Analysis of Airborne Laser Scanning Data with Regional Shape Descriptors

Zehra Shah[1], Stewart He[2], Peter Tittmann[3] & Nina Amenta[4]

[1]Computer Science, UC Davis - zshah@ucdavis.edu
[2]Computer Science, UC Davis - sthe@ucdavis.edu
[3]Center for Forestry, UC Berkeley - pwt@berkeley.edu
[4]Computer Science, UC Davis - amenta@cs.ucdavis.edu

Paper Number: SL2013-079

1 Introduction

Inferring forest parameters such as biomass, coverage, or basal area from airborne laser scanning (ALS) data with ground-truth training data is a well-established practice (Hyyppä et al., 2008). Generally forest parameters are estimated at the plot level, often for circular plots with a radius of 10-30 meters. Segmenting individual tree crowns from the canopy and using tree-level statistics to infer parameters at a finer scale is a topic of considerable research interest (eg. Popescu et al. 2004, Lee et al., 2010, Tittmann et al. 2011, Li et al. 2013). But it remains quite difficult to segment trees reliably from discrete-return LiDAR data, especially in the dense forests that account for most of the planet's terrestrial biomass.

In this study we analyze discrete-return ALS data at higher-than-plot resolution without attempting to identify individual trees. We use a strategy popular in computer vision, the computation of *regional shape descriptors*. Our shape descriptor is a particular binning of points within a local area, intended to combine data at a point with context data from its neighborhood. In computer vision, regional shape descriptors have been used to discover and match features in images (Belongie, et al., 2002) and in three-dimensional point clouds representing small objects generated by triangulation laser scanners (Johnson and Hebert, 1999, Frome, et al., 2004). In this paper we show how regional shape descriptors can be used as input variables to infer forest parameters via regression. Also, we observe that regional shape descriptors can be used as points in a high-dimensional space with a natural metric, allowing the use of some unsupervised analyses such as clustering. This is in contrast to disparate collections of local LiDAR statistics, for which the choice of metric would be rather arbitrary (eg. Hastie, et al., 2009). The inherent scarcity of field data makes robust unsupervised methods appealing.

We also use visualization to evaluate the quality of our results. Evaluating a linear model or clustering on shape descriptors computed at a dense grid of points over an entire tile of LiDAR, shows spatial variation and lets us compare it to features visible in the LiDAR.

2 Shape descriptors

We define a local shape descriptor for any point in the x-y plane as a way of summarizing the local point distribution. We divide a vertical cylinder centered at the point into bins of equal volume. The shape descriptor then is a vector, each element of which is the number of points falling into the corresponding bin. We would like the bins to be invariant to rotation around the z axis, on the assumption that small patches of forest that differ only by a rotation should be similar. To produce rotationally symmetric bins, first, the cylinder is sliced horizontally into uniformly sized disks, and then each of those smaller disks is divided into annular bins; see Figure 1. To keep the bin volumes equal, the difference between the two radii determining a ring decreases as we go out from the core. The height of the cylinder is determined by the maximum height value for the entire dataset.

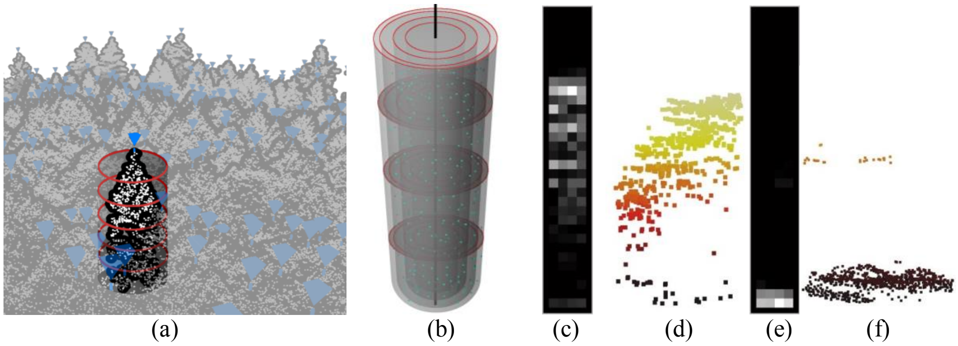

Figure 1: The shape descriptor bins the LiDAR points within a cylinder. On the left (a), we highlight the cylinder of a shape descriptor of radius 6.5 meters centered on a peak in the CHM; all peaks are indicated by blue arrows. The cylinder is divided into height levels, each of which is binned into concentric rings (annular bins), as shown on the right (b). Some example shape descriptors are depicted in (c) and (e), with corresponding LiDAR points shown in (d) and (f). Brighter squares represent higher bin counts, with left-to-right representing inner-to-outer radius bins and bottom-to-top representing lower-to-higher height bins.

This descriptor, being a histogram, can be embedded naturally into a high-dimensional metric space. This allows any kind of unsupervised learning, such as clustering, to be performed. Distance metrics for histograms are well-studied. Bin-to-bin measures compare only the corresponding bins of two histograms. Examples are Euclidean distance and the other Minkowski norms, Chi-squared distance, KL divergence, and Histogram Intersection distance. Bin-to-bin distances do not account for the relationships between different bins (see Figure 2). In contrast, cross-bin histogram dissimilarity measures, besides comparing individual bin counts, also incorporate some measure of the similarity *between* histogram bins. They include quadratic form distance (Puzicha, Buhmann, Rubner, & Tomasi, 1999), Earth Mover's Distance (Rubner, Tomasi, & Guibas, 1998) and Diffusion Distance (Ling & Okada, 2006). For this study, we chose Diffusion Distance, which is convenient and efficient to compute.

To compute Diffusion Distance, we augment each histogram with two levels of lower-resolution histograms. To compute each of these histogram we blur the previous level with a Gaussian filter and then downsample. The resulting augmented vectors, containing original and diffused bins, can then be compared using any bin-to-bin similarity measure; we use the L^1 norm.

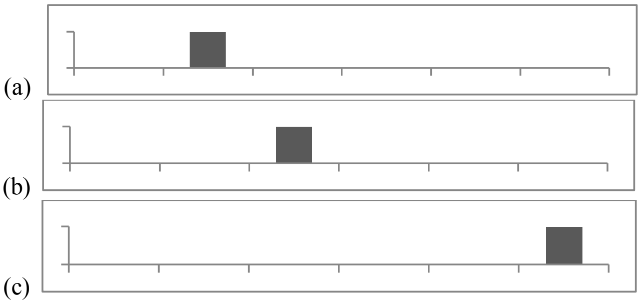

Figure 2: The one dimensional histogram in 2(a) is more similar to the one in 2(b) than to the one in 2(c), but a bin-to-bin distance measure assigns the same distance in both of these cases.

3 LiDAR and field data

We used data from the Panther Creek study area as an example of analysis using regional shape descriptors.

3.1 LiDAR survey

The Panther Creek study area was established by the US Bureau of Land Management with participation by other interested parties as a LiDAR laboratory. The approximately 2580ha (6100ac) area is located in western Yamhill county, northwest of McMinnville, Oregon. We are using a 2009 LiDAR survey of the Panther Creek study area, close in time to the collection of the field data. It utilized a Leica ALS60 sensor mounted in a Cessna Caravan 208B. The specifications for this survey are detailed in Table 1.

Table 1: Panther Creek LiDAR survey specifications

Survey Altitude	900m
Pulse rate	>105kHz
Pulse mode	Single
Mirror scan rate	52 Hz
Field of view	28°(±14° from nadir)
Overlap	≥100% (≥50% side-lap)
Avg. pulse density	10.18 pt/m^2

3.2 Field data collection

Stands were delineated and stratified into three classes: conifer, mixed and riparian. Three 16.05m radius plots are located within each stand; one is fully sampled and must be entirely within the stand. Two supplementary plots have plot centroids within the stand, but the plot can extend into neighbouring stands. All trees with Diameter at Breast Height (DBH) of 2.5cm and larger in the plots were measured. We use the following field data in our analysis: Species, DBH, Total Height (recorded to the nearest 0.1m from ground level to the highest green point), and location. Information on height-to-live-crown, and additional indicators for dead trees, etc, are included in the dataset but not used in this study.

3.3 Computing shape descriptors

We experimented with two ways of choosing locations for the shape descriptor centers. First, we chose centers on a grid covering each plot. This kind of sampling would be useful for estimating basal area or biomass across a region by combining higher-resolution estimates at the grid points. Second, we computed peaks in the Canopy Height Map (CHM) using the Fusion/LDV software package (McGaughey, 2013) and placed a shape descriptor at each peak. This approximates a per-tree sampling. In this case the shape descriptors of nearby peaks overlap arbitrarily, and since some areas are not covered by any shape descriptor, we cannot form per-plot parameter estimates by adding up the per-peak parameter estimates.

On the Panther Creek data, we computed shape descriptors using both kinds of sampling; the grid we used was 5x5 meters. For each kind of sampling, we tried shape descriptors at two radii, summarized in Table 2.

Table 2: Shape Descriptor Parameters

Radius (m) / parameter	5	6.5
Number of annular bins	6	8
Number of vertical bins	32	32
Radius of inner bin (m)	2.041	2.298
Volume of each bin (m^3)	28.66	36.33

Larger shape descriptors have more bins, to retain roughly the same resolution per bin for comparison purposes. The vertical height of every bin is 2.19m.

3.4 Normalizing shape descriptors

LiDAR data acquisition involves a significant amount of flight line overlap, meaning there are regions covered by multiple swathes of LiDAR. In order to avoid having higher values in the shape descriptor bins in regions of higher coverage, we normalized each shape descriptor by dividing each bin count by the total number of first returns falling within the shape descriptor. This approach is similar, but not quite equivalent, to the more common practice of normalizing height bins by reporting them as percentiles of the total number of returns; normalizing only by first returns preserves local differences in the number of returns per laser pulse.

4 Regression

We used the shape descriptors from the field plots of the Panther Creek data as independent variables for estimating basal area (BA), volume (Vol) and biomass (Mass). For each shape descriptor, we selected the trees whose trunk centers lay within the cylinder defined by the shape descriptor. To estimate the value of the variable on the shape descriptor, we computed basal area and the volume statistic from the field data:

$BA = \Sigma_{trees} (DBH)^2$ (1) and $Vol = \Sigma_{trees} (DBH)^2 x Height$ (2)

We also used DBH to estimate the biomass covered by each shape descriptor, using the formula:

$$Mass = \exp[\ \beta_1 + \beta_2 \ln(DBH)\ ] \qquad (3)$$

taking β_1 and β_2 from Table 3 in (Jenkins, et al., 2004) based on the tree type in the field data.

All of these forest parameters are strongly and positively correlated with maximum height. But height alone, as measured by the LiDAR data, is only a moderately good predictor, as measured by the adjusted R^2 statistic. For instance using maximum measured height on 6.5m radius regions gives models predicting basal area with R^2 0.37, volume with R^2 0.45, and biomass with R^2 0.43. The regional shape descriptors produced linear models which fit the field data with much greater accuracy. We tried combining maximum height with the regional shape descriptors, but it produced no significant improvement.

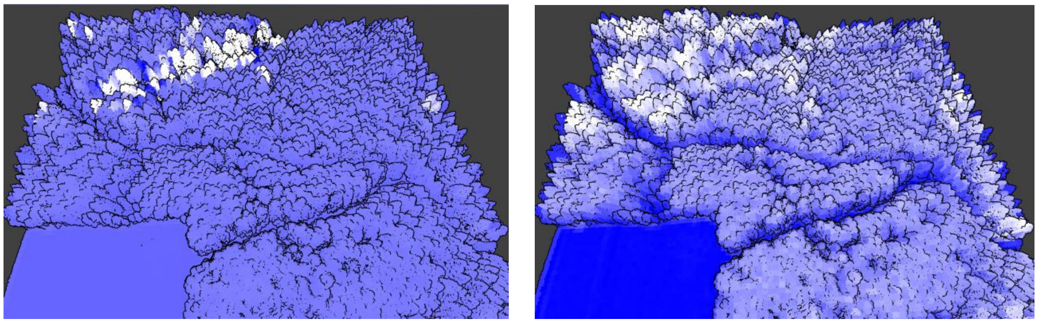

Figure 3: Visualizations of linear models of basal area as predicted by height bins and by PCA components of regional shape descriptors, on a tile of the Panther Creek dataset. Deep blue is the lowest BA prediction and white is the highest. The model on the left has an R^2 of 0.590, but this is due to over-fitting, and the model assigns both very large and very small values to the atypical tall trees in the upper left, washing out mostly reasonable predictions on the rest of the tile. The model on the right, using the first 50 PCA components of the shape descriptor data as input, makes reasonable predictions everywhere with an R^2 of 0.572.

Unfortunately visualization (left in Figure 3) reveals that much of the improvement in the models produced from shape descriptors is due to overfitting using the increased degrees of freedom in the higher-dimensional input space, resulting in wildly incorrect predicted values on other parts of the Panther Creek LiDAR in which the shape descriptors do not resemble that in the field data plots. Binning the LiDAR only by height, a more traditional method of representing local LiDAR data as a histogram, also suffered from overfitting (we used 32 height bins, again normalized by number of first returns). For example, the binning-by-height linear model applied to the ground truth resulted in a minimum basal area of -869.5 cm^2 and maximum of 29,583 cm^2 with an average of 8,363 cm^2 and standard deviation of 4,597 cm^2. However, applying the same model to the 2500 m^2 region in Figure 3 gives a minimum of -277,750 cm^2 and maximum of 831,210 cm^2 with an average of 17,108 cm^2 and standard deviation of 44,573 cm^2. The huge increase in the standard deviation suggests that the linear model is overfitted.

Reducing the dimensionality by computing the Principal Components of the 198 or 256 shape descriptor bins, and only using the most significant components, gave much more satisfactory results. We found that even using 50 components, accounting for 94% of the variability of the input, we produced models that generalized well to other parts of the Panther Creek dataset; the lower R^2 and MSE scores reflect the reduction in overfitting. In this case the PCA based linear model applied on the ground truth results in a standard deviation of 15,805 cm^2 and a standard deviation of 5,970 cm^2 on a larger plot. Notice that the Principal Components depend on the metric applied to the input data, so this technique is more appropriate with a representation of the local LiDAR distributions that comes with a natural metric, such as our shape descriptors or height bins.

Table 3: Adjusted R^2 values for regressions using grid sampling of local regions.

Radius / Variables	5	6.5
Shape Descriptors – BA	0.631	0.710
Height Bins – BA	0.505	0.590
PCA – BA	0.467	0.572
Shape Descriptors –Vol	0.762	0.831
Height Bins – Vol	0.620	0.699
PCA – Vol	0.570	0.679

Shape Descriptors – Mass	0.705	0.764
Height Bins – Mass	0.570	0.631
PCA – Mass	0.496	0.602

We tried fitting each of the three parameters using the three different representations of the local distribution information. The shape descriptors and height bin models show overfitting when applied to entire tiles, although their R^2 values on the training data are better.

Table 4: Adjusted R^2 values for regressions using peak sampling

Radius / Adjusted R^2	5	6.5
Shape Descriptors – BA	0.784	0.827
Height Bins – BA	0.647	0.733
PCA – BA	0.637	0.673
Shape Descriptors –Vol	0.893	0.931
Height Bins – Vol	0.745	0.827
PCA – Vol	0.719	0.744
Shape Descriptors – Mass	0.843	0.897
Height Bins – Mass	0.681	0.788
PCA – Mass	0.668	0.688

Clearly the larger shape descriptors produce better models than smaller ones, by averaging together more ground-truth data and separating the LiDAR data into more spatial bins. The volume statistic is easier to infer than basal area, probably because it includes height, which is readily apparent in LiDAR data. It is interesting that biomass is also better inferred than basal area, which suggests that the shape descriptors might be capturing some of the variation due to different tree species. While shape descriptors evaluated at peaks in the CHM give more accurate linear models than shape descriptors placed at grid points, it is less clear to us how to interpolate these local values to produce estimates of forest parameters over larger regions.

5 Clustering

Any Minkowski metric applied to the augmented shape descriptors (vectors with original and diffused bins) forms a well-justified metric space, within which unsupervised learning algorithms such as clustering can be performed. As an example, we used the k-means clustering algorithm, under the L^1 norm. We found that there was no obvious best choice for the number of clusters k, so we chose k to be six, balancing number of clusters against the quality score.

Clustering a tile of the Panther Creek dataset with disparate land cover types yields clusters with very good spatial locality, picking out distinct regions of vegetation. We used the shape descriptors with a cylinder radius of 6.5m. We also performed the clustering using only the thirty-two height bins. The two representations produced very nearly identical clusterings. The clusters based on the height bins are shown in Figure 4.

6 Discussion

We used the simplest possible methods for parameter prediction and clustering. More powerful learning methods will likely give more useful results.

Using a representation with a natural metric space makes unsupervised methods such as PCA and clustering more justifiable. Possibly unsupervised methods and visualization applied to the LiDAR data might be used to guide field work, identifying areas with unusual LiDAR distributions that should be included in the field data.

Semi-supervised learning combines supervised and unsupervised learning, to leverage the structure inherent in a dataset to better generalize from a small number of training items. Encoding the LiDAR data as points in a high-dimensional space with a natural metric should enable some semi-supervised as well as unsupervised learning techniques.

Level-of-detail analysis is another direction of research in computer vision that could be applied to a shape descriptor representation. Analyzing the same point cloud with shape descriptors of different sizes could reveal different features; features consistent across several levels of detail are considered more significant.

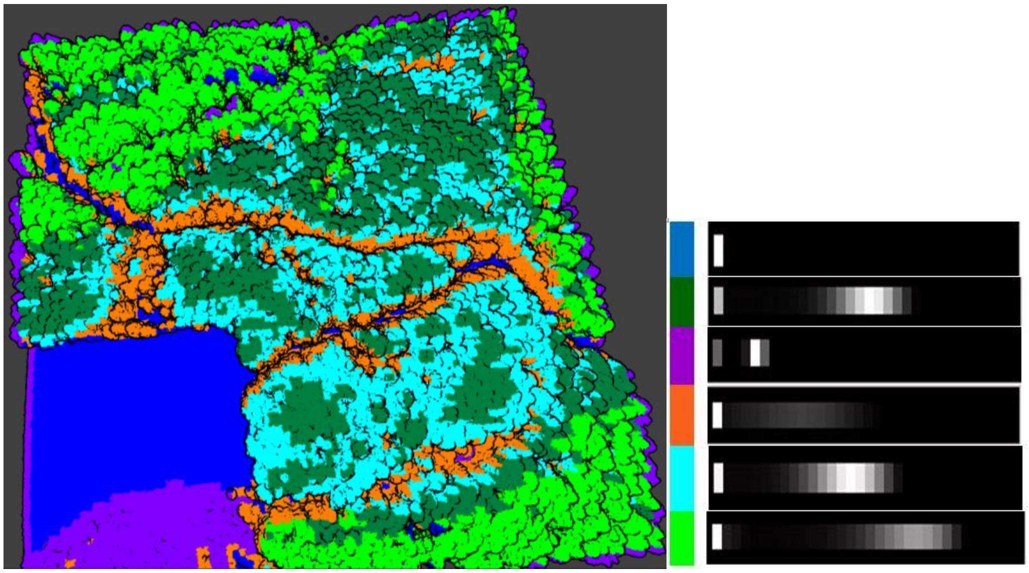

Figure 4: A tile from the Panther Creek data (left) containing a variety of ground covers. Clustering the 32 height bins does a good job of separating different stand types and revealing features. The orange cluster picks out roads or gaps, the deep blue is cleared, the purple is brush, the bright green is the tallest trees, followed by dark green and then light blue, with denser canopies. The figure on the right shows a visualization of the height clusters; brighter areas represent greater LiDAR density, with height increasing from left to right.

Acknowledgments

Many thanks to Jim Flewelling of Seattle Biometrics for providing the field and LiDAR survey data from the Panther Creek Cooperative Watershed Research project. Grateful thanks to NSF grants IIS-0964357 and IIS-1117663 for support. .

References

Belongie, S., Malik, J. & Puzicha, J., 2002. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 24(4), pp. 509-522.

Frome, A. et al., 2004. *Recognizing objects in range data using regional point descriptors*. Prague, Proceedings of the Eighth European Conference on Computer Vision.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. Object dissimilarity. In: *The elements of statistical learning*. Stanford: Springer, pp. 505-507.

Hyyppä, J. et al., 2008. Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. *International Journal of Remote Sensing,* 29(5), pp. 1339-1366.

Jenkins, J., Chojnacky, D., Heath, L. & Birdsey, R., 2004. *Comprehensive database of diameter-based biomass regressions for North American tree species,* Newtown Square: USDA Forest Service.

Johnson, A. E., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3D scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *21*(5), 433-449.

Lee, H., Slatton, K., Roth, B. & Cropper Jr., W., 2010. Adaptive clustering of airborne LiDAR data to segment individual tree crowns in managed pine forests. *International Journal of Remote Sensing,* 31(1), pp. 117-139.

Li, J., Hu, B. & Noland, T., 2013. Classification of tree species based on structural features derived from high density LiDAR data. *Agricultural and Forest Meteorology,* Volume 171-172, pp. 104-114.

Ling, H. & Okada, K., 2006. *Diffusion distance for histogram comparison*. s.l., IEEE Computer Society, pp. 246-253.

McGaughey, R. J., 2013. *FUSION/LDV: Software for LiDAR analysis and visualization,* Seattle: US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Popescu, S. C. & Wynne, R. H., 2004. Seeing the trees in the forest: Using LiDAR and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering & Remote Sensing,* 24061(0324).

Puzicha, J., Buhmann, J., Rubner, Y. & Tomasi, C., 1999. *Empirical evaluation of dissimilarity measures for color and texture*. Kerkyra, s.n., pp. 1165-1172.

Rubner, Y., Tomasi, C. & Guibas, L., 1998. *A metric for distributions with applications to image databases*. Washington DC, IEEE Computer Society, pp. 59-66.

Tittmann, P., Shafii, S., Hartsough, B. & Hamann, B., 2011. *Tree detection and delineation from LiDAR point clouds using RANSAC*. Proceedings of SilviLaser.

Yu, X. et al., 2011. Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS Journal of Photogrammetry and Remote Sensing,* 66(1), pp. 28-37.