Performance metrics for the evaluation of hyperspectral chemical identification systems

Citation: Truslow, Eric, Steven Golowich, Dimitris Manolakis, and Vinay Ingle. "Performance Metrics for the Evaluation of Hyperspectral Chemical Identification Systems." Opt. Eng 55, no. 2 (February 10, 2016): 023106. ©2016 SPIE.

As Published: http://dx.doi.org/10.1117/1.oe.55.2.023106

Publisher: SPIE--Society of Photo-Optical Instrumentation Engineers

Persistent URL: http://hdl.handle.net/1721.1/108594

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Performance metrics for the evaluation of hyperspectral chemical identification systems

Eric Truslow
Steven Golowich
Dimitris Manolakis
Vinay Ingle

SPIE.

Performance metrics for the evaluation of hyperspectral chemical identification systems

Eric Truslow,[a,*] Steven Golowich,[a] Dimitris Manolakis,[a] and Vinay Ingle[b]
[a]MIT Lincoln Laboratory, 244 Wood Street, Lexington, Massachusetts 02420-9185, United States
[b]Northeastern University, Department of Electrical and Computer Engineering, 360 Huntington Avenue, Boston, Massachusetts 02115, United States

Abstract. Remote sensing of chemical vapor plumes is a difficult but important task for many military and civilian applications. Hyperspectral sensors operating in the long-wave infrared regime have well-demonstrated detection capabilities. However, the identification of a plume's chemical constituents, based on a chemical library, is a multiple hypothesis testing problem which standard detection metrics do not fully describe. We propose using an additional performance metric for identification based on the so-called Dice index. Our approach partitions and weights a confusion matrix to develop both the standard detection metrics and identification metric. Using the proposed metrics, we demonstrate that the intuitive system design of a detector bank followed by an identifier is indeed justified when incorporating performance information beyond the standard detection metrics. [DOI: 10.1117/1.OE.55.2.023106]

Keywords: hyperspectral; remote sensing; detection; identification; performance estimation; matched filtering.

Paper 150739P received Jun. 1, 2015; accepted for publication Jan. 13, 2016; published online Feb. 10, 2016.

1 Introduction

Passive hyperspectral sensors operating in the long-wave infrared (LWIR) provide high-resolution measurements in a region of the electromagnetic spectrum, where many chemicals have unique absorption profiles. The high spectral and spatial resolution of these sensors allows for the identification of the individual chemicals within a gaseous plume.[1–4] A library of chemical absorption signatures is often all the prior knowledge available, and it is the job of the signal processing algorithms to decide if chemicals are present and also which ones; the process of finding which chemicals are present in the plume is known as identification. A simple but effective system design consists of separate detection and identification algorithms.[5] This type of system is shown in Fig. 1. In order to assess the performance of this system, and to compare the performances of different algorithms, it is necessary to define metrics that address both the detection and identification tasks. In this paper, we propose a performance evaluation methodology based on confusion matrices. Furthermore, we employ this methodology to demonstrate that the design of a state-of-the-art detection algorithm followed by an identification algorithm is superior to that of the detection algorithm alone.

The difference between detection and identification is somewhat subtle. We categorize problems depending on the number of chemicals in the library and whether mixtures of chemicals can be present or not. These distinctions can be summarized as:

1. looking for a specific chemical;
2. looking for a single chemical from a library of L chemicals;
3. looking for mixtures of up to m chemicals from a library of L chemicals.

The first case is a detection problem where the library contains a single gas whose presence or absence needs to be decided; detection is inherently a binary hypothesis problem. When the library contains L chemicals and we are trying to determine which one is present, we no longer have a pure detection problem as there are multiple hypotheses to choose from. In case 2, not only do we have the presence or absence of the plume to decide, but also which single chemical is actually present. In case 3, instead of picking a single chemical, a mixture of chemicals can be present.

Chemical identification can be formulated as a multiple hypothesis testing problem where each hypothesis represents a subset of the library chemicals. Each pixel in the scene has an associated true hypothesis and an output hypothesis. A natural way to represent the performance of such a system is through a confusion matrix (CM), also known as an error matrix, where each pair of truth and output hypotheses is represented by a single entry of the matrix. A particular dataset and set of algorithm parameters produce a single realization of the CM, which can then be summarized by performance metrics.

For detection problems, a single threshold determines the operating point of the system and the CM can be summarized using the correct detection rate (CDR) and the false alarm rate (FAR). Sweeping a range of thresholds leads to a plot of CDR versus FAR, called a receiver operating characteristic (ROC) curve. An ROC curve fully characterizes the performance for detection problems.[6] In cases 2 and 3, the CM becomes larger and more difficult to interpret and may not be governed by a single threshold. When multiple thresholds are used, the construction of an ROC surface which characterizes performance is possible but is not easily visualized and is difficult to interpret.[7] To simplify our analysis, we consider algorithms that require a single threshold. Evaluating the system for a range of thresholds then produces a performance

*Address all correspondence to: Eric Truslow, E-mail: eric.truslow@ll.mit.edu

0091-3286/2016/$25.00 © 2016 SPIE

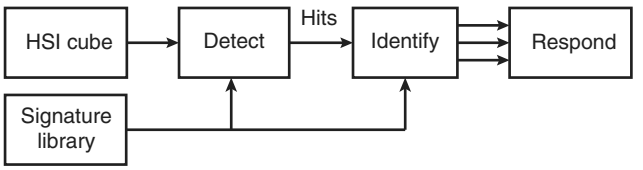

Fig. 1 Practical chemical detection–identification system. The identifier has one output for each chemical in a known library.

curve for each metric used from which performance trends can be assessed.

The appropriate metrics to use in summarizing a CM depend on the particular application. In hyperspectral chemical plume detection and identification problems, the number of background (nonplume) pixels is far larger than the number of plume pixels, making the FAR an important consideration when operating a real system. Therefore, our approach partitions the CM into two parts: one involving the background pixels and the other involving the pixels that contain plume. For the portion of the CM containing plume pixels, we utilize two different metrics, the CDR and the F-metric which is computed using the Dice index.[8,9] Both performance metrics can be computed by weighting the CM and then averaging. The choice of weights determines the performance metrics.

The CDR is the fraction of plume pixels where at least one chemical is correctly detected, and it is the performance metric we use when considering the detection problem. The identification metric we choose is known as the F-metric and is computed using the Dice weighting, which is based on the amount of agreement between the truth and output hypotheses.[10] Unlike the weighting for the CDR, the Dice index considers both the number of correct and incorrect chemicals in the output. Since identification deals with mixtures of chemicals and the CDR does not incorporate mixtures, the F-metric is the weighting we choose for evaluating identification performance. Ultimately, a good system should have a low FAR, a high CDR, and high identification performance as measured by the Dice index.

The system we evaluate is a detector bank followed by an identifier. The detector bank is composed of adaptive coherence (cosine) estimator (ACE) detectors,[11] while the identifier is the Bayesian model averaging (BMA) approach.[12,13] Intuitively, using a bank of detectors as an identifier should perform worse as an identifier than an algorithm designed for identification; similarly, an identifier should perform worse at detection than a detector. We use the proposed identification metric and standard detection metrics to demonstrate that the ACE detector bank has a better detection performance than BMA, but also has a lower identification performance than BMA. These results suggest using the detector bank followed by the identifier for improved performance over either individually, which we demonstrate to be the case.

For single-chemical problems, a variety of algorithms have been used including classical detection algorithms such as the MF, matched filter variants, and the ACE detector.[6] The spectral angle mapper is another statistic commonly used in determining spectral similarity.[14] We choose the ACE algorithm as a detector since it is a very effective and popular detection algorithm for hyperspectral imagery.[15] Identifiers are designed for when chemical mixtures are permissible. Several identification techniques have been proposed including a bank of detectors,[11] linear regression models with significance testing,[16,17] stepwise regression,[18,19] and Bayesian techniques.[12,20,21] BMA was selected as the identifier because it is considered the state-of-the-art for identification problems.

To understand how performance is affected by the parameters of the plume, we used a plume-embedding procedure to produce synthetic plume data that preserves the variability of the background data. Both algorithms were tested individually on embedded data for a range of thickness parameters. The cascaded system was then tested for the same parameter ranges.

The key results of our study show that using a cascaded detector and an identifier can achieve an overall good performance, a result that has not been demonstrated in the literature before. To our knowledge, using a CM to develop a series of performance metrics and the use of the F-metric as a performance measure have not been done in this field. Our approach provides a starting place for comparative analysis of other system designs.

The remainder of this paper is organized as follows. Section 2 presents background material on the phenomenology of the data and explains the simplifications used in deriving the at-sensor radiance models. In Sec. 3, we discuss confusion matrices and the proposed identification performance metric. In Sec. 4, the two identification algorithms are defined, and the key equations are presented. In Secs. 5.2 and 5.3, we compare the detection and identification performance of the two identification algorithms individually. The effect of plume thickness on the identification performance for each algorithm is explored in Sec. 5.4. The performance of the cascaded system with respect to plume thickness is examined in Sec. 5.5. Finally, in Sec. 6, we provide a short summary of the paper and discuss future work.

2 At-Sensor Radiance Signal Model

A simple but useful model for the at-sensor radiance in the LWIR can be developed from a full radiative transfer model with a few key assumptions:

1. The plume is optically thin and the distance between the plume and background are small enough to neglect the atmospheric transmittance in that region.
2. The plume is homogeneous in temperature and composition across the instantaneous field of view of the pixel under inspection.
3. Scattering and reflections can be neglected.

These simplifications allow the use of the three-layer model of Fig. 2 which can be used to derive our primary measurement equations using Kirchoff's law.[22] From Fig. 2, the measured radiance for an off-plume pixel is

$$L_{\text{off}}(\lambda) = [1 - \tau_{\text{a}}(\lambda)]B(\lambda, T_{\text{a}}) + \tau_{\text{a}}(\lambda)L_{\text{b}}(\lambda), \quad (1)$$

where λ is typically in units of μm, T_{a} is the temperature of the atmosphere, τ_{a} is the transmittance of the atmosphere, L_{b} is the background radiance, and $B(\lambda, T)$ is the Planck function, which describes a black body at temperature T.

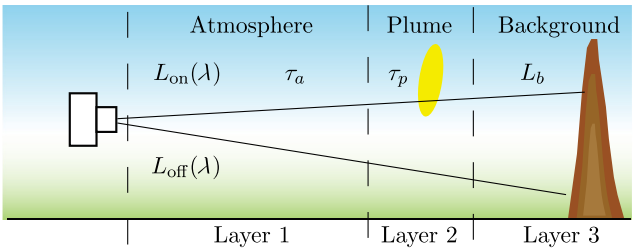

Fig. 2 Simplified three-layer radiance model for thin plumes.

The measured radiance for a pixel with plume becomes

$$L_{on}(\lambda) = [1-\tau_a(\lambda)]B(\lambda,T_a) + \tau_a(\lambda)[1-\tau_p(\lambda)]B(\lambda,T_p) + \tau_a(\lambda)\tau_p(\lambda)L_b(\lambda), \tag{2}$$

where τ_p is the transmittance of the plume. In terms of L_{off}, we instead have

$$L_{on}(\lambda) = \tau_a(\lambda)[1-\tau_p(\lambda)][B(T_p,\lambda) - L_b(\lambda)] + L_{off}(\lambda). \tag{3}$$

The model in Eq. (3) is useful for analysis and gives a clear method for generating synthetic data under certain circumstances; namely, that the plume and atmosphere are in equilibrium.

The transmittance of the plume τ_p is governed by Beer's law:[22]

$$\tau_p(\lambda) = \exp\left[-\sum_{i=1}^{m}\alpha_i s_i(\lambda)\right], \tag{4}$$

where m is the number of gases in the plume, α_i is the concentration path length (CL) for gas i, and s_i is the gas's absorption spectrum. The product of the CL and the absorption spectrum $\alpha_i s_i(\lambda)$ is known as the optical depth (OD) and it is a unitless quantity.

When the thermal contrast $\Delta_T = T_p - T_b$ between the plume and background is small and the background radiance is slowly varying with respect to wavelength, the difference $B(T_p,\lambda) - L_b(\lambda)$ is approximately proportional to Δ_T. Using the approximation $(1-e^x) \approx x$, we obtain

$$x(\lambda) \approx \Delta_T \tau_a(\lambda)\sum_{i}^{m}\alpha_i s_i(\lambda) + L_{off}(\lambda), \tag{5}$$

which is linear in terms of the signatures. Assuming the atmospheric transmission $\tau_a(\lambda)$ is known, the signatures are multiplied by the atmospheric transmission τ_a before further processing is done. The problem of estimating the temperature and emissivity quantities separately is known as temperature emissivity separation. For our purposes, we consolidate the product $\Delta_T\alpha_i$ into a single quantity b_i.

The input signal gets convolved with the sensor response function and is then sampled by the sensor at a set of band centers $[\lambda_1,\ldots,\lambda_p]$ to produce a measurement vector $\mathbf{x} = [x_1,\ldots,x_p]^T$ where p is the number of sensor channels. The library signatures $s_i(\lambda)$ are multiplied by the assumed atmospheric transmission $\tau_a(\lambda)$ and the product is then sampled to the sensor's resolution using the sensor's estimated spectral response to obtain sampled signatures $\mathbf{s}_i$. Organizing the signatures as a matrix $\mathbf{S}$ and the b_i's as a vector $\mathbf{b}$, we have

$$\mathbf{x} = \sum_{i=1}^{m}\mathbf{s}_i b_i + \mathbf{v} = \mathbf{S}\mathbf{b} + \mathbf{v}, \qquad \mathbf{v} \sim \mathcal{N}(\mathbf{m}_b, \mathbf{C}_b), \tag{6}$$

where $\mathbf{m}_b$ is the background clutter mean and $\mathbf{C}_b$ is the clutter covariance. The assumption in Eq. (6) is that the noise and background clutter $\mathbf{v}$ are well modeled by a multivariate Gaussian distribution, which may not hold in reality, but it is a useful model for many practical algorithms. Defining the whitening matrix $\mathbf{C}_b^{-1/2}$ and whitened vectors

$$\tilde{\mathbf{x}} = \mathbf{C}_b^{-1/2}(\mathbf{x}-\mathbf{m}_b), \quad \tilde{\mathbf{S}} = \mathbf{C}_b^{-1/2}\mathbf{S}, \quad \tilde{\mathbf{v}} = \mathbf{C}_b^{-1/2}(\mathbf{v}-\mathbf{m}_b), \tag{7}$$

yields the standard regression model

$$\tilde{\mathbf{x}} = \sum_{i=1}^{m}\tilde{\mathbf{s}}_i\tilde{b}_i + \tilde{\mathbf{v}} = \tilde{\mathbf{S}}\tilde{\mathbf{b}} + \tilde{\mathbf{v}}, \qquad \tilde{\mathbf{v}} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{8}$$

where the clutter and noise is zero mean and has identity covariance. The linear model of Eq. (8) is useful for developing the identification algorithms we consider, but it is a good approximation only for thin plumes and fairly uniform backgrounds. While the sensor noise may be well modeled by a multivariate Gaussian distribution, real hyperspectral backgrounds may be multimodal and have heavy-tails which make the Gaussian model inadequate for some applications, but useful in deriving practical algorithms.[23]

3 Plume Identification Performance Metrics

The framework we present for evaluating plume identifiers relies on using the CM for multiclass or multilabel problems.[24–29] For a single dataset, the CM has all the performance information available in detail. However, for ease of interpretation and comparing algorithms, the CM is summarized using several scalar performance metrics that are appropriate for detection and identification.

Given a library of L chemical signatures, it is the goal of the identifier to find which chemicals are present in each pixel. For each pixel, the identifier produces a binary variable g_k for each library signature, with $k \in \{1,\ldots,L\}$. Similarly, for each pixel there is a ground truth value for each chemical in the library denoted t_k. These variables indicate the presence or absence of each library chemical as follows:

$$g_k = \begin{cases} 1, & \text{Chemical } k \text{ identified in pixel} \\ 0, & \text{Chemical } k \text{ not identified in pixel} \end{cases}$$

$$t_k = \begin{cases} 1, & \text{Chemical } k \text{ is present in pixel} \\ 0, & \text{Chemical } k \text{ is absent from pixel} \end{cases}.$$

To represent these values compactly, we arrange the binary variables into binary vectors as follows: $\mathbf{g} = [g_1, g_2, \ldots, g_L]^T$ and $\mathbf{t} = [t_1, t_2, \ldots, t_L]^T$. The binary vectors $\mathbf{g}$ and $\mathbf{t}$ have M unique configurations depending on the maximum number of chemicals allowed to be present. The allowed configurations are denoted $\mathbf{g}_i$ and $\mathbf{t}_j$, with $i,j \in [1,\ldots M]$. Each truth vector $\mathbf{t}_j$ is assigned to a column of the CM, while every possible identifier output is assigned to a row of the CM; each cell

of the CM corresponds to a particular pair of truth and output vectors, and each pixel is assigned to a particular cell based on the vectors associated with it. In summary, the CM contains in element (i, j) a tally of the number of pixels with output $\mathbf{g}_i$ and truth $\mathbf{t}_j$.

Operationally, the CM is constructed by tallying each pixel in the correct entry of the CM based on the identifier output and truth, as shown in Fig. 3. The CM varies in size depending on both the size of the library and whether mixtures are allowed. In terms of hypothesis testing or classification, each hypothesis H_k has a corresponding binary indicator vector $\mathbf{g}$ or $\mathbf{t}$ depending on which library chemicals are present and on whether the hypothesis is the true one or the output from the system. When looking for only a single gas, the CM is only 2×2 as in Fig. 4(a) and a single threshold controls whether a pixel is assigned to the null-hypothesis (H_0) or the gas present hypothesis (H_1). When looking for one gas out of a library of size L, the CM is size $[L + 1] \times [L + 1]$ with the possibility of both correct identifications and incorrect identifications, as in Fig. 4(b).

Incorrect identifications occur when one chemical is mistaken for another or when there is no overlap between the chemicals in the truth and output. In Fig. 4(b), the hypotheses H_1 and H_2 represent each chemical from a library of size two; the corresponding binary vectors are $\mathbf{g}, \mathbf{t} = [1 \quad 0]$ or $[0 \quad 1]$. Hypothesis H_3 represents the presence of both chemicals in the plume. When looking for up to m of L, the CM is of size $\sum_{k=0}^{m} \binom{L}{k}$, but, in general, when any mixture of chemicals is allowed, the CM is size 2^L. In Fig. 4(c), the full CM for a library of size two is shown; the hypothesis H_3 represents the mixture of both chemicals; when there is some overlap between the chemicals that are detected and the chemicals actually present, we have a partially correct identification. In general, since the CM and the number of hypotheses to test grow exponentially with the size of the library, it is impractical to fill out the full CM or to test all possible models. For even moderately sized threat libraries, the CM becomes difficult to interpret because of its size necessitating summarization.

The CM can be summarized by partitioning, weighting, and then averaging. For plume identification applications, the CM can be partitioned into several submatrices that contain false alarms, misses, correct IDs, and incorrect IDs as shown in Fig. 4. Broadly, the false alarm section is the column where no gases are present ($\mathbf{t}_j = 0$) and the other cases occur when at least one gas is present. Performance metrics can be calculated using portions of the CM as follows: choose a partition of the CM; sum the elements of the partition; apply a weight matrix $\mathbf{W}$ with weights $w_{i,j}$ to the CM; sum the weighted elements of the partition; and take the ratio

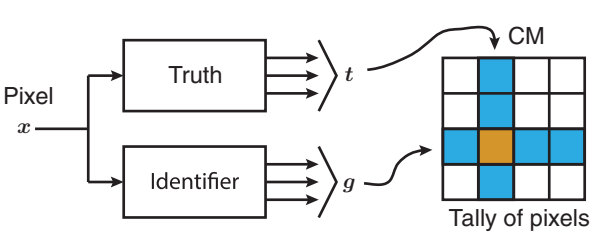

Fig. 3 CM construction.

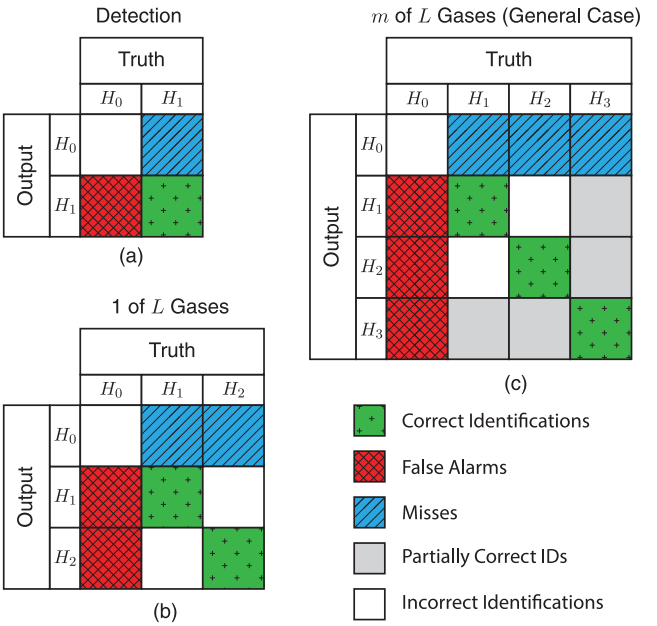

Fig. 4 Different confusion matrices depending on the particular problem. (a) Detection type problem. (b) 1 of L gases with library of size 2. (c) Up to two gases with library of size 2. H_0 is the null-hypothesis; H_1 and H_2 contain gas 1 and 2, respectively, while H_3 has both.

of the weighted and unweighted sums. These operations can be written succinctly as

$$d_{\text{perf}} = \frac{\sum_{(i,j)\in\mathcal{S}} [\text{CM} \odot \mathbf{W}]_{i,j}}{\sum_{(i,j)\in\mathcal{S}} [\text{CM}]_{i,j}}, \qquad 0 \le d_{\text{perf}} \le 1, \tag{9}$$

where $\odot$ denotes elementwise multiplication, and the set S represents the submatrix to sum over. The brackets [.] with subscripts indicate a single element of the matrix within the brackets.

Although there are a substantial number of different metrics that can be derived from the CM depending on the weights used, the metric we use for identification performance comes from the family of indices defined by

$$w_{i,j} = \frac{(\mathbf{g}_i^T \mathbf{t}_j)}{\beta|\mathbf{g}_i| + (1-\beta)|\mathbf{t}_j|}, \tag{10}$$

with $\beta \in [0,1]$. The numerator in Eq. (10) indicates the number of agreements between the output and truth, while the denominator has the sum of the number of chemicals in the truth and the number in the output. The resulting weights incorporate the truth and output vectors to varying degrees depending on the value of β used. Setting $\beta = 1/2$ in Eq. (10) we obtain the Dice index

$$w_{i,j} = \frac{2(\mathbf{g}_i^T \mathbf{t}_j)}{|\mathbf{g}_i| + |\mathbf{t}_j|} \tag{11}$$

or F-metric, not to be confused with the F-test from linear regression.[8,30] We choose the Dice index for identification because it incorporates both the number of chemicals in the identifier's output and the number of chemicals actually in the pixel. Other choices of β lead to metrics that weigh the

Table 1 Weights derived from Eq. (10) using several values of β.

β	$w_{i,j}$	Name(s)	Description
0	$\frac{(\mathbf{g}_i^T\mathbf{t}_j)}{\lvert\mathbf{t}_j\rvert}$	Sensitivity, recall	Fraction of chemicals in pixel that are detected
1/2	$\frac{2(\mathbf{g}_i^T\mathbf{t}_j)}{\lvert\mathbf{g}_i\rvert+\lvert\mathbf{t}_j\rvert}$	Dice index, F-metric	Incorporates both number of chemicals identified and number actually present Harmonic mean of precision and recall
1	$\frac{(\mathbf{g}_i^T\mathbf{t}_j)}{\lvert\mathbf{g}_i\rvert}$	Precision	Fraction of chemicals detected that are present in the pixel

importance of $\mathbf{g}$ and $\mathbf{t}$ in different proportions. For example, when $\beta = 0$, the number of incorrect outputs is not taken into account. Several common weights derived from Eq. (10) are presented in Table 1.[31]

The three metrics we use are the FAR, the CDR, and identification performance as measured using the F-metric of Eq. (11). Table 2 lists these metrics along with the partitions of the CM used in the weighting and summarization processes. The FAR is very important in both detection and identification systems since it determines how much background data will have to undergo additional scrutiny. In standoff systems, since the vast majority of data does not contain any plume, a low FAR with a high CDR and identification performance are desirable system characteristics.

4 Plume Identification Techniques

The two algorithms we compare are a detector bank approach and a model averaging algorithm. The detector bank has a set of single-chemical detectors, one for each library signature. Each detector produces a score for a single chemical, which is then thresholded to make a decision about the presence or absence of that chemical. Similarly, model averaging produces a score for each library chemical which is then thresholded to make a decision. There are two main reasons we evaluate only these two algorithms: both produce a score for each chemical which is easy to interpret; there are only a few parameters that need to be picked. In this section, we give overviews of the ACE detector bank and BMA and provide the relevant equations for each.

4.1 *Detector Bank for Identification*

Perhaps the most well-known detection algorithm used in hyperspectral imagery is the MF.[6] The MF for the k'th signature is defined as

Table 2 Detection and identification performance weights.

Statistic	Weights used	Portion of CM used
False alarm rate	$w_{i,j} = \begin{cases} 1, & \lvert\mathbf{g}_i\rvert > 0 \\ 0, & \lvert\mathbf{g}_i\rvert = 0 \end{cases}$	$j = 0$ (Gas absent)
Correct detection rate	$w_{i,j} = \begin{cases} 1, & \lvert\mathbf{g}_i^T\mathbf{t}_j\rvert > 0 \\ 0, & \lvert\mathbf{g}_i^T\mathbf{t}_j\rvert = 0 \end{cases}$	$j > 0$ (Gas present)
Identification performance	$w_{i,j} = \frac{2(\mathbf{g}_i^T\mathbf{t}_j)}{\lvert\mathbf{g}_i\rvert+\lvert\mathbf{t}_j\rvert}$	$j > 0$ (Gas present)

$$y_{\mathrm{MF}} = \frac{\tilde{\mathbf{x}}^T\tilde{\mathbf{s}}_k}{\|\tilde{\mathbf{s}}_k\|}, \tag{12}$$

where $\tilde{\mathbf{s}}_k$ and $\tilde{\mathbf{x}}$ are the whitened signature and whitened measurement of Eq. (7), respectively. In practice, the MF has a higher FAR when compared to similar algorithms which are often used instead. The normalized matched filter (NMF) is a simple modification where the MF is normalized by the measurement length $\|\tilde{\mathbf{x}}\|$. The MF and NMF can take both positive and negative values depending on the orientation of the measurement relative to the signature. In the LWIR, the relative direction of the measurement and signature depends on the thermal contrast Δ_{T} from Eq. (5), which may be positive or negative. To create a sign-insensitive detector, the NMF can be squared to obtain the ACE. We define ACE for the k'th library signature as

$$y_k = \frac{(\tilde{\mathbf{x}}^T\tilde{\mathbf{s}}_k)^2}{\|\tilde{\mathbf{x}}\|^2\|\tilde{\mathbf{s}}_k\|^2} = \cos^2(\tilde{\theta}_k), \tag{13}$$

where $\tilde{\theta}_k$ is the angle between the whitened signature $\tilde{\mathbf{s}}_k$ and the whitened pixel $\tilde{\mathbf{x}}$. We use the term ACE for the squared detector, although various terms are used in the literature.[32] Each of the ACE detectors in the bank is tuned to a particular library signature and is based on the linear model of Eq. (8), which is given as

$$\tilde{\mathbf{x}} = \tilde{\mathbf{s}}_k\tilde{\mathbf{b}}_k + \tilde{\mathbf{v}}, \tag{14}$$

where $\tilde{\mathbf{s}}_k$ is the whitened signature and $\tilde{\mathbf{v}}$ is the background clutter and noise. Each linear model considers a single chemical and excludes mixtures. However, thresholding the ACE scores may result in a mixture.

To use ACE as an identifier, a bank of detectors can be constructed, where each detector is tuned to a particular library signature. For detection of a specific chemical, only one ACE detector is needed; for detecting one of L gases, a bank of L detectors can be used and the maximum taken; for the m of L problem, instead of taking the maximum, the outputs can be thresholded to obtain a list of gases.

When only the maximum of the detector bank is considered, mixtures are excluded from consideration, which can be problematic when mixtures are present in the data. Picking the maximum may be appropriate in applications where only a single target is allowed in any pixel; e.g., in the reflective regions of the electromagnetic spectrum, the ground resolution can be small enough that only a single target can be in any particular pixel.[33] We use the thresholding

approach instead of taking a maximum because of the possible presence of mixtures.

4.2 *Bayesian Model Averaging*

BMA is a technique for estimating parameters using the construction of a set of models that are fitted to the data. In our case, each model is a linear model for $\tilde{\mathbf{x}}$ using a unique subset of chemicals from the library $\mathbf{S}$. Specifically, model j is in the form of Eq. (8) but with a particular library subset in $\mathbf{S}_j$ and the estimate of $\mathrm{CL} \times \Delta_{\mathrm{T}}$ in $\mathbf{b}_j$. Each model M_j is defined as

$$M_j : \tilde{\mathbf{x}} = \tilde{\mathbf{S}}_j \tilde{\mathbf{b}}_j + \tilde{\mathbf{v}}, \tag{15}$$

where the index j refers to the model, and it is a separate index from other sections. The null model is denoted M_0 and consists of clutter-only. Clearly, the models defined for the ACE detector bank in Eq. (14) are a subset of the models defined here for BMA. Defining A_k as the event that gas k is present, BMA computes the probability of the event A_k as the average over all models

$$y_k = \Pr\{A_k|\tilde{\mathbf{x}}\} = \sum_j \Pr\{A_k|M_j, \tilde{\mathbf{x}}\} \Pr\{M_j|\tilde{\mathbf{x}}\}, \tag{16}$$

where M_j is the j'th model being considered. The quantity in Eq. (16) is then thresholded to make a decision about each chemical in the library.

The probability $\Pr\{A_k|M_j, \tilde{\mathbf{x}}\}$ is an indicator of whether or not gas k is in the model. The model probabilities can be calculated using Bayes' rule

$$\Pr\{M_j|\tilde{\mathbf{x}}\} = \frac{\Pr\{\tilde{\mathbf{x}}|M_j\} \Pr\{M_j\}}{\sum_i \Pr\{\tilde{\mathbf{x}}|M_i\} \Pr\{M_i\}}, \tag{17}$$

where $\Pr\{\tilde{\mathbf{x}}|M_j\}$ is the likelihood of the data given in the model, and $\Pr\{M_i\}$ are the prior probabilities of the models. Typically, the likelihood of the data depends on model parameters that make it difficult to find expressions for the likelihood. Instead, the likelihood can be approximated using the model's Bayesian information criterion (BIC) as

$$\Pr\{\tilde{\mathbf{x}}|M_j\} \approx \exp\{-\mathrm{BIC}_j/2\}. \tag{18}$$

For linear regression models, as in Eq. (15), the BIC is

$$\mathrm{BIC}_j = n \ln(\mathrm{RSS}_j/n) + d_j \ln(n), \tag{19}$$

where n is the number of spectral bands, d_j is the number of gases in the model, and j is essentially a model complexity penalty. The residual sum of squares (RSS) is defined as

$$\mathrm{RSS}_j = \tilde{\mathbf{x}}^T (\mathbf{I} - \mathbf{P}_{\tilde{S}_j}) \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^T \mathbf{P}^{\perp}_{\tilde{\mathbf{S}}_j} \tilde{\mathbf{x}}, \tag{20}$$

where $\mathbf{P}_{\tilde{\mathbf{S}}_j} = \tilde{\mathbf{S}}_j (\tilde{\mathbf{S}}_j^T \tilde{\mathbf{S}}_j)^{-1} \tilde{\mathbf{S}}_j^T$.

The standard approach to BMA is to set all models equally likely, but considering all models of the form of Eq. (15) is prohibitively costly.[34,35] One way to keep the number of models manageable is to limit the maximum number of chemicals in a model. This effectively sets the prior probabilities of large models to zero. Excluding the empty model we denote the number of remaining models with nonzero prior probability by N_m, where m is the maximum number of chemicals in a model. Assuming all models are equally likely, Eq. (17) becomes

$$\Pr\{M_j|\tilde{\mathbf{x}}\} \approx \frac{\exp\{-\mathrm{BIC}_j/2\}}{\sum_{i=0}^{N_m} \exp\{-\mathrm{BIC}_i/2\}}, \tag{21}$$

which is the standard equation for BMA found in the chemical identification literature.[12] Equation (21) is similar in form to the softmax equation for normalizing classifier outputs.[36–38]

The main assumption in Eq. (21) is that the empty or null model has the same probability as the non-null models. However, in standoff applications, and the data we examine, the number of background pixels is orders of magnitude larger than the number of plume pixels. Therefore, it is reasonable to choose a nonuniform prior that takes into account the relative frequency of encountering plume. The prior probability of the null model can be made more likely than the other models as

$$\Pr\{M_0\} = Q \Pr\{M_i\}, i \neq 0, \tag{22}$$

where Q is the tunable parameter. Assuming all the other models are equally likely, we have

$$\Pr\{M_0|\tilde{\mathbf{x}}\} = \frac{\Pr\{\tilde{\mathbf{x}}|M_0\} Q}{\sum_{i=1}^{N_m} \Pr\{\tilde{\mathbf{x}}|M_i\} + \Pr\{\tilde{\mathbf{x}}|M_0\} Q}, \tag{23}$$

$$\Pr\{M_j|\tilde{\mathbf{x}}\} = \frac{\Pr\{\tilde{\mathbf{x}}|M_j\}}{\sum_{i=1}^{N_m} \Pr\{\tilde{\mathbf{x}}|M_i\} + \Pr\{\tilde{\mathbf{x}}|M_0\} Q} \quad (j \neq 0), \tag{24}$$

where N_M is the number of non-null models under consideration and Q adjusts the importance of the null model relative to the others. We select this type of prior because of its simplicity and we only have prior knowledge about the null model. Setting $Q = 1$, we obtain the formulation in Eq. (21). Equations (24) and (16) together define the probabilities of each model and each chemical occurring.

4.3 *Algorithm Complexity Comparison*

There are a number of different algorithm categories and many different flavors of algorithm within each category. A major problem in designing practical identification algorithms is that it becomes impractical to construct every possible model. Most approaches limit the number of models tested in some way. Algorithms can be broadly categorized as detection, hybrid, or exhaustive depending on how models excluded.

As in Sec. 1, detectors only consider single-chemical models, making them the simplest algorithms. Exhaustive techniques with all possible models are usually limited to a certain maximum size. Our implementation of BMA is exhaustive in this sense. Hybrid algorithms limit the number of models further than the exhaustive approaches usually through a testing procedure. For example, stepwise regression techniques iteratively search the model space by adding or removing signatures from an initial model.[39] Similarly,

Table 3 Identification algorithm complexity overview.

Algorithm type	Models	Common algorithms	Computational demand
Detector	Single chemical	ACE, MF	Least
Hybrid	Some models	Stepwise regression	Moderate
Exhaustive	All models	BMA, best model selection	Highest

in BMA several different approaches have been proposed for reducing the number of models actually tested.[34,40–42] Table 3 presents a summary of common identification algorithms and their computational demands.

In the naive approach we take for BMA, each model that is constructed is tested for each pixel under consideration. Therefore, the computational demand of each algorithm is determined by the total number of models under consideration. The large computational demand of BMA makes it desirable to either limit the number of models being tested or limit the number of pixels being tested. To compare the computational demand of different algorithms, comparing run times is one possible approach. However, run times are very implementation and platform dependent. Instead, computational complexity is measured in terms of the size of the input.[43] The number of non-null models up to size m out of a library of L chemicals is

$$N_m = \sum_{i=1}^{m} L_k = \sum_{k=1}^{m} \frac{L!}{(L-k)!k!} = \sum_{k=1}^{m} \frac{L(L-1)\ldots(L-k+1)}{k!}. \tag{25}$$

Evaluating Eq. (25) for several values of m we have

$$N_1 = L = O(L), \tag{26}$$

$$N_2 = L + \frac{L(L-1)}{2} = O(L^2), \tag{27}$$

$$N_3 = L + \frac{L(L-1)}{2} + \frac{L(L-1)(L-2)}{6} = O(L^3), \tag{28}$$

where $O(.)$ denotes the asymptotic upper bound. A detection bank considers N_1 models, while a BMA limited to models of size 3 considers N_3 models. For a library of size 8, $N_3 = 96$ which indicates that about 10 times as many computations are required for BMA as compared to the detection bank.

Another way to limit the computational cost associated with using an exhaustive algorithm is to limit the number of pixels that are processed. The solution we propose uses a detection bank as a first pass and only runs BMA on pixels that exceed the detection threshold. Dual thresholding approaches have been used for real-time processing, but two detection banks were used instead of BMA as a second pass.[44]

5 Performance Evaluation of Identification Algorithms

5.1 *Experimental Setup*

To evaluate and compare different plume detection and identification systems, ground truth for the dataset is a necessity. In data with real plumes, the spatial extent of the plume, the constituent chemicals, the concentrations, and temperature of the plume are typically unknown. To obtain a performance estimate with known plume parameters, a synthetic plume embedding technique was used. The embedding technique is based on Eq. (3) and requires a background-only cube and a signature library.[45,46]

For the embedding procedure, a background-only cube was selected from a set of cubes taken by a side-looking hyperspectral LWIR sensor.[47–49] The data had 128 channels with centers ranging from 7.6 to 13.5 μm; the plume was embedded over a ground portion of the image where the embedding model is most appropriate. Embedding over the sky portion of the image may be an avenue of future work, but it is challenging due to changes in atmospheric depth over the image and unknown atmospheric composition. However, the ground portion of the image can be considered fairly uniform in temperature and atmospheric depth, and the background materials are well modeled as blackbody radiators.

The signature library was created using a subset of eight chemicals taken from the PNNL LWIR spectral library and a secondary set of signatures.[50–53] The library consisted of TMP, BBR3, TFAA, F116, TEP, DIPF, Styrene, and TEPTO, which were selected because of their spectral characteristics in the region of study. We consider TEP alone and then a mixture of TEP and TMP.

The signatures were normalized to their maxima prior to embedding, since it is the product of the CL and the absorbance, also known as the OD, that appears in the exponent of Beer's law of Eq. (4). The reported values are the maximum OD, or $\max(\alpha s_i)$, since this value determines the maximum absorption of the plume.

Based on an estimate of the background temperature, the plume was simulated to be about 10 K colder than the background. Keeping the temperature contrast and embedding region constant, we varied the OD of the plume and the plume's constituent chemicals. Thus, the study we conducted is not exhaustive, but can be used as a starting point for future investigations.

To compare the ACE and BMA algorithms fairly, the embedding region was excluded from background mean and covariance estimates, which are substituted into Eq. (7). The inclusion of the plume in these estimates can lead to substantial performance degradation. Since the background estimates remained constant for each set of embedding parameters, the FARs reported do not depend on these parameters. Both techniques processed the entire cube and produced scores for each pixel and each library chemical. BMA considered mixtures of up to three chemicals. For BMA, the parameter Q of Eq. (22) was varied also.

In the following sections, we present results for several experiments using the same embedding region in each experiment, but vary the parameters of the embedded plume. In Secs. 5.2 and 5.3, we consider a single chemical at a fixed OD of 0.027 and examine the distribution of ACE

and BMA scores with respect to a given threshold. In Secs. 5.4 and 5.5, we examine system performance over a range of ODs and embed both a single chemical and a mixture of two chemicals in the same plume region. In Sec. 5.5, we propose a cascaded system and examine the performances for both the single chemical and two chemical embeddings over the same CL ranges as in the other sections.

5.2 *Detection Performance*

Our analysis presents histograms that are intended to complement Fig. 4 but were constructed as follows. For a background pixel, if the maximum score among all outputs exceeds a specified threshold, the result is a false alarm. For a plume pixel, there is one chemical that is the correct chemical embedded, the other chemicals are incorrect. If the maximum score among incorrect chemicals is above the threshold, then an incorrect identification was made. The histograms presented used only the maximum scores described here. The goal is to illustrate how the threshold determines the performance of the system, and to highlight the difference between good detection performance and good identification performance.

For perfect detection, there should be perfect separation between the background and plume scores. As shown in Fig. 5(a), ACE separates the background (red) and plume (green) quite well. In Fig. 5(a), the red histogram is composed of the maximum scores for each background pixel since only the maximum score needs to exceed the threshold for the pixel to be a false alarm. The background pixel scores are distributed close to 0, with very few exceeding a value of 0.1. The green histogram contains the scores for the correct gas only, and consists only of the pixels within the embedded plume region; the plume pixel scores are distributed away from 0 with very little overlap between the background and plume scores.

Using the formulation of BMA given in Eq. (17) ($Q = 1$), BMA does not separate the background and plume as well as ACE. In Fig. 5(b), the background scores (red) of BMA are not tightly distributed near 0, and span the full range of thresholds. There is significantly more overlap between the background and plume scores (green) of BMA than for the detector bank. The overlap between the distributions means that for any threshold, the number of false alarms and missed detections will be higher for BMA than for the ACE bank, as shown in Fig. 6(a) in blue when $Q = 1$.

The curves in Fig. 6(a) were constructed by sweeping a range of thresholds to produce a series of confusion matrices from which the CDR and FAR were computed using the partitions and weights of Table 2. Varying the prior probability of the null model through the parameter Q of Eq. (22) is one way to control the detection performance of BMA. In the naive implementation of BMA, Q was set to 1, which led to poor separation between the plume and background. In Fig. 6(a), the parameter Q was varied from 1 to 10^{10} and the resulting ROC curves plotted. Making the null model more likely by increasing Q improved the detection performance of BMA. For very large values of Q, the BMA and ACE ROC curves were nearly identical. For smaller values of Q, the detection performance of BMA was worse than that of ACE. For reference, Fig. 6(b) shows the FAR versus threshold for both ACE and BMA.

5.3 *Identification Performance*

Detection performance was measured using the background scores for each chemical and the plume scores for the correct chemical only. For identification, the scores for the incorrect chemicals determine whether we made a correct identification or not. Since a single threshold is applied to each output of the identifier, scores for chemicals that are not actually present may exceed the specified threshold. Multiple thresholds could be used, but selecting a threshold for each chemical individually is not practical without prior knowledge of which chemicals will be present. To have good identification performance, the distributions of scores for the correct chemical and the incorrect chemicals should be well separated.

In Fig. 7, the same histograms as in the previous section are presented, but the scores for the incorrect gases are also included (blue). For each pixel in the plume, the maximum score among the incorrect gases was used to construct the histogram; since a single threshold is used to determine which gases were present, if the maximum exceeded the threshold then an incorrect or partial identification occurred. If only the correct gas passes the threshold then a correct

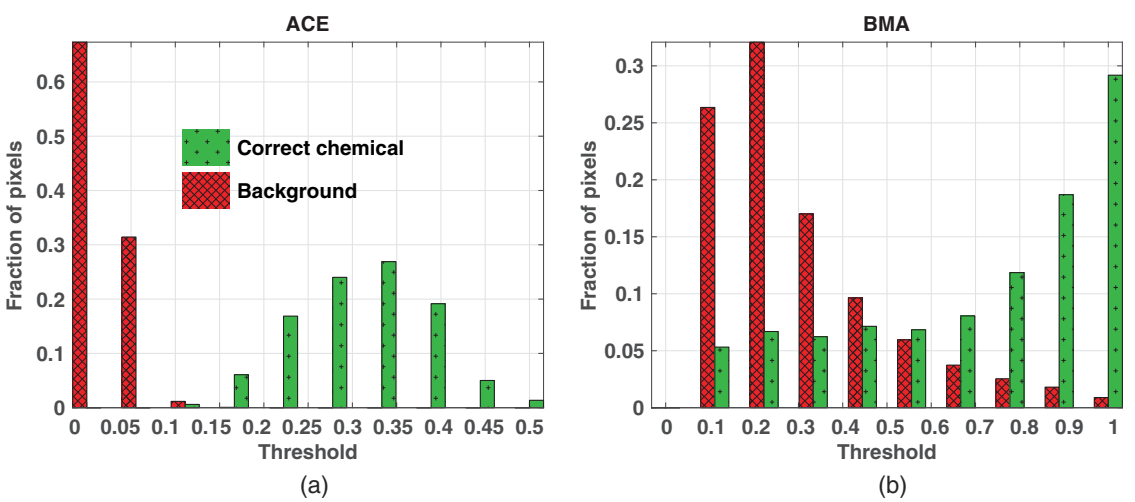

Fig. 5 Scaled histograms of representative outputs for (a) ACE and (b) BMA using embedded data. Background and plume pixels are easily separated using ACE, but less so by BMA with L set to 1.

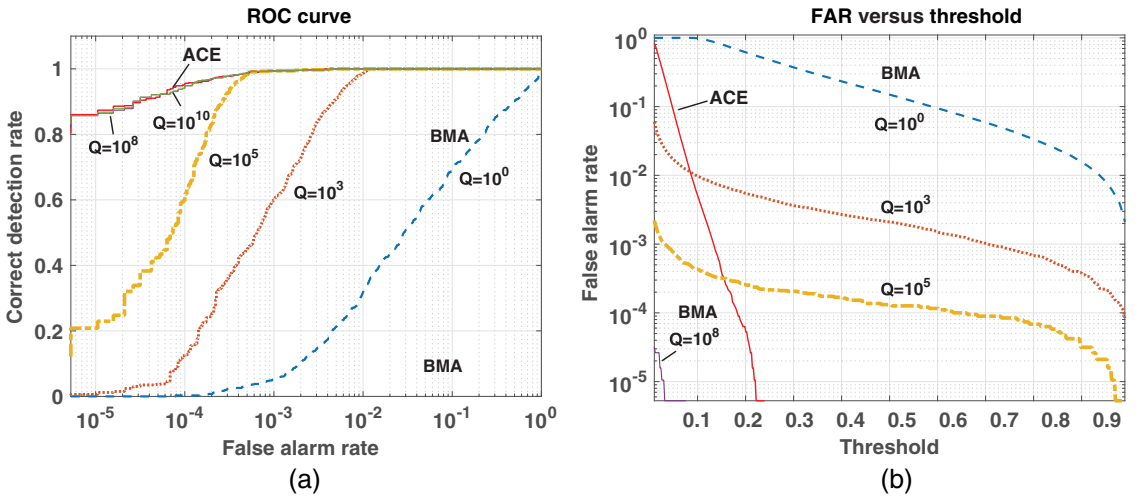

Fig. 6 (a) ROC curves for ACE and BMA based on the scores for the correct gas and for the background. (b) FARs of the two identifiers versus threshold.

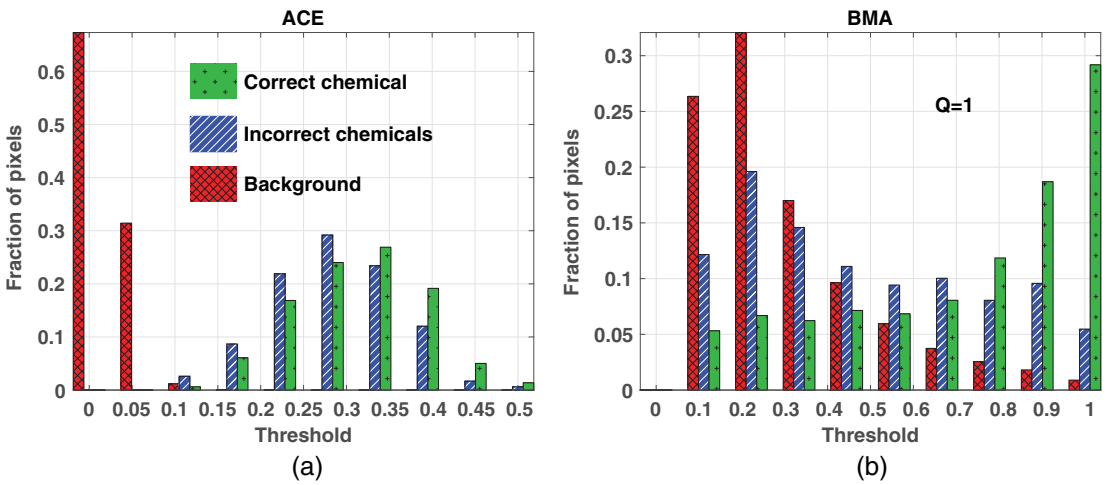

Fig. 7 Scaled histogram of representative outputs for (a) ACE and (b) BMA using embedded data. Incorrect chemicals are not easily separated from the correct chemical using ACE but are better separated by BMA.

identification occurred. In Fig. 7(a), the ACE detector bank shows good separation between the background and the plume for both the correct and incorrect chemicals; however, with the chosen embedding parameters, several chemicals have similar ACE scores because the embedded chemical's signature is similar to two of the other library signatures. In this case, the ACE bank does a poor job of identifying exactly which chemical is present when compared to BMA's results in Fig. 7(b). Based on the green and blue histograms of Fig. 7, poor separation of different chemicals should lead to poor identification performance when compared to BMA.

As in the previous section, increasing the parameter Q in BMA leads to a better detection performance. Figure 8(a) shows the resulting histogram when $Q = 1000$. The separation between the background and plume improves dramatically, while the blue and green histograms remain largely unchanged.

The identification performance curves shown in Fig. 8(b) were created by constructing a CM for a range of thresholds and then using the partitioning and weighting scheme described in Sec. 3 with the Dice weighting of Eq. (11). As shown in Fig. 8(b), the resulting performance curves demonstrate that ACE achieves a lower maximum than BMA for small choices of the parameter Q. Increasing Q leads to a better detection performance, but can degrade identification performance as in the purple $(Q = 10^8)$ and green curves $(Q = 10^{10})$ in Fig. 8(b). For $Q = 1$, BMA has a higher identification performance over a wide range of thresholds, achieving its maximum near a threshold of 0.5. Overall, the performance curves indicate that BMA selects the correct chemical more often than ACE for a wide range of thresholds.

The identification performance curve has an unusual characteristic compared to the CDR and FAR curves, which monotonically decrease with threshold. The identification performance curve is the average Dice score for the pixels within the plume at a number of thresholds. Performance initially increases with respect to threshold, reaches a maximum, and then decreases to zero at the maximum threshold. For small thresholds, several chemicals pass the threshold

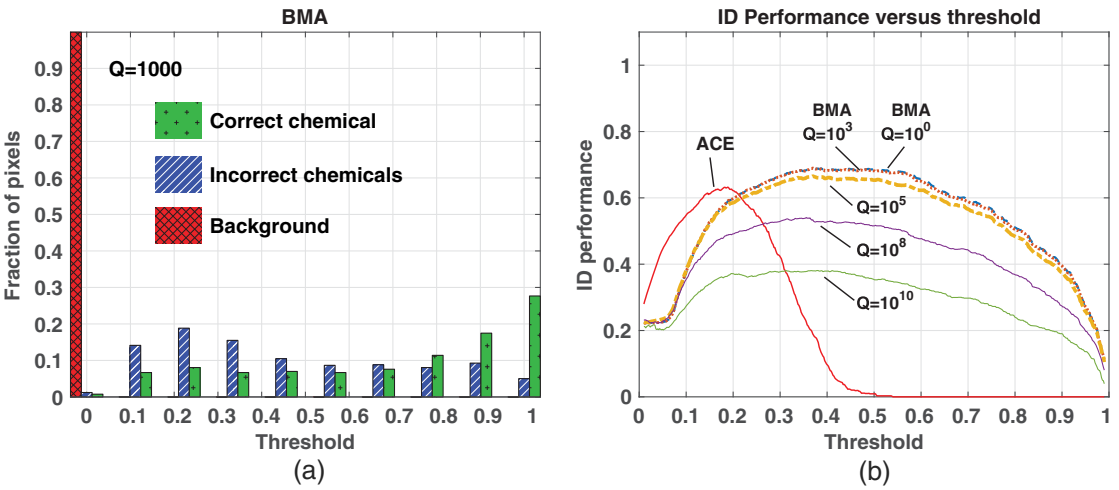

Fig. 8 (a) Effect of increasing null model probability on output histogram. (b) Identification performance for ACE and BMA over a range of thresholds.

for a large portion of the plume pixels. In the low threshold regions, many of the plume pixels are partial identifications, while for high thresholds there are more misses as in the confusion matrices of Fig. 4(c).

The number of chemicals in the output of each pixel determines its Dice weight. For a library of eight signatures with one chemical embedded and eight chemicals in the output, the weight is about 0.22, which is approximately the score BMA achieves at the lowest thresholds. For high thresholds, many plume pixels do not pass the threshold at all, leading to many pixels with a score of 0, causing the performance curve to deteriorate. Performance near one indicates that the majority of plume pixels has been correctly identified, i.e., all of the chemicals actually present are correctly identified and there are no extras.

Since the maximum identification performance of BMA is higher than ACE in this case, BMA can perform better as an identifier given an appropriate threshold and value for Q. Selecting a threshold is usually accomplished by assuming that each threshold produces a constant FAR (CFAR). With a single threshold, the selected FAR determines the identification performance, but this choice may not maximize identification performance. Instead, dual threshold approaches may be a good alternative.

5.4 *Effects of Plume Thickness on Performance*

The peak OD and temperature are the major drivers for how easily detected and identified the plume is. In the previous sections, a single OD was used for embedding; in this section, we examine the identification performance for a range of ODs. We focus on the effects of OD instead of temperature because the measured signal is approximately linear in terms of the temperature contrast, as in Eq. (5). The same background cube and embedding region as in Sec. 5.2 were used for this section. First, a single chemical was embedded for a range of ODs and both algorithms run on the data. For consistency, the same chemical as Sec. 5.3 was used. Second, a mixture of two chemicals was embedded and the experiment repeated. The second chemical used in the mixture was one that was spectrally similar to the first one. In both cases, the background statistics were the same for each OD. Consequently, at any particular threshold, the FAR at a particular threshold for each algorithm remained the same for all ODs.

Since we expected the identification performance to deteriorate for sufficiently thick plumes, embedding was done for two different ranges of ODs. The first range simulated thin plumes, while the second range was chosen to show identification performance reduction with thick plumes. Figures 9(a) and 11(a) have maximum OD of 0.1, which corresponds to $\max(\alpha s(\lambda)) = 0.1$ in Beer's law of Eq. (4) and the measurement equation of Eq. (3). This corresponds to a minimum transmission of $\min(\tau_p) = e^{-0.1}$ or about 90% in this region. The change in signal is approximately linear with respect to OD for this region; however, in Figs. 9(b) and 11(b), a larger range of ODs is shown. At an OD of 10, the plume is almost completely opaque $(\tau_p \approx 0)$ at its maximum absorption channel.

For the embedded plume with a single chemical, the identification performance for both BMA and ACE for a range of ODs is shown in Fig. 9. Performance is plotted with respect to OD for thresholds from 0.1 to 0.99 for BMA and from 0.1 to 0.9 for ACE. For most of the selected thresholds, BMA's performance increases with OD and is generally higher than ACE's performance for low ODs. After peak performance, BMA's performance slowly decreases, while ACE reaches a peak quickly and then decreases quickly. As the plume gets even thicker, as in Fig. 9(b), the performance of BMA drops off from the maximum, but ACE's performance has several peaks, depending on the threshold. The plume becomes too thick for BMA to distinguish individual gases and performance degrades substantially. However, with a very high ACE threshold, a decent performance for very thick plumes can be achieved. Using ACE, it is possible to set a high enough threshold to separate the correct chemical from the others when the plume is very thick. However, this sacrifices performance at small ODs, which is generally where standoff systems are expected to operate.

For both algorithms, the general trend is that as OD increases identification performance increases, then reaches a peak, and then begins to degrade before completely failing. The poor performance of both algorithms for very thick plumes indicates that multiple chemicals have similar scores and cannot be separated by a single threshold, or that

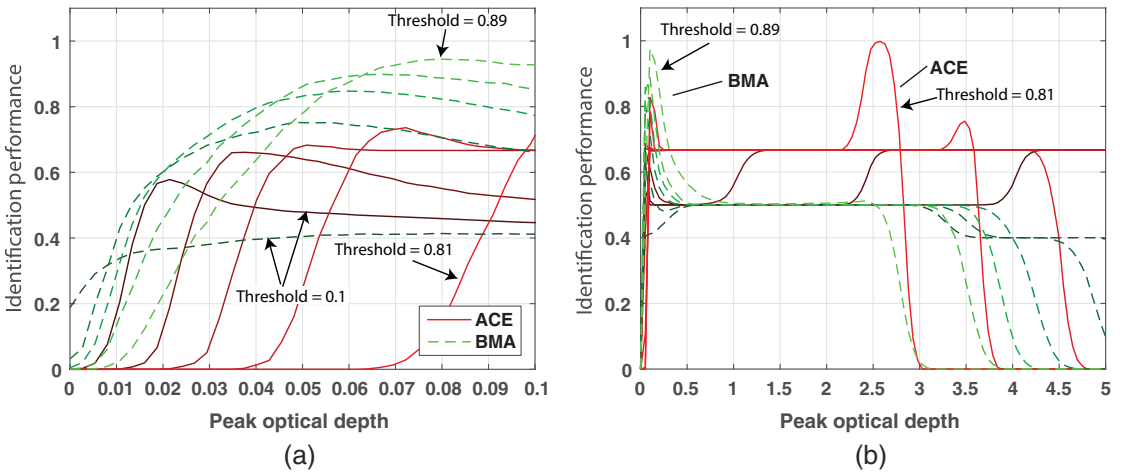

Fig. 9 Identification performance of ACE and BMA for a single-chemical embedding at various thresholds. (a) ODs close to zero. (b) Larger range of ODs. Thresholds were uniformly spaced between 0.1 and 0.81 for ACE and 0.1 and 0.89 for BMA.

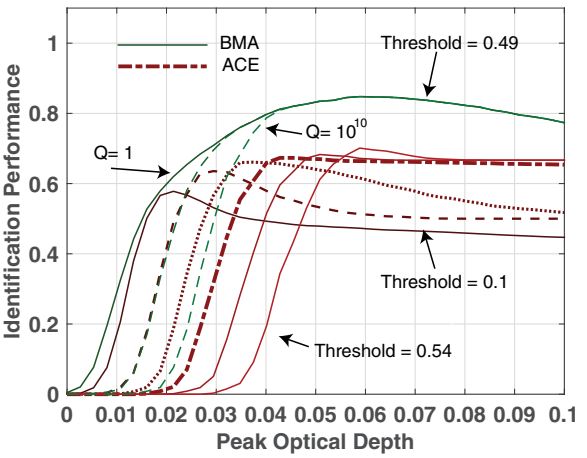

Fig. 10 Identification performance of ACE and BMA for a single-chemical embedding. The L parameter for BMA was set to 1, 10^3, and 10^{10}. The BMA threshold is fixed at 0.49 while ACE thresholds between 0.1 and 0.54 are shown.

incorrect gases are being identified. When the plume becomes very thick, both techniques fail in identifying the plume and should not be used.

The effect of the choice of Q in the BMA algorithm is demonstrated in Fig. 10. At low ODs, the factor has a larger effect on performance than at higher ODs. As the OD increases, alternatives to the null model fit better and the prior distribution becomes less important. At small ODs, there is a trade between the prior on the null model and the identification performance at small ODs.

To test performance for mixtures, two chemicals were embedded in the same location as in the previous experiment. The ODs of both chemicals were varied. In Fig. 11, the performance of both algorithms is shown for a range of ODs. The trends are similar to the previous experiment except that the performance of the detector bank is uniformly worse than BMA. Again, the problem ACE has is that multiple chemicals have similar scores. BMA gives both correct gases high scores because the model containing the mixture has a relatively high probability. The result is that BMA

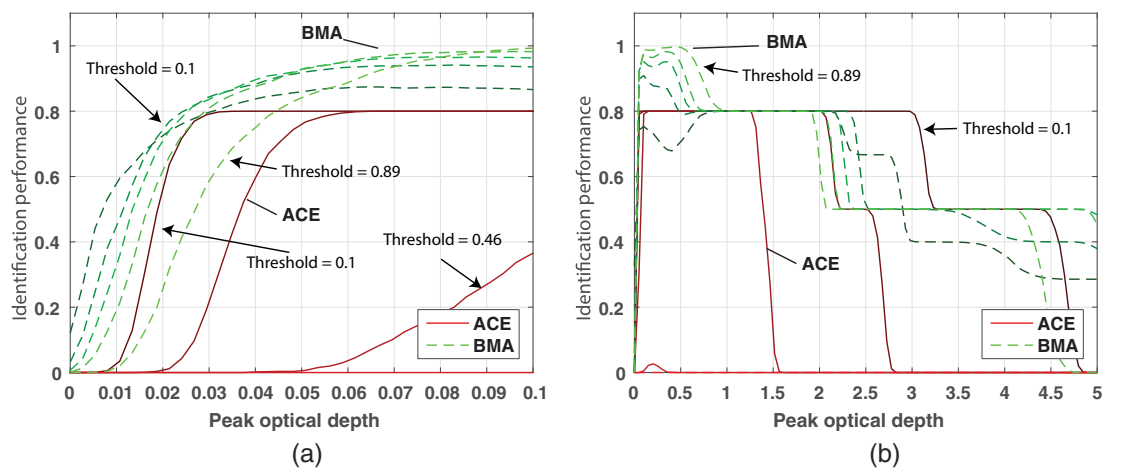

Fig. 11 Identification performance of ACE and BMA for a mixture of two chemicals at various ODs and thresholds. (a) Thin plume regime. (b) Larger range of ODs. Thresholds were uniformly spaced between 0.1 and 0.8 for ACE and 0.1 and 0.89 for BMA. For some ACE thresholds, no plume was detected.

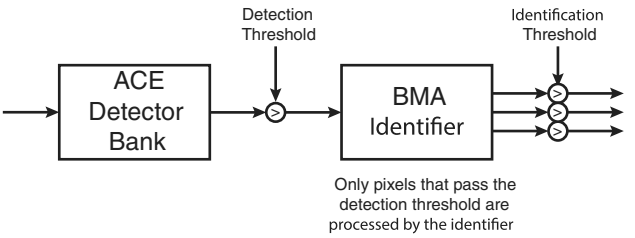

Fig. 12 The cascaded system design.

performs well for this mixture relative to ACE. In Fig. 11(b), a similar degradation in identification performance is seen as with the single gas embedding. However, as compared to the single-chemical embedding, BMA performs better for the mixture over a wider range of ODs than ACE.

5.5 *Detection Followed by Identification*

For the single-chemical embedding, the detection performance of the ACE detector bank was relatively high compared to BMA, but the identification performance of ACE was lower than BMA. We argue that combining the two algorithms in cascade leads to a system with desirable performance characteristics compared to either algorithm individually. The cascaded system uses the ACE detector bank as a first pass and then passes only the hits to the BMA identifier as shown in Fig. 12. Using the ACE bank as a first pass for the data can yield a high plume detection rate at a low FAR. Each pixel that passes the threshold is then passed to BMA for identification, which makes a final identification decision about those pixels. In this section, the cascaded system is evaluated using the same embedding scenarios as before. The cascaded system was run on the embedded data using two ACE thresholds and several BMA thresholds. For this system, the number of false alarms for a particular threshold is constant with respect to OD, and it is determined primarily by the ACE threshold. We selected two ACE thresholds: 0.1 and 0.36. The corresponding FARs for both ACE and the cascaded system are 3×10^{-3} and zero; the second threshold is high enough so that no background pixels pass the ACE threshold. The system was tested at evenly spaced BMA thresholds ranging from 0.11 to 0.89.

The resulting identification performance curves for the single-gas embedding are shown for the low ACE threshold in Fig. 13(a) and for the higher threshold in Fig. 13(b). The blue-dashed curve shows the performance of the ACE detector bank alone; the solid green curves show BMA's performance alone; the red-dotted curves are the cascaded system's performance.

The cost of cascading the two algorithms compared to ACE alone is a generally worse identification performance at low CLs, which is more pronounced in Fig. 13(b). However, at higher ODs, the cascaded system achieves better performance than the ACE bank for most choices of BMA threshold. Selecting the lowest BMA threshold of 0.1 actually results in worse performance than the ACE system alone for the single-gas embedding. Selecting the lower ACE threshold leads to a smaller difference between the green and red curves at smaller ODs but leads to a higher FAR. In practice, the maximum operational FAR will dictate what ACE threshold to select. However, it is unclear how best to set a BMA threshold in the cascaded system. From our experiments, the trends show that low thresholds lead to higher identification performance when the plume is very thin, but become comparatively worse as the plume thickens.

In Fig. 14, we compare the BMA algorithm with a large value for Q to the cascaded design. In the cascaded system, Q was set to 1. The choice of prior for the null model is a very important consideration for the BMA algorithm, especially at low ODs. At higher ODs, it is clear that using BMA for identification is superior to using ACE, but the cascaded system has equivalent performance in this region. One advantage of using the cascaded system is that the threshold for the detector bank can be used to select the FAR, while the BMA threshold can be used to select a good identification operating point.

It is tempting to set a threshold for the identifier that maximizes identification performance; however, the maximum depends on the properties of the plume, which are known for synthetic data, but are unknown in real data. Even with good plume models, it is impractical to try to find a good threshold over all possible simulated scenarios. Based on our results, BMA thresholds greater than 0.5 showed decent performance over a range of ODs.

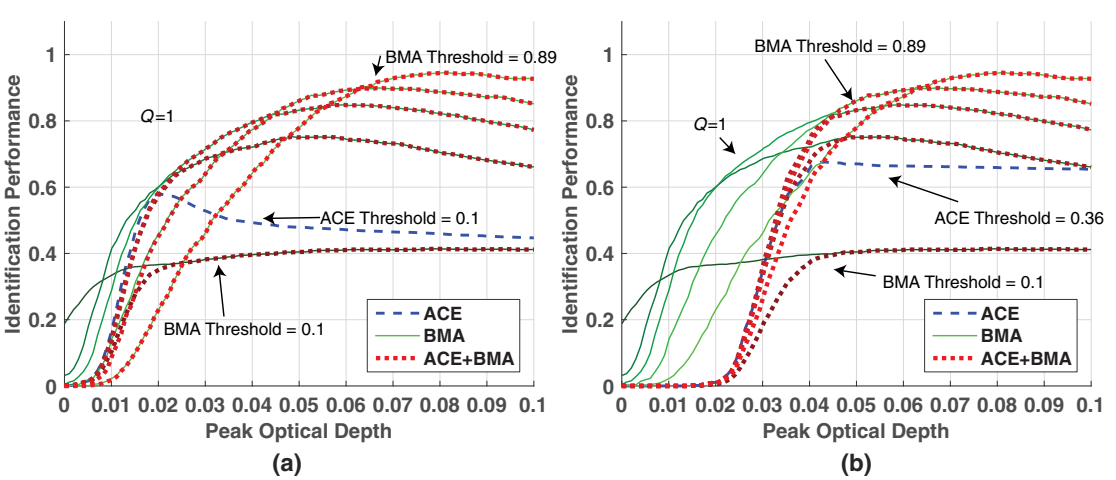

Fig. 13 Performance of the cascaded system for BMA thresholds from 0.11 to 0.89. Performance for the single-chemical data with ACE thresholds of (a) 0.1 and (b) 0.36.

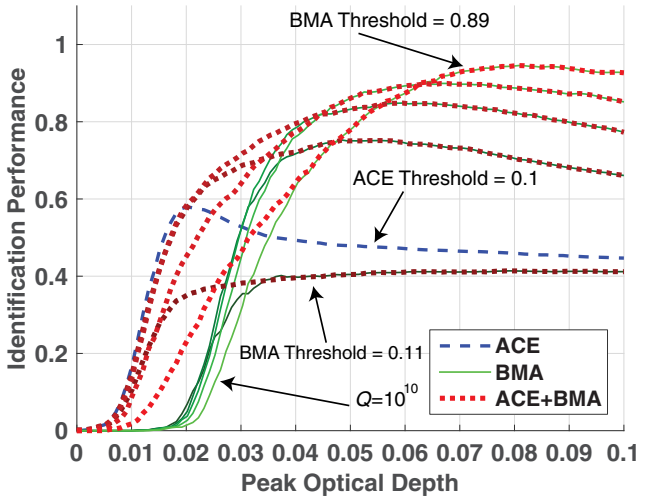

Fig. 14 Performance for the cascaded system for single-chemical data with an ACE threshold of 0.1, and BMA with L set to 10^{10}.

6 Conclusions and Future Work

The two main contributions of this work are the development of a performance evaluation framework for the evaluation of chemical plume detection and identification algorithms, and a demonstration that an ACE detector followed by a BMA identifier yields a better performance than the ACE alone, and can outperform the BMA algorithm, depending on the prior selected. The reduced computational burden of using ACE as a first step can be highly beneficial when processing power is limited. However, for good identification performance, it should be followed by the BMA identifier. The approach to performance evaluation using a weighted CM and performance evaluation using the F-metric are unique in this area of remote sensing. We applied the metric to quantitatively demonstrate that a cascaded detector and identifier can attain a high correct detection rate and low false alarm rate, while also having a good identification performance.

In the future, other types of algorithms should be evaluated and the number of datasets expanded. In addition, other types of priors for BMA may be examined. Our study here is not exhaustive but provides a framework place for further investigations. A study of all possible combinations of parameters is impractical, but smart experimental design can indicate what the greater trends are. In particular, a wider range of parameters and background clutter data will give insight into the problem of threshold selection for the cascaded system.

Acknowledgments

This work was sponsored by the Defense Threat Reduction Agency under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and not necessarily endorsed by the United States Government.

References

1. D. Manolakis, S. Golowich, and R. DiPietro, "Long-wave infrared hyperspectral remote sensing of chemical clouds: a focus on signal processing approaches," *IEEE Signal Process. Mag.* **31**(4), 120–141 (2014).
2. T. Burr and N. Hegarther, "Overview of physical models and statistical approaches for weak gaseous plume detection using passive infrared hyperspectral imagery," *Sensors* **6**, 1721–1750 (2006).
3. S. J. Young, "Detection and quantification of gases in industrial-stack plumes using thermal-infrared hyperspectral imaging," Aerospace Report ATR-2002 (8407)-1 (2002).
4. N. B. Gallagher, B. M. Wise, and D. M. Sheen, "Estimation of trace vapor concentration-pathlength in plumes for remote sensing applications from hyperspectral images," *Anal. Chim. Acta* **490**(1), 139–152 (2003).
5. B. Basener et al., "A detection-identification process with geometric target detection and subpixel spectral visualization," in *2011 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–4, IEEE (2011).
6. D. Manolakis et al., "Detection algorithms in hyperspectral imaging systems: an overview of practical algorithms," *IEEE Signal Process. Mag.* **31**(1), 24–33 (2014).
7. D. Edwards, C. Metz, and M. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imaging* **23**, 891–895 (2004).
8. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology* **26**(3), 297–302 (1945).
9. F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Ling.* **29**(1), 19–51 (2003).
10. X. Shen et al., "Multilabel machine learning and its application to semantic scene classification," *Proc. SPIE* **5307**, 188–199 (2003).
11. P. Tremblay et al., "Standoff gas identification and quantification from turbulent stack plumes with an imaging Fourier-transform spectrometer," *Proc. SPIE* **7673**, 76730H (2010).
12. T. Burr et al., "Chemical identification using Bayesian model selection," in *Proc. of 2002 Spring Research Conf. on Statistics in Industry and Technology* (2002).
13. T. Burr et al., "Performance of variable selection methods in regression using variations of the Bayesian information criterion," *Commun. Stat. Simul. Comput.* **37**, 507–520 (2008).
14. F. Kruse et al., "The spectral image processing system (SIPS) interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.* **44**(2), 145–163 (1993).
15. D. Manolakis et al., "The remarkable success of adaptive cosine estimator in hyperspectral target detection," *Proc. SPIE* **8743**, 874302 (2013).
16. R. Harig, G. Matz, and P. Rusch, "Scanning infrared remote sensing system for identification, visualization, and quantification of airborne pollutants," *Proc. SPIE* **4574**, 83–94 (2002).
17. R. Harig and G. Matz, "Toxic cloud imaging by infrared spectrometry: a scanning FTIR system for identification and visualization," *Field Anal. Chem. Technol.* **5**(1–2), 75–90 (2001).
18. D. Pogorzala et al., "Gas plume species identification in airborne LWIR imagery using constrained stepwise regression analyses," *Proc. SPIE* **5806**, 194–205 (2005).
19. D. Pogorzala et al., "Gas plume species identification by regression analyses," *Proc. SPIE* **5425**, 583–591 (2004).
20. S. Higbee et al., "A Bayesian approach to identification of gaseous effluents in passive LWIR imagery," *Proc. SPIE* **7334**, 73341T (2009).
21. P. Heasler et al., "Nonlinear Bayesian algorithms for gas plume detection and estimation from hyper-spectral thermal image data," *Sensors* **7**(6), 905 (2007).
22. R. M. Goody and Y. L. Yung, *Atmospheric Radiation: Theoretical Basis*, Oxford University Press, (1995).
23. D. Manolakis, "Realistic matched filter performance prediction for hyperspectral target detection," *Opt. Eng.* **44**, 116401 (2005).
24. S. Nowak et al., "Performance measures for multilabel evaluation: a case study in the area of image classification," in *Proc. of the Int. Conf. on Multimedia Information Retrieval*, pp. 35–44, ACM (2010).
25. M. R. Boutell et al., "Learning multi-label scene classification," *Pattern Recognit.* **37**(9), 1757–1771 (2004).
26. G. Tsoumakas and I. Katakis, "Multi-label classification: an overview," *Int. J. Data Warehousing Min.* **3**, 1–13 (2007).
27. A. P. Streich and J. M. Buhmann, "Classification of multi-labeled data: a generative approach," in *Machine Learning and Knowledge Discovery in Databases*, pp. 390–405, Springer (2008).
28. G. Madjarov et al., "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognit.* **45**, 3084–3104 (2012).
29. G. Tsoumakas, I. Katakis, and I. Vlahavas, "A review of multi-label classification methods," in *Proc. of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD 2006)*, pp. 99–109, CiteSeer (2006).
30. S. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Models Methods Appl. Sci.* **1**(4), 300–307 (2007).
31. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.* **45**(4), 427–437 (2009).
32. S. Kraut, L. L. Scharf, and L. T. McWhorter, "Adaptive subspace detectors," *IEEE Trans. Signal Process.* **49**, 1–16 (2001).
33. M. L. Pieper et al., "Hyperspectral detection and discrimination using the ACE algorithm," *Proc. SPIE* **8158**, 815807 (2011).
34. J. A. Hoeting et al., "Bayesian model averaging: a tutorial," *Stat. Sci.* **14**, 382–401 (1999).

35. A. E. Raftery, D. Madigan, and J. A. Hoeting, "Bayesian model averaging for linear regression models," *J. Am. Stat. Assoc.* **92**(437), 179–191 (1997).
36. L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Hoboken, New Jersey (2004).
37. E. Alpaydin, *Introduction to Machine Learning*, MIT Press, Cambridge, Massachusetts (2014).
38. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York (2001).
39. G. Seber and A. Lee, *Linear Regression Analysis*, Wiley Series in Probability and Statistics, 2nd ed., John Wiley & Sons, Hoboken, New Jersey (2003).
40. D. Madigan and A. E. Raftery, "Model selection and accounting for model uncertainty in graphical models using Occam's window," *J. Am. Stat. Assoc.* **89**(428), 1535–1546 (1994).
41. G. M. Furnival and R. W. Wilson, "Regressions by leaps and bounds," *Technometrics* **16**(4), 499–511 (1974).
42. R. E. Kass and L. Wasserman, "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *J. Am. Stat. Assoc.* **90**, 928–934 (1995).
43. T. H. Cormen et al., *Introduction to Algorithms*, 3rd ed., MIT Press , Cambridge, Massachusetts (2009).
44. A. Vallières et al., "Algorithms for chemical detection, identification and quantification for thermal hyperspectral imagers," *Proc. SPIE* **5995**, 59950G (2005).
45. M. Griffin et al., "A procedure for embedding effluent plumes into LWIR imagery," *Proc. SPIE* **5806**, 78–87 (2005).
46. S. Niu, S. E. Golowich, and D. G. Manolakis, "Algorithms for remote quantification of chemical plumes: a comparative study," *Proc. SPIE* **8390**, 83902I (2012).
47. J. L. Hall et al., "Characterization of aerosol-containing chemical simulant clouds using a sensitive, thermal infrared imaging spectrometer," *Proc. SPIE* **8018**, 801816 (2011).
48. J. A. Hackwell et al., "LWIR/MWIR imaging hyperspectral sensor for airborne and ground-based remote sensing," *Proc. SPIE* **2819**, 102–107 (1996).
49. T. Gerhart et al., "Detection and tracking of gas plumes in LWIR hyperspectral video sequence data," *Proc. SPIE* **8743**, 87430J (2013).
50. T. J. Johnson, R. L. Sams, and S. W. Sharpe, "The PNNL quantitative infrared database for gas-phase sensing: a spectral library for environmental, hazmat, and public safety standoff detection," *Proc. SPIE* **5269**, 159–167 (2004).
51. S. W. Sharpe et al., "Gas-phase databases for quantitative infrared spectroscopy," *Appl. Spectrosc.* **58**(12), 1452–1461 (2004).
52. T. S. Spisz et al., "Field test results of standoff chemical detection using the FIRST," *Proc. SPIE* **6554**, 655408 (2007).
53. K. C. Gross et al., "Instrument calibration and lineshape modeling for ultraspectral imagery measurements of industrial smokestack emissions," *Proc. SPIE* **7695**, 769516 (2010).

Eric Truslow is a technical staff member at MIT Lincoln Laboratory. His research interests include hyperspectral imagery, detection theory, machine learning, and signal processing. He received his BS degree from the Union College in 2010. He received his MS degree in 2012 and his PhD in 2015 from Northeastern University, both in electrical engineering. This work was performed as part of his PhD research.

Steven Golowich received his PhD in physics from Harvard University and his AB in physics and mathematics from Cornell University. He is a technical staff member at MIT Lincoln Laboratory, where his research interests include statistical signal processing, remote sensing, and optical fiber systems. Previously, he was a member of technical staff at Bell Labs and taught at Princeton University. He received the American Statistical Association Outstanding Application Award and the Wilcoxon Prize.

Dimitris Manolakis is a senior member of the technical staff at MIT Lincoln Laboratory. He is coauthor of the textbooks D*igital Signal Processing: Principles, Algorithms, and Applications* (Prentice-Hall, 2006, 4th ed.), *Statistical and Adaptive Signal Processing* (Artech House, 2005), and *Applied Digital Signal Processing* (Cambridge University Press, 2011). His research experience and interests include the areas of digital signal processing, adaptive filtering, array processing, pattern recognition, remote sensing, and radar systems.

Vinay Ingle is currently an associate professor in the Department of Electrical and Computer Engineering at Northeastern University. He has coauthored several textbooks on signal processing including digital signal processing using Matlab (Cengage, 2016, 4th ed.), applied digital signal processing (Cambridge University Press, 2011), and statistical and adaptive signal processing (Artech House, 2005). His research is in the areas of signal/image processing, stochastic processes estimation theory, and hyperspectral imaging applications.