A Novel Detection Paradigm and its Comparison to Statistical and Kernel-Based Anomaly Detection Algorithms for Hyperspectral Imagery

Colin C. Olson and Timothy Doster

U.S. Naval Research Laboratory

4555 Overlook Ave., SW, Washington, DC

`colin.olson@nrl.navy.mil`

Abstract

Detection of anomalous pixels within hyperspectral imagery is frequently used for purposes ranging from the location of invasive plant species to the detection of military targets. The task is unsupervised because no information about target or background spectra is known or assumed. Some of the most commonly used detection algorithms assume a statistical distribution for the background and rate spectral anomalousness based on measures of deviation from the statistical model; but such assumptions can be problematic because hyperspectral data rarely meet them. More recent algorithms have employed data-driven machine learning techniques in order to improve performance. Here we investigate a novel kernel-based method and show that it achieves top detection performance relative to seven other state-of-the-art methods on a commonly tested data set.

1. Introduction

Typical three-color cameras lack the spectral sensitivity required for accurate material detection and/or identification in remote sensing applications. In response, rather than coarsely dividing the incoming light into three bands, imagers that capture multispectral and hyperspectral imagery (MSI/HSI) seek to generate finer samplings of the spectrum such that useful identifying information about the underlying materials is not averaged out by the sampling process.

It is frequently the case that information pertaining to specific targets is unavailable or unknown. In addition, knowledge of background statistics is usually unknown for real-world applications. This leads to an unsupervised anomaly detection problem for which a variety of algorithms have been proposed in the remote sensing and MSI/HSI literature. The goal is to classify pixels as belonging to a background or anomalous class given a set, Ω, comprised of the N L-dimensional pixels, $\mathbf{x}_i \in \mathbb{R}^L$, $i =$ $1, \ldots, N$, that make up an image. Most of the pixels are assumed to represent background with just a very few anomalous pixels present as well.

The most commonly employed algorithm is the RX detector introduced by Reed and X. Yu [16] which assumes that each spectral channel is Gaussian-distributed and thus the pixels can be modeled as L-dimensional multi-variate Gaussian distributions. Statistics governing the RX model are calculated globally from the entire image but can also be considered locally about each pixel under test. A test pixel is considered anomalous if it exceeds a threshold corresponding to a certain number of standard deviations from the data mean. Details regarding calculation of global RX detection statistics are provided in Section 2 along with variants of the basic RX algorithm.

Additional detection algorithms exploiting the kernel trick and graph-theoretic techniques have also been proposed. Kwon and Nasrabadi [9] saw improved performance by introducing a kernelized version of the RX algorithm under the assumption that background and target would be described by Gaussian distributions in the high-dimensional feature space. Basener *et al* [2] introduced the Topological Anomaly Detector (TAD) which builds a graph from the data and measures the connectedness of graph components to describe the background then measures nearest neighbor distances to the background components as a measure of anomalousness .

Our "skeleton kPCA" method builds a better background model by combining kernel PCA (kPCA) [18] with a subsampling of the image and calculates a detection statisic by employing a measure of reconstruction error introduced by Hoffmann [8]. We compare the performance of this algorithm to seven other algorithms using Forest Radiance imagery from the HYDICE sensor (described in Section 4), which is a standard data set common throughout the hyperspectral literature. We begin with a description of the comparison algorithms in Section 2, detail our detection paradigm in Section 3, describe the parameter optimizations performed for each algorithm and subsequent experimental

results in Section 4, and conclude in Section 5.

2. A Summary of Anomaly Detection Algorithms

Here we provide a description of the tested anomaly detection algorithms. As described previously, $x \in \mathbb{R}^L$ is a test vector (or hyperspectral pixel) drawn from the image data set Ω.

2.1. RX

The Reed-Xiaoli [16] or RX anomaly detector (sometimes referred to as global RX) is based upon the Mahalanobis distance between x and the background:

$$D_{RX}(x) = (x - \mu)\Sigma^{-1}(x - \mu)^T, \quad (1)$$

where the mean, μ, and covariance, Σ, are calculated globally from all pixels in the image. Values of D_{RX} that fall below some user-defined threshold, γ, belong to background pixels while pixels that yield larger values are classified as anomalous.

2.2. SSRX

Subspace RX (SSRX) [17] performs global RX on a subset of principal components calculated from the data covariance matrix. Principal Component Analysis (PCA) [14] is applied to the hyperspectral data cube in order to find the directions of maximum variation. Background information is presumed to be represented by the eigenvectors corresponding to the largest principal components. Dropping the largest principal components and performing RX on the retained eigenvectors is assumed to improve anomaly detection because confounding information pertaining chiefly to background spectra has been removed from the analysis.

2.3. UTD

The Uniform Target Detector (UTD) [7] supposes that because we don't know anything about the anomalous data we should assume a uniform spectra as the matched signal such that the detector becomes

$$D_{UTD}(x) = (\mathbf{1}_L - \mu)\Sigma^{-1}(x - \mu)^T, \quad (2)$$

where $\mathbf{1}_L$ is a unity vector of length L.

2.4. UTD-RX

The UTD and RX detectors can be combined, as proposed in [7]:

$$\begin{aligned} D_{UTD-RX}(x) &= D_{UTD}(x) - D_{RX}(x) \\ &= (x - \mathbf{1}_L)\Sigma^{-1}(x - \mu)^T. \end{aligned} \quad (3)$$

2.5. OSPRX

Orthogonal Subspace Projection RX (OSPRX) [5], or RX after OSP, is a variant of the RX algorithm which seeks to better model the background of a scene. OSPRX conditions the data by projecting it onto the several of the leading singular vectors of the singular value decomposition (SVD) of the covariance matrix, $C = U\Lambda U^T$, where U is a matrix of eigenvectors of the decomposition and Λ is a diagonal matrix of decreasing eigenvalues. The eigenvectors associated with the largest eigenvalues are assumed to expresses the background. The detector can be formulated as:

$$D_{\text{OSPRX}}(x) = (x - \mu)^T(I_{b'} - MM^T)(x - \mu), \quad (4)$$

where $I_{b'}$ is the $b' \times b'$ identity matrix, $M = [U_1, \ldots, U_{b'}]$ is the truncated matrix of singular vectors, and $1 \leq b' \leq b$ is the number of retained singular vectors.

2.6. TAD

The Topological Anomaly Detection (TAD) algorithm [2, 3] does not assume the same Gaussian distribution as the previous detectors but instead measures the connectedness of graph components. The corresponding detection statistic is given by Algorithm 1.

Algorithm 1 TAD Algorithm

1: Perform optional normalization to the data.
2: For a sampling $\mathcal{S}$, construct a graph representation of the subsampled data by connecting the closest 10% of all edge pairs.
3: Define the background set, $\mathcal{B}$, as all pixels belonging to graph components containing at least 2% of the samples.
4: Calculate the anomalous rank of each pixel,

$$D_{TAD}(x) = \sum_{i=3}^{5} \delta_i(x, \mathcal{B}), \quad (5)$$

where $\delta_i(x, \mathcal{B})$ is the i-th smallest distance between points in $\mathcal{B}$ and x.

2.7. KRX

Kernel RX (KRX) [9] first transforms the pixels into a high-dimensional feature space, $\mathcal{F}$, using the kernel trick. Both the transformed background pixels and transformed anomalies are assumed to be Gaussian-distributed in the feature space but each with different means. Anomalousness is determined by calculating the Mahalanobis distance of each test pixel in the feature space. A spherical decision surface in the feature space corresponding to a selected

threshold, γ, yields a highly nonlinear decision surface in the original spectral (ambient) space. A nonlinear decision surface calculated in this manner is expected to yield improved performance because it better conforms to the actual data distribution.

Given the transformation $\mathbf{x}_i \rightarrow \mathbf{\Phi}(\mathbf{x_i})$ that maps a pixel into the feature space we can conceptually represent the detection statistic as:

$$D_{KRX}(x) = (\mathbf{\Phi}(\mathbf{x}) - \mu_k)^T \Sigma_k^{-1} (\mathbf{\Phi}(\mathbf{x}) - \mu_k)), \quad (6)$$

where μ_k and Σ_k represent, respectively, the mean and covariance of all the features. Given a kernel function such as:

$$k(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2 / \sigma^2), \quad (7)$$

that quantifies the similarity between pixels, the actual transformation $\mathbf{\Phi}(\cdot)$ need never be calculated and all calculations requiring inner products in the feature space can be determined using kernel (adjacency) matrices. Thus Equation 6 may be determined as a function of various adjacency matrices but we leave the details to reference [9].

3. Skeleton Kernel PCA

We provide background on and a description of kPCA and then detail our skeleton kPCA method.

3.1. Kernel PCA

kPCA was introduced by Schölkopf *et al.* [18] and adapted to the anomaly detection problem by Hoffmann [8]. The idea is to exploit the "kernel trick" and map data that are not linearly separable in the original (ambient) space into a high-dimensional (possibly infinite-dimensional) feature space, $\mathcal{F}$, via the transformation

$$\mathbf{x}_i \rightarrow \mathbf{\Phi}(\mathbf{x_i}), \quad (8)$$

whereupon linear decision surfaces can be constructed in $\mathcal{F}$ that yield highly nonlinear decision surfaces in the ambient space.

kPCA is a nonlinear version of PCA, based on calculating the principal components of the data after the nonlinear mapping has been applied. If a nonlinear PCA model of the training data has been learned, then it is possible to declare a test point anomalous by comparing it against the learned model. Hoffmann showed that constructing a "reconstruction error" between a transformed point in $\mathcal{F}$ and its reconstruction in $\mathcal{F}$ as synthesized from the learned model yields an anomaly detection statistic that outperforms PCA, one-class SVM, and the Parzen window density estimator [13] on a number of toy problems and real-world data sets [8].

In this case the learned model is a reduced linear model (as in standard PCA) but one that is applied in the high-dimensional feature space. High-dimensional feature vectors are synthesized as a linear combination of eigenvectors calculated from the high-dimensional covariance matrix. More specifically, what is required are the eigenvectors $\mathbf{\Psi} = (\psi_1, \psi_2, \ldots, \psi_N)$ and corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ of the high-dimensional data covariance matrix:

$$\tilde{\mathbf{C}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{\Phi}}(\mathbf{x}_i) \tilde{\mathbf{\Phi}}(\mathbf{x}_i)^T. \quad (9)$$

The tilde used throughout refers to mean-centered data:

$$\tilde{\mathbf{\Phi}}(\cdot) = \mathbf{\Phi}(\cdot) - \mathbf{\Phi_0}, \quad (10)$$

where,

$$\mathbf{\Phi_0} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{\Phi}(\mathbf{x_i}) \approx \mathbf{E}[\mathbf{\Phi}(\mathbf{x_i})], \quad (11)$$

is the centroid of the transformed data in the feature space and $E[\cdot]$ is the expected value. As with standard PCA, the data in the feature space are modeled as a linear combination of the eigenvectors, e.g.,

$$\mathbf{\Phi}(\mathbf{x}) = \mathbf{\Psi}_m \theta_m, \quad (12)$$

where $\mathbf{\Psi}_m$ denotes the first $m < N$ eigenvectors corresponding to the largest eigenvalues, and the associated model coefficients are $\theta_m \in \mathbb{R}^m$. Error between this reduced model and the full coordinate transformation $\theta = \mathbf{\Psi}^T \mathbf{\Phi}(\mathbf{x}) \in \mathbb{R}^N$ is captured by the Euclidean norm which can be written as:

$$D_K(\mathbf{x}') = \tilde{\mathbf{\Phi}}(\mathbf{x}') \cdot \tilde{\mathbf{\Phi}}(\mathbf{x}') - \mathbf{\Psi}_m^T \tilde{\mathbf{\Phi}}(\mathbf{x}') \cdot \mathbf{\Psi}_m^T \tilde{\mathbf{\Phi}}(\mathbf{x}'), \quad (13)$$

for test point $\mathbf{x}'$. Large errors indicate anomalous test points. The reason for writing the model error as in Eq. (13) is that inner products in the high-dimensional space can be easily computed without having to form the high-dimensional vectors themselves.

Hoffman showed that Eq. (13) could be written solely as a function of the $N \times N$ kernel (adjacency) matrix, $\mathbf{K} \equiv K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x_i}) \cdot \mathbf{\Phi}(\mathbf{x_j})$, associated with an underlying kernel function, $k(\cdot, \cdot)$, that quantifies the similarity between data points. In short, Eq. (13) can be written as:

$$D_K(\mathbf{x}') = D_S(\mathbf{x}') - \sum_{l=1}^{m} g_l(\mathbf{x}')^2, \quad (14)$$

where $m \leq N$ is the number of retained eigenvectors, $D_S(\mathbf{x}')$ is the spherical potential (discussed below), and where

$$g_l(\mathbf{x}') = \tilde{\mathbf{\Phi}}(\mathbf{x}') \cdot \boldsymbol{\psi}_l = \sum_{i=1}^{N} \alpha_{il} \left[k(\mathbf{x}', \mathbf{x}_i) - \frac{1}{N} \sum_{s=1}^{N} k(\mathbf{x}_i, \mathbf{x}_s) \right.$$

$$-\frac{1}{N} \sum_{s=1}^{N} k(\mathbf{x}', \mathbf{x}_s)$$

$$\left. +\frac{1}{N^2} \sum_{r,s=1}^{N} k(\mathbf{x}_r, \mathbf{x}_s) \right], \qquad (15)$$

which computes the projection of $\tilde{\mathbf{\Phi}}(\mathbf{x}')$ on $\boldsymbol{\psi}_l$, is itself a function of $k(\cdot, \cdot)$ and the eigenvectors, $\boldsymbol{\alpha}_l \in \mathbb{R}^N$, $l = 1, \ldots, N$, associated with the centered inner-product matrix given by:

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{N} \sum_{s=1}^{N} K_{is} - \frac{1}{N} \sum_{r=1}^{N} K_{rj} + \frac{1}{N^2} \sum_{r,s=1}^{N} K_{rs}. \tag{16}$$

The spherical potential of a test point $\mathbf{x}'$ is the squared distance of that point in $\mathcal{F}$ from the centroid of $\mathbf{\Phi}(\mathbf{\Omega})$:

$$D_S(\mathbf{x}') = \|\mathbf{\Phi}(\mathbf{x}') - \mathbf{\Phi_0}\|^2 = \tilde{\mathbf{\Phi}}(\mathbf{x}') \cdot \tilde{\mathbf{\Phi}}(\mathbf{x}'). \tag{17}$$

In terms of the kernel function, the spherical potential is given by:

$$D_S(\mathbf{x}') = k(\mathbf{x}', \mathbf{x}') - \frac{2}{N} \sum_{i=1}^{N} k(\mathbf{x}', \mathbf{x}_i) + \frac{1}{N^2} \sum_{i,j=1}^{N} k(\mathbf{x}_i, \mathbf{x}_j). \tag{18}$$

Given the Gaussian kernel,

$$k_G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right), \tag{19}$$

where σ is a bandwidth parameter, the first and last terms in (18) become constant and can be ignored. Thus, the potential is governed by the middle term, which is proportional to the Parzen window density estimator [8].

3.2. Skeleton kPCA Description

Hoffman used all available data to test anomaly detection performance in [8]. However, the expense of calculating an adjacency matrix is a primary limitation of kernel-based methods. A uniform sub-sampling of the scene (a data "skeleton") was proposed in [12] to reduce the computational cost of building an adjacency matrix with reliance on out-of-sample extension [4, 10] as a means of developing a detection statistic for the remaining unsampled points. In particular, the Nyström extension [1, 15] was used to extend the manifold to the unsampled points, diffusion map [6] was used to learn the background model, and a nearest-neighbor-based detection statistic was constructed in the manifold space. This same skeleton sub-sampling scheme was used in [11] to build a background

model but kPCA was used to learn the model and Hoffman's reconstruction error was used as the detection statistic. More specifically, the background model is learned from a uniformly sampled "skeleton" subset of the original data, $\mathcal{S} = \{\boldsymbol{x}_{s_i} : s_i \sim U(1, N), \ i = 1 \cdots N_\mathcal{S}\}$ where $N_\mathcal{S} < N$. The underlying assumption is that for the purpose of anomaly detection, the randomly chosen $N_\mathcal{S}$ points comprising a skeleton set, $\mathcal{S}$, are sufficient to build an accurate description of the background and are unlikely to contain a large fraction of anomalies. Thus a kernel matrix of size $N_\mathcal{S} \times N_\mathcal{S}$ comprises the bulk of computational cost.

Building an adjacency matrix from an extracted subset of the data is a simple concept, but it enables the realistic application of a wide variety of kernel-based and manifold-learning techniques to the anomaly detection problem. The trade-off relative to approaches based on the assumption of a statistical model is the introduction of a new set of considerations such as the fraction of the data set that must be sampled in order to guarantee sufficient background model fidelity and the robustness of background embeddings to alternate samplings of the scene.

4. Algorithm Comparison

We test the performance of our algorithm against the anomaly detection algorithms described in Section 2 on a hyperspectral image that is commonly tested in the literature

4.1. Data

The Hyperspectral Digital Imaging Collection Experiment (HYDICE) imaging spectrometer was used during te Forest Radiance I data collect to acquire signature data from 210 equally-spaced, contiguous spectral bands from a total bandwidth covering approximately 400-2500 nm. Some of the bands in the image are removed from the analysis due to low signal-to-noise ratios (SNRs) which leaves 158 total bands. The original image is 1280 x 308 but we only consider a 600 x 308 segment of the image in order to limit processing times. An approximate RGB image of *run05* is shown in Figure 1. We created a truth-mask (not shown) describing anomalous pixels corresponding to targets of interest in the scene.

4.2. Parameter Selection

RX, UTD, and UTD-RX, are parameter-free by design, but the other algorithms require tuning of free parameters to acquire upper bounds on their performance. To measure the change in algorithmic performance based on different parameters we utilize the area under the curve (AUC) for our hyperspectral scene and set of anomalies. AUC numerically integrates the receiver operator characteristic (ROC) curve which measures the probability of detection (P_d) for each probability of false alarm (P_{fa}). A distinct (P_d)-(P_{fa}) pair

Figure 1: Approximate RGB image of *run05* from HY-DICE Forest Radiance I collect.
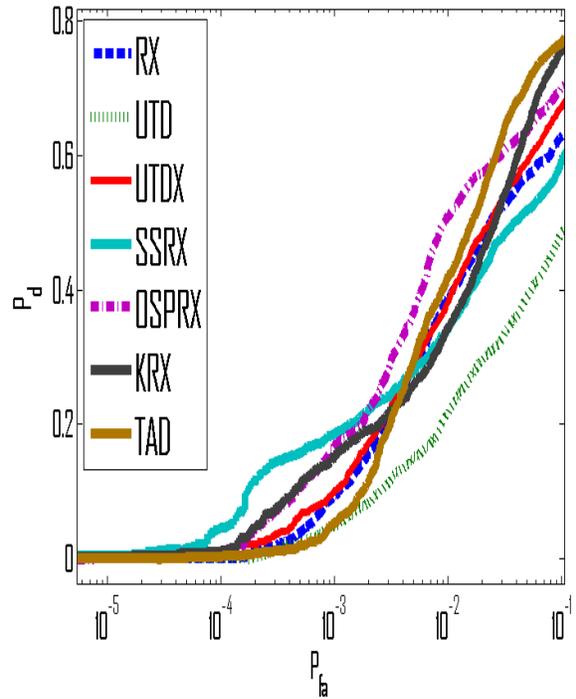

Figure 2: ROC curves corresponding to best discovered free parameter settings for comparison algorithms. SSRX outperforms the other algorithms over the range of small P_{fa} that is of most practical interest.

was computed for every pixel in the scene. We performed grid searches for each algorithm to reveal the optimal free parameter choices and increased the grid size if the maximum occurred near a grid edge.

SSRX and OSPRX have a free parameter which controls the number of principle components or number of eigenvectors of the covariance matrix which are retained respectively. For both of these methods we analyzed all possible contiguous ranges of the principle components or eigenvectors sorted in descending order of the largest corresponding eigenvalues (i.e., $\{1\}, \{1, 2\}, \{1, 2, \ldots, L\}, \{2\}, \{2, 3\}, \ldots, \{L\}$, where L is dimensionality of the hyperspectral scene) and chose the range which provided the maximum AUC.

KRX has three free parameters, the type of kernel function, kernel function parameter, and sampling size, $|\mathcal{S}|$. We chose the Gaussian kernel based on past literature which requires only one parameter, the kernel bandwidth, σ. We then performed a grid-search over $\sigma = \{1, 3, 5, 8, 10, 15, 20, 25, 40, 60, 80, 100, 200\}$ and $|\mathcal{S}| = \{0.3, 0.5, 0.8, 1, 1.5, 2, 3, 5, 10\}\%$ of total pixels N and chose the pair which provided the maximum AUC.
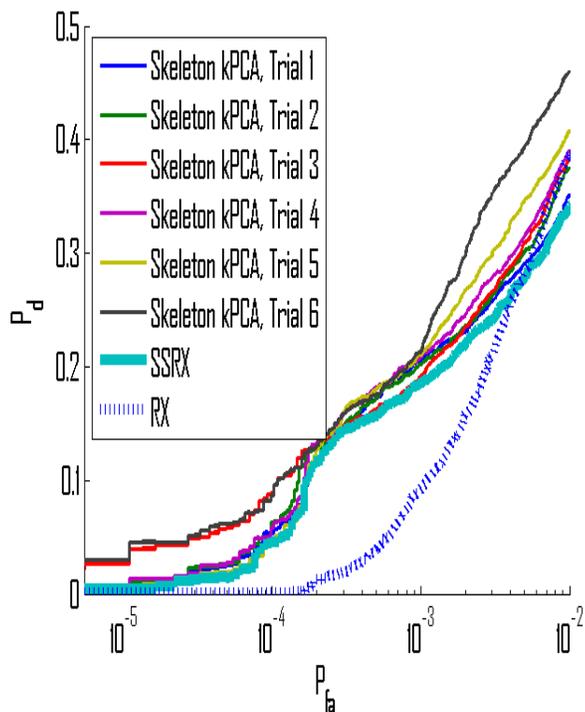
TAD has several parameters which the authors provided

Figure 3: ROC curves corresponding to six different skeleton samplings of our method (thin solid lines) as well as RX (dotted dark blue line) and the top-performing comparison algorithm, SSRX (solid thick blue line). Our method outperforms the other algorithms over the range of small P_{fa} that is of most practical interest. Note the smaller range of P_{fa} displayed here relative to the range in Figure 2

in their algorithmic description such as percent of edges to connect, size of graph components to count as background, and number of neighbors to consider for anomalousness ranking; the authors, however, give a range for the number of pixels, $|\mathcal{S}|$, to sample for building the graph and whether to perform the normalization procedure. We varied $|\mathcal{S}| = \{0.3, 0.5, 0.8, 1, 1.5, 2, 3, 5, 10, 20\}\%$ of total pixels N with and without the optional normalization procedure (the first option seemed to produce better results) and chose the pair which provided the maximum AUC.

Skeleton kPCA is governed by the choice of kernel, the parameters governing the kernel, the number of retained eigenvectors, and the number of points sampled for the skeleton. Here we have selected the Gaussian kernel with bandwidth σ. Hoffmann provided guidance for setting the bandwidth and the number of retained eigenvectors by examining performance on a toy problem [8], but he performed his analysis on the entire data set and did not consider performance on a skeleton subsample of the data. A study of said parameter settings as a function of subsam-

ple size was considered in [11] and showed that a relatively small skeleton sample ($|\mathcal{S}| = 0.001 \cdot N$) provided consistent results. Larger bandwidths coupled with a larger number of retained eigenvectors were found to provide good results on an 8-band multispectral data set. Performance consistency between different skeleton samplings was attempted by setting the bandwidth as a multiple of the largest Euclidean distance, D_{max}, calculated for a given sampling. In this study we set $\sigma = 16 \cdot D_{max}$ and retain 32 of N total eigenvectors corresponding to the largest eigenvalues.

ROC curves for all algorithms other than kPCA are provided in Figure 2. SSRX is the best performing algorithm over the range of low P_{fa} that are of the most interest. Figure 3 shows skeleton kPCA ROC performance for six separate skeleton samplings using the described settings relative to the best performing comparison algorithm (SSRX). Although there is variation in ROC performance corresponding to different sub-samplings, the skeleton kPCA method outperforms or is equivalent to SSRX over the range of low P_{fa} values that are of practical interest. Significant improvement at very low false alarm rates is shown for some of the skeleton samplings.

5. Conclusion

We have shown that our skeleton kPCA anomaly detection algorithm outperforms seven other state-of-the-art detection algorithms on a common data set. We have shown that the algorithm is reasonably robust to performance variation driven by variability in the selection of a random subsample. Although the study was performed on a commonly tested dataset, additional analysis using a variety of other images is required to better understand performance robustness among the competing algorithms. Future work will consider the fusion of both spatial and spectral information using the top algorithms.

References

[1] C. T. H. Baker. *The Numerical Treatment of Integral Equations*. Clarendon Press, Oxford, 1977.

[2] B. Basener, E. J. Ientilucci, and D. W. Messinger. Anomaly detection using topology. In *Defense and Security Symposium*, pages 65650J–65650J. International Society for Optics and Photonics, 2007.

[3] W. F. Basener and D. W. Messinger. Enhanced detection and visualization of anomalies in spectral imagery. In *SPIE Defense, Security, and Sensing*, pages 73341Q–73341Q. International Society for Optics and Photonics, 2009.

[4] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In *Advances in Neural Information Processing Systems*, volume 16. The MIT Press, Cambridge, MA, USA, 2004.

[5] C.-I. Chang. Orthogonal subspace projection (OSP) revisited: A comprehensive study and analysis. *IEEE trans-*

actions on geoscience and remote sensing, 43(3):502–518, 2005.

[6] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.

[7] J. C. Harsanyi, W. Farrand, and C.-I. Chang. Detection of subpixel spectral signatures in hyperspectral image sequences. In *Proc. Amer. Soc. Photogram. Remote Sens.*, pages 236–247. Amer. Soc. Photogram. Remote Sens., 1994.

[8] H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40:863–874, 2007.

[9] H. Kwon and N. M. Nasrabadi. Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery. *IEEE Trans. on Geosci. Remote Sens.*, 43(2):388–397, 2005.

[10] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on pattern analysis and machine intelligence*, 28(11):1784–1797, 2006.

[11] C. C. Olson and T. Doster. A parametric study of unsupervised anomaly detection performance in maritime imagery using manifold learning techniques. In *SPIE Defense+ Security*, pages 984016–984016. International Society for Optics and Photonics, 2016.

[12] C. C. Olson, J. M. Nichols, J. V. Michalowicz, and F. Bucholtz. Improved outlier identification in hyperspectral imaging via nonlinear dimensionality reduction. *Proceedings SPIE*, 7695, 2010.

[13] E. Parzen. On estimation of probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.

[14] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1988.

[16] I. S. Reed and X. Yu. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(10):1760–1770, 1990.

[17] A. P. Schaum. Spectral subspace matched filtering. In *Aerospace/Defense Sensing, Simulation, and Controls*, pages 1–17. International Society for Optics and Photonics, 2001.

[18] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.