.

What still needs to be described is how we obtain a contiguous set of pixels belonging to a seed point. The seed specifies the plane (with its depth and slopes), and so all points within a grid cell can be evaluated by their perpendicular distance to this plane. A mask is created for all pixels belonging to this plane. Then starting at the seed point a dilation operator is repeatedly used to expand the seed point, and at each step the dilated pixels are constrained to lie on the mask as well as not cross any gradient peaks above a threshold, $t_g$, or to diverge by more than $t_p$ in depth from the depth pixels. After a fixed number of iterations the dilation stops and the set of pixels reached from within the mask are the points on the plane. The energy of this planar facet is evaluated with Eq. (1). By repeating this procedure for a randomly chosen set of good seed points, the pixels belonging to the highest scoring plane are chosen for that

![Figure 3. A slice illustration of the challenge in fitting facets to data. The red data points (one per pixel) are the same in (a) and (b). In (a) a robust line is fit to the data, but the true surface is really a step function closer to (b). Our method seeks to grow facets that do not span these boundaries.](image-url)
grid cell, and plane is fit to them using least squares to represent the superpixel.

B. Facet Expansion

Up to now we have each grid cell modeled as a single facet covering a subset of the pixels. For instance when there is a discontinuity across a grid cell, the facet will ideally cover the region to one side of the discontinuity. We would like a way to explain the remainder of the pixels in each facet. Now assuming that the structures in the have dimension at least the size of the superpixel, then the planar region not covered is likely to extend out over adjacent superpixels. If this plane is modeled in one of these adjacent superpixels we can simply extend that plane to cover the uncovered pixels of the current superpixel. The expansion algorithm works as follows.

We apply the contiguous planar facet fitting algorithm described in Section IV-A with the following modifications. We do not need to search for a seed as we start with the fitted plane for that superpixel. We increase the region in which we grow the plane from a single grid cell to a $3 \times 3$-cell region centered at that cell. Our mask is set to those pixels in the new cell region lying within an expanded distance threshold of $1.6 t_p$ from the surface, and that do not belong to any of the other facets. In this way unclaimed pixels from adjacent grid cells are added to a superpixel if they fall on the plane and are contiguous. This procedure is applied to all of the superpixels in a random order.

The final result is a set of superpixels modeling the depth image as a set of planar facets. There remain pixels not claimed by any superpixel. These are primarily of two types neither of which is well approximated by a plane. At depth discontinuities the depth pixels may have mixed values between a surface and its occluding surface with high variance. Also pixels at long range or with low intensity have large noise and do not pass the thresholds we use for fitting planes. We could adjust the threshold based on the pixel noise, but for our application these pixels are not useful and excluding them from the superpixels does not impact the foliage surface estimation.

V. RESULTS

To test our method we labeled individual leaves on depth images collected from six plants, three of which are shown in Fig. 4 with their labeling. The leaves are densely packed presenting a challenge to segmentation and superpixel algorithms.

A goal of superpixel models is that the superpixels boundaries include the leaf boundaries. We evaluate the results based on two standard evaluation criteria [16] for superpixel methods: the Corrected Undersegmented Error (CUE) and the Boundary Recall (BR) score.

To calculate CUE we first assign each superpixel to a ground truth segment or the background based on its maximum overlap. The error for a superpixel is the sum of pixels falling outside its assigned segment. Then the CUE is the sum of the errors for all the superpixels divided by the sum of the ground truth segments. The denominator intentionally excludes the background region since adding more background should not affect the CUE.

Boundary Recall, on the other hand, means the fraction of boundary pixels that can be traced by the superpixels. Here it is the fraction of ground truth boundary pixels that are within a 1.5 pixel distance to an edge of a superpixel.

We set the following parameters for all instances of our algorithm: $\alpha = 0.5$ from Eq. (1), gradient peak threshold: $t_g = 5 mm$, and plane error threshold, $t_p = 5 mm$. Based on the ground-truth, the results of our algorithm are compared with two other methods: one baseline method and the other is SLIC [15]. The baseline method is based on splitting the image into regular grid where each grid is a superpixel. SLIC is modified by replacing the LAB color-space image with the depth image. Our method outperforms the other methods both in terms of CUE and BR scores in all the six canopies tested.

For subjective evaluation, we plot the facet superpixel boundaries on the plant canopies as shown in Fig. 5. It shows the superpixel boundaries overlapping quite well the boundaries of the depth images.

VI. CONCLUSION

We propose a new superpixel approach that operates purely on depth images. It approximates depths images as collections of planar facets that correspond to superpixels. It is simple to implement, generates a mostly regular grid of superpixels that approximate the depth image as a set of planar facets, and achieves good fidelity to object boundaries. Our results show improvements over a baseline method for fidelity to ground truth. They also show practicality for using depth-image superpixels for foliage analysis.

As far as we are aware this is the first depth-only method for superpixel segmentation. This opens the door to applications where color is not available or where color provides poor segmentation cues as in some of the foliage data. Since the depth superpixels are do not depend on color superpixels, they could be fused with a color-based approach; a future direction for our work.

ACKNOWLEDGMENT

This research was supported by an MSU start-up grant, the U.S. Department of Energy, Office of Science, Basic Energy Sciences [award number DE-FG02-91ER20021], the National Science Foundation [award number 1458556] and the MSU Center for Advanced Algal and Plant Phenotyping.
Figure 4. Ground-truth labeled segments for three of the six plants: (a) Plant A, (b) Plant B also shown in Fig. 1, and (c) Plant C.

REFERENCES


Figure 6. (a) CUE metric and (b) BR score for six plants. Plants A, B and C are shown in Figs. 4 and 5.


