The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology

http://dms.sagepub.com/

Outlier Detection in Hyperspectral Imagery using Closest Distance to Center with Ellipsoidal Multivariate Trimming

Ryan F Caulk, Kevin B Reyes and Kenneth W Bauer, Jr

The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology published online 21 April 2011
DOI: 10.1177/1548512911403520

The online version of this article can be found at:
http://dms.sagepub.com/content/early/2011/04/21/1548512911403520

Published by:

http://www.sagepublications.com

On behalf of:

The Society for Modeling and Simulation International

Additional services and information for *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* can be found at:

Email Alerts: http://dms.sagepub.com/cgi/alerts

Subscriptions: http://dms.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
21 APR 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Outlier Detection In Hyperspectral Imagery Using Closest Distance To Center With Ellipsoidal Multivariate Trimming	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Institute of Technology,2950 Hobson Way,Wright-Patterson AFB,OH,45433**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The Journal of Defense Modeling and Simulation, 2011

14. ABSTRACT

In this paper we examine the efficacy of using the closest distance to center algorithm in conjunction with ellipsoidal multivariate trimming (MVT) to find outliers in a hyperspectral image. MVT is applied here as a global anomaly detector on images that are pre-processed into clusters using a technique called X-means. Under the assumption that there are no more than 5% outliers in any given cluster set, we develop a method, based upon principal component analysis preprocessing to create a flexible threshold for determining the percentage of data to retain with MVT. Using a retention percentage that more adequately reflects the actual number of outlier-free observations allows one to form estimates of the mean and covariance matrix that more effectively decrease the effects of swamping and masking as compared to using a set percentile for retention. These ideas are tested against real and synthetically generated hyperspectral imagery.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

Original article

JDMS

Journal of Defense Modeling and Simulation: Applications, Methodology, Technology
1–10
© 2011 The Society for Modeling and Simulation International
DOI: 10.1177/1548512911403520
dms.sagepub.com
SAGE

Outlier Detection in Hyperspectral Imagery using Closest Distance to Center with Ellipsoidal Multivariate Trimming

Ryan F Caulk,[1] Kevin B Reyes[1] and Kenneth W Bauer Jr[2]

Abstract

In this paper we examine the efficacy of using the closest distance to center algorithm in conjunction with ellipsoidal multivariate trimming (MVT) to find outliers in a hyperspectral image. MVT is applied here as a global anomaly detector on images that are pre-processed into clusters using a technique called X-means. Under the assumption that there are no more than 5% outliers in any given cluster set, we develop a method, based upon principal component analysis pre-processing, to create a flexible threshold for determining the percentage of data to retain with MVT. Using a retention percentage that more adequately reflects the actual number of outlier-free observations allows one to form estimates of the mean and covariance matrix that more effectively decrease the effects of swamping and masking as compared to using a set percentile for retention. These ideas are tested against real and synthetically generated hyperspectral imagery.

Keywords

blocked adaptive computationally efficient outlier nominators (BACON), closest distance to center, clustering, ellipsoidal multivariate trimming, hyperspectral imagery, outlier detection, principal component analysis

1. Background

This paper deals with the military application of hyperspectral imagery (HSI). A succinct, well-written summary of the military utility of HSI is found at GlobalSecurity.org.[1] An excerpt follows:

> Hyperspectral imaging technology uses hundreds of very narrow wavelength bands to 'see' reflected energy from objects on the ground. This energy appears in the form of 'spectral fingerprints' across the light spectrum and enables collection of much more detailed data and produce a much higher spectral resolution of a scene than possible using other remote sensing technologies.
>
> Once these fingerprints are detected, special algorithms—repetitive, problem-solving mathematical calculations—then assess them to differentiate various natural and manmade substances from one another. 'Signature' libraries may also be used to identify specific materials—e.g., rooftops, parking lots, grass, or mud—by comparing a library's pre-existing reference catalogs with freshly taken hyperspectral images of the battlefield from space.
>
> Image processing equipment then portrays the various types of terrain and objects upon it in different colors forming a 'color cube,' each based on the wavelength of the reflected energy captured by the image. These colors are subsequently 'translated' into maps that correspond to certain types of material or objects to detect or identify military targets such as a tank or a mobile missile launcher. Algorithms can also categorize types of terrain and vegetation (useful, for example, in counter-narcotic operations), detecting features such as disturbed soil, stressed vegetation, and whether the ground will support the movement of military vehicles.
>
> Once this technology is mature, theater commanders can use mobile ground stations to process in real-time information transmitted by the satellite, critical to theater commanders for them to keep pace with rapidly changing conditions.

[1] Randolf AFB, 151 J St East, TX 78150, USA
[2] Air Force Institute of Technology, Wright-Patterson AFB, 2950 Hobson Way, OH 45433, USA

Corresponding author:
Kenneth W Bauer Jr, Air Force Institute of Technology, Wright-Patterson AFB, 2950 Hobson Way, OH 45433, USA.
Email: kenneth.bauer@afit.edu

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

More detailed discussions are found in Ardouin et al.[2] and Briottet et al.[3] The background material given above reflects a general application area known as signature matching.[4] The application of such algorithms is complicated by the need to convert the collected sensor data into the spectra of the material of interest. The sensor collects what is known as spectral radiance. Radiance is modulated by atmospheric effects, such as the absorption of the energy of certain spectral bands and the superposition of solar energy scattered by the atmosphere on to the light reflected by an object. The spectra of the material of interest are measured in terms of what is called emissivity or reflectance. It can be thought of as the spectral signature of the material as collected under laboratory conditions where the effects of the atmosphere have been factored out. The application of signature-matching algorithms requires the conversion of the radiance into reflectance. This process is known as atmospheric calibration. Now, if the complications due to weather are compounded with the interest in several different objects, each of which may be associated with multiple signatures, there may be a desire to apply what are known as anomaly detectors. These seek to find observations that are different from typical background materials without using specific target signatures.[4]

In this paper, we propose a new anomaly detector. The method can be described as a global version of the localized RX algorithm.[5,6] The new method incorporates robust estimates of the filter's parameters. Where the RX algorithm involves the movement of a window through the pixels of an image while computing localized statistics, the proposed method computes its scores relative to a robust parameter set computed for clusters of pixels within the image. Related work can be found in Taitano et al.[7] In Section 2, a little more background is given for readers unfamiliar with hyperspectral imaging.

2. Introduction

Digital photographs taken from aircraft or satellites can be used for a wide range of military and civilian applications, such as locating a tank in a field or establishing the presence of a certain type of foliage. Several methods exist to locate anomalies in an image; for instance, highly trained individuals view the photograph with the human eye, or data from the image is analyzed using either local or global anomaly detectors. The first method can prove to be very difficult, especially in a highly cluttered area. It can also be extremely time consuming, because the area of interest has to be photographed and sent to an imagery expert, who then manually analyzes the image to determine if there are anomalies. The second, and possibly more effective method, is to analyze the data from the image using an image-processing algorithm known as an anomaly detector.

A hyperspectral image is similar to a photograph taken from an ordinary digital camera; however, a hyperspectral image may contain data from more than 250 wavelength bands from the electromagnetic (EM) spectrum, which includes some non-visible bands, whereas a standard digital camera collects data from only three bands in the visible spectrum, that the human eye sees as red, green, and blue. These images are made by specialized cameras placed on, say, an aircraft within the Earth's atmosphere or on a satellite in space. The image is divided up into pixels and the magnitude of the signal for each band is recorded for each pixel. The number of pixels in an image depends on the resolution of the camera. An image that captures fewer than 20 bands of the spectrum referred to as a multispectral image, and an image with 20 or more bands is called a hyperspectral image. All of this data is then stored in a three-dimensional *hypermatrix*,[8] referred to as a data cube, with the first two dimensions of the hypermatrix, x and y, being the location of the pixel in the image, and the third dimension, z, being the magnitude at each of the recorded EM bands. The image can be thought of as a series of vectors, one for each pixel location, that contains the wavelength magnitudes for each of the bands.

Consequently, HSI data can be analyzed using standard multivariate statistical techniques, and anomalies may be found by locating outliers within the data. Certain techniques specific to locating anomalies in an image, such as global anomaly detectors, work most efficiently when applied to homogenous datasets. Therefore, if data are being analyzed for the presence of anomalies in an image containing more than one main feature, such as a field with a road running through it, cluster analysis must be accomplished prior to using a global anomaly detector, or the detector may determine the road is the anomaly in the image, and true anomalies may be overlooked. When performed properly, cluster analysis splits the data into the requested number of subsets, known as clusters, allowing global anomaly detectors to analyze each cluster individually to produce the best results. In this paper, we will propose the use of the closest distance to center (CDC) algorithm[9] in conjunction with the ellipsoidal multivariate trimming (MVT) algorithm[10] as a method for finding anomalies in hyperspectral images. We call this new method 'screened MVT'. The purpose of this paper is to demonstrate that some standard tools used in process control can be readily adapted to a new problem area.

The paper is organized as follows. Firstly, we present a brief overview of the area of HSI. Next, we discuss the need for dimensionality reduction and begin the development of a CDC/MVT anomaly detector, screened MVT. The proposed method is tested on a set of hyperspectral images, and the paper is closed with a summary.

3. Hyperspectral imagery

To gain a basic understanding of HSI, we can begin with a discussion of the common digital camera that has become

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

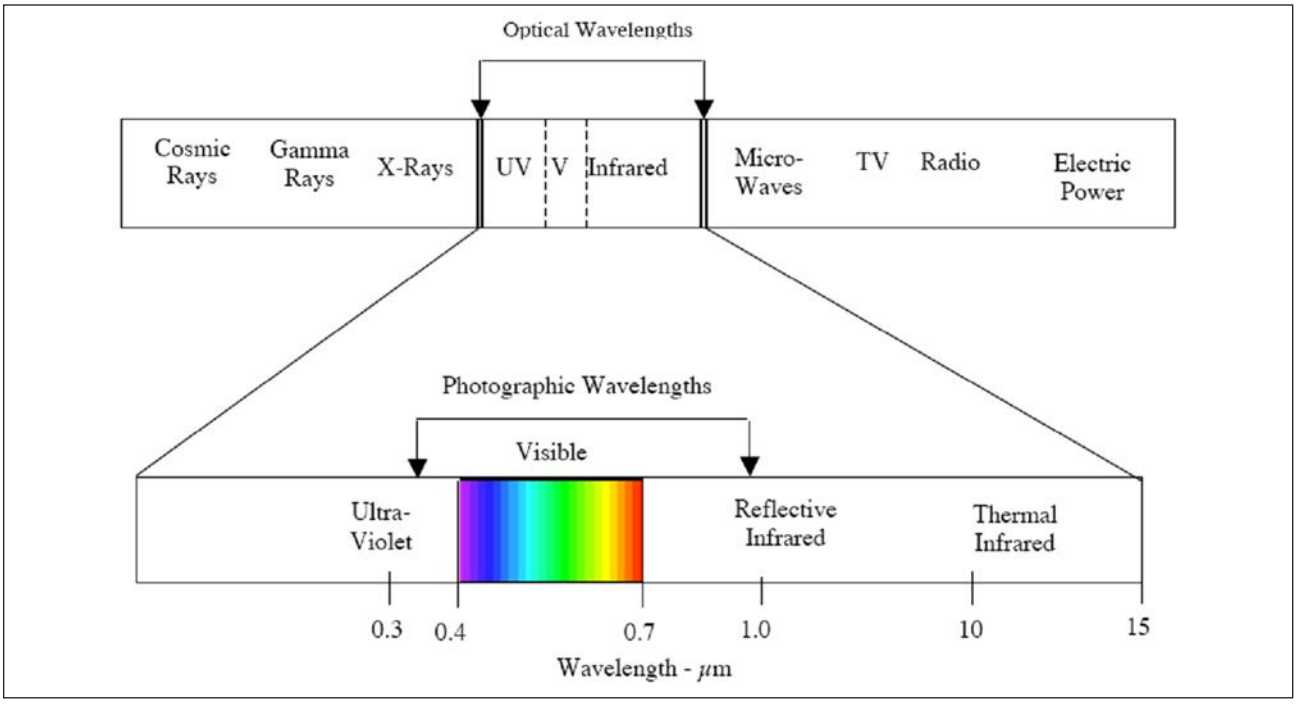

Figure 1. Example of the bands of the electromagnetic spectrum used in hyperspectral imagery.[11]

ubiquitous in modern society. Conceptually, when we use a digital camera to take a color photograph, the camera divides the imaged scene into a two-dimensional grid of pixels. For each pixel, three pieces of information are collected. These are, respectively, the amount of energy emanating from the pixel in the red, green, and blue portions of the EM spectrum. This information is stored in three separate two-dimensional arrays. For any given pixel, combining its respective red, green, and blue information produces the true color of the pixel. Of course, viewing the array of colored pixels on a computer screen or in its printed form reveals the scene originally photographed.

If we image a scene for the purpose of identifying different objects that it may contain, a simple color image produced by a digital camera may suffice; however, a true-color image has its limitations. For example, vegetation and camouflage nets may both appear green, making it very difficult for the human eye – or worse, for the computer – to discriminate one from the other. As seen in Figure 1, it is important to note that the visible spectrum of light is only a small fraction of the total EM spectrum that may be detected.

To address this limitation of true-color imagery, hyperspectral sensors collect information beyond the visible region of the EM spectrum. Just as a digital camera produces three images for wavelength bands corresponding to red, green, and blue light, a hyperspectral sensor produces images for many different contiguous wavelength bands, typically spanning the visible to near-infrared regions of the EM spectrum. The number of image bands collected by a sensor can range from 20 to over 500.

Consider the $M \times N$ pixelated scene of Figure 2. The hyperspectral sensor can be thought of as producing P different images, one for each band it collects. This collection of pixel-by-band information is often called an 'image cube'. For $m = 1, \ldots, M$ and $n = 1, \ldots, N$ the pixel in row m, column n of band 1 refers to the same spatial location of the scene as the pixels in row m, column n of every other band in the image cube. The sensor reading for a pixel in row m, column n, and band $\lambda = 1, \ldots, P$, can be referred to by the variable $x_{mn\lambda}$. For a given pixel address (m,n), we can form the vector

$$\begin{Bmatrix} x_{mn1} \\ x_{mn1} \\ \vdots \\ x_{mnP} \end{Bmatrix}. \quad (1)$$

This vector is often referred to as a pixel vector. If we take the transpose of all the pixel vectors in the image and place them in an $(M \times N) \times P$ array, we form the data matrix, $\mathbf{X}$, which is commonly used in multivariate statistical analysis.

Using $\mathbf{X}$, we are free to analyze the image data using multivariate analysis methods such as principal component analysis (PCA), cluster analysis, maximum likelihood classification, discriminant analysis, and others.

In this paper, all *real* test images are taken from the COMPact Airborne Spectral Sensor (COMPASS) and Hyperspectral Digital Imagery Collection Equipment (HYDICE) sensors. The COMPASS sensor is able to receive data on an area at 255 different wavelengths of light across

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

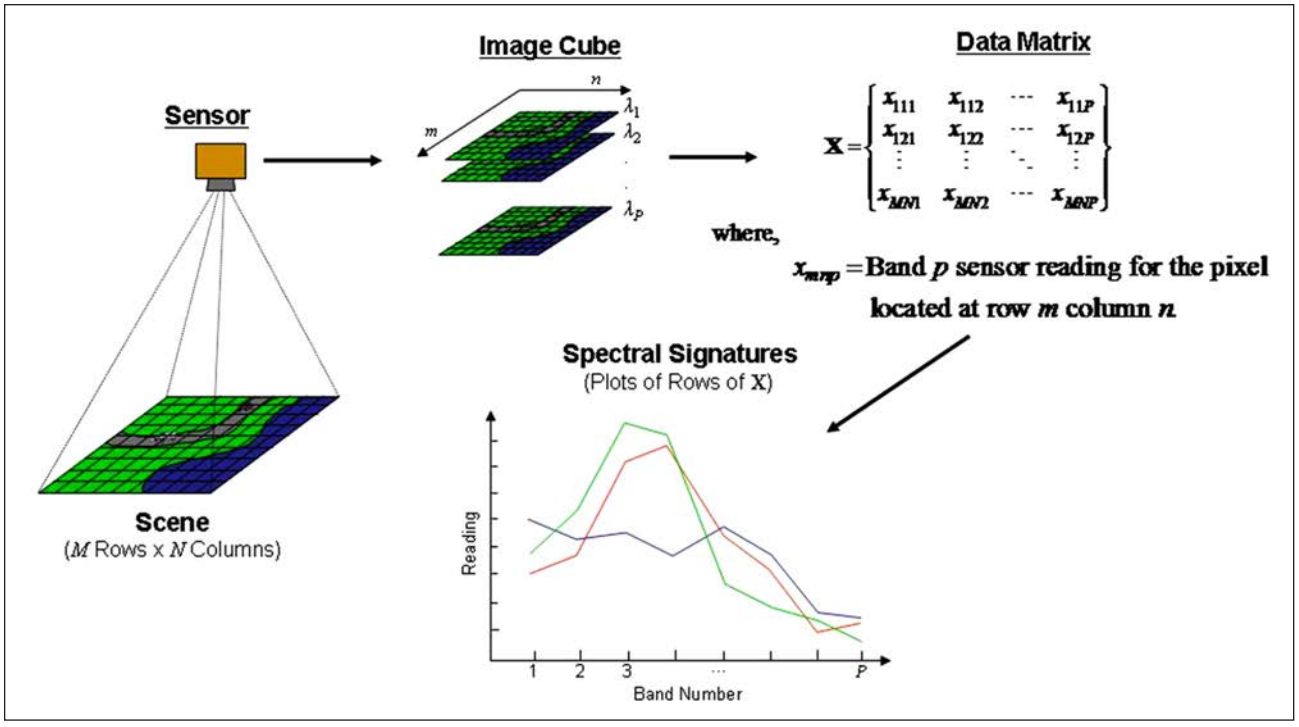

Figure 2. The basic hyperspectral imaging process and data representation.

the EM spectrum. Synthetic images were also employed in this research. They are described in a subsequent section.

4. Decreasing the dimensionality of the dataset

Since HSI is characterized by large volumes of data (over 28,800 pixels taken at over 200 wavelengths in the smallest example used in this paper), it is of practical necessity to decrease the dimensionality of the dataset. This is accomplished by PCA.[11,12] Other compression methods, such as the use of wavelets, have been proposed.[13,14] Here, PCA is employed since it is relatively simple to apply and, arguably, it is the standard compression method used in practice. It is well known that PCA can decrease the dimension of the data significantly while still maintaining the ability to explain variability in the dataset.[15] PC scores are found by projecting the data onto eigenvectors of their correlation/covariance matrix. For the purposes of this paper, we determined the number of PC components to retain by using Kaiser's criteria[15] on all images so that each set that was originally of dimension $(M \times N) \times P$ was decreased to dimension $(M \times N) \times r$, where $r << P$. The use of PCA for finding outliers in multivariate data is surveyed by Gnanadesikan and Kettenring[16] and Rao.[17]

As alluded to earlier, rather than attempting to find anomalies across entire images, the images were first clustered into homogeneous spectral groups using an algorithm called X-means.[18] X-means is a clustering technique that uses an iterative scheme to find the proper number of clusters and in turn perform the cluster binning based upon Bayesian information criterion (BIC) scores.[19]

5. Outlier detection overview

Certain outlier detection methods, such as MVT are known to be unreliable due to their use of the Mahalanobis distance in determining an initial mean vector and covariance matrix estimate.[9,10] The CDC algorithm has been employed to alleviate this problem by determining a more robust initial starting point for mean vector and covariance estimation;[9] the starting point being more compatible to the set of good data (without the outliers present) and with the object being to use these estimates to begin MVT. Applying MVT with the CDC algorithm as an initial starting point should perform significantly better than simply using MVT when there are multiple outliers in the data.[9] CDC/MVT seeks to trim out the bad data points to obtain more robust estimates of the covariance matrix and the mean vector. Such a procedure provides a more accurate Mahalanobis distance measurement that can be used to advantage to spot outlying observations.

6. The CDC/MVT algorithm

Herein the data have been pre-processed using PCA and a transformed dataset of significantly lower dimensionality is generated. The initial starting point for MVT is found by performing the CDC algorithm on the transformed dataset

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

to find 'good' estimates for the mean and the covariance. These estimates are rendered by finding the $n/2$ observations that are closest to the median vector using either Euclidean distance (2-norm) or the largest component absolute value difference from the centroid (max norm) and determining the mean and covariance from this subset.[9] Once these estimates are found we then begin MVT.

MVT is an iterative process that based upon a percentile (50% for Chiang's algorithm[9]) of the smallest observations of Mahalonobis distance within a sample. These observations are used to determine a new mean vector and covariance matrix. This process is repeated using the most-recent mean vector and covariance matrix until the mean vector and covariance matrix have stabilized. Once the iterations are complete, the resulting Mahalanobis distance for each data point is then used for outlier determination.[9] The literature varies slightly in one regard during this process. While Chiang et al.[9] state that the mean vector and covariance matrix must stabilize, Delvin et al.[20] recommend using the stabilization of only the correlation matrix as the stopping criterion. In addition, Delvin et al.[20] recommend using a difference of 10^{-3} as the stabilization criteria within the correlation matrix or a maximum of 25 total iterations. Now, the 50th percentile for retention used in MVT is due to the low breakdown point of 50% outliers for the algorithm. Devlin et al.[20] propose a different value for retention in MVT in which the percentile is equal to $100 \times (1-1/(p+1))\%$, where p is equal to the dimensionality of the dataset. In this paper, we also propose using a higher percentile for retention in MVT based upon the assumption that there are significantly fewer outliers than background pixels in any given hyperspectral image cluster. We call this procedure *screened MVT*. The percentage to retain is based upon first taking the Mahalanobis distances found after the CDC algorithm and computing a conservative percentile from a fitted gamma distribution (maximum likelihood parameter estimates)[21] with $\alpha_{\text{retain}} = 10^{-1}$. Next, the *percentage* to retain in MVT is determined by the number of observations that fell beyond this percentile.

The CDC/MVT algorithm is very similar to the blocked adaptive computationally efficient outlier nominators (BACON) algorithm described by Billor et al.[22] and adapted for hyperspectral image processing by Smetek and Bauer[23] in that its overall goal is to trim the dataset so that the true covariance structure of the data can be determined. Observations that are outliers will subsequently have much larger distance estimates and should be found easily in outlier determination after iterative estimates have stabilized. The CDC algorithm and our 'screened MVT' algorithm are detailed below.

6.1 The CDC algorithm

Input: An $n \times r$ matrix $\mathbf{X}$ of PC scores from hyperspectral data.

Output: An initial estimate of the mean, μ_0, and covariance, Σ_0, of the cluster set based upon the closest $n/2$ observations to the median.

Step 1: The median vector of the data is obtained.

Step 2: Determine the $n/2$ observations that are closest to this median vector using either Euclidean distance or the max norm distance, where n is equal to the size of the dataset.

Step 3: From the $n/2$ observations find estimates for the mean vector, μ_c, and covariance matrix, Σ_c The mean vector and covariance matrix are then used as a starting point for MVT.

6.2 The screened MVT algorithm

Input: An $n \times r$ matrix $\mathbf{X}$ of PC scores from hyperspectral data and initial estimates of μ and Σ from CDC.

Output: Mahalanobis distance calculations for the corresponding n data points in the cluster set.

Step 1: Determine the μ and Σ for the dataset via CDC (as shown above). These are μ_c and Σ_c, respectively.

Step 2: Compute Mahalanobis distances for each observation, x_i, $i = 1, \ldots, n$, using μ_c and Σ_c:

$$d_i(\mu, \Sigma) = \sqrt{(x_i - \mu_c)^{\mathrm{T}} \Sigma_c^{-1} (x_i - \mu_c)}$$

Step 3: Determine the percentage to retain in MVT by:

[A] fitting a Gamma distribution to the $d_i(\mu, \Sigma)$ Let

$F(z) = \Pr\{d_i(\mu, \Sigma) \le z\}$ for all real z

denote the Gamma c.d f. that is fitted to the$\{d_i(\mu, \Sigma): 1 = 1, \ldots, n\}$ of Mahalanobis distances;

[B] find the quantile, $d_{1-\alpha}^{\text{retain}}$, associated with a conservative

$\alpha_{\text{retain}} = 10^{-1}$, that is,

$d_{1-\alpha}^{\text{retain}} = F^{-1}(1 - \alpha_{\text{retain}})$;

[C] let m/n be the percentage to retain in MVT where m is the number of $d_i(\mu, \Sigma) < d_{1\ \alpha}^{\text{retain}}$.

Step 4: Take the corresponding observations that fall below the retention percentile, $d_{\alpha}^{\text{retain}}$, as determined in Step 3 for computation of new estimates for μ and Σ.

Step 5: Compare the new estimates for μ and Σ to that of the previous iteration. Return to Step 4 if the maximum absolute difference between estimates is above a user-defined threshold and MVT has iterated fewer than 25 times. Else, proceed to Step 6.

Step 6: Declare observations as outliers by comparing distances to an empirical distribution function of testing data. The cutoff α_{outlier} is typically chosen to be of the order 10^{-6} or, as will be seen in subsequent results, α_{outlier} can be varied to produce operating characteristic (OC) curves.

In application, these algorithms are processed sequentially and for each cluster set within an image.

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

Figure 3. Synthetic image using only red (dark grey), green (light grey), and blue (black) wavelengths.

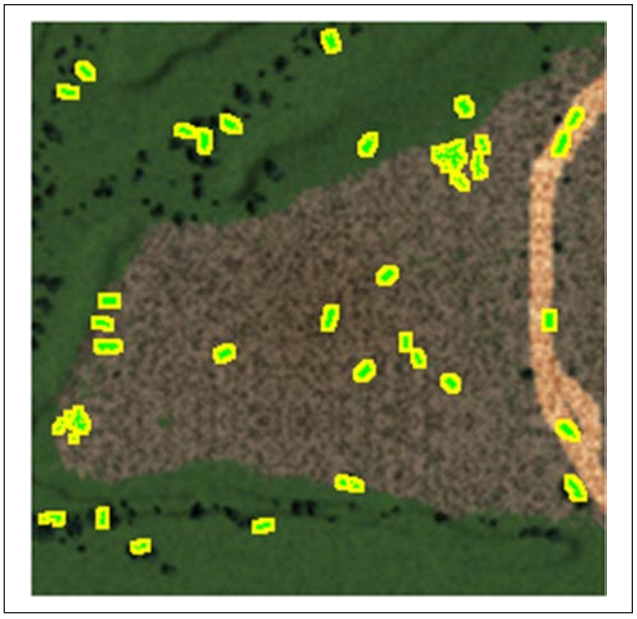

Figure 4. Synthetic image target outlier location mask.

7. Application to anomaly detection in hyperspectral imagery

In this research, we are looking for anomalies in a hyperspectral image. The assumption is that these anomalies are manmade and constitute spectral outliers in a statistical sense. To this end, we are not concerned with natural anomalies that may appear to be outliers when compared to their surroundings/clusters. As seen in the synthetic hyperspectral image (Figure 3), there are many observations that may be considered outliers in the picture (e.g. trees in a grass field).

To alleviate the problem of having many natural outliers dispersed throughout the image, the images are first clustered using X-means.[18] Once the image has been clustered, we are in a position to look for outliers within relatively homogenous datasets. As seen in Figure 4, the outliers in the image are shown in green (black) with a yellow (white) border.

The task is to identify as many of the green target pixels as possible (true positives), while minimizing the identification of non-target pixels as targets (false positives).

8. Testing results and analysis

We compared the results for 21 different hyperspectral images using three forms of MVT retention (detailed below). The results from these three algorithms were compared to output for the BACON algorithm[22,23] for the same images to determine the efficacy of each MVT algorithm against a robust baseline algorithm.

The images used for testing were taken from the Air Force's Airborne Remote Sensing Program (ARES), and synthetic images created using the Digital Imaging and Remote Sensing Image Generation (DIRSIG) program.[24] The ARES images were acquired through testing of the HYDICE sensor during the Forest Radiance I and Desert Radiance II data collection efforts. The images consist of manmade objects such as vehicles, panels, camouflage nets, and tables. For all real images, the locations of objects of interest were determined during collection. The synthetic images employed here were created at the same hypothetical geographical location with differences in time of day, sensor view angle, visibility, and target size. The reason for using synthetic images in our testing is that currently there are not a great number of 'truthed' hyperspectral images available to the general research community. The DIRSIG program is able to produce hyperspectral images that are representative of real-world images, and afford the advantage of allowing the user to specify the exact nature and location of all the anomalies in the image. The synthetic images used were all different variations of Figure 3.

OC curves were found by processing the resulting Mahalanobis distances and plotting the estimates for the true positive rate (at the pixel level) against the false positive rate. An additional measure was taken as the area underneath the OC curve. An example OC curve for the Air Force image is given in Figure 5.

In Figure 5, screened MVT has the largest area under the OC curve (AUC). The AUCs for all algorithms are given in Table 1.

The OC curves were obtained by varying $\alpha_{outlier}$. Each OC curve was only considered up to a false positive rate of 0.05, since rates higher than 0.05 would render target detection worthless based upon the preponderance of background pixels in any given image.

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

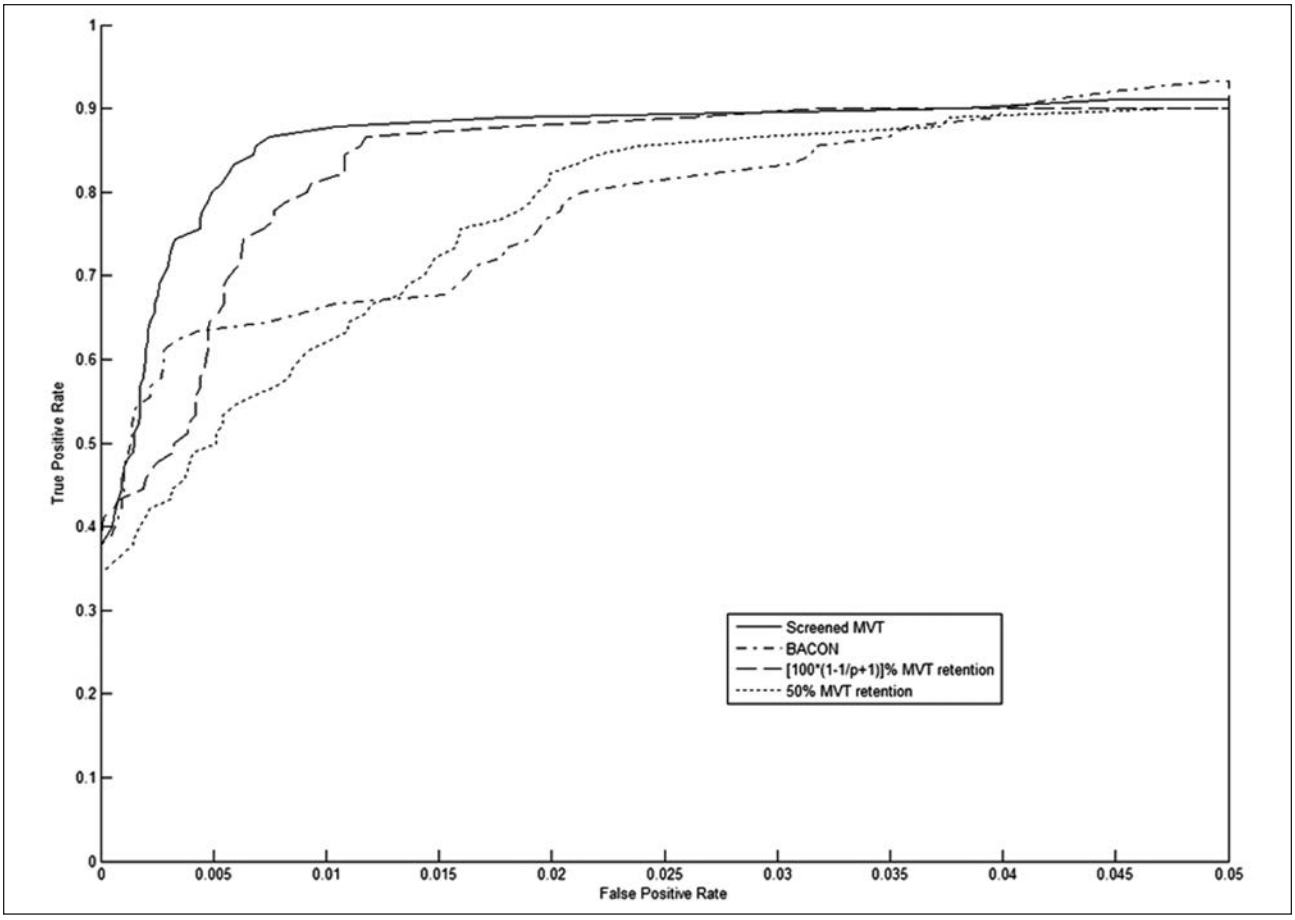

Figure 5. Operating characteristic curve example for Air Force image. MVT: ellipsoidal multivariate trimming, BACON: blocked adaptive computationally efficient outlier nominators.

Table 1. Area under the operating characteristic curve (AUC) values from Figure 5

Area under the OC curve			
Screened MVT	50% MVT retention	100 x (1 – 1/(p + 1))% retention	BACON
0.8646	0.7685	0.8351	0.7792

MVT: ellipsoidal multivariate trimming, BACON: blocked adaptive computationally efficient outlier nominators

Table 2. Summary of the repeated measures analysis of variance procedure with the Holm–Sidak test for multiple comparisons

Image type by performance measure	Significant treatments?	Significant contrasts
Real/AUC	No	
Real/Time	Yes	1–2/2–4/2–3
Synthetic/AUC	Yes	1–3/1–2/1–4
Synthetic/Time	Yes	1–2/2–3/2–4/1–4

AUC: area under the operating characteristic curve

A repeated measures analysis of variance (ANOVA) procedure, supplemented by the Holm–Sidak test for multiple comparisons was conducted for each of four datasets.[25] There were two distinct types of imagery (synthetic and real) and two performance measures of interest: AUC and computation time (in seconds with all processing on the same physical system). The four treatments are the algorithms as numbered below:

1. Screened MVT;
2. MVT with Chiang's 50% retention;
3. MVT with 100 × (1–1/(p+1))% retention;
4. BACON.

The subjects are the images. There were 15 synthetic images and six real images. Table 2 summarizes the analysis.

As is evident from Table 2, the methods, as applied to the real imagery, showed significant differences due to the treatments only where time was concerned. Examination of Figure 6 shows, as expected, that using a more flexible percentile for retention within MVT, in general, resulted in a larger AUC. As seen in Figure 6, screened MVT performed

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

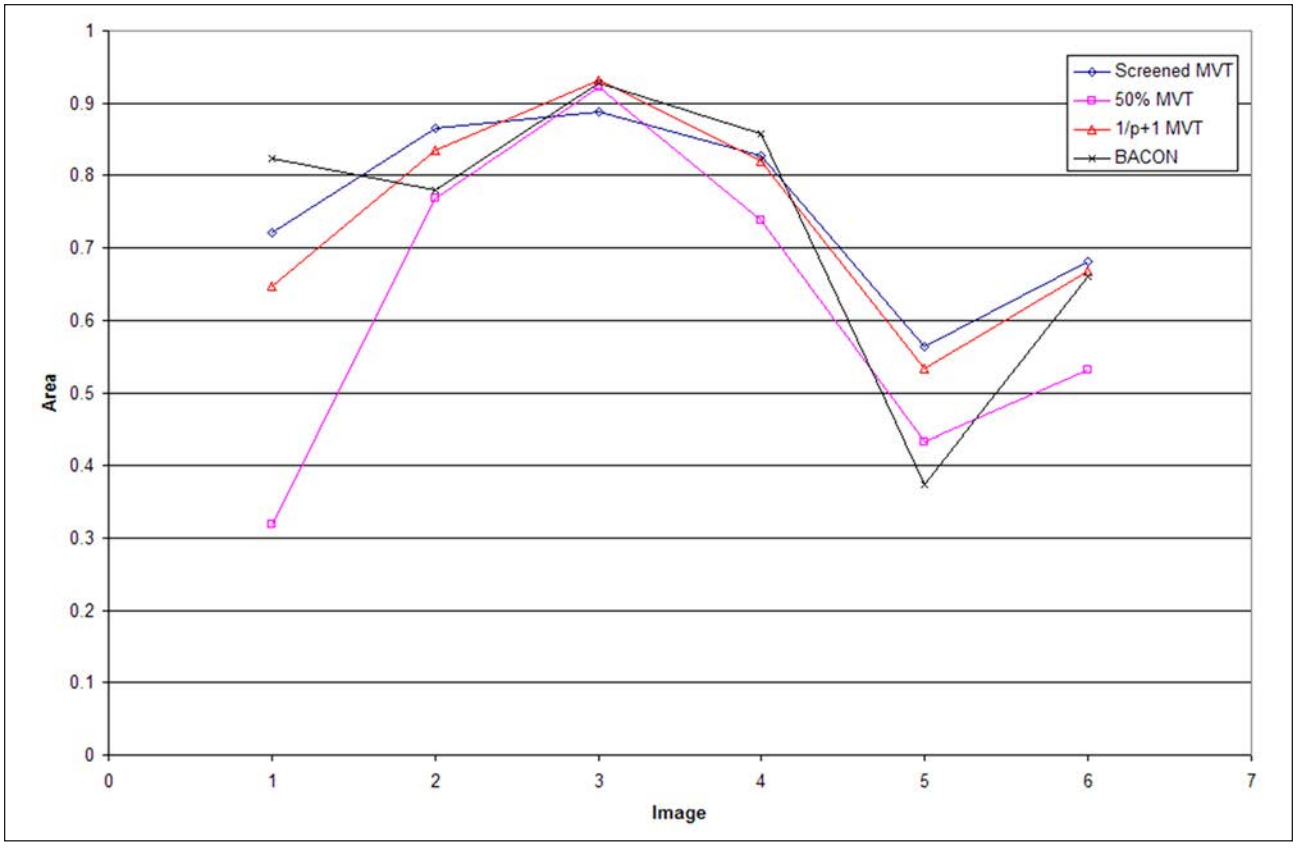

Figure 6. Area under the operating characteristic curve output for real images using four algorithms.
MVT: ellipsoidal multivariate trimming, BACON: blocked adaptive computationally efficient outlier nominators.

Table 3. Mean performance of the procedures for the real images

Treatment	Average AUC	Average time
Screened MVT	0.76	63
MVT with Chiang's 50% retention	0.62	150
MVT $100\times\left(1-\frac{1}{p+1}\right)\%$ retention	0.74	72
BACON	0.74	64

AUC: area under the operating characteristic curve, MVT: ellipsoidal multivariate trimming, BACON: blocked adaptive computationally efficient outlier nominators

Table 4. Mean performance of the procedures for the real images

Treatment	Average AUC	Average time
Screened MVT	0.34	67
MVT with Chiang's 50% retention	0.16	197
MVT $100\times\left(1-\frac{1}{p+1}\right)\%$ retention	0.15	93
BACON	0.23	116

AUC: area under the operating characteristic curve, MVT: ellipsoidal multivariate trimming, BACON: blocked adaptive computationally efficient outlier nominators

the best for three of the six real images. It was also observed that in most cases it took less time to complete the algorithm. Average responses are recorded in Table 3.

A more distinct separation of algorithms was observed for the synthetic images. As depicted in Figure 7, screened MVT performed the best for 11 of the 15 images tested. It must be noted that the procedures performed poorly, in general, across the synthetic images. This was largely due to the presence of many natural anomalies in the image that were not clustered into their own cluster set. Although these observations were outliers within their cluster set, they were not the anomalies of interest.

Significant contrasts were noted for screened MVT versus the other procedures in terms of AUC (Tables 2 and 4). Significant differences in processing times were also found, as indicated in Table 2.

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

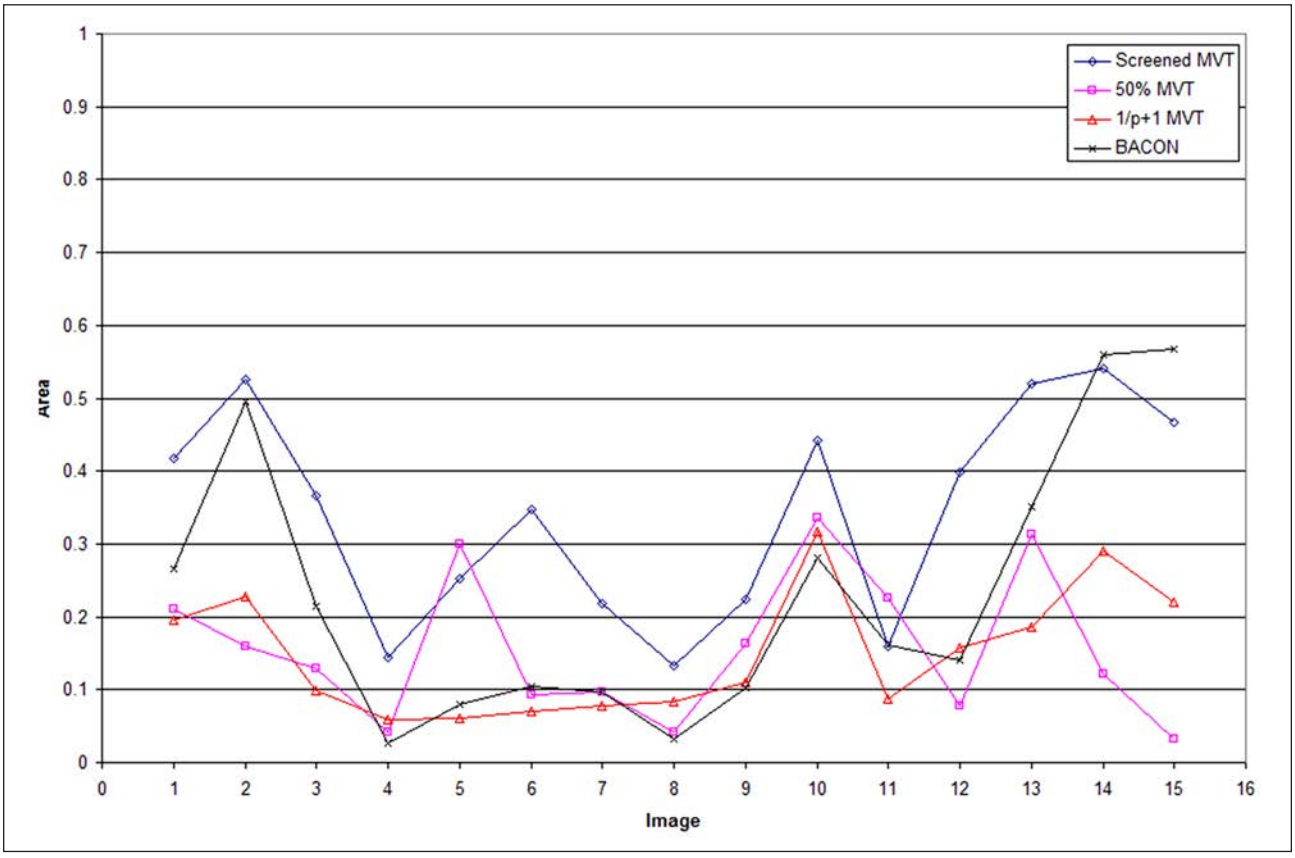

Figure 7. Area under the operating characteristic curve output for real images using the four algorithms.
MVT: ellipsoidal multivariate trimming, BACON: blocked adaptive computationally efficient outlier nominators.

9. Summary

By utilizing a more flexible estimate for MVT retention via PCA screening we were able to improve upon the algorithms that use the standard set retention percentiles of 50% and $100 \times (1–1/(p+1))\%$ within MVT. This was accomplished while maintaining an analysis approach that was relatively 'hands off'. We believe the approach is promising largely because no two images are alike in all aspects. By maintaining a more flexible trimming percentage, we were able to avoid some of the swamping and masking effects that were present in the rigid MVT settings that use fewer of the available 'good' observations for MVT retention.

There were many natural anomalies in the DIRSIG-created images that resulted in a fairly significant swamping effect for both the BACON and MVT algorithms. The BACON algorithm suffered more from the presence of these anomalies, though, due to the trimming process used. This is because these natural anomalous observations were not included in the estimates of the mean and covariance in the trimming process since they were considered outliers for the clusters in which they were located. Even though these observations should have been trimmed since they do not fit into the clusters, they are not targets of interest. Furthermore, not including these observations in the dataset tended to tighten the estimates for the covariance and the mean so that their Mahalanobis distances were inflated and, thus, they tended to have distance estimates that looked like targets.

In this paper we examine the efficacy of using the CDC algorithm in conjunction with MVT to find outliers in a hyperspectral image. A method is advanced to create a flexible retention percentage that more adequately reflects the actual number of outlier-free observations, thereby allowing one to form robust estimates of the mean and covariance matrix that may more effectively decrease the effects of swamping and masking as compared to using a set percentile for retention. The effectiveness of these ideas is demonstrated against real and synthetically generated HSI.

Acknowledgements

Special thanks to Mr Chuck Sadowski, ACC/A8SP, for his continuing support of our research. We are also indebted to four anonymous reviewers for their helpful comments.

Funding

This work was supported by ACC/A8SP and AFRL/RYZT.

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011

Conflict of interest statement

None declared.

References

1. GlobalSecurity.Org. '300 N', Washington St. Arlington VA, http://www.globalsecurity.org/intell/library/imint/hyper.htm. Accessed: 17/9/2010
2. Ardouin J-P, Levesque J and Rea TA. A demonstration of hyperspectral image exploitation for military applications. In: *Proceedings of the 10th Conference on Information Fusion*, Quebec, 9–12 July 2007 pp.1–8.
3. Briottet X, Boucher Y, Dimmeler A, Malaplate A, Cini A, Diani M, et al. Military applications of hyperspectral imagery, targets and backgrounds XII: characterization and representation. *Proc SPIE* 2006; 6239: 62390B.
4. Stein DWJ, Beaven SG, Hoff LE, Winter EM, Schaum AP and Stocker AD. Anomaly detection from hyperspectral imagery. *IEEE Signal Process Mag* 2002; 19: 58–69.
5. Reed IS and Yu X. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Trans Acoust Speech Signal Process* 1990; 38: 1760–1770.
6. Yu X, Hoff LE, Reed IS, Chen AM and Stotts LB. Automatic target detection and recognition in multiband imagery: a unified ML detection and estimation approach. *IEEE Trans Image Process* 1997; 6: 143–156.
7. Taitano YP, Geier BA and Bauer KW Jr. A locally adaptable iterative RX detector. *EURASIP J Adv Signal Process* 2010; Article ID 341908.
8. Wilson JR. Antithetic sampling with multivariate inputs. *Am J Math Manage Sci* 1983; 3: 121–144.
9. Chiang LH, Pell RJ and Seasholtz MB. Exploring process data with the use of robust outlier detection algorithms. *J Process Contr* 2003; 13: 437–449.
10. Egan, WJ and Morgan SL. Outlier detection in multivariate analytical chemical data. *Anal Chem* 1998; 77: 2372–2379.
11. Landgrebe DA. *Signal theory methods in multispectral remote sensing*. Hoboken, NJ: John Wiley & Sons, 2003.
12. Richards JA and Jia X. *Remote Sensing Digital Image Analysis: An Introduction*. Berlin: Springer, 1999, p. 363.
13. Bruce LM, Koger CH and Li J. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *IEEE Trans Geosci Remote Sens* 2002; 40: 2331–2338.
14. Kaewpijit S, Le Moigne J and El-Ghazawi T. Automatic reduction of hyperspectral imagery using wavelet spectral analysis. *IEEE Trans Geosci Remote Sens* 2003; 41:863–871.
15. Dillon WR and Goldstein M. *Multivariate analysis: methods and applications*. New York: John Wiley & Sons, Inc., 1984.
16. Gnanadesikan R and Kettenring JR. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* 1972; 28: 81–124.
17. Rao CR. The use and interpretation of principal component analysis in applied research. *Sankhya A* 1964; 26: 329–358.
18. Williams JP. Robustness of multiple clustering algorithms on hyperspectral images. *MS thesis*, School of Operation Sciences, Air Force Institute of Technology (AU), Wright-Patterson AFB, OH, March 2007.
19. Pelleg D and Moore A. X-means: extended K-means with efficient estimation of the number of clusters. In: *Proceedings of the 17th International Conference on Machine Learning*, 2000, pp.727–734.
20. Devlin SJ, Gnandesikan R and Kettenring JR. Robust estimation of dispersion matrices and principal components. *J Am Stat Assoc* 1981; 76: 354–362.
21. Caulk RF. Outlier detection in hyperspectral imagery using closest distance to center with ellipsoidal multivariate trimming. *MS thesis*, School of Operation Sciences, Air Force Institute of Technology (AU), Wright-Patterson AFB, OH, March 2007.
22. Billor N, Hadi AS and Velleman PF. BACON: blocked adaptive computationally efficient outlier nominators. *Comput Stat Data Anal* 2000; 34: 279–298.
23. Smetek TE and Bauer KW. A comparison of multivariate outlier detection methods for finding hyperspectral anomalies. *Mil Oper Res* 2008; 13: 19–44.
24. Bellucci JP, Smetek TE and Bauer KW. Improved hyperspectral image processing algorithm testing using synthetic imagery and factorial designed experiments. IEEE *Trans Geosci Remote Sens* 2010; 48: 1211–1223.
25. Cardillo G. 'Anovarep: Compute the Anova for Repeated Measures and Holm–Sidak Test for Multiple Comparisons if Anova is Positive', http://www.mathworks.com/matlabcentral/fileexchange/18746 (2008). Access dates: 17/9/2010

Author Biographies

Captain Ryan F Caulk is a 2007 graduate of the Masters of Science in Operations Research degree program at the Air Force Institute of Technology, Wright-Patterson Air Force Base (AFB), Ohio, USA. He is currently assigned to the Studies and Analysis Squadron, Randolph AFB.

Captain Kevin B Reyes is a 2007 graduate of the Masters of Science in Operations Research degree program at the Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA. Upon graduation from that program he was assigned to the Defense Threat Reduction Agency in Alexandria, Virginnia, USA.

Kenneth W Bauer Jr is a Professor of Operations Research at the Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA. where he teaches classes in applied statistics and pattern recognition. His research interests lie in the areas of automatic target recognition and multivariate statistics.

Downloaded from dms.sagepub.com at AFIT/ENWL on July 5, 2011