LIDAR DETECTION ALGORITHM BASED ON HYPERSPECTRAL ANOMALY DETECTION

Avishai Ben-David [(1)], Richard G. Vanderbeek [(1)], Charles E. Davidson [(2)]

[(1)]RDECOM, Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland, 21010, USA

E-mail: avishai.bendavid@us.army.mil and richard.vanderbeek@us.army.mil

[(2)] Science and Technology Corporation, 500 Edgewood Rd., Suite 205, Edgewood, Maryland 21040, USA

E-mail: charlie.davidson@us.army.mil

ABSTRACT

A new detection algorithm for lidar applications has been developed. The algorithm employs an option for a preprocessing stage where undesired oscillations and artifacts is filtered out with a low-rank orthogonal projection technique. The filtering technique is useful for visualization of the lidar signals when the signal-to noise ratio is low and the presence of the target is difficult to see. The detection is based on hyperspectral anomaly detection where m discrete ranges between R_1 and R_2 is regarded as an m-dimensional hyperspectral vector. A detection score is computed from each lidar measurement for the hyperspectral range vector. A detection score significantly different in magnitude from the detection scores for the background measurements suggests that an anomaly (interpreted as the presence of a target-signal) exists within R_1 and R_2. A Gaussian-mixture probability model for two hypotheses (anomaly is present or absent) is computed with an expectation–maximization algorithm to produce a detection threshold, probability of detection, and probability of false alarm. Results of the algorithm for CO_2 lidar bioaerosol *Bacillus subtilis* subsp. *niger* (BG) measurements are shown and discussed.

1. INTRODUCTION

Most lidar analysis can be viewed as parametric analysis where the lidar range dependent backscattering and extinction are deduced from the analytical lidar equation and the range dependent concentration is deduced from the backscattering (or extinction). Many methods have been developed for solving the lidar equation. In most field experiments vast amount of data is catalogued and often the question "is a target cloud of interest present between ranges R_1 and R_2" is asked. For each of the n lidar measurements we produce a score $s(n)$; a scalar quantity that results from transformation from m-D space onto a 1-D space via the anomaly detector [1] of the m dimensional lidar signal (m ranges between R_1 and R_2).

To assess the probabilities of detection and false alarm we assume the detection scores follow a Gaussian mixture model (normally distributed statistics). A detection threshold computed with the probability model is used to divide the scores into two distinct classes: (1) cloud is present or (2) cloud is absent.

Detection models are based on constructing a test statistic ($s(n)$ in this study) derived from the probability density function (*pdf*) of the data, which is contaminated with noise. The noise may consist of random noise and interferences (structured noise) from "look alike" targets (e.g., ambient atmospheric backscattering). The objective of the test statistic is to decide between two hypotheses, H_0 and H_1, regarding the data. For the null hypothesis H_0 the interpretation is that the data does not contain the target-signal of interest and consists of only noise and interference. The alternative hypothesis H_1 is that the data does contain the target-signal in addition to noise and interference. A *pdf* model for the two hypotheses enables us to compute probabilities of detection and false alarm and to set a threshold for detection under the constraint of maintaining a constant false alarm rate.

In this paper we implement concepts from hyperspectral data analysis where the range vector $\boldsymbol{r}$ serves the same role as a spectral vector in hyperspectral analysis; (1) Data filtering for unwanted signatures and interferences, (2) Anomaly detection, and (3) Expectation-maximization algorithm for producing probabilities of detection and false alarm to enable setting a threshold for detection.

2. DATA FILTERING

It is not uncommon for lidar measurements to contain unwanted signatures; some due to measurements from

atmospheric interferences and some due to hardware noise. In our CO_2 lidar measurements, large oscillations were introduced due to amplifier "ringing". Traditional filtering techniques and simple subtraction of a mean measurements of the oscillations (i.e., with the lidar telescope blocked) were not successful to eliminate these stochastic signatures that obscured the sought after signals.

Let the lidar data be an $m \times n$ column matrix $\mathbf{X}$ with n columns for the n received lidar measurements (e.g., lidar backscattering pulses), where each column is the range-dependent lidar signal $X(r)$ at m ranges between the R_1 and R_2. We construct a filter $\mathbf{F}$ ($m \times m$ matrix) from a small set of signals measured within the H_0 condition ($\mathbf{X}_{H0}$) to filter the data $\mathbf{X}$ The filtered lidar signals (superscript f) are given by

$$\mathbf{X}^f = \mathbf{F}\,\mathbf{X} \tag{1}$$

The filter $\mathbf{F}$ projects the data onto a subspace orthogonal to $\mathbf{X}_{H0}$ and thus annihilates the unwanted signatures $\mathbf{X}_{H0}$, $\mathbf{F}\,X_{H_0} = \mathbf{0}$. The filter $\mathbf{F}$ is constructed via an orthogonal subspace projection operator [2, 3] from the $m \times q$ column matrix $\mathbf{X}_{H0}$ of the unwanted range-dependent signatures:

$$\mathbf{F} = \mathbf{I} - \mathbf{X}_{H_0}(\mathbf{X}_{H_0}^T\mathbf{X}_{H_0})^{-1}\mathbf{X}_{H_0}^T \tag{2}$$

where superscript T denotes a matrix-transpose operation, $\mathbf{I}$ is the $m \times m$ identity matrix, and $(\mathbf{X}_{H_0}^T\mathbf{X}_{H_0})^{-1}$ is the inverse matrix of $\mathbf{X}_{H_0}^T\mathbf{X}_{H_0}$ approximated with a low-rank inverse operation (performed with only a few eigenvectors from a singular-value decomposition; see section 3.A in [4]).

In Fig. 1a, 1250 raw lidar data signals X(r,n) are shown at wavelength 10.247 μm. This corresponds to the 10R20 band of the CO_2 laser that is commonly used to monitor water vapor in the atmosphere [5]. A description of the CO_2 lidar system is given in [6]. The oscillations in the raw data are clearly seen together with the target-signal from a *Bacillus subtilis* subsp. *niger* (BG) bio-aerosol cloud at range ~ 2350 m. In Fig. 1b, the filtered data X^F is shown when a rank 3 approximation of $\left(\mathbf{X}_{H_0}^T\mathbf{X}_{H_0}\right)^{-1}$ was used to create $\mathbf{F}$. X_{H0} contained the first 150 lidar measurements taken prior to the official start of the experiment. At the 2 Hz lidar pulse repetition frequency, 150 measurements take ~75 s. In Fig. 2, lidar signals are shown for wavelength 9.695 μm (9P36) that is used [6] for water vapor DIAL measurement. At this wavelength the lidar transmit power is lower, the signal-to-noise ratio (SNR) is low, and thus the target signal is obscured by the noise. The figure shows the effectiveness of the filtering technique to enhance the visualization (lower panel) of the presence of the BG cloud at 2350 m.

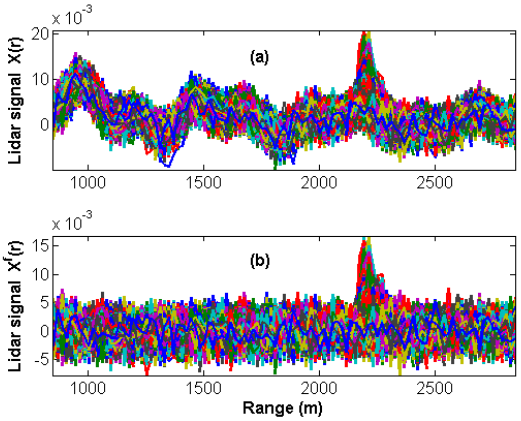

Fig. 1. Lidar signals at 2 Hz at wavelength 10.247 μm (10R18) as a function of range. (a) Raw lidar data where oscillations are shown and (b) Filtered lidar data where the BG cloud at ~2200 m is clearly seen and the oscillations are absent.

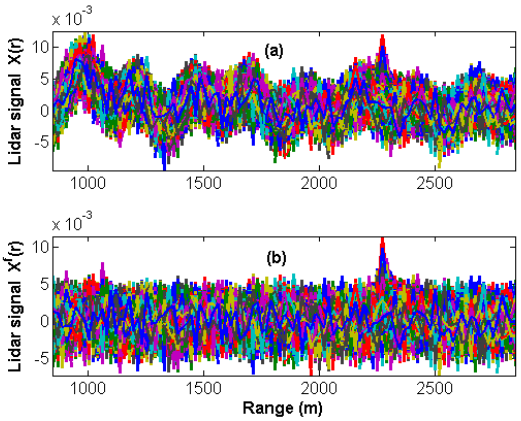

Fig. 2. Same as Fig. 1 but for a weak CO_2 wavelength 9.695 μm (9P36) for which the SNR is low and the signal from The BG cloud is less visible.

3. ANOMALY DETECTION AND THRSHHOLD FOR DETECTION

We use the classical anomaly detection algorithm [1] where a detector score $s(n)$ is computed from $X(r,n)$. A high detection score indicates that a pattern significantly different from the background is present. The anomaly detector is a version of the matched filter algorithm [3] where the shape of the target-signal is unknown. Both algorithms are based on whitening the range content (decorrelate range information) of the measurements under the H_0 hypothesis. Any signal $X(r)$ that is outside the whitened sphere, where "outside" is defined by a threshold distance γ, is declared to be a target-signal under the H_1 hypothesis

(i.e., detection). A good estimate of the range covariance $C(r_i,r_j)$ under the H_0 hypothesis is the basis of the algorithm. When $C(r_i,r_j)$ is constructed from a set of n_0 measurements, we implicitly assume that this covariance structure is valid for all the measurements $X(r,n)$.

The anomaly detection score is given by

$$s(n) = [X(n) - E(X_0)]^T C^{-1}[X - E(X_0)] \quad (3)$$

where $E(X_o)$ is an $m \times 1$ vector consisting of the mean of all n_0 measurements $X(r,n_0)$. The inverse covariance matrix C^{-1} is computed with the full rank of the matrix C in order to capture all the information available in $x|H_0$ (where the notation $x|H_0$ reads: measurement x conditional to the H_0 scenario). The operation $C^{-1/2}(\cdot)$ is a whitening process within a co-ordinate system centered at the origin (because in the covariance $C(r_i,r_j)$ the mean $E(X_o)$ is removed). The score $s(n)$ is a dot product of whitened mean-subtracted signals $C^{-1/2}[X(n) - E(X_0)]$.

Given the detection scores $s(n)$ we compute a threshold for detection γ with the Expectation-Maximization (EM) algorithm [(see section 2 in [7]) and estimate probabilities for detection and false alarm. We assume a mixture model for the *pdf* of the scores given by

$$pdf(s) = w_0\, pdf(s \mid H_0) + (1 - w_0)\, pdf(s \mid H_1) \quad (4)$$

where $pdf(s \mid H_0)$ and $pdf(s \mid H_1)$ are the conditional probabilities for scores under the H_0 and H_1 assumptions, and ω_0 is the prior probability $P(H_0)$. We assume normal (Gaussian) *pdf*

$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ where μ and σ are the mean and standard deviation of a normal *pdf*. Thus, the *pdf* for the two hypotheses in the Gaussian mixture model are $pdf(s \mid H_0) = N(s;\mu_0,\sigma_0^2)$ and $pdf(s \mid H_1) = N(s;\mu_1,\sigma_1^2)$. The parameters $(w_0,\mu_0,\sigma_0,\mu_1,\sigma_1)$ are computed with the EM algorithm.

The threshold γ is computed from the solution of

$$w_0 N(\gamma;\mu_0,\sigma_0^2) = (1 - w_0) N(\gamma;\mu_1,\sigma_1^2) \quad (5)$$

This value of γ minimizes the sum of the probability of false alarm (i.e., deciding H_1 when H_0 is true) and the probability of a missed detection (i.e., deciding H_0 when H_1 is true). The probabilities of detection P_D and false alarm, P_{FA}, for any arbitrary threshold value γ are given by

$$P_D(\gamma) = \int_\gamma^\infty pdf(s \mid H_1)ds = 2^{-1}[1 - erf(\frac{\gamma - \mu_1}{2^{1/2}\sigma_1})]$$
$$P_{FA}(\gamma) = \int_\gamma^\infty pdf(s \mid H_0)ds = 2^{-1}[1 - erf(\frac{\gamma - \mu_0}{2^{1/2}\sigma_0})] \quad (6)$$

where $erf(\cdot)$ is the error function.

4. RESULTS

The covariance matrix was constructed from the first 150 measurements (~ 75 seconds) prior to the official start of the experiment. A continuous acquisition of data was performed at 19 CO_2 lidar wavelengths for 1100 measurements that spans ~ 13 minutes. In Fig. 3 1100 scores at the 10R18 CO_2 lidar wavelength are shown, where a 2.5 s moving average process (i.e., 5 measurements) was used to enhance the signal-to-noise ratio of the derived scores. In the experiment the concentration of the cloud varied between ~ $5 \cdot 10^3$ to 5 10^4 particles per liter.

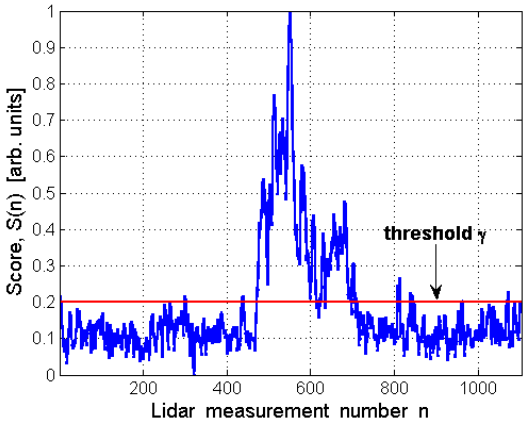

Fig. 3. Anomaly detection scores for a BG cloud (varied concentration) at lidar wavelength 10R18. A threshold line for detection probability P_d=0.81 and P_{fa}=0.006 false alarm is noted.

The five parameters for the Gaussian mixture model were found with the EM algorithm to be: ω=0.7471, μ_0=0.1135, σ_0=0.0349, μ_1=0.3709 and σ_1=0.1897. The pdf of the scores, the Gaussian mixture model (Eq. 4), the two Gaussians for the conditional scores $pdf(s|H_1)$ and $pdf(s|H_2)$, and the threshold γ (Eq. 5), are shown in Fig. 4.

A receiver operating characteristic (ROC) that describes the effect of an arbitrary choice of a threshold γ (each point on the curve is characterized by a different value of γ) on the probabilities of detection and false alarm (Eq. 6) is shown in Fig. 5. The ROC curve is computed for BG cloud at 2.2 km with concentration of $27 \cdot 10^3$ particles per liter for which we

estimated the parameters (from Figs. (3-4) to be μ_0=0.1135, σ_0=0.0349, μ_1=0.35 (the approximate location for cloud concentration of $27 \cdot 10^3$ particles per liter) and $\sigma_1=\sigma_0$ (assuming additive noise)

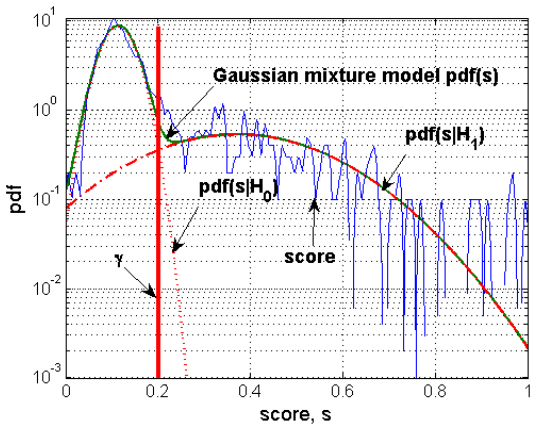

Fig. 4. The Gaussian mixture model (Eq. 4) and the location of the threshold γ (Eq. 5) computed for Fig. 3 with the EM solution parameters: ω=0.7471, μ_0=0.1135, σ_0=0.0349, μ_1=0.3709 and σ_1=0.1897. Note that the y-axis is logarithmic.

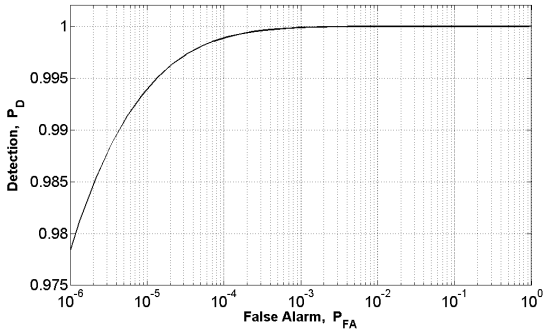

Fig. 5. Lidar receiver operating characteristic (ROC) for BG cloud with concentration of $27 \cdot 10^3$ particles per liter at ~ 2.2 km. Note that the x-axis is logarithmic.

5. SUMMARY

A new lidar detection algorithm that is based on hyperspectral anomaly detection was developed where the range vector ***r*** serves the same role as a spectral vector in hyperspectral analysis. The detection algorithm addresses the question "is a target of interest present within ranges R_1 to R_2" where the target in this paper is a biological cloud consists of BG bioaerosol particles. Anomaly of lidar signals for $R_1 < r < R_2$ (with respect to "other" lidar signals that are assumed not to contain the presence of the target) is interpreted as a presence of the target. The anomaly detector whitens (spheres) the lidar measurements that contain no target (the H_0 condition) and gives a high score to signals that contain anomaly (the H_1 condition). Assuming Gaussian statistics for the scores and a mixture model for the H_0 and H_1 scenarios, a probability model was developed for the detector. With the analytical probability model the probability of detection, probability of false alarm for a given threshold of detection and ROC curves are computed.

A filtering technique based on a low-rank orthogonal projection was developed to filter out unwanted oscillations (caused by hardware problems in our measurements). The filtering technique is useful for visualization of the lidar signals when the signal-to noise ratio is low and the presence of the target is difficult to see.

An example of CO_2 lidar measurements of BG aerosol cloud demonstrated the anomaly detection and the data filtering technique. The Roc curve computed for BG aerosol cloud of $27 \cdot 10^3$ particles per liter at ~ 2.2 km. shows excellent (high) detection probability and low false alarm.

6. RFERENCES

1. I. S. Reed and X. Yu, Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution, IEEE Trans. on Acoustic, Speech and Signal Processing,. Vol. 38, 1760-1770, 1990.

2. J. Harsanyi and C-I Chang, Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach, IEEE Trans. Geocsi. Rem. Sens., Vol. 32**,** 779-785, 1994.

3. L. L. Scharf, *Statistical Signal Processing, Detection, Estimation and Time Series Analysis*, Addison-Wesley, New York, New York, 1991.

4. A. Ben-David and H. Ren, Detection, Identification and estimation of biological aerosols and vapors with Fourier transform infrared spectrometer, Appl. Opt. Vol. 42, 4887-4900, 2003.

5. A. Ben-David, Temperature dependence of water vapor absorption coefficients for CO_2 differential absorption lidars, Appl. Opt., Vol. 32, 7479-7483, 1993

6. A. Ben-David, Backscattering measurements of atmospheric aerosols at CO_2 laser wavelengths: implications of aerosol spectral structure on differential absorption lidar retrieval of molecular species, Appl. Opt., 38, 2616-2624, 1999.

7. A. Ben-David., Remote detection of biological aerosols at a distance of 3 km with a passive Fourier transform infrared (FTIR) sensor, Opt. Express, Vol. 11, 418-429, 2003.