A Fuzzy Mean-Shift Approach to Lidar Waveform Decomposition

Qinghua Li, Serkan Ural, John Anderson, and Jie Shan, *Senior Member, IEEE*

***Abstract*—Waveform decomposition is a common step for exploitation of full-waveform lidar data. Much effort has been focused on designing algorithms based on the assumption that the returned waveforms follow a Gaussian mixture model where each component is a Gaussian. However, many real examples show that the waveform components can be neither Gaussian nor symmetric even when the emitted signal is Gaussian or symmetric. This paper proposes a nonparametric mixture model to represent lidar waveforms without any constraints on the shape of the waveform components. A fuzzy mean-shift algorithm is then developed to decompose the waveforms. This approach has the following properties: 1) It does not assume that the waveforms follow any parametric or functional distributions; 2) the waveform decomposition is treated as a fuzzy data clustering problem and the number of components is determined during the time of decomposition; and 3) neither peak selection nor noise floor filtering prior to the decomposition is needed. Experiments are conducted on a dataset collected over a dense forest area where significant skewed waveforms are demonstrated. As the result of the waveform decomposition, a highly dense point cloud is generated, followed by a subsequent filtering step to create a fine digital elevation model. Compared with the conventional expectation–maximization method, the fuzzy mean-shift approach yielded practically comparable and similar results. However, it is about three times faster and tends to lead to slightly fewer artifacts in the resultant digital elevation model.**

***Index Terms*—Classification, fuzzy algorithm, LiDAR, mean-shift, waveform decomposition.**

I. Introduction

LIGHT detection and ranging (LiDAR) has been emerging as a direct 3-D topographic data collection technique and is extensively used in routine topographic mapping [1]. Recently, more advanced lidar measurement techniques were introduced and became available. One of the most common techniques is the full-waveform digitization lidar [2]–[5]. Instead of individual ranges, a full-waveform lidar system features sampling and recording of the whole backscattered signal at a temporal resolution of nanoseconds or subnanoseconds [6], [7]. The recorded signal, referred to as a (returned) waveform, consists of a series of temporal waves, with or without overlap, where each corresponds to an individual reflection from an object [8]. Compared with the process of analyzing traditional discrete return lidar systems, working with full-waveform data often needs one additional yet important step, i.e., modeling and decomposing the waveforms. In this step, the waveforms are decomposed into a number of independent components, each of which corresponds to a detected target, i.e., a point in the point cloud.

Different functions, including Gaussian, generalized Gaussian, lognormal, Weibull, Nakagami, and Burr, have been proposed to model lidar waveforms [9], [10]. Among them, the Gaussian mixture model (GMM) has prevailed for many years and is widely adopted [3], [11]–[15]. It models the returned waveform $y(t)$ as a weighted sum of a number of Gaussian components, i.e.,

$$y(t) = \sum_{i=1}^{C} w_i \cdot \exp\left[-\frac{(t-\mu_i)^2}{2\sigma_i^2}\right] \tag{1}$$

where t is the sampling time, y the intensity of the waveform, (μ_i, σ_i) the mean and the standard deviation of ith Gaussian component, and C the total number of Gaussian components. Both C and $\theta = \{\mu_1, \sigma_1, w_1, \mu_2, \sigma_2, w_2, \ldots, \mu_C, \sigma_C, w_C\}$ are parameters to be estimated.

Many algorithms were proposed to decompose the waveform data based on GMM. For example, in [11], the Levenberg–Marquardt optimization algorithm was used. The main challenge of such nonlinear optimization approach is its convergence being sensitive to the initial values of the unknown parameters. Another widely used and cited approach is the expectation–maximization (EM) method [13], [16], [17]. In EM, the number of waveform components needs to be predefined and may vary from waveform to waveform. Recently, there were efforts to apply wavelet methods to resolve the GMM model [14], [15], [17]; however, they cannot deal with cases where multiple waveform components are overlapped, which are common for complex land covers.

The popular GMM is based on two assumptions: 1) the emitted lidar signal is a Gaussian waveform; and 2) the returned waveform is also a Gaussian, although its mean, standard deviation, and the number of Gaussian waveform components may vary from waveform to waveform. However, this is not always true. In fact, although the laser sensors can generate an electromagnetic wave that ideally leads the intensity of the emitted signals to a Gaussian [18], the intrinsic noise in the electrical device will introduce flat tails to the actual signal, as shown in Fig. 1(a). The tails make the curve look more like a Cauchy curve than a Gaussian [19]. Moreover, the transmission path can also alternate the signal in a way that deviates from Gaussian. For example, one of the properties of Gaussian waves is symmetry, but the multipath effect will yield

Manuscript received February 16, 2016; revised May 30, 2016; accepted June 23, 2016. Date of publication September 6, 2016; date of current version September 30, 2016. This work was supported in part by the U.S. Army Research Office. *(Corresponding author: Jie Shan.)*

Q. Li, S. Ural, and J. Shan are with the Lyles School of Civil Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: li975@purdue.edu; sural@purdue.edu; jshan@purdue.edu).

J. Anderson is with the Engineer Research and Development Center, U.S. Army Corps of Engineers, Washington, DC 20314 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2016.2596105

0196-2892 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

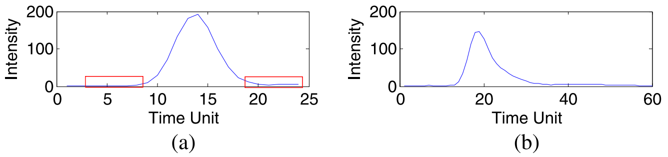


Fig. 1. Non-Gaussian signal in lidar. (a) Emitted lidar signal of Riegl Q680i. The flat tails (noninformative) are highlighted in boxes. (b) Right-skewed returned waveform of (a).

a Rayleigh channel [20], which makes a single waveform component asymmetric, skewed, and non-Gaussian [10], [21]–[23], as shown in Fig. 1(b).

In addition to this unrealistic GMM assumption, the existing decomposition techniques, e.g., EM and Levenberg–Marquardt methods, have the following limitations. First, the number of total decomposed components C is usually prescribed in the stage of initialization. To determine the optimal C, peak detection should be applied, and the algorithms need to run multiple times to compare the results from different C's. Second, a filtering process is desired prior to the waveform decomposition because the involvement of noise will deteriorate the performance of decomposition. However, there is no standard way to filter out the noise floor [24] in waveforms. One simple and popular way is to apply a constant noise threshold [25], although it is subjective. As a result, the previously described decomposition algorithms are sensitive to the threshold value, and the waveform is often either over or under filtered. Some recent works do consider modeling the waveforms with more complex distributions [22], [26]. They build a library of models and choose the most suitable ones with Bayesian analysis. However, their effectiveness is constrained by the creation of a comprehensive model library and the complexity in selecting one or more best models for a particular application. As a result, such methods are not practical to handle large volume lidar data and do not meet the needs of diverse applications.

Unlike the existing methods that use a mathematical parametric function, we introduce a nonparametric model to describe the waveforms. The nonparametric model does not constrain the shape of the waveform components so that asymmetric, non-Gaussian waveform components can be included. Specifically, we model a waveform as a histogram of a collection of random samples of x, $X = \{x_n : 0 \leq x_n \leq T, 1 \leq n \leq N\}$. Under this nonparametric model, the procedure of waveform decomposition will be realized by clustering techniques. Considering the number of waveform components C as an unknown parameter in addition to the waveform components themselves, we introduce a fuzzy mean-shift (FMS) algorithm so that this number can be simultaneously estimated during the process of waveform decomposition. As a result, the waveform is decomposed into several clusters, either symmetric or asymmetric and either informative (useful signal) or noninformative (noise). Compared with existing methods, the significant properties of our development are that it does not need to assume the waveform components to be a Gaussian or of any parametric form, and the number of waveform components be known beforehand. Moreover, the noise floor can be filtered out during the same time of waveform decomposition. Finally, the peak time is the cluster center, i.e., the mass center of a determined component.

The remaining part of this paper is structured as follows. Section II will describe the proposed method, including the nonparametric mixture model (NMM) and the FMS algorithm. The study area and the dataset are described in Section III. The waveform decomposition with the FMS approach is implemented and discussed in Section IV; as an application of waveform decomposition, a digital elevation model (DEM) using waveform decomposition results is created, and a preliminary evaluation on its quality is conducted. Finally, this paper is concluded in Section V.

II. Methods

Our objective is to design a waveform model that: 1) takes possibly asymmetric waveform components into account; 2) allows the popular GMM model to be its special case; and 3) yields a high-quality high-efficiency decomposition. To address the first two properties, we introduce a NMM, whereas for achieving the third property, we propose an FMS clustering algorithm.

A. NMM

Let a (returned) waveform be collected along time t from tag 1 to T. We regard the waveform $Y = \{y(t) : y(t) \geq 0, 1 \leq t \leq T\}$ as the histogram of samples of a random variable x, $X = \{x_n : 0 \leq x_n \leq T, 1 \leq n \leq N\}$, where n is the sample index, and N the total number of samples. The total number of samples will be $N = \sum_{t=1}^{T} y(t)$. Thus, the waveform $y(t)$ can be modeled by a nonparametric function as

$$y(t) = \frac{1}{N \times h} \sum_{n=1}^{N} k\left(\frac{t - x_n}{h}\right). \tag{2}$$

Equation (2) is nonparametric because the waveform is not assumed to follow any specific distribution. Instead, each sample of x contributes to the waveform through a kernel function $k(\cdot)$, which satisfies

$$\begin{gathered} \sup |k(x)| < \infty \\ \int_R |k(x)| < \infty \\ \lim_{|x| \to \infty} |x| \cdot k(x) = 0 \\ \int_R |k(x)|\, dx = 1 \end{gathered} \tag{3}$$

and the parameter h is the bandwidth of the kernel function, which determines the density of the lidar point cloud to be generated through this decomposition. Examples of the kernel functions include a rectangle function, a triangle function, or a Gaussian function. A complex kernel function such as a Gaussian can make the histogram smoother, whereas a simple kernel function such as a rectangle may consume less computation time.

When more than one target is encountered by an emitted lidar signal, there will be multiple components $y_1(t), y_2(t), \ldots, y_C(t)$ in the (returned) waveform. Each waveform

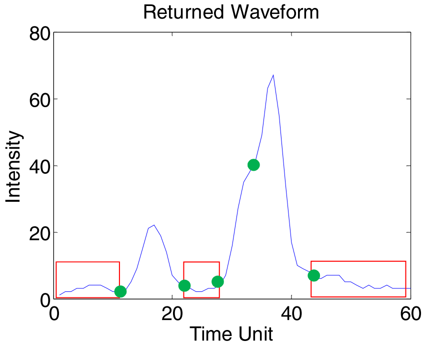


Fig. 2. Returned waveform of Riegl Q680i. The noninformative noise is highlighted in boxes, and the possible mixture positions of two clusters are in dots.

component $y_i(t)$ is a cluster of x samples. This phenomenon can be described by a mixture model as

$$y(t) = \sum_{i=1}^{C} y_i(t) = \sum_{i=1}^{C} \left[\frac{1}{N_i \times h} \sum_{n_i=1}^{N_i} k\left(\frac{t - x_{n_i}}{h}\right) \right] \tag{4}$$

where $y_i(t)$ is the ith component and C is the total number of components. Here, $y_i(t)$ is modeled as a summation of a number of kernel functions evaluated at each point in cluster i. Depending on the clustering result, $y_i(t)$ could fit a Gaussian or any other forms, either symmetric or asymmetric. It is noted that $y_i(t)$ could be either informative laser energy or noninformative noise, as shown in Fig. 2. It is desired that an algorithm can determine and filter the noise levels at the same time as decomposing the waveforms.

In (4), the samples of x are clustered in a way that each sample belongs to a single component y_i; therefore, all the waveform components are separated without overlap. However, in most cases, lidar waveform components do convolute and are mixed. For example, the waveform intensity at the dots in Fig. 2 is contributed by two adjacent, mixed waveform components. To model such phenomenon, a mixture weight w_n is introduced, and (4) is evolved to a mixture model, i.e.,

$$y(t) = \sum_{i=1}^{C} y_i(t) = \sum_{i=1}^{C} \left[\frac{1}{N_i \times h} \sum_{n_i=1}^{N_i} w_{n_i} \cdot k\left(\frac{t - x_{n_i}}{h}\right) \right] \tag{5}$$

where a sample x_n can be assigned to multiple clusters with the weight w_{n_i}. The NMM in (5) is the one we introduce for waveform decomposition. We will design an algorithm to determine C, w_{n_i}, and $y_i(t)$, with a preset bandwidth h.

B. FMS Algorithm

With NMM, the problem of waveform decomposition becomes a kernel clustering problem. Some popular kernel-based clustering methods include the kernel k-means [27], spectral clustering [28], support vector machine clustering [29], and mean-shift clustering [30]. Among the different clustering methods, the mean-shift algorithm offers the advantage of being simple and able to estimate the number of clusters at the same time the clustering is performed. This is a desired property for waveform decomposition because the number of waveform components is unknown, varying from waveform to waveform, and most of the existing methods have to put extra effort to estimate it. Moreover, the traditional mean-shift algorithm is a hard or crispy clustering algorithm suitable for problems described by (4), but not suitable for the mixture problems defined in (5). Here, we will adapt the traditional mean-shift algorithm to an FMS so that it can be utilized to decompose waveforms with NMM.

The traditional mean-shift algorithm estimates the local modes of the histogram from a random variable x, as we described in (2) and (4). At any t, the gradient of the waveform function (2) and (4) can be calculated using the following:

$$\frac{\partial y(t)}{\partial t} = \sum_{i=1}^{C} \left[\frac{1}{N_i \times h^2} \sum_{n_i=1}^{N_i} k'\left(\frac{t - x_{n_i}}{h}\right) \right]. \tag{6}$$

Next, we define $\overrightarrow{m}(t)$, as the mean-shift vector shown in (7). Its sign indicates the direction of the gradient of the waveform function, i.e., it points toward the region where the majority points of a waveform component reside. In practice, the kernel function $k(\cdot)$ is chosen in such a way that only the data points within the neighborhood (∂t) of x_n contribute to $\overrightarrow{m}(t)$. Thus, $\overrightarrow{m}(t)$ can be written as

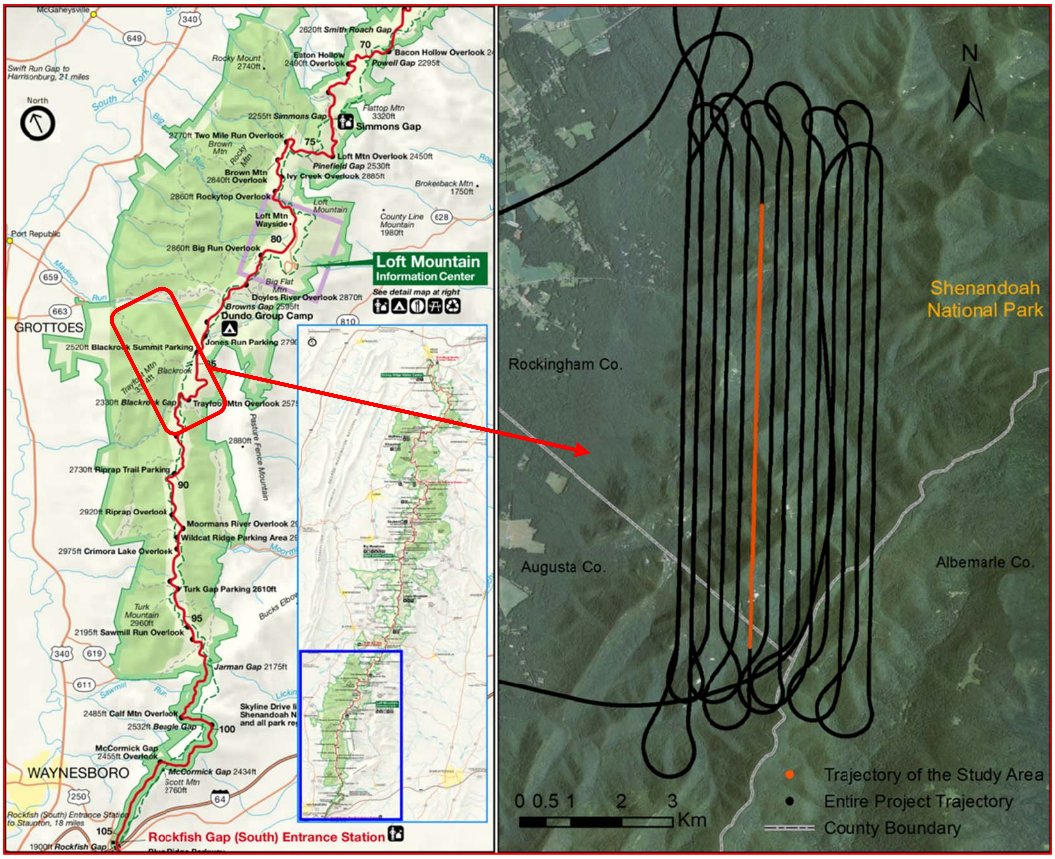
$$\overrightarrow{m}(t) = \frac{\sum_{x_n \in \partial t} x_n k\left(\frac{t-x_n}{h}\right)}{\sum_{x_n \in \partial t} k\left(\frac{t-x_n}{h}\right)} - t. \tag{7}$$

The algorithm starts from an initial point x_0 and moves along the mean-shift vector until it converges at $\overrightarrow{m}(t) = 0$. The center of the first waveform component (either informative or noise) is found at this point. During the process, all the data points of x_n that are once located within the search neighborhood will contribute to the vector $\overrightarrow{m}(t)$ and form the first waveform component $y_1(t)$. After that, the process will be conducted on the rest of the data points and search for the second waveform component $y_2(t)$. The algorithm continues until all the x_n samples are used to form one of the waveform components. In the end, the algorithm returns all the waveform components $y_1(t), \ldots, y_C(t)$ together with C, i.e., the number of found components.

As pointed earlier, the traditional mean-shift algorithm conducts a hard clustering by assigning each data point x_n to a single component and returns a result where all the waveform components are separated without overlap or mixture. To resolve the mixture of waveform components, we adapt the traditional mean-shift to an FMS to conduct soft clustering. We notice that the sample points at the border of adjacent waveform components are visited multiple times by different clusters during the mean-shift process. We define such a point as a mixture point and assign it through a fuzzy weight function to all the clusters that once visited the point, i.e.,

$$w_j(x_n) = \frac{v_j}{\sum_{i=1}^{C} v_i} \tag{8}$$

where v_j is the number of times that x_n is visited by cluster j. In most cases, a mixture point will be visited by only two adjacent components.



(a) (b)

Fig. 3. Shenandoah National Park and flight trajectory. (a) Park map (National Park Services: http://www.nps.gov/shen/planyourvisit/maps.htm) and (b) flight trajectory of the sensor platform over the study area under Microsoft Bing Maps. The black lines are the trajectory of the airplane, and the single orange line in the middle highlights the section where lidar sensor collected the data used in this paper.

III. Study Area and Data

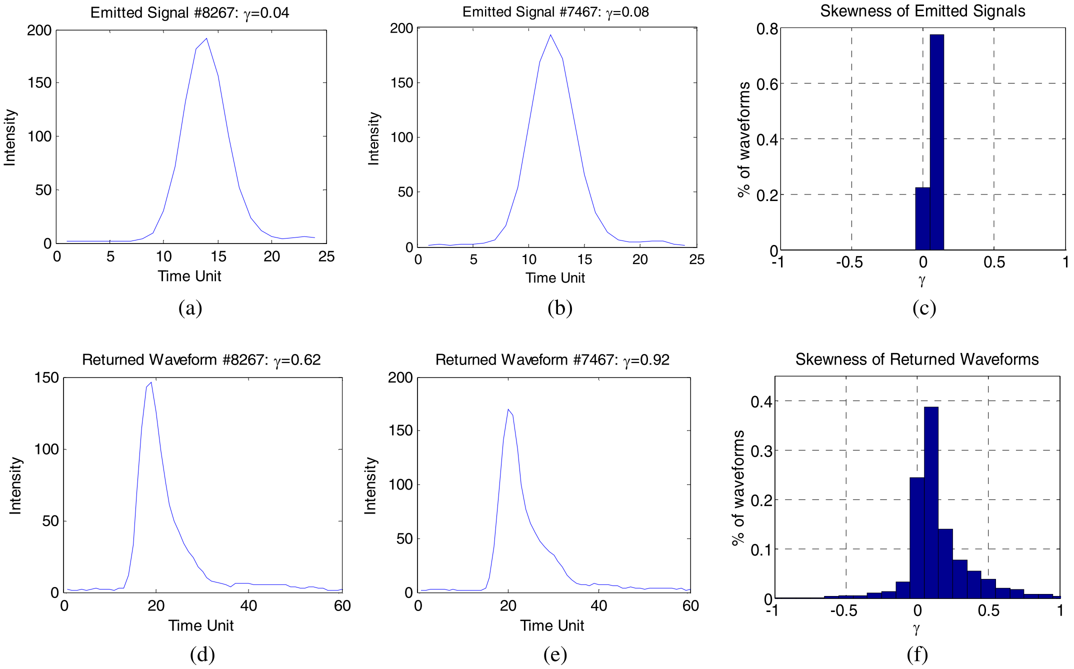
The study area is over the southern district of Shenandoah National Park of Virginia, USA. It is a dense forest area with limited visible ground exposure. The district is full of 132 species of trees and has an elevation range of 171–1234 m relative to the North American Vertical Datum 1988 (NAVD 88) [31]. The maps of the Shenandoah National Park are shown in Fig. 3. The black lines are the trajectory of the airplane, and the single orange line in the middle highlights the section where lidar sensor collected the data used in our tests.

The data were collected by a Riegl Q680i airborne laser scanner that recorded all the emitted signals in this flight with a sampling rate of 1 ns. Some emitted signals are plotted in Figs. 1(a) and 4(a) and (b). All the emitted signals have similar characteristics of symmetry, i.e., a Gaussian signal. The dataset we used for implementation has two main files. A binary full-waveform scan data file in ".sdf" format was collected by the sensor. This file was accompanied with a binary smoothed best estimate of trajectory (SBET) file in ".out" format, which stores the trajectory and flight dynamics of the airplane. The waveform file contains approximately 50 000 000 recorded waveforms and is about 8 GB in size. The part of the trajectory file that corresponds to the acquisition of this laser-scan data set consists of approximately 45 000 trajectory points and other related records, e.g., speed and orientation of the platform for each trajectory point. Fig. 3(b) shows the trajectory of the platform over the study area and the complete trajectory of the whole flight [25]. The SBET file contains a 136-B standard navigation record, which includes the time, position, speed, orientation, platform heading, wander angle, and acceleration information at a sampling rate of 200 Hz. We access this binary file by using the "IceBridge Applanix SBET file Perl reader" provided by the National Snow and Ice Data Center (NSIDC).

The full-waveform binary file is accessible via the RiWaveLib C++ waveform extraction library provided by Riegl as an interface to extract information from the waveform data. A specific waveform dataset consists of all sample blocks of the waveforms detected by the receiver and additional information, e.g., the beam direction. Data from two different channels, namely, high power and low power channels, are recorded together with the reference channel for the emitted pulse. A timestamp for the start of each sample block is recorded, which allows to determine the time of all samples in the same sample block using the sample interval. The dataset consists of only one swath of flight, 770 m wide (West–East) and 6700 m long (North–South). The flight height above ground is about 500 m.

IV. Discussion and Evaluation

This section will examine the asymmetric properties of the waveform, discuss the selection of bandwidth h of the kernel function, explore the mixture of waveform components, and evaluate the waveform decomposition by comparing its



(a) (b) (c)

(d) (e) (f)

Fig. 4. Skewness of lidar emitted signals and returned waveforms. (a)–(c) Emitted signals: (a) #8267, $\gamma = 0.04$. (b) #7467, $\gamma = 0.08$. (c) the distribution of skewness of the emitted signals. (d)–(f) Returned waveforms: (d) #8267, $\gamma = 0.62$. (e) #7467, $\gamma = 0.92$. (f) Distribution of skewness of the one-component returned waveforms.

resultant DEM with reference to the USGS National Elevation Dataset (NED) 1/3″ DEM.

A. *Asymmetric Waveforms*

The symmetry of the returned waveform is measured by the Pearson's moment coefficient of skewness γ [32], i.e.,

$$\gamma = \frac{m_3}{m_2^{\frac{3}{2}}} \tag{9}$$

where

$$\bar{t} = \frac{\sum_{t=1}^{T} t \cdot y(t)}{\sum_{t=1}^{T} y(t)}$$
$$m_i = \frac{\sum_{t=1}^{T} (t - \bar{t})^i \cdot y(t)}{\sum_{t=1}^{T} y(t)}. \tag{10}$$

A positive γ suggests a right skewed waveform and a negative γ a left skewed one; $\gamma = 0$ represents a symmetric waveform. To study the asymmetry of the waveforms, we conduct an experiment on the first 10 000 waveforms of the dataset. The MDL-EM method [25] is applied to the subset and finds 2084 (20.8%) waveforms with only one component. As examples, two emitted signals and their returned waveforms are shown in Fig. 4(a), (b), (d), and (e), respectively. Fig. 4(c) exhibits the distribution of γ's for all the 10 000 emitted signals, whereas Fig. 4(f) is the distribution of γ's for the one-component returned waveforms. Clearly, the emitted signals are statistically symmetric, where their skewness has a mean of 0.0563 and standard deviation of 0.0194. In contrast, the returned waveforms are considerably skewed with a mean of 0.1489 and a standard deviation of 0.239. Compared with Fig. 4(c) for the emitted signals, Fig. 4(f) demonstrates that the skewness of 34% of the returned waveforms are greater than 0.15. In other words, right-skewed returned waveforms are prevalent in this subset.

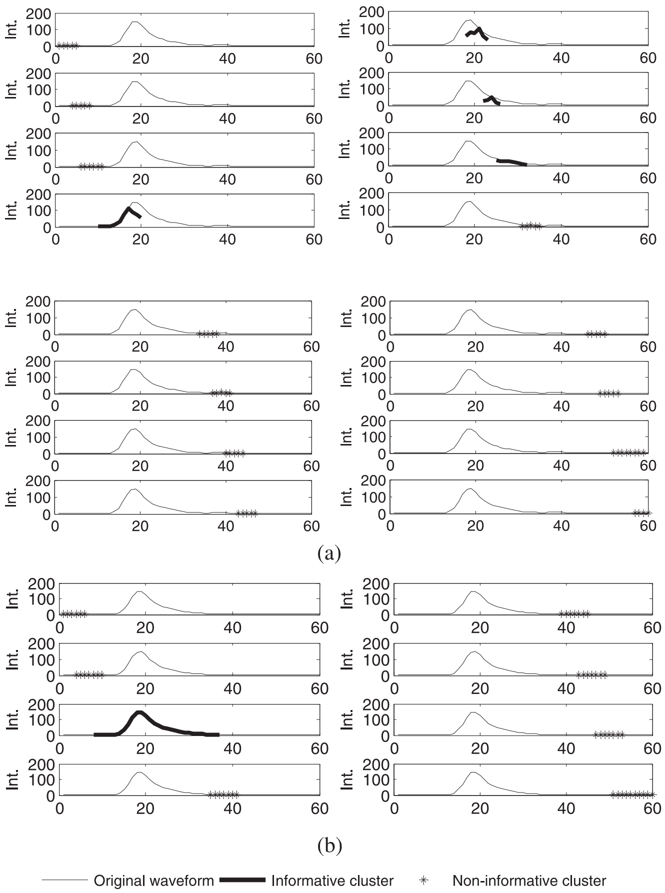
Due to the noticeable existence of asymmetry in our dataset, the waveforms should be modeled by a method that takes the asymmetric components into account and be decomposed by appropriate algorithms. This demonstrates the need for this paper. The nonparametric mixture model and previously described FMS algorithm will be utilized for waveform decomposition.

B. *Selection of the Kernel Bandwidth h*

Application of the kernel density model in the waveforms starts from choosing a kernel function $k(\cdot)$ in (2). At this point, we have no evidence that any specific kernel function is superior to others for decomposing lidar data. Hereafter, a rectangle kernel function is chosen in this paper to demonstrate the algorithm since the use of a complex kernel (e.g., cosine or Gaussian kernels) is primarily to smooth the estimated density function. Moreover, the simplicity of the rectangle kernel function is desirable when dealing with a large volume of waveform data.

The kernel bandwidth h is an important factor that controls the density of the point clouds to be generated. Some of the techniques for bandwidth selection include asymptotic expansion (AMISE) minimization [33], stability maximization, and isolation-connectivity optimization. In the application of lidar waveform processing, a larger bandwidth will yield a smaller volume of point cloud and a narrower one will lead to a denser point cloud.

In our experiment, we choose $h = 3.3$ as the bandwidth, where h is measured by the number of sampling time intervals or data points. The following facts are considered for selecting such value. First, the bandwidth essentially determines the minimum width of a waveform component. Since we need at



(a)

(b)

Fig. 5. Effects of the bandwidth of the kernel function on waveform #8267 [see Fig. 4(d)] decomposition. (a) Inappropriately small bandwidth $h = 2.6$ wrongly leads to four informative clusters and 12 noninformative clusters. (b) Proper bandwidth $h = 3.3$ correctly leads to one informative cluster and seven noninformative clusters.

least three data points to determine a bell-shaped or second-order waveform component so that a meaningful peak can exist, the bandwidth should satisfy $h \geq 3$, i.e., a waveform component should have minimum three data points. Second, the bandwidth also determines the minimum separable distance between targets in ranging by $\Delta r = h \times c \times \Delta t/2$, where Δr is the minimal separable distance, c is the speed of light, and $\Delta t = 1$ ns is the sampling time interval. We let $\Delta r \leq 0.5$ m, which leads to $h \leq 2\Delta r/(c \times \Delta t) = 3.33$. Finally, an inappropriate small bandwidth may mistakenly generate artifacts. Combing all above considerations, $h = 3.3$ is chosen. Cautions should be taken when waveforms are stretched by the high slope of the terrain and the multipath effect. We address this phenomenon by demonstrating the decomposition result of one such broadened waveform. The original waveform is shown in Fig. 4(d) and shown as dashed lines in Fig. 5. Its decomposition results under two different bandwidths are also shown in Fig. 5. The plot plates in Fig. 5 are ordered horizontally, and plots within each plate are ordered vertically. It depicts the time sequence the clusters are found through the FMS algorithm. Different cluster types are shown in different line styles. In Fig. 5(a), a bandwidth $h = 2.6$ yields 16 waveform components, among which four are informative (in solid lines) and 12 noninformative (star lines). Apparently, the algorithm mistakenly generates false components for this waveform. A more realistic decomposition result can be obtained when $h = 3.3$, as shown in Fig. 5(b). It is noticed that adjacent clusters often overlap, i.e., one waveform data point can belong to more than one cluster, as suggested by the principle of the FMS clustering approach.

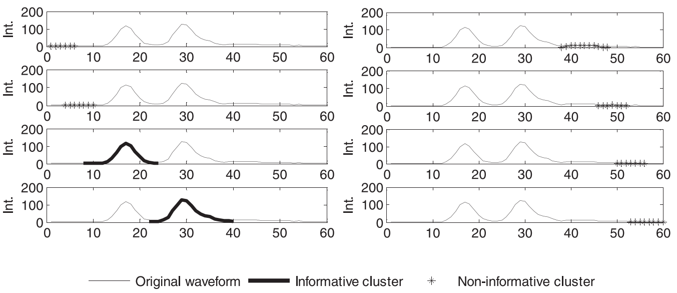
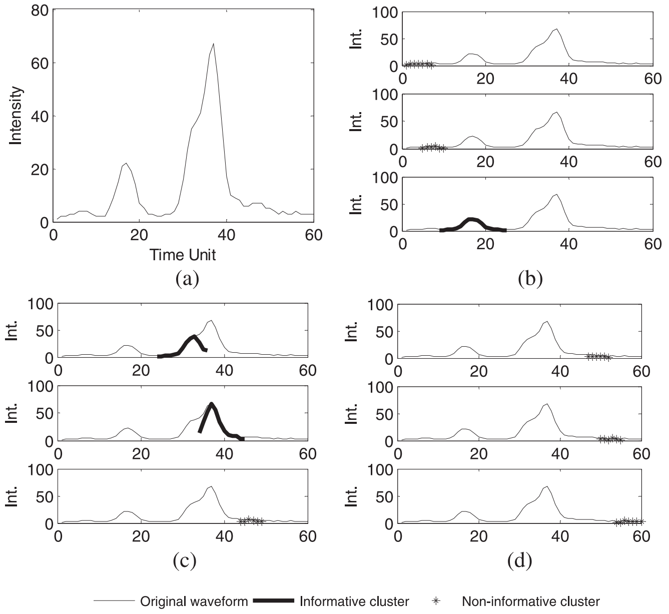


Fig. 6. Waveform decomposition and noise floor filtering on waveform #8269 with the FMS method. There are eight components found, starting from the upper left ordered in vertical direction first.



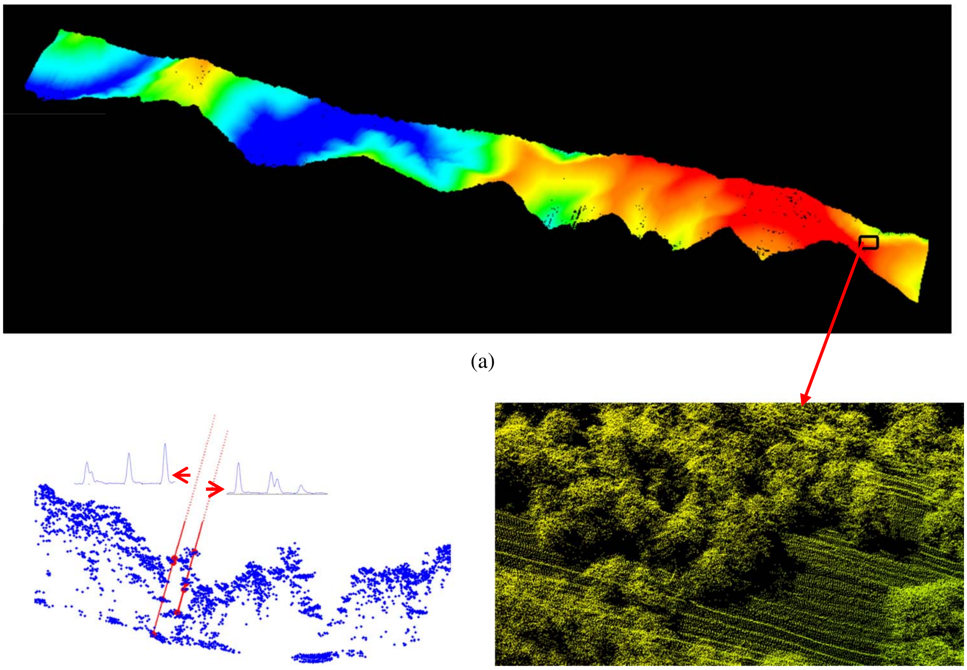
(a) (b)

(c) (d)

Fig. 7. FMS decomposition of a waveform where the mixtures exhibit (a) the waveform and (b)–(d) the decomposed components.

C. Waveform Decomposition

The position of the initial point x_0, as described in Section II-B, does not affect the decomposed result. We start the FMS algorithm from $t_0 = 1$. The algorithm moves toward right until the mean-shift vector equals to zero. All the points visited during the moving process are plotted as stars, as shown in the upper left plot in Fig. 6. FMS then restarts from the left most point of the remaining points and finds out the other clusters. We can see that there is overlap between the adjacent clusters because the point in the overlapped area contributes to the mean-shift vectors for both clusters. As an example, the decomposed waveform (#8269) is shown in Fig. 6. There are altogether eight clusters (components) as plotted in stars or solid lines. The informative (useful signal) and noninformative (noise) clusters can be distinguished by setting a threshold on the cluster size N_i in (5), which is chosen as 100 in this paper.



(a)

(b)

(c)

Fig. 8. Point clouds generated from the FMS approach. (a) Entire profile of the point cloud. (b) Two sample waveforms, each with four components, corresponding to four points (in red dots). (c) Blow-up section of the point cloud.

The informative clusters $(\text{size} >= 100)$ are plotted in solid lines and the noninformative ones $(\text{size} < 100)$ in stars.

Decomposing a waveform where mixture exists is the most meaningful use of waveform decomposition. Here, we illustrate another example of FMS applied on such case. The original waveform is plotted in Fig. 7(a). There may be three waveform components, one isolated and two mixed, as would be suggested by the EM algorithm. The FMS algorithm returns the decomposed results in Fig. 7(b)–(d). The mixed component is plotted in solid bold lines in Fig. 7(c). Each component could be fine-tuned by applying a threshold on the intensities so that only the points above the threshold are utilized to generate the point clouds. Since the waveform is decomposed, the threshold will be much lower than the one used in traditional prefiltering. There are nine waveform components found by the FMS method; their weighted sum will be the original waveform.

Once the decomposition is completed, the decomposed waveform components are used to determine the time that the signal is returned from the target. We take the mass center of each waveform component to determine its time

$$t_i = \frac{\sum_{n=1}^{N_i} w_n \cdot x_n}{\sum_{n=1}^{N_i} w_n}. \tag{11}$$

Similarly, for the corresponding emitted signal, its outgoing time is

$$t_{\text{out}} = \frac{\sum_{n=1}^{N_{\text{out}}} z_n}{N_{\text{out}}} \tag{12}$$

where z_n in (12) is defined the same way as we define x_n, except that z_n is for the emitted signal. t_i and t_{out} are the time of the waveform component returned from the target and the time the laser signal is emitted, respectively. N_i and N_{out} are, respectively, the number of samples within the ith waveform component and the emitted signal. The range between the target and the sensor is calculated by using the time the laser shot takes to travel to and from the target and the speed of light, i.e.,

$$\rho_i = \frac{v}{2}(t_i - t_{\text{out}}) \tag{13}$$

where v is the speed of light, and ρ_i is the range between the sensor and the target. ρ_i will be utilized to determine the coordinates of point clouds.

D. Point Clouds and DEM

Having the ranges and the directions of the emitted laser pulses, we can calculate the coordinates of the targets in the sensor coordinate frame. Using the positions of the platform recorded in the SBET file, a series of coordinate transforms [34] are then carried out to calculate the coordinates of the targets. It should be noted that the flight trajectory is not available at the same frequency as the pulse repetition rate of the lidar sensor. The pulse repetition rate of the lidar system is at about 400 kHz, whereas the trajectory is available at 200 Hz. Only one trajectory position is available for approximately every 2000 laser pulses due to this difference in data frequency. The position and the orientation of the sensor platform are calculated by linear interpolation of the trajectory at each instance of laser pulse emission.

As shown in Fig. 8, we generated a point cloud by using the FMS approach. For comparison, we ran EM on the same data set. It is noted that there are various implementations of EM, and each of them can tune parameters differently. Moreover, since EM itself cannot determine the number of components in the waveform, a peak detection procedure or a model selection method has to be applied. In our experiment, a peak detection initialization and a model selection method based

TABLE I
NUMBER OF POINTS GENERATED BY WAVEFORM DECOMPOSITION AND THE SUBSEQUENT GROUND FILTERING. ALL COMPUTATIONS WERE COMPLETED ON A DELL PRECISION WORKSTATION WITH EIGHT-CORE 3.7-GHZ CPU AND 16-GB MEMORY

		EM with GMM	FMS with NMM	FMS/EM
Decom-position	Total CPU Time (sec.)	498,052.85	170,349.52	0.34
	CPU Time per waveform (ms.)	9.996	3.406	0.34
	#total points	112,769,843	115,543,824	1.02
	#points/sq. m	21.9	22.4	1.02
Filtering	#ground points	7,817,965	8,208,448	1.05
	#ground points/sq. m	1.5	1.6	1.05

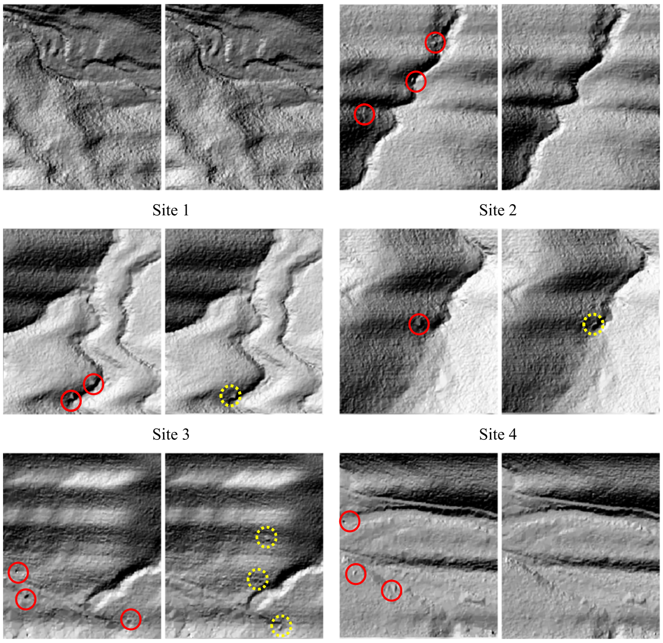

Site 1
Site 2
Site 3
Site 4
Site 5
Site 6

Fig. 9. Hillshading of six (6) DEM samples (195 m × 230 m) generated by EM (left) and FMS (right).

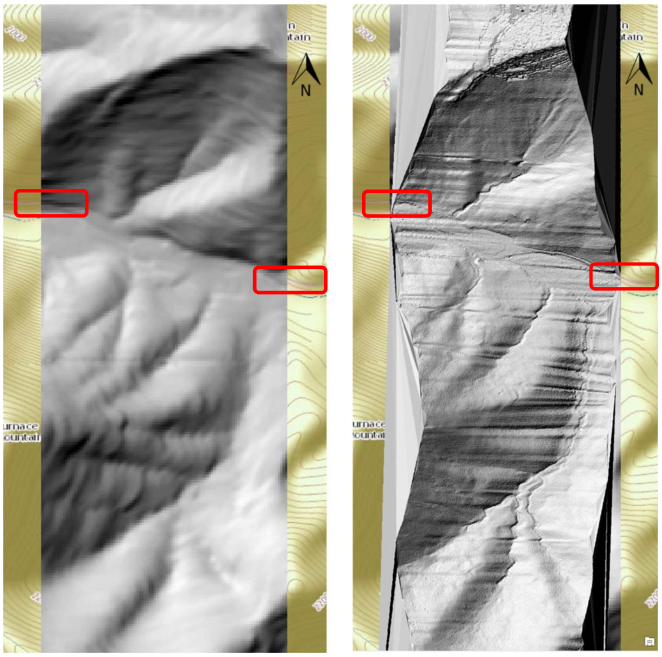


Fig. 10. Hillshading generated from NED 1/3″ DEM (left) and FMS DEM (right). A road should exist as indicated by the red boxes. Many fine topographic features are also clearly visible in FMS DEM. The background map is USTopo map. The size of this area is about 770 m wide and 2600 m long.

on the minimum description length (MDL) [25], [35] were used to determine the number of components in a waveform. The penalty parameter of the MDL-constrained EM method was adjusted such that it generated about the same number of waveform components as the FMS did. In this way, the results of the two methods are comparable. The number of points and computing time of both EM and FMS are summarized in Table I.

Since one of the applications for waveform decomposition is to create DEM using the waveform-generated discrete point clouds, we extracted the ground points using a filter implemented in the LAStools (http://rapidlasso.com/) software. It uses a variation of the Axelsson's [36] triangulated irregular network (TIN) refinement algorithm that avoids some of the trigonometry overhead. We used the default parameter settings to filter the point clouds resultant from the FMS and EM, respectively. Notice from Table I that the ground point density after filtering is 1.5 (EM) and 1.6 (FMS) points per square meter, respectively. Both are suitable to generate a DEM at 1-m resolution. DEMs at a resolution of 1 m were then created with ArcGIS 10.3.1. The ground elevation of the DEM ranges from 422 to 1008 m.

As shown in Fig. 9, hillshading of six 195 m × 230 m sites in the study area were selected to highlight the terrain details from the two decomposition methods. Circles of solid and dashed lines are, respectively, used to label locations of apparent artifacts in the EM and FMS results. As a general observation, the two DEMs have very comparable qualities. Both can very satisfactorily reflect the terrain morphology and almost all topographical details. On the other hand, both results have a few noticeable small scale pits. These artefacts, both in size and in number, are actually minimal, considering the fact that the terrain is under heavy canopy. Examining the hillshading in a closer view, the FMS approach created slightly fewer artifacts (Sites 2, 3, 5, and 6) than EM. It should be noted that the artifacts in both DEMs were likely introduced by imperfect filtering. As many studies reported [37]–[39], high-quality DEM generation under dense forest canopy is still a challenging task for lidar data processing.

As a primary study to the vertical accuracy of the resultant FMS DEM, its hillshading is stacked over the USTopo in Fig. 10. The waveform DEM significantly reveals more detailed terrain features. For example, a road under the forest can be successfully detected in the waveform DEM, whereas it is not visible in the NED 1/3″ DEM, which has a resolution of approximate 8 m. Other topographic features, including ridges, valleys, and gullies, are also clearly presented. On the other hand, suspicious striping effect in west–east (perpendicular to the flight) direction is noticed. Such artifacts are likely from the jitter of the airplane and/or the sensor, which has not been considered or fully corrected by the data vendor. The pixel by pixel differences between the two DEMs show a very good normal distribution with a mean of 0.89 m and a standard deviation of 6.99 m. This result basically shows that there is no significant systematic bias existing in our waveform DEM.

According to the specification of this NED 1/3″ DEM dataset, it was from compilation of the NED 1″ DEM, which has a nominal vertical accuracy of 7 m. Therefore, we can only reason the vertical errors of the waveform DEM are negligible compared with the quality of the NED 1/3″ DEM. Comprehensive and more convincing evaluation on the waveform DEM's accuracy remains to be a future effort.

V. Conclusion

Effective waveform decomposition needs both simplicity and accuracy. On one hand, the simplest and widely adopted decomposition method is peak detection [40], [41]. It is attractive because the waveform components can simply be detected by applying a threshold. It has the desired high efficiency for coping with large volumes of lidar data but does not take mixtures into consideration. As a result, it sacrifices the high accuracy and fidelity that come with advanced lidar systems in exchange for such simplicity. On the other hand, the GMM and corresponding algorithms are among the most popular waveform decomposition methods. They are supposed to be more accurate since they take into account the mixture of waveform components. However, the complexity of the algorithms are greatly increased not only because the algorithms themselves are iterative but because a preprocessing for estimating the number of components and/or a postprocessing for optimizing this number are often needed. In lidar waveform decomposition, any additional preprocessing and/or postprocessing are computationally expensive because they ought to apply to each and every individual waveform. Moreover, the GMM is unable to precisely model non-Gaussian waveform components, particularly asymmetric waveform components that have been frequently reported by researchers.

This paper introduced a nonparametric mixture model that describes asymmetric lidar waveform components and leads to a general approach to waveform decomposition. Compared with the GMM, the nonparameter mixture model successfully models a variety of waveform components, regardless of whether they are Gaussian or non-Gaussian and symmetric or asymmetric. The FMS algorithm essentially is a density-based data clustering approach, which does not assume that waveforms follow any functional or parametric distribution. Unlike many existing practices, the FMS algorithm does not need peak detection prior to decomposition and can simultaneously determine the number of waveform components during the decomposition process. Furthermore, the point density of the decomposed waveforms largely relies on one single parameter, i.e., the kernel bandwidth, for which a value of 3–4 times the waveform sampling interval has been shown suitable in this paper. Our tests with small footprint lidar data over a dense forest area have validated the noticeable asymmetry of returned waveforms and demonstrated satisfactory performance of the proposed FMS method. A detailed DEM with minimum artifacts can be produced through the subsequent filtering. Compared with the conventional EM method under an optimal implementation, the FMS approach is about three times faster, whereas the resultant DEM is very similar and tends to have slightly fewer artifacts. This paper not only develops a novel theoretical model and general solution to the waveform decomposition problem but practically provides a promising satisfactory approach to terrain generation under heavy canopy. This is useful for studies in geomorphology, hydrology, and other Earth science subjects. The decomposed waveform components can be further utilized for vegetation classification, biomass estimation, and single tree detection.

Future studies can be carried out in a wide range of topics. They may include optimal kernel selection and bandwidth selection, and alternative fuzzy clustering algorithms. There is also a need to comprehensively evaluate the proposed approach with reference to the popular EM method and its variations. To be specific, further studies may look into the necessity of the consideration of asymmetry and the adoption of nonparametric mixture model in urban areas since, in such scenarios, the symmetry in the emitted signals may not be greatly changed in the returned waveforms. For vegetation, forestry, and bathymetry studies, further exploration is needed to examine the ability of the FMS method to extract near-ground returns under low vegetation and dense canopy for subsequent construction of DEMs and its comparison to higher order accuracy measures.

Acknowledgment

The authors would like to thank the critical and constructive comments and suggestions from the anonymous reviewers, which greatly helped us improve the initial manuscript.

References

[1] J. Shan and C. Toth, *Topographic Laser Ranging and Scanning: Principles and Processing*. Boca Raton, FL, USA: CRC Press, 2009.

[2] C. Mallet and F. Bretar, "Full-waveform topographic LiDAR: State-of-art," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, pp. 1–16, 2009.

[3] C. E. Parrish and R. Nowak, "Improved approach to LiDAR airport obstruction surveying using full-waveform data," *J. Surv. Eng.*, vol. 135, no. 2, pp. 72–82, 2009.

[4] Y. Qin, T. T. Vu, and Y. Ban, "Toward an optimal algorithm for LiDAR waveform decomposition," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 482–486, May 2012.

[5] P. J. Hartzell, C. L. Glennie, and D. C. Finnegan, "Empirical waveform decomposition and radiometric calibration of a terrestrial full-waveform laser scanner," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 162–172, Jan. 2015.

[6] A. Jalobeanu and G. R. Goncalves, "The full-waveform LiDAR Riegl LMS-Q680i: From reverse engineering to sensor modeling," in *Proc. ASPRS Annu. Conf.*, Sacramento, CA, USA, Mar. 19–23, 2012, pp. 264–274.

[7] J. Castorena and C. D. Creusere, "Sampling of time-resolved full-waveform LIDAR signals at sub-Nyquist rates," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3791–3802, Jul. 2015.

[8] C. Wang, Q. Li, Y. Liu, G. Wu, P. Liu, and X. Ding, "A comparison of waveform processing algorithms for single-wavelength LiDAR bathymetry," *ISPRS J. Photogramm. Remote Sens.*, vol. 101, pp. 22–35, Mar. 2015.

[9] C. Mallet, F. Lafarge, M. Roux, U. Soergel, F. Bretar, and C. Heipke, "A marked point process for modeling LiDAR waveforms," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3204–3221, Dec. 2010.

[10] M. Słota, "Advanced processing techniques and classification of full-waveform airborne laser scanning data," *Geomatics Environ. Eng.*, vol. 8, no. 2, pp. 85–95, 2014.

[11] M. Hofton, J. Minster, and B. Blair, "Decomposition of laser altimeter waveforms," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 4, pp. 1989–1996, Jul. 2000.

[12] W. Wagner, A. Ullrichb, A. Ducica, T. Melzera, and N. Studnickab, "Gaussian decomposition and calibration of a novel small-footprint full-waveform digitising airborne laser scanner," *ISPRS J. Photogramm. Remote Sens.*, vol. 60, no. 2, pp. 100–112, 2006.

[13] J. Jung and M. M. Crawford, "A two-stage approach for decomposition of ICESat waveforms," in *Proc. IEEE IGARSS*, Jul. 2008, pp. 680–683.

[14] C. K. Wang, "Exploring exploring weak and overlapped returns of a lidar waveform with a wavelet-based echo detector," in *Proc. 12th Congr. Int. Soc. Photogramm. Remote Sens.*, Aug. 2012, XXXIX-B7, pp. 529–534.

[15] C. Wang, F. Tang, L. Li, G. Li, F. Cheng, and X. Xi, "Wavelet analysis for ICESat/GLAS waveform decomposition and its application in average tree height estimation," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 115–119, Jan. 2013.

[16] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[17] Z. Pan, C. Glennie, P. Hartzell, J. C. Fernandez-Diaz, C. Legleiter, and B. Overstreet, "Performance assessment of high resolution airborne full waveform LiDAR for shallow river bathymetry," *Remote Sens.*, vol. 7, no. 5, pp. 5133–5159, 2015.

[18] J. Verdeyen, *Gaussian Beams, Laser Electronics*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1995, pp. 63–84.

[19] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Boston, MA, USA: Cengage Learning, 2001.

[20] B. Sklar, "Rayleigh fading channels in mobile digital communication systems Part I: Characterization," *IEEE Commun. Mag.*, vol. 35, no. 7, pp. 90–100, Jul. 1997.

[21] A. Chauve, C. Mallet, F. Bretar, S. Durrieu, M. Deseilligny, and W. Puech, "Processing full-waveform LiDAR data: Modelling raw signals," in *Proc. ISPRS Workshop Laser Scanning SilviLaser*, Espoo, Finland, Sep. 12–14, 2007, pp. 102–107.

[22] S. Hernandez-Marin, A. Wallace, and W. Gibson, "Bayesian analysis of LiDAR signals with multiple returns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2170–2180, Dec. 2007.

[23] M. A. Montes-Hugo, J.-S. Bailly, N. Baghdadi, and A. Bouhdaoui, "Modeling the effects of surface and bottom geometries on LiDAR bathymetric waveforms," in *Proc. IEEE IGARSS*, Jul. 2014, pp. 2706–2708.

[24] P. M. Page-Jones, "Notes on the RSGB observations of the HF ambient noise floor," Radio Soc. Great Britain, Bedford, U.K., 2003. [Online]. Available: http://rsgb.org/main/files/2012/12/EMC_RSGB_HF_Ambient_Noise_Floor_2003.pdf

[25] Q. Li, S. Ural, J. Anderson, and J. Shan, "Minimum description length constrained LiDAR waveform decomposition," in *Proc. IEEE IGARSS*, Jul. 2014, pp. 165–168.

[26] C. Mallet, F. Lafarge, F. Bretar, U. Soergel, and C. Heipke, "Lidar waveform modeling using a marked point process," in *Proc. IEEE 16th ICIP*, Nov. 2009, pp. 1713–1716.

[27] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel K-means: Spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2004, pp. 551–556.

[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[29] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, 2001.

[30] K. Fukunaga, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 32–40, Jan. 1975.

[31] "Shenandoah fact sheet, U.S. Dept. the Interior," Nat. Park Service, Washington, DC, USA, 2013.

[32] C. E. Parrish, J. N. Rogers, and B. R. Calder, "Assessment of waveform shape features for lidar uncertainty modeling in a coastal salt marsh environment," *Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 569–573, 2014.

[33] S. Sheather and M. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *J. Roy. Statist. Soc. B*, vol. 53, pp. 683–690, 1991.

[34] N. El-Sheimy, *Topographic Laser Ranging and Scanning: Principles and Processing*, J. Shan and C. K. Toth, Eds. Boca Raton, FL, USA: CRC Press, 2009, pp. 195–214.

[35] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 417–431, 1983.

[36] P. Axelsson, "DEM generation from laser scanner data using adaptive TIN model," *Int. Arch. Photogramm. Remote Sensi.*, vol. 33, no. B4/1, pp. 110–117, 2000.

[37] W. Mücke, B. Deák, A. Schroiff, M. Hollaus, and N. Pfeifer, "Detection of fallen trees in forested areas using small footprint airborne laser scanning data," *Can. J. Remote Sens.*, vol. 39, no. S1, pp. S32–S40, 2013.

[38] R. M. Langridge, W. F. Ries, T. Farrier, N. C. Barth, N. Khajavi, and G. P. De Pascale, "Developing sub 5-m LiDAR DEMs for forested sections of the Alpine and Hope faults, South Island, New Zealand: Implications for structural interpretations," *J. Struct. Geol.*, vol. 64, pp. 53–66, 2014.

[39] Z. Lin, H. Kaneda, S. Mukoyama, N. Asada, and T. Chiba, "Detection of subtle tectonic-geomorphic features in densely forested mountains by very high-resolution airborne LiDAR survey," *Geomorphology*, vol. 182, pp. 104–115, 2013.

[40] F. Bretar, A. Chauve, C. Mallet, and B. Jutzi, "Managing full waveform LiDAR data: A challenging task for the forthcoming years," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 37, pp. 415–420, 2008.

[41] W. Wagner, A. Roncat, T. Melzer, and A. Ullrich, "Waveform analysis techniques in airborne laser scanning," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 36, no. 3/W52, pp. 413–418, 2007.

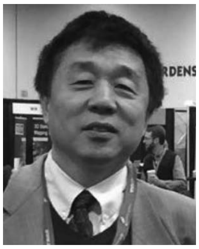
Qinghua Li received the B.S. and M.S. degrees in electromagnetic field and microwave technology from the School of Electronic Engineering, Xidian University, Shaanxi, China. He is currently working toward the Ph.D. degree with the Lyles School of Civil Engineering, Purdue University, IN, USA.

His doctoral research involves the nonparametric modeling of lidar waveforms, and the study of Geiger mode, single photon, and full-waveform lidar systems. He is also a Graduate Research Assistant with Purdue Libraries, working on text mining of geo-referenced tweets and mining of geographic information system articles. Prior to attending Purdue University, he was a Research Assistant with the Institute of Space and Earth Information Science, the Chinese University of Hong Kong, where he focused on the long-series SAR change detection and urban applications of persistent-scatterer interferometric SAR techniques.

Serkan Ural received the B.S. degree in geodesy and photogrammetry engineering from Yildiz Technical University, Istanbul, Turkey, and the M.S. degree in geodetic and geographic information technologies from the Middle East Technical University, Ankara, Turkey. He is currently working toward the Ph.D. degree from Lyles School of Civil Engineering, Purdue University, IN, USA.

In his Ph.D. research, he studies the classification and segmentation of point clouds acquired with airborne lidar systems. From 2001 to 2006, he worked as a Surveying Engineer, a Land Acquisition Supervisor, and a Manager at BOTAS Petroleum Pipeline Corporation. Since 2006, he has been a Research Assistant with the Department of Geomatics Engineering, Hacettepe University, Ankara. His research interests include optical and lidar remote sensing, pattern recognition, and spatial analysis.

John Anderson has worked for the U.S. Army Corps of Engineers for 30 years as a Research Biologist and a Spectroscopist. He has numerous journal publications and articles related to spectral sensing in the life sciences, including two book sections on biofilm spectral characterization. His area of expertise is remote sensing biological phenomena and characterization of terrain using reflectance and fluorescence. His current research interests include unmanned aerial vehicles for high-resolution remote sensing and lidar sensing and rapid construction of geospatial analytical products.



Jie Shan (SM'14) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China.

He has held faculty positions at universities in China and Sweden, and has been a Research Fellow in Germany. He is currently with the Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, USA. His research interests include sensor geometry, pattern recognition from images and light detection and ranging (lidar) data, object extraction and reconstruction, urban remote sensing, and automated mapping.

Dr. Shan serves as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He has received multiple academic awards, including the Talbert Abrams Grand Award and the Environmental Systems Research Institute Award for Best Scientific Paper in Geographic Information Systems (First Place).