Forest or fumarole? Photogrammetric point cloud classification from UAV images taken over an active volcano

Robert Sare
Stanford University
Department of Geological Sciences
rmsare@stanford.edu

Abstract

Classifying ground and vegetation points is an essential step in producing elevation datasets for many natural hazards applications. This project applies image segmentation and classification methods to improve estimates of ground surface extent using unmanned aerial vehicle (UAV) survey photographs of an active volcano. Supervised (support vector machine) and unsupervised (k-means and mean shift) classification methods were tested to determine how accurately vegetation can be filtered from these data. Results suggest a linear SVM can identify ground pixels with better than 90% accuracy using coarsely labelled training data on the basis of color features. This classifier estimates that about 65% of the survey images consist of bare ground, compared to a true coverage of 63%. Unsupervised classification results (k-means) have lower recall rates for ground pixels, but also identify fewer false negative shadow pixels. The distribution of predicted pixel labels is used to classify points in a dense structure-from-motion (SfM) reconstruction of the scene. Estimation of the ground plane of the survey area by RANSAC allows for filtering of false positive ground points by point-to-plane distance from the ground plane. Unlike conventional airborne laser returns, SfM data cannot be filtered to distinguish points on the basis of reflection return characteristics, making this an important open problem in stereo photogrammetry for Earth science applications. Future work may extend the results presented here by applying a conditional random field model or recurrent neural network to hazards segmentation tasks.

1. Introduction

Classifying photogrammetric datasets to reconstruct the ground surface is an important open problem for natural hazards researchers. This project compares the performance of several classifiers for identifying vegetation and ground pixels. The class labels are projected onto points from a dense SfM reconstruction and ground and shadow points are used to estimate the ground plane orientation in the survey scene.

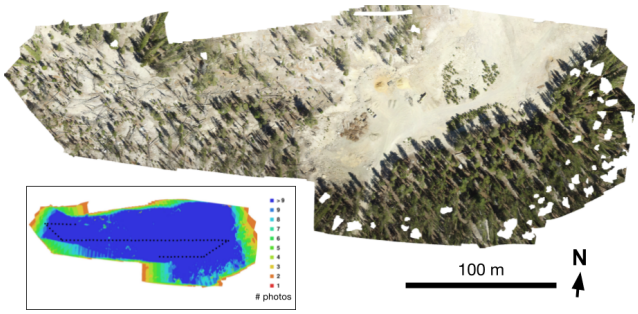


Figure 1: Orthophoto of Horseshoe Lake tree kill survey area formed by rectifying and stitching UAV survey photographs. Inset shows overlap between photographs.

The site surveyed for this project is the Horseshoe Lake tree kill area on the southern flank of Mammoth Mountain in eastern California (Figure 1). Venting of volcanic carbon dioxide from beneath the volcano poses a hazard to forest service personnel and outdoor enthusiasts in this area [6, 11]. The Horseshoe Lake tree kill is predominately covered by white, bare soil resulting from active degassing of CO_2, and green and brown trees and treefall (Figure 1). The extent of this and other tree kills and changes in the soil topography are important indicators of volcanic activity.

The area was surveyed by UAV, and several hundred nadir images were collected. Images were taken by a 12.4 Mpixel camera using a lens with 20 mm focal length and f-ratio of f/2.8. The camera was mounted on a mobile gimbal with a pitch range of -125° to +45° and ±330° pan. The gimbal was kept in an approximate nadir viewing geometry by the operator with some variation due to high winds. Time-lapse images were collected at a rate of 1 Hz at an altitude of approximately 2400 m, flying at a constant height above ground level. Along-track overlap between images is

approximately 90%. Geo-referenced point clouds and elevation models of the site were also produced from the full set of survey images using a commercial dense SIFT matching, structure-from-motion, and bundle adjustment pipeline (Agisoft PhotoScan API).

2. Background and Related Work

Pixel and point classification in hazards research is typically reserved for aerial surveys or lidar (light detection and ranging) elevation datasets. Pixels are classified by a geographic information systems (GIS) approach, requiring some manual operator intervention using a sequence of photos from roughly the same perspective [2]. Lidar returns are typically classified by return number, intensity, and using a progressive morphologic filtering algorithm that discards points at high elevation relative to a ground surface [12]. Point color and full-waveform lidar spectra are increasingly incorporated into reconstruction of elevation models for Earth science applications as photogrammetric techniques and improved lidar units have been adopted.

Deep learning and probablistic graphical models are increasingly applied to 2D-3D segmentation problems in other domains. The state-of-the-art in joint 2D-3D segmentation uses conditional random fields (CRFs) to segment point clouds and images into multiple classes, particularly for autonomous vehicle vision tasks and lidar return classification [5, 8]. Several papers show good segmentation results for large point clouds using CRFs [5, 10]. These methods have seen recent re-formulation as recurrent neural networks for pixel-level labelling of humans and other objects [13]. Convolutional neural networks based on RGB-depth data have also given good results for specific small-scale object recognition tasks on point clouds [4]. In these applications, point clouds are typically segmented into supervoxels and voxel normal orientations, color features, and geometric features derived from the voxel scatter matrix can be used to represent segments and their corresponding image points for classification or learning [9].

.

3. Technical Approach

3.1. Pixel classification

This project used different classification methods to identify vegetation and ground points in a images of a scene from a photogrammetric survey. Two unsupervised clustering techniques, k-means and mean-shift, were tested on segmented images. k-means clustering was performed with 2-4 clusters, and the bandwidth of the mean-shift density search was varied from 0.25 to 0.6. Training data were produced from several over-segmented images and a supervised Support Vector Machine (SVM) classifier was trained on 1000 training data. Minimizing the false negative rate for ground

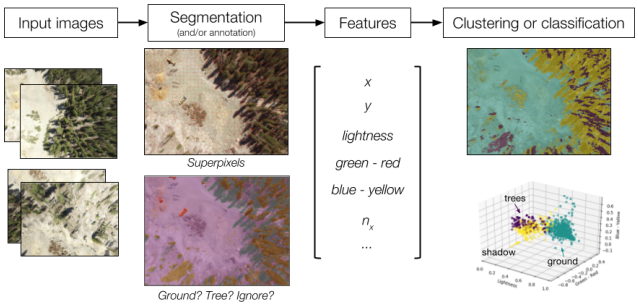


Figure 2: Image segmentation and classification pipeline.

pixels is the primary goal of this image segmentation and pixel classification task. These false negatives negatively bias estimates of tree kill area and contribute to data gaps in reconstructed point clouds and derivative scientific data products.

3.2. Ground plane estimation *via* RANSAC

A dense reconstruction of the surveyed scene was independently constructed from dense SIFT features using the API for a commercial SfM implementation. Camera matrices were estimated for each image and the histogram of pixel classifications calculated for each reconstructed point by projection onto 2D classification results. The ground plane of the scene was estimated by a random sample consensus algorithm (RANSAC) using only ground and shadow points as determined from the mode of the label distribution.

4. Data Processing and Feature Representation

4.1. Image segmentation

The data for this project consist of 100 testing images and labelled training images of the Horseshoe Lake tree kill site. The images were obtained during an aerial survey in JPEG format at compression level 9; each image is 4000×3000 pixels. Images were segmented and color features extracted for each segment and at each pixel of testing images (Figure 2).

All images were segmented into superpixels using a simple linear iterative clustering (SLIC) segmentation algorithm [1, 3]. This algorithm defines a specified number of cluster centers in the image by performing a restricted k-means clustering in which only points in a a neighborhood about each cluster center are compared to the center [1]. This divides the image into roughly equal-area segments depending on the neighborhood dimensions. The cluster centers are intialized on a regular grid over the image. SLIC can also be run on 3D points to produce supervoxels, although other voxelization methods are often used [1, 9].

SLIC requires two parameters: the number of superpixels and a Gaussian smoothing width σ. A smoothing factor of 5 pixels was chosen based on the smallest resolvable features (bushes and boulders) in these images. Several segmentation levels were tested to determine the appropriate number of segments. For these data, a successful segmentation captures trees and shadows in individual segments, as well as swaths of relatively uniform ground (Figure 8). Segmentations are defined in terms of downsampling factor f, where the total number of segments is $n_{seg} = 4000 \times 3000/f$. Factors from 100 to 10^5 were tested, and $f = 10^4$ was chosen by visual inspection of the segmented images.

4.2. Training data

A simple image labelling utility was written to label pixels in training data. Images are labelled by a user by filling a portion of the image with the operator's chosen label (Figure 3). Over-segmented versions of the training images consisting of 120K superpixels were labelled for efficiency. Operator time for a typical training image was about 45 minutes.

Pixel/segment classes for this classification are "ground", "shadow", "vegetation", or null. Including a shadow class is a key improvement over a binary classification in which both ground and vegetation classes are contaminated by uniformly low-lightness ($L \approx 0$) shadow points.

The classifiers presented here were trained on color features and pixel position from each image. The original RGB channels of the images are converted to CIELAB color space and normalized by channel to better capture perceptual differences between pixels. Additional features from the dense reconstruction such as normal orientation and elevation were also extracted for correspondence points in the images.

4.3. Challenges and feature engineering

Although image points clearly differ in color, additional features could better capture the rough texture of the trees compared to smooth ground surface. Normal orientations and other 3D features are also available from a dense reconstruction of the scene. These were projected onto the images using the camera matrix associated with each view. Results with more features generally did not improve recall rates above color-based classification.

Finally, an alternate approach was used to identify the bounding boxes of trees from histograms of oriented gradient (HOG) descriptors. However, the the camera view is top-down, and these descriptors only identify trees from a particular viewpoint. As the camera moves along the survey track it captures tilted oblique views of stationary trees that are poorly fit by a single descriptor. A pose-invariant descriptor focused on edge gradients, like the gradient descriptor proposed in [7], which would be more robust to overhead pose changes, is likely to yield better results.

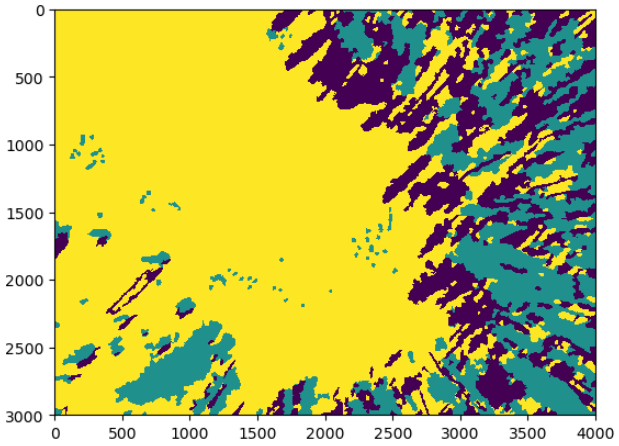


Figure 3: Labelled ground truth image. Ground pixels shown in yellow, shadow pixels in black, and vegetation pixels shown in green.

Class	Precision	Recall
Ground	93%	96%
Shadow	76%	85%
Vegetation	87%	67%

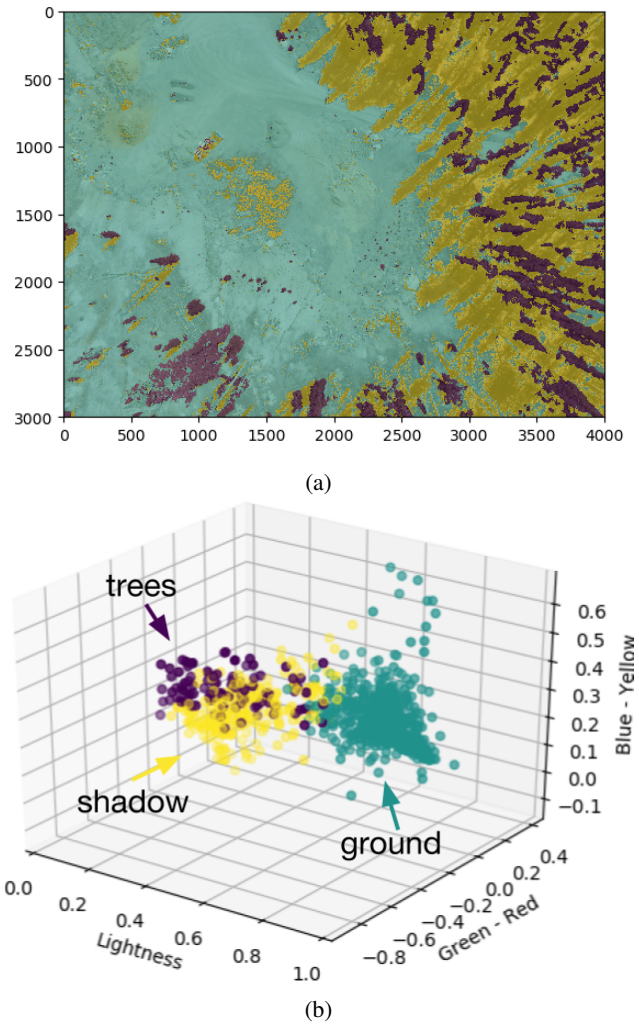
Table 1: Error rates from predictions on a single image from an SVM trained on 1000 segments of training data.

5. Results

5.1. SVM classification results

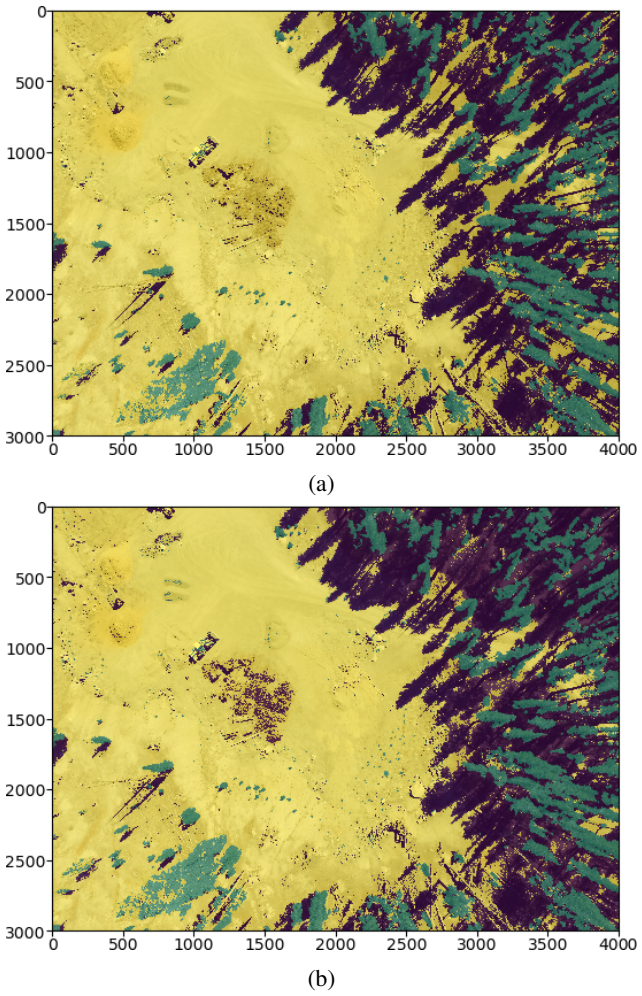
A linear SVM trained on 1200 training data gave the best classification results for ground pixels in these images (Table 1). Qualitatively, the SVM identifies ground pixels better in areas of high shadowing and occlusion which are most difficult to image (Figure 5). Only about 4% of the labels predicted by the SVM are false negatives in the ground category, or just under 3×10^5 pixels (Figure 6). These classification errors might be ameliorated using some contextual information, like that available to a higher-order conditional random field model, or textural information captured by a different descriptor.

The SVM classification out-performed unsupervised classifiers in occluded areas (Figure 5), but had poorer recall rate for shadow pixels. This many reflect the ambiguity of the shadow category or a under-sampling of shadows in the randomly split training data. Likewise, the good performance on ground pixels may reflect an over-sampling



(a) (b)

Figure 4: Relevance analysis of color features in L-a-b color space. a) Example of k-means clustering of image pixels. b) Interpreted clusters (k-means) in color feature space. Ground pixels shown in blue, shadow pixels in yellow, and vegetation pixels in purple. Point density thinned by 50% for ease of visualization.



(a) (b)

Figure 5: Predictions of a) SVM and b) k-means on a single image. Note mis-classification of ground pixels in debris pile and upper right forested area by k-means.

of ground training points due to the over-representation of ground in the whole dataset. More than 50% of the labelled pixels are ground, whereas the vegetation and shadow categories are less well-represented.

Other supervised methods, including CRFs, suffer from poor performance due to training data imbalances in a similar way. For example, CRF class segmentation of the Leuven dataset, a state-of-the-art benchmark dataset for autonomous driving, achieves relatively low accuracies for under-represented "person" and "pedestrian" classes [5, 10]. Imbalances in training data can be addressed by masking out under-represented categories, as in [5], or increasing the number and coverage of the training images.

5.2. Unsupervised classification results

The unsupervised methods performed reasonably well on the input images. k-means with three clusters had a recall rate of 90% for ground pixels and 93% for shadow pixels (Table 2). The k-means recall rate for vegetation pixels was much lower at 59%, indicating a large number of false negatives in the cluster associated with this category. This could be due to a large amount of textural and color variation between light green and dark green trees in the tree kill, or the inherent ambiguity of shadows cast on trees. Intra-class variability for shadows is also potentially responsible for the low precision (60%) of k-means in this category.

The best performance was achieved with the natural

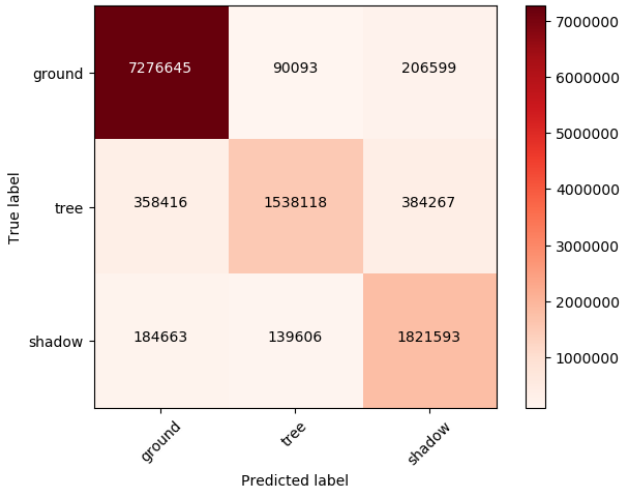


Figure 6: Confusion matrix for SVM classification pictured in Figure 5 and described in Table 1. Units are pixels in ground truth image.

Class	Precision	Recall
Ground	96%	90%
Shadow	60%	93%
Vegetation	90%	59%

Table 2: Error rates from predictions on a single image from k-means clustering.

number of cluster in these images, $k = 3$. A binary clustering under-segmented the image into ground and an amalgamation of trees and shadows. Specifying four cluster centers differentiated between vegetation with slightly different color values but degraded the quality of the ground cluster with false negatives at pixels with high lightness values.

Mean-shift underperformed substantially due to the presence of distinctive non-tree or ground features such as construction or monitoring equipment (Figure 9). Small but dense clusters of unique colors from these objects were identified when the search bandwidth was reasonably small (0.25), but increasing the bandwidth under-segmented the image and blurred cluster boundaries.

5.3. Reconstruction results

Point clouds were generated using 80 of the testing images without any masking (Figure 10). RANSAC was used to estimate the orientation of a plane going through the ground and shadow points as determined by projection of SVM-classified pixels onto image points. The modal label for each reconstructed point was assigned to that point as a heuristic classification.

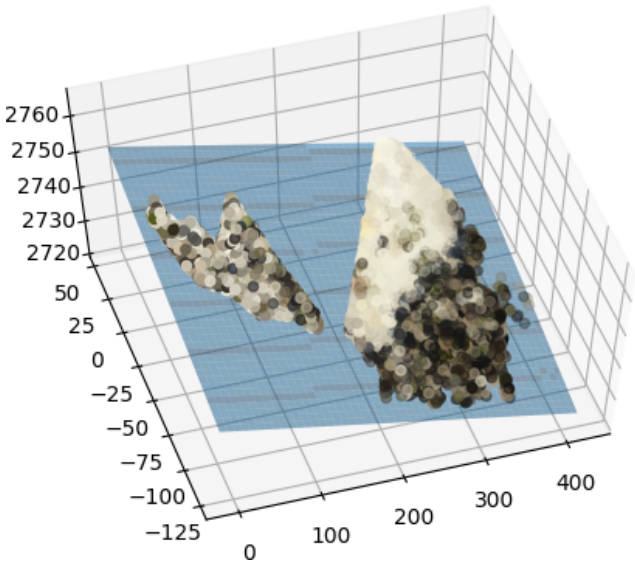


Figure 7: Best-fit plane through ground and shadow pixels determined by RANSAC regression on all ground and shadow points. Only 10^4 points shown for clarity.

In an arbitrary local coordinate system (with metric units in meters), the RANSAC estimate for the plane is defined by the linear equation

$$z = mx + ny + d$$
$$z = -0.043x + 0.086y + 2750.49$$

This corresponds to a scalar equation for the plane with normal vector

$$n = \begin{bmatrix} 0.043 & -0.086 & 1 \end{bmatrix}^\top$$

or orientation angles of

$$\theta_x = 87.6^\circ$$
$$\theta_y = 94.9^\circ$$
$$\theta_z = 5.46^\circ$$

with respect to the local Cartesian coordinates. The local z-axis is approximately aligned with the Up direction in geographic coordinates (UTM), making this estimated ground plane nearly horizontal. Images of the estimated ground plane through a small subset of ground points are shown in Figure 7. Local slopes near the eastern and western edges of the survey contributed to excess tilt of the RANSAC ground plane. The horizontal orientation of the ground plane relative to the reconstructed points will allow for point cloud features to be projected onto rectified images in subsequent stages of this project. These 3D features, such as supervoxel covariance properties or bounding cube facet normal

vectors, will likely improve segmentation and classification of more complex scenes.

6. Conclusions

This project shows promising results for rapid estimation of the extent of affected ground in a volcanically active area. Although supervised methods performed best, unsupervised clustering segmented the images in a qualitatively similar way with a higher ground pixel false negative rate. k-means could be used as an alternative for more rapid clustering if the training time of an SVM is prohibitively long when using a larger amount of training data. While these results are promising, experiments will be conducted used UAV surveys over more densely-vegetated areas with different degrees of color contrast to assess the generalizability of these results. High vegetation density will likely result in fewer reconstructed ground points in forested areas, and color contrasts between vegetation types will either require more careful cluster parameter tuning, or additional training data capturing multiple vegetation classes.

The products derived from these classifications have immediate relevance for hazards monitoring. Most directly, the classification can be used to quantify tree kill area at the site and consistently make this measurement across surveys. Additionally, the ground plane estimate from this survey can be used for change detection. This will allow researchers to measure changes in regional slope, and detect points in future surveys that may have moved relative to a stable ground plane.

Monitoring and change detection applications of computer vision techniques put a premium on efficiency and on-line processing (*e.g.*, using a un-networked laptop in the field). Training data for specific surveys is also difficult to produce in real time. These practicalities make standard linear classification methods like k-means or SVM attractive. Although conditional random field models achieve much more accurate label predictions, and the time to inference is relatively small, training of these models is computationally expensive and requires a large number of annotated training images [5]. Consequently, a CRF approach may be successful after several tree kill surveys have been conducted and an archive of high-quality training data is available. Such a model will allow more precise measurements of changes at the tree kill over time and may be more appropriate for an integrated airborne sensor system.

7. Future Work

As airborne monitoring systems are developed for volcano hazards, efficient change detection over tree kill areas will be an important monitoring tool. If crack or fissures form in the soil, as observed during previous volcanic unrest, corner detection, possibly using the Harris corner detector, or line detection using a Hough transform, could be used to track crack development over multiple images. Likewise, spatially variable changes in tree kill extent, measureable by differences between classified points or pixels, could indicate trends in the direction of CO_2 diffusion during volcanic unrest.

These classification results will be compared to subsequent surveys in August 2017 to estimate the change in affected ground area over multiple surveys during periods of volcanic activity or quiescence. The SVM classifier described here will also be applied to historical air photographs to estimate tree kill extent over the period from 1951 to 1983. As the archive of training images grows, a conditional random field model will be formulated to improve the joint segmentation of these images and point clouds by incorporating pixel-wise and neighborhood similarity (unary and higher-order clique potentials) and 3D features.

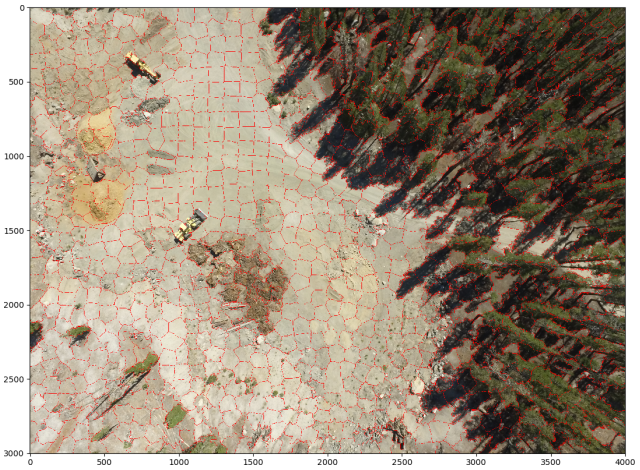
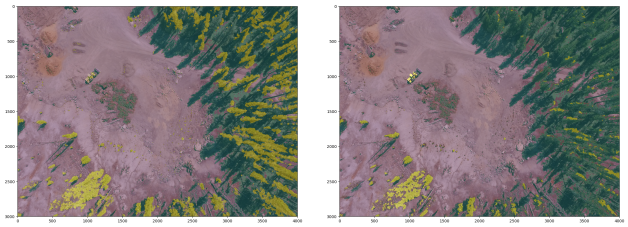


Figure 8: Example of segmented image with segmentation factor $f = 10^4$ and smoothing width $\sigma = 5$.



(a) (b)

Figure 9: Predictions of a) k-means and b) mean-shift (bandwidth 0.25) on a single image. Note mis-classification of tree pixels in forested area by mean-shift.

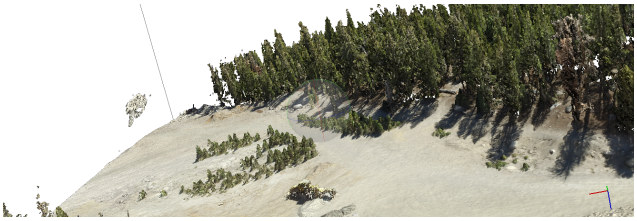


Figure 10: Oblique view of forested area in point cloud reconstructed from un-masked images.

8. Code Repository

Code for this project is available at https://github.com/rmsare/cs231a-project. This implementation uses point clouds saved as plain text files, but additional wrappers for point cloud I/O with point RGB values and normal orientations will be available at https://github.com/rmsare/python-pcl, which is a fork of an existing Point Cloud Library (PCL) Python wrapper.

9. Supplementary Material

A time lapse of SVM classification results is available in a Stanford Box folder. Examples of an annotated ground truth image, predicted labels, and several test images are also included.

Acknowledgments

Thanks to Boris Ivanovic, Helen Jiang, and Prof. Savarese for discussions about joint 2D and 3D segmentation methods, and Dr. Jennifer Lewicki of the USGS and Prof. George Hilley for discussions about and access to the Horseshoe Lake tree kill site.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. 2

[2] L. Clor, J. Barefoot, S. Hurwitz, and A. Diefenbach. A photogrammetric approach to measuring temporal change in tree kill areas at Mammoth Mountain and Long Valley Caldera, California. In *American Geophysical Union Fall Meeting Abstracts*, 2015. 2

[3] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *IEEE 12th International Conference on Computer Vision*, pages 670–677. IEEE, 2009. 2

[4] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014. 2

[5] L. Ladickỳ, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. Torr. Joint optimization for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision*, pages 1–12, 2012. 2, 4, 6

[6] J. Lewicki, G. Hilley, T. Tosha, R. Aoyagi, K. Yamamoto, and S. Benson. Dynamic coupling of volcanic CO2 flow and wind at the Horseshoe Lake tree kill, Mammoth Mountain, California. *Geophysical Research Letters*, 34(3), 2007. 1

[7] Z. Lin and L. Davis. A pose-invariant descriptor for human detection and segmentation. *European Conference on Computer Vision*, pages 423–436, 2008. 3

[8] J. Niemeyer, J. D. Wegner, C. Mallet, F. Rottensteiner, and U. Soergel. Conditional random fields for urban scene classification with full waveform LiDAR data. In *Photogrammetric Image Analysis*, pages 233–244. Springer, 2011. 2

[9] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2027–2034, 2013. 2

[10] P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, P. H. Torr, et al. Joint optimisation for object class segmentation and dense stereo reconstruction. In *Proc. BMVC*, pages 1–11, 2010. 2, 4

[11] C. Werner, D. Bergfeld, C. D. Farrar, M. P. Doukas, P. J. Kelly, and C. Kern. Decadal-scale variability of diffuse CO_2 emissions and seismicity revealed from long-term monitoring (1995–2013) at Mammoth Mountain, California, USA. *Journal of Volcanology and Geothermal Research*, 289:51–63, 2014. 1

[12] K. Zhang, S.-C. Chen, D. Whitman, M.-L. Shyu, J. Yan, and C. Zhang. A progressive morphological filter for removing nonground measurements from airborne lidar data. *IEEE transactions on geoscience and remote sensing*, 41(4):872–882, 2003. 2

[13] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 2