Modified EM algorithm and its application to the decomposition of laser scanning waveform data

MA Hong-chao, LI Qi

School of remote sensing, Wuhan University, Hubei Wuhan 430079, *China*

Abstract: Small footprint airborne LIDAR systems now possesses the capability to sample the whole returned waveform rather than to extract discrete 3D coordinate values (discrete point cloud), thanks to the improvement of data storage hardware and data processing speed. One merit to analyze waveform data is that the end-user can extract point cloud by him/herself from the raw waveform data in the post processing, instead of being provided by the LIDAR system. The first step to analyze waveform data is to decompose the waveform into individual components. Conventional methods for waveform decomposition are usually polynomial fitting by non-linear least square algorithm, or simply thresholding with the threshold value provided by system vendor. Literature has pointed out that it is impossible to get higher accurate decomposition results by such conventional methods. The paper modifies the Expectation Maximum (EM) algorithm in the context of laser scanning waveform decomposition. Experiments with data from both airborne and space borne LIDAR systems show the high reliability and accuracy of the proposed method for waveform decomposition.

Key words: LIDAR, EM algorithm, full waveform digitizing, waveform decomposisiton, Gaussian decomposition

CLC number: TN957.52 **Document code**: A

1 INTRODUCTION

Airborne Light Detection and Ranging (LIDAR) technique has been witnessed widely applied for rapid 3D mapping in the past decade. Though the essential of LIDAR is similar to that of laser ranging technology, the integrated LIDAR system with Positioning and Orientation System (POS) and CCD camera (with 20—40 million pixels) makes it irreplaceable for mapping in areas such as heavily vegetation covered area, coastal zones, beaches and islands, *etc.*, where it is very difficult for mapping by conventional photogrammetric means. Airborne LIDAR as a new type of remote sensing sensor is becoming familiar to surveying and mapping community (Ackmann, 1999; Baltsavias, 1999; Gamba & Housh mand, 2000).

Airborne LIDAR system could date back to 1980's when some experimental systems emerged. It was matured in the mid to late 1990's. The earlier generation of LIDAR systems records single echo, i.e., the first echo is recorded and only the Digital Surface Model (DSM) over the surveyed area can be obtained. In the later commercially available systems, both first and last echoes could be recorded, which can be input to specific algorithms to remove non-ground points so as to establish the Digital Terrain Model (DTM) over the surveyed area. Though such a simple working flow seems to be out of any problem, it is impossible for the user to get any equipment-related information, such as: how to geo-locate the echo pulse? How does the ground objects influence the shape and amplitude of the return signal? How are the return signals quantified? The detection techniques and quantification methods for echo pulses are usually kept as commercial secrets by system vendors. Literature pointed out that the final achievements should be error prone if different detection techniques and quantification methods are employed (Wehr & Lohr, 1999).

One solution for the above mentioned problem is to sample and record the transmitted and returned signals at an infinitesimal interval, rather than to record several discrete returns only. Such a sampling and recording manner is defined as full waveform digitizing, and the system possessing such a capability is referred to as full waveform digitizing LIDAR. Users know why and how the discrete echoes are generated by analyzing the waveform data, and application-oriented methods can be employed to process these data.

As a matter of fact, airborne LIDAR systems were developed by NASA and characterized by full waveform digitizing capability as early as 1990's, such as SLICER (Scanning Lidar Imager of Canopies by Echo Recover) and LVIS (Laser Vegetation Imaging Sensor). Some space-

Received date: 2007-06-01; **Accepted date**: 2007-09-13

Foundation: National 973 program (contract No. 2009CB724007) and the 11th 863 program (contract No. 2006AA12Z101)

First-author Biography: MA Hong-chao (1969—), male, professor. He achieved his PhD degree in geophysical prospecting and information technology, and did post-doctor research in photogrammetry and remote sensing at Wuhan University. He focuses on algorithm research and software development of data processing. He has published 26 papers and a book. E-mail: hchma@whu.edu.cn, hongchao_ma@263.net.cn.

borne LIDAR systems such as GLAS also possess full waveform digitizing capability. However, none of them is for commercial purpose. The technology was firstly adopted by RIEGL in 2004 for its commercial system. Though only a few years later, most of mature LIDAR systems available in the current market have integrated full waveform digitizer as a standard configuration component, such as Falcon III by Toposys, ALS50-II by Leica Geosystem and ALTM 3100EA by Optech.

Study on waveform data analysis and processing is relatively behind the development of full waveform digitizer hardware. This due partially to the fact that the application and analysis of waveform data are still limited to research community, seldom is used in engineering projects. Furthermore, methodology for waveform data analysis is usually application-oriented, leading to the lack of generalized methods and algorithms to process these data.

The paper discusses one of the key steps of waveform data analysis: waveform decomposition. A modified Expectation Maximum (EM) algorithm is applied, using the publicly accessed waveform data provided by NASA SLICER and RIEGL as the experimental data sets.

2 THE GENERATION AND DECOMPOSITION OF WAVEFORM DATA

Though there is a technical term "full waveform", we should pay much attention to that the waveform mentioned in airborne LIDAR technology is not actually a mathematically smooth curve. It is in fact the infinitesimal time interval sampling of the returned echo makes the recorded data seem to be a continuous curve if they are plotted in a planar coordinate system spanned by time and amplitude as xy-axes. We firstly investigate the generation of waveform which is closely related to the latter decomposition. This should start with introducing the radar equation. The relationship between the transmitted and the received power can be formulated as equation (1), according to the derivation suggested by Wagner *et al*. (2006)

$$P_r(t) = \sum_{i=1}^{N} \frac{D_r^2}{4\pi R_i^4 \beta_t^2} P_t(t) * \sigma_i'(t) * \Gamma(t) \quad (1)$$

where N denotes the object number encountered by the laser pulse in the trip it traverses forth and back, under the condition that the range the pulse visited is larger than the minimum distance the system required, the so called range resolution. Only if the distance separated by two objects along the laser pulse's forward and back path is larger than the range resolution, could the two objects be distinguished. $P_t(t)$ denotes the transmitting power, $P_r(t)$ the receiving power, D_r the aperture diameter of the receiver optics, R_i the range between laser transmitter and object i, β_t is the transmitter beamwith, $\sigma_i'(t)$ the differential backscatter cross-section which is defined as the backscatter cross-section within a infinitesimal range interval dR, and t the time variable. Since range R and time t is related by $t = 2R_i/v_g$, where v_g is the velocity of light, they are used interchangeably in the LIDAR community. $\Gamma(t)$ is the receiver impulse function, and $*$ denotes the convolution operator. Waveform data are recorded by sampling the return signal with predefined interval. Since sampling frequency satisfies the Nyquest theorem, the sampled data can be restored. Fig. 1 illustrates the basic idea of generating waveform and discrete point data.

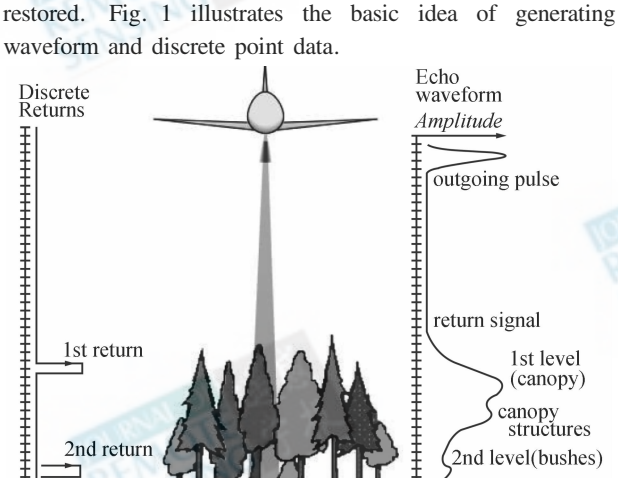

Fig. 1 Full waveform data and discrete echo. (Optech, 2006)

In practice, $P_t(t)$ and $\Gamma(t)$ cannot be easily determined independently. Therefore it is advantageous to rewrite the convolution term by making use of the commutative property of the convolution operator: $P_t(t) * \Gamma(t) * \sigma_i'(t)$, where we introduce the system waveform $S(t)$ of the laser scanner, defined as the convolution of the transmitted pulse and the receiver response function. It can be measured experimentally and is shown in Fig. 2 for the RIEGL LMS-Q560. It can be seen that it is well described by a Gaussian function: $S(t) = \hat{S}\, e^{-\frac{t^2}{2s_s^2}}$ where $\hat{S}$ is the amplitude and s_s the standard deviation.

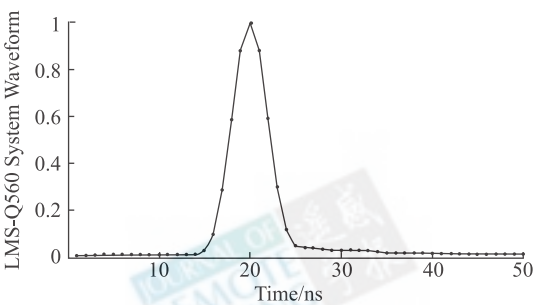

Fig. 2 System waveform of RIEGL LMS-5600

In radar remote sensing theory, it is assumed that the scattering properties of a cluster of scatterers can be described by a Gaussian function $\sigma_i'(t) = \hat{\sigma}_i e^{-\frac{(t-t_i)^2}{2s_i^2}}$

where $\hat{\sigma}_i$ is the amplitude and s_i the standard deviation of the cluster i. Since the convolution of two Gaussian curves gives again a Gaussian function, so that we obtain: $P_r(t) = \sum_{i=1}^{N} \hat{P}_i e^{-\frac{(t-t_i)^2}{2s_{p,i}^2}}$, where $s_{p,i} = \sqrt{s_s^2 + s_i^2}$, $\hat{P}_i = \frac{D_r^2}{4\pi R_i^4 \beta_t^2} \hat{S}\sigma_i \frac{s_s}{s_{p,i}}$, which shows that the return waveform is actually superimposed by several Gaussian functions. If we try to decompose the return waveform into individual Gaussian function and estimate its parameters such as amplitude, mean value and standard deviation, then it does not only provide preprocessed data for later processing, but also can estimate the backscatter cross-section of the pulsed area.

3 ALGORITHM FOR WAVEFORM DECOMPOSITION

Since the return wave can be described as the superimposition of several Gaussian functions, it is of great value to estimate parameters of them. The amplitude, the position of wave peak, the width of the wave and the distance between two continuous wave peaks, among many others, are parameters of most importance. It is mainly dependent on the determination of individual waveform to estimate these parameters; therefore it is one of significant steps for waveform decomposition in analyzing waveform data.

Wagner *et al.* (2006) and Hoton *et al.* (2000) have already developed Gaussian function based waveform decomposition algorithms, where the Gaussian function was only regarded as an aim function to which the waveform data were fitted by using non-linear least square method. Some constraint conditions should be given upon their algorithms and the solution tended to be local optimal. From the other point of view, however, the problem of waveform decomposition is actually a problem of decomposition for mixture Gaussian distribution since the waveform can be described as the superimposition of several Gaussian functions. Decomposing mixture Gaussian distribution is a problem often occurred in fields such as pattern recognition and statistical inference the Expectation Maximum (EM) algorithm developed by Dempster, etc. in 1977 can perform parameter estimation for Gaussian mixture distribution (Dempster *et al.*, 1977). An overview of the algorithm is given in the following while detailed description deserves to be referenced by Olive *et al.* (1996).

The original formula for estimating p_j, μ_j and σ_j by EM is as listed below:

$$Q_{ij} = \frac{p_j f_j(x_i)}{\sum_{j=1}^{k} p_j f_j(x_i)} \tag{2}$$

$$p_j = \frac{\sum_{i=1}^{n} Q_{ij}}{n} \tag{3}$$

$$\mu_j = \frac{\sum_{i=1}^{n} Q_{ij} i}{p_j \times n} \tag{4}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} Q_{ij} (i-\mu_j)^2}{p_j \times n}} \tag{5}$$

Suppose that the waveform is superimposed by individual Gaussian distributions, then the result mixed distribution, that is, the small-time-interval sampled data, can be formulated as:

$$f(x) = \sum_{j=1}^{k} p_j \times f_j(x)$$

$$f_j(x) \in N(\mu_j, \sigma_j^2)$$

k denotes the number of Gaussian distributions taking part in the superimposition; $f_j(x)$ is the Gaussian probability density function; p_j is the weight of $f_j(x)$ describing the percentage of the j^{th} component occurred in the mixture distribution, and satisfies: $0 < p_j < 1$, $\sum_{j=1}^{k} p_j = 1$; μ_j and σ_j are the mean value and standard deviation of Gaussian distribution respectively. For each component j, the estimated μ_j gives the position of the return wave in the abscissa, while σ_j represents the width of the wave. All the parameters p_j, μ_j, and, σ_j can be estimated by EM algorithm. When applied in practice, however, data preprocessing should be carried out and initial parameter values for formula (2)—(5) should be given. The initial values are usually predicted from raw data, and then the constructed $f(x)$ is used to calculate Q_{ij}. Though such a workflow is feasible, it will not get a satisfactory outcome if formula (2)—(5) are adopted directly without considering the amplitude of the waveform. A modified EM algorithm is derived if amplitude N_i is taken into consideration. Details of derivation are given in the following:

Substituting (3) into formula (4)—(5), it reads:

$$\mu_j = \frac{\sum_{i=1}^{n} Q_{ij} i}{n \times p_j} = \frac{n \times \sum_{i=1}^{n} Q_{ij} i}{\sum_{i=1}^{n} Q_{ij} \times n} = \frac{\sum_{i=1}^{n} Q_{ij} i}{\sum_{i=1}^{n} Q_{ij}} \tag{6}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} Q_{ij} (i-\mu_j)^2}{n \times p_j}} = \sqrt{\frac{n \times \sum_{i=1}^{n} Q_{ij} (i-\mu_j)^2}{\sum_{i=1}^{n} Q_{ij} \times n}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} Q_{ij}(i-\mu_j)^2}{\sum_{i=1}^{n} Q_{ij}}} \quad (7)$$

where the amplitude N_i is added to nominator and denominator simultaneously. N_i is equivalent to a weight constraining p_j, μ_j and σ_j.

$$\mu_j = \frac{\sum_{i=1}^{n} N_i Q_{ij} i}{\sum_{i=1}^{n} N_i Q_{ij}} \quad (8)$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} N_i Q_{ij}(i-\mu_j)^2}{\sum_{i=1}^{n} N_i Q_{ij}}} \quad (9)$$

If formula (8) and (9) are restored to the form coinciding with EM algorithm, then:

$$Q_{ij} = \frac{p_j f_j(i)}{\sum_{j=1}^{k} p_j f_j(i)} \quad (10)$$

$$p_j = \frac{\sum_{i=1}^{n} N_i Q_{ij}}{n \times \sum_{i=1}^{n} N_i} \quad (11)$$

$$\mu_j = \frac{\sum_{i=1}^{n} N_i Q_{ij} i}{n \times p_j \times \sum_{i=1}^{n} N_i} \quad (12)$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} N_i Q_{ij}(i-\mu_j)^2}{n \times p_j \times \sum_{i=1}^{n} N_i}} \quad (13)$$

where n denotes the number of points for sampling raw waveform data, and N_i the sampling interval of the i^{th} sample. Since the initial values of p_j, μ_j and σ_j are usually suboptimal, iteration is carried out using formula (10)—(13) to obtain the optimal values by adjusting these parameters with amplitude.

The difference between two consecutive mean values can be calculated after the optimal μ_j is estimated by EM algorithm. An additional μ_{j+1} can be inserted between the two consecutive mean values if the difference exceeds a given threshold in order to decompose as many individual Gaussian functions as possible, and then re-iterate (10)—(13). Otherwise, if there is no more individual Gaussian functions can be decomposed, the iteration is ended and the final result is obtained. Fig. 3 illustrates the flow chart of decomposition EM algorithm. Point cloud can be generated directly via the process if raw waveform data are imported. Theoretically, accuracy of point clouds obtained by the method of waveform decomposition is higher than that generated by LIDAR system itself.

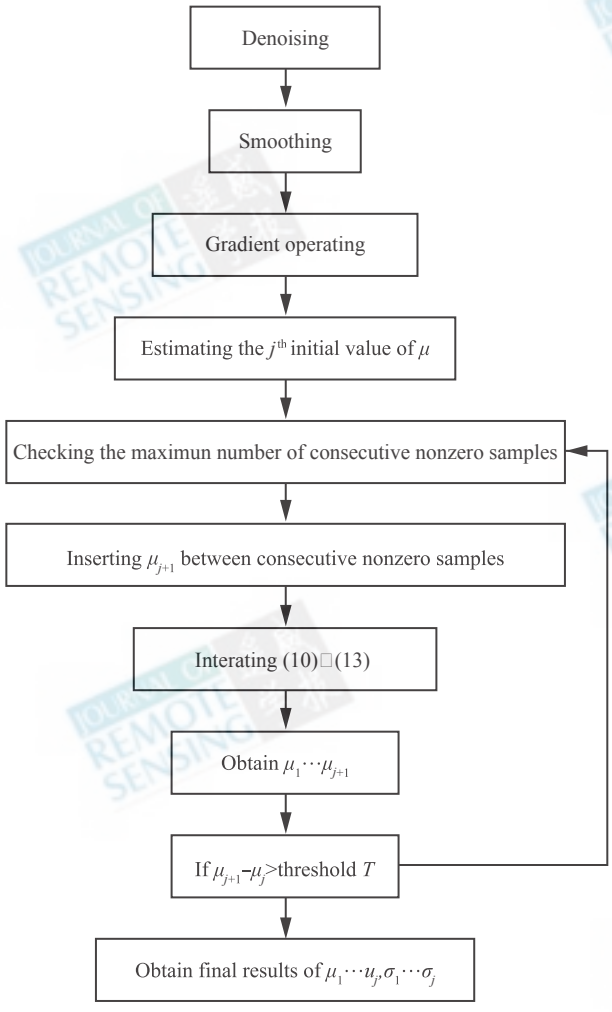

Fig. 3 Flow chart of waveform decomposition by EM algorithm

4 EXPERIMENTS AND DISCUSSION

4.1 SLICER

The waveform data used in our first experiment was acquired by SLICER over a vegetation area and stored in a binary format file with suffix. dat. Each returned signal is sampled with an interval of 0.1112 me and there are totally 600 sampling points. The inclination angle and azimuth of the transmitted pulse, Lat/Lon and elevation of the highest surface detected are also contained in the data.

4.1.1 Pre-processing of the raw data

Fig. 4 shows the raw waveform data, where the abscissa represents the time interval (bearing in mind the interchangeability of time and range) for sampling while the ordinate the amplitude of the return signal. Noises from many sources contaminate the sampled data, leading the sampled curve to a jittered one where the jittered parts with small amplitudes illustrating noises. These noises should be

removed before the EM algorithm is applied. Cutting 5% of the waveform tail off since the tail jitters around a small value, and then calculates the mean value from it as threshold σ_{noise}. All samples with amplitude less than threshold σ_{noise} are set to zero. Fig. 5 illustrates the final result of waveform after smoothing.

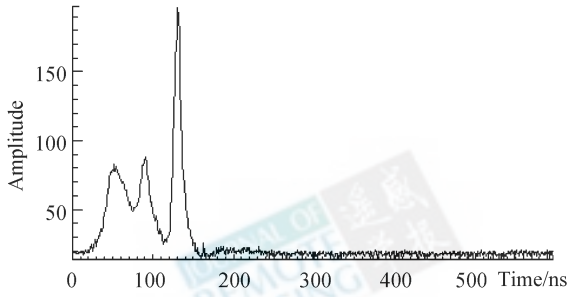

Fig. 4 Raw waveform data acquired by SLICER

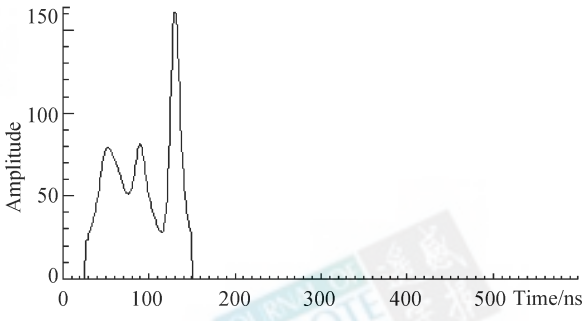

Fig. 5 Pre-processed waveform of Fig. 4

4.1.2 *Parameter initialization*

The initial values of μ_j, p_j and σ_j can be given randomly in principle when the EM algorithm is carried out, though a better prediction for these values will greatly improve the calculating speed. The initial value of μ_j can be determined by local maximum since it is obvious that the local maximum would be a candidate for optimal μ_j, so the first derivative of the waveform is calculated, as shown in Fig. 6. The initial values of p_j are set so that each component has an equal weight and σ_j is set as 7 in this experiment. Fig. 7 illustrates the pre-processed waveform (solid curve) and its initial curve (dashed curve).

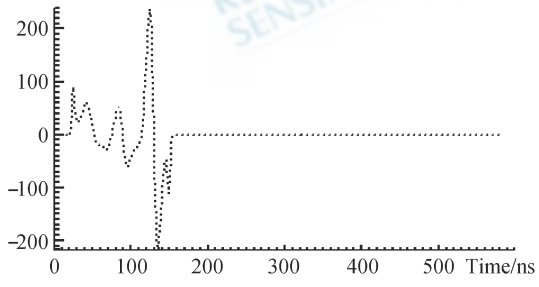

Fig. 6 First derivative of waveform

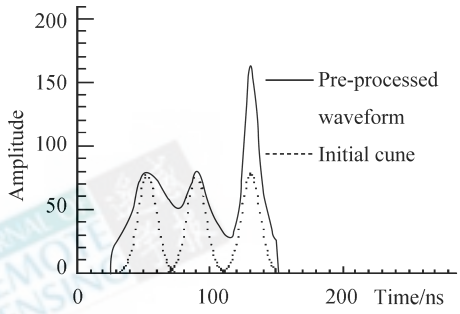

Fig. 7 Waveform with initial values estimated

Fig. 8 shows the Waveform (solid) with three Gausssian components (dashed) decomposed through iteration (10)—(13). The estimated parameters of these distributions are: $\mu_1 = 54$、$\sigma_1 = 9.0$、$\mu_2 = 89$、$\sigma_2 = 8.5$、$\mu_3 = 131$、$\sigma_3 = 7.2$.

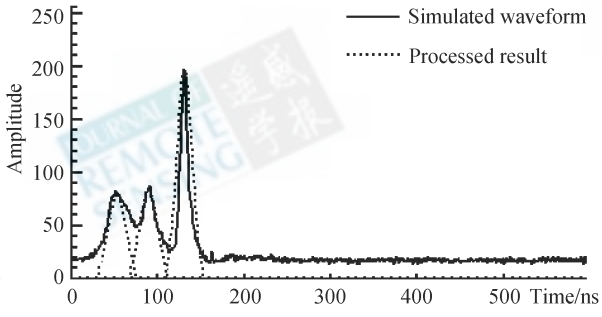

Fig. 8 A simulated waveform and processed result by the proposed algorithm

An automated last return detection software provided by SLICER is applied to the return waveform to identify the start, peak and end of a return, which is shown in Fig. 9. The vertical lines Grstart、Grpeak and GrEnd (the first, second and third line from the left to right) in the figure illustrate the positions of the ground extracted by the SLICER. The peak between Grpeak and GrEnd is inferred to be from the ground. In Fig. 9 the laser pulse hits the canopy first and creates one echo pulses; a fraction of the laser pulse also hits the ground giving rise to a second echo pulses. The two echo pulses overlap as a result of the distance between vegetation and ground is short.

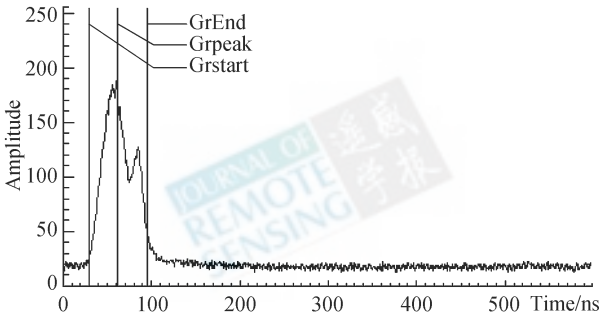

Fig. 9 Ground location of SLICER

Using EM algorithm for decomposing the same data, two fitted Gaussian distributions can be obtained as shown in Fig. 10. The two vertical lines show the positions of two return peaks. Estimated parameters are: $\mu_1 = 57$、$\sigma_1 = 7.3$、$\mu_2 = 83$, $\sigma_2 = 7.3$. It is obvious by comparing Fig. 9 and Fig. 10 that EM decomposition algorithm outperforms the method provided by SLICER.

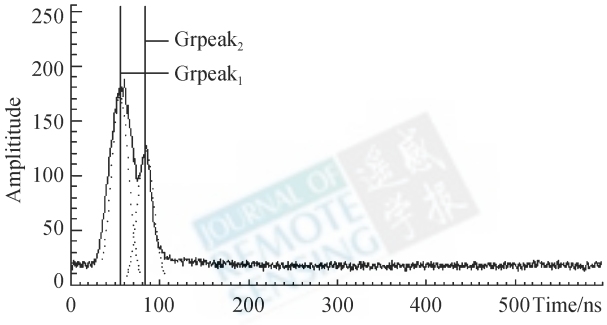

Fig. 10 Simulated waveform

Dotted curve shows two Gaussian distribution decomposed from the same data set shown in Fig. 9

Similarly, though the algorithm provided by SILCER can detect the start, peak and end of the waveform well, it fails in detecting the first and second return, as shown in Fig. 11. However, three Gausssian components (dashed) are decomposed by EM algorithm as shown in Fig. 12 EM algorithm performs more accurately.

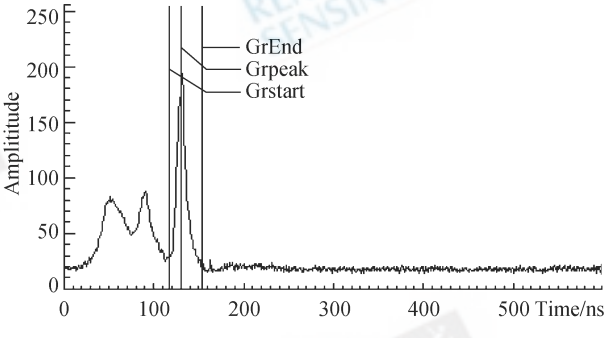

Fig. 11 Ground location of SLICER for another data set

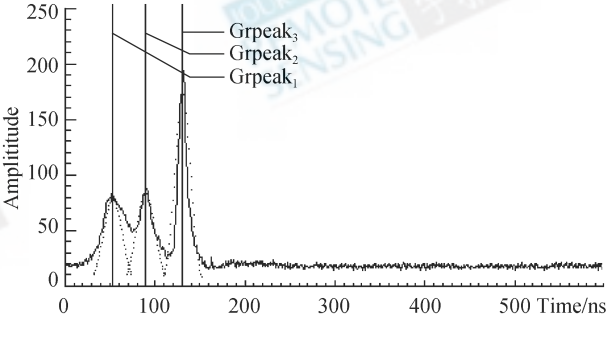

Fig. 12 Simulated waveform by EM algorithm

* Dotted curve shows two Gaussian distributions decomposed from the same data set shown in Fig. 11

4.2 RIEGL

RIEGL waveform data are stored in two separate files: *.LWF file contains the calibrated waveform sample data, *. LGC file contains the geocoding and indexing information for each laser shot. Each waveform consists of a byte array of STRTWFLEN (start waveform length) samples representing the emitted pulse waveform of a laser shot for reference, followed by a byte array or ushort array of WFLEN (waveform length) samples representing the surface return waveform. The distance from one sample to the next is 0.149855 m, shown as Fig. 13.

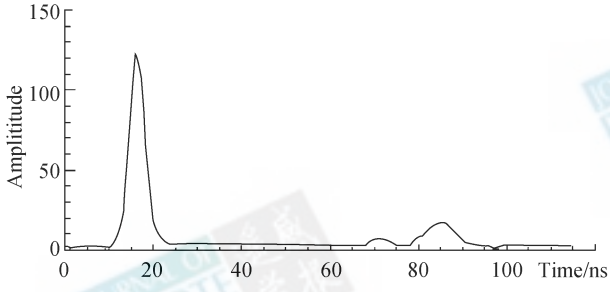

Fig. 13 Raw waveform data acquired by RIEGL LM5600

The data preprocessing of RIEGL is different from that of SLICER. The threshold for denoising in RIEGL is determined by calculating the mean value of the tail of the emitted waveform since the emitted waveform is recorded and provided by the system, again 5% of the last part of the tail is cut off. There are less samples in comparison with SLICER. The initial value of σ_j is set to 1. Fig. 14 shows the preprocessing of the backscatrrering waveform, in which the horizontal line represents the threshold value σ_{noise} for smoothing.

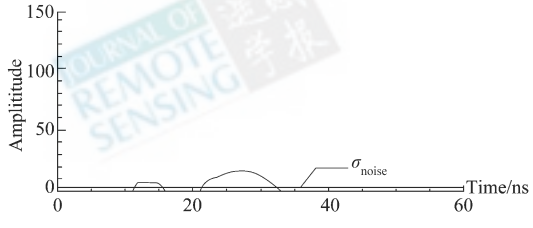

Fig. 14 Pre-processed waveform

Fig. 15 shows the two Gaussian components decomposed from the data shown in Fig. 14. Two vertical lines in the figure illustrate the location of two return echoes. The parameters estimated are $\mu_1 = 13.5$, $\sigma_1 = 1.5$, $\mu_2 = 27.0$ and $\sigma_2 = 2.5$.

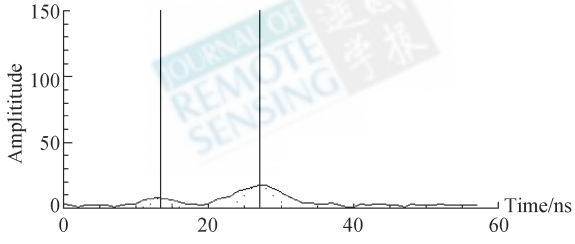

Fig. 15 Simulated waveform by EM algorithm

5 CONCLUSION

At present, full waveform digitizing technology is not only adopted by space-borne LIDAR systems, but also becoming a standard configuration component of airborne LIDAR systems. Analyzing and processing waveform data will provide additional surface information than discrete echoes, though equipment vendors and researchers believe that the lack of mature algorithm and analyzing workflow is a bottleneck. Waveform decomposition is one of the key steps of waveform data analysis. Works in this paper show that laser scanning waveform decomposition based on modified EM algorithm provides better waveform fitting precision compared with conventional methods.

REFERENCES

Ackmann, F. 1999. Airborne laser scanning-present status and future expectations. *ISPRS Journal of Photogrammetry & Remote Sensing*, **54**:64—67

Baltsavias E P. 1999. A comparison between photogrammetry and laser scanning. *ISPRS Journal of Photogrammetry & Remote Sensing*, **54**:83—94

Bilmes J A. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Department of Electrical Engineering and Computer Science. U. C. Berkeley TR-97-021

Dempster A P, Laird N M, Rubin DB, 1977. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, **39**(1), 1—38

Gamba P, Houshmand B. 2000. Digital surface models and building extraction: a comparison of IFSAR and LIDAR data. *IEEE Transaction on Geoscience and Remote Sensing*, **38**:(4):1959—1968

Hofton M A, Blair J B, Minster J. 2000. Decomposition of laser altimeter Waveforms. *IEEE Transactions on Geoscience and Remote Sensing*, **38**:(4)1989—1996

Oliver J J, Baxter R A, Wallace C S. 1996. Unsupervised learning using MML. Machine Learning. Proceedings of the Thirteenth International Conference (ICML 96). Morgan Kaufmann Publishers, San Francisco CA USA

Optech: ALTM Waveform Digitizer Operation and Processing Manual ALTM 3100. 2006

Wagner W, Ullrich A, Ducic, V, *et al.* 2006. Gaussian decomposition and calibrati on of a novel small-footprint full-waveform digitizing airborne laser scanner. *ISPRS Journal of Photogrammetry and Remote Sensing*, **60**(2): 100—112

Wagner W, Ullrich A, Melzer T, *et al.* 2004. From single-pulse to full waveform airborne laser scanners: potential and practical challenges. ISPRS XXth Congress. Istanbul **35**: Part B/3. 201—206

Wehr A, Lohr U. 1999. Airborne laser scanning: an introduction and overview. *ISPRS Journal of Photogrammetry & Remote Sensing* **54**: 68—82

改进的 EM 模型及其在激光雷达全波形数据分解中的应用

马洪超,李　奇

武汉大学 遥感学院,湖北 武汉　430079

摘　要： 随着数据存储能力和处理速度的提高,小光斑机载激光雷达系统已经可以通过数字化采样来存储整个反射波形,而不仅仅是由系统提取出来的三维坐标(即离散点云)。分析波形数据最重要的优点之一是可以在后处理过程中让使用者自己来提取三维坐标。一般的分解方法基于非线性最小二乘的多项式拟合,或者有设备厂商提供的简单阈值法,无法获得高精度的分解结果。本文使用改进的 EM 脉冲检测算法得到回波脉冲的位置和宽度,证明是一种性能可靠、精度较高的波形分解算法。

关键词： LIDAR,EM 算法,波形数字化,波形分析,高斯分解