ITERATIVE DIFFUSION-BASED ANOMALY DETECTION

Gal Mishne and Israel Cohen

Faculty of Electrical Engineering
Technion - Israel Institute of Technology, Haifa 32000, Israel
{galga@tx, icohen@ee}.technion.ac.il

ABSTRACT

Diffusion maps, when applied to large datasets, are typically constructed by a process of sampling and out-of-sample function extension. However, the performance of anomaly detection in large data when using diffusion maps is sensitive to the chosen samples. In this paper we propose an iterative data-driven approach to improve the sample set and diffusion maps representation. By updating the sample set with suspicious points detected in the previous iteration, the constructed diffusion maps better separate the anomaly from the normal points in each iteration. Experimental results in side-scan sonar images demonstrate the improvement gained by our iterative sampling compared to random sampling and other competing detection algorithms.

Index Terms— anomaly detection, diffusion maps, dimensionality reduction, manifold learning, automated target detection

1. INTRODUCTION

In image anomaly detection, the goal is to identify an object in the image and separate it from the background based on its different appearance or statistical properties. Automatic detection algorithms are of practical importance in military target detection applications, medical image analysis and automation of quality assurance processes, given the large amount of images produced in such applications. A robust solution will present the user only with suspicious objects, saving valuable time and effort, as suspicious objects occur very rarely by nature.

Anomaly detection in images is challenging due to several factors: the large size of the dataset (images of up to millions of pixels), high dimensionality of the features used for image representation and the presence of noise, especially in remote sensing imagery. It is also difficult to obtain labeled data, and the datasets tend to be unbalanced due to the sparseness of anomalies compared to normal data [1]. There are many approaches to solving this problem based on statistical models, machine learning, saliency based methods, sparse representations and more [2–10]. Statistical approaches estimate a statistical model for the data and use this information to determine whether a test sample comes from the distribution describing the normal data points [2, 4, 6]. The problem with statistical approaches is that the choice of the distribution to model the image is not obvious, it is application-dependent and the parameter estimation becomes complex in multi-class backgrounds. Detection methods using machine learning require training data, which is not always available and time-consuming to label. Both approaches may fail when the query images differ from the training data [1] and they are not easily adapted to new application and image characteristics.

Image features are typically high-dimensional, but can be shown to lie on a low-dimensional manifold. Manifold learning techniques, such as diffusion maps [11], find a new low-dimensional representation for the data, which reveals meaningful structures. Such techniques can provide a representation that separates the anomaly from the background, making the detection easier in the new representation space. In addition, such approaches are data-driven and do not depend on a-priori model for the data. Related work proposing dimensionality reduction methods for anomaly detection employ statistical tools for detection in the reduced dimension and relied on training data [5, 8]. We propose an approach which is completely data-driven, with the detection score based on nearest neighbors in the embedding space.

Due to the large size of the dataset, applying diffusion maps to images typically requires performing sampling and out-of-sample function extension in order to calculate an embedding for all image patches [12, 13]. In a previous work [14] we analyzed that this process can limit the success of the dimensionality reduction in revealing the presence of anomalies in the data. To overcome these limitations, we proposed an unsupervised multiscale algorithm for anomaly detection using diffusion maps [14], and an anomaly detection score relying on the noise-robust diffusion distance.

In this paper, we propose an alternative iterative approach on the full-scale image. Starting with a random subset of images patches, we apply a diffusion-maps based anomaly detector at each iteration, and then extract a set of suspicious patches to be used in the next iteration. Thus, the iterative process updates the subset of samples to ensure better separability of the anomaly from the background clutter. The advantage of this approach compared to the multiscale approach is that it is not limited to images and it requires less parameters. In addition, we provide motivation for this approach based on theoretical analysis of spectral clustering, where we view anomaly detection as a vastly unbalanced clustering problem.

The paper is organized as follows. Sec. 2 reviews the diffusion map framework for dimensionality reduction. In Sec. 3, we provide motivation for the proposed approach and present the iterative algorithm. Finally, Sec. 4 demonstrates the application of the proposed algorithm to automatic target detection in real images of side-scan sonar where the anomalies are sea-mines.

2. DIFFUSION MAPS

In recent years, a large number of non-linear techniques for dimensionality reduction have been developed, based on the underlying geometry of the data and intended as a precursor to other types of processing. These include ISOMAP [15], locally linear embedding

The work was supported by the Israel Science Foundation (grant no. 576/16), Qualcomm Research Fund and MAFAAT-Israel Ministry of Defense.

(LLE) [16], Laplacian Eigenmaps [17], Hessian Eigenmaps [18], and Diffusion Maps [11]. We focus on the diffusion maps method, which is based on the construction of the graph Laplacian of the data. In addition to capturing the main structure of the data in few dimensions, this method is equipped with a noise-robust computationally efficient metric, the diffusion distance, which is the Euclidean distance in the low-dimensional embedding space.

Let $\Gamma = \{\mathbf{x}_i\}_{i=1}^M$ be a high-dimensional dataset of M samples of dimensionality N. A weighted graph is constructed on Γ with the data points as nodes and the weight of the edge connecting two nodes is a measure of the similarity between the two data points. Let $w(\mathbf{x}_i, \mathbf{x}_j)$ be a kernel representing the pairwise affinity between two samples, which conveys the local geometry of the data. The choice of the specific kernel function should be application-driven. For all $\mathbf{x}_i, \mathbf{x}_j \in \Gamma$, the kernel function has the following properties: (1) symmetry; (2) non-negativity; (3) fast decay. A popular choice which satisfies these properties is the Gaussian kernel:

$$w(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2\right\}. \tag{1}$$

We construct the normalized graph Laplacian on the dataset as follows. For each $\mathbf{x}_i$ the kernel is normalized by the degree $d(\mathbf{x}_i)$:

$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{k(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)}, \quad d(\mathbf{x}_i) = \sum_{j=1}^{M} k(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

Since $p(\mathbf{x}_i, \mathbf{x}_j) > 0$ and $\sum_{j=1}^{M} p(\mathbf{x}_i, \mathbf{x}_j) = 1$, $p(\mathbf{x}_i, \mathbf{x}_j)$ can be interpreted as the probability for a random walker to jump from x_i to x_j in a single time step. Therefore, the row-stochastic matrix $\mathbf{P}[i, j] = p(\mathbf{x}_i, \mathbf{x}_j)$ can be seen as the transition matrix of a Markov chain on the dataset Γ. Taking powers of the matrix, $\mathbf{P}^t$, is equivalent to running the Markov chain forward t steps. The eigendecomposition of the matrix $\mathbf{P}$ yields a sequence of left and right biorthogonal eigenvectors $\{\varphi_j, \psi_j\}$ and eigenvalues, written in a descending order: $1 = |\lambda_0| > |\lambda_1| \geq |\lambda_2| \geq \ldots$.

Coifman and Lafon [11] propose a new metric, termed the diffusion distance, based on the construction of the random walk:

$$d_{\text{DM}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l \geq 1} \lambda_\ell^{2t} (\psi_\ell(i) - \psi_\ell(j))^2. \tag{3}$$

This distance measures the similarity of two points according to the evolution of the transition probability distribution $p_t(\mathbf{x}_i, \mathbf{x}_j)$ in the Markov chain. It is robust to noise, since the distance depends on all possible paths of length t between two points. Due to the fast spectrum decay of $\{\lambda_j\}$, the diffusion distance can be well approximated by only the first few ℓ eigenvectors.

Thus, the right eigenvectors of the transition matrix P and corresponding eigenvalues can be used to define a new data-driven representation of $\{\mathbf{x}_i\}$ termed diffusion maps. The mapping $\Psi_t(\mathbf{x}_i)$ embeds the data points $\{\mathbf{x}_i\}$ in a low-dimensional Euclidean space $\mathbb{R}^\ell$, defined as

$$\Psi_t(\mathbf{x}_i) = \left[\lambda_1^t \psi_1(i), \cdots, \lambda_\ell^t \psi_\ell(i)\right]^T, \tag{4}$$

where we set $t = 1$.

The scale parameter σ is of great significance in constructing the affinity kernel (1). Setting σ to be too small results in a disconnected graph, where many points are connected only to themselves. Setting σ to be too large results in all the points of the graph being connected. This is especially undesirable in the setting of anomaly detection, where setting σ to be too large will connect the anomalies with the cluttered background. We expect the anomaly to be in a low density neighborhood and the background to belong to a high density neighborhood. Therefore a local scale factor $\sigma = \sigma_i \sigma_j$ is beneficial, such as the one proposed by Zelnik-Manor and Perona [19].

3. ITERATIVE ANOMALY DETECTION

Our aim is to detect a connected group of pixels in the image that constitute an anomaly. We choose to represent the pixels by their surrounding patches, but other features can be used. Let Γ be the set of all $\sqrt{N} \times \sqrt{N}$ overlapping patches from the image $\Gamma = \{\mathbf{p}_i\}_{i=1}^M$. The size of the dataset for images, M, is very large, since even for small images there is a large number of patches. Therefore, it can be computationally inefficient to construct a diffusion map using all the patches in the image, in terms of both calculating the affinity and performing an eigen-decomposition, especially for high-resolution images. Instead, it is a common approach [12, 13] to construct the diffusion map for an image using a subset of $m < M$ random patches, denoted $\overline{\Gamma} = \{\overline{\mathbf{p}}_j\}_{j=1}^m \subseteq \Gamma$. Then the diffusion map coordinates Ψ are extended to the set of all patches in the image Γ using an out-of-sample extension (OOSE) method [20–22]. As discussed in [14], OOSE methods can cause anomaly detection to fail, depending on the set of random samples $\overline{\Gamma} \subseteq \Gamma$ used to construct the diffusion map. To overcome this limitation of the OOSE for anomaly detection, we propose an iterative approach which drives the sampling process and ensures the inclusion of samples from the anomaly region in $\overline{\Gamma}$. We first provide motivation for our approach based on spectral clustering and then present the iterative approach.

3.1. Spectral Clustering

Spectral clustering methods propose to cluster data into k clusters with the first dominant eigenvectors of an affinity matrix on the data [19, 23–25]. Anomaly detection can be seen as a special case of clustering in which there is a vast imbalance in the size of clusters, background vs. anomaly, and the density of each cluster. Several works in spectral clustering analyze the limitations and success of spectral clustering in the setting of a diffusion process in a multi-well potential or a mixture of Gaussians model [25–27]. They present two characteristic times that determine the success of spectral clustering. The first is the relaxation time for each cluster τ_{R} and the second is the mean first passage time τ_{exit} between two clusters. These depend on the covariance matrix of each cluster and its density. Spectral clustering succeeds when the fastest exit time from any of the clusters $\mathcal{C}_i$ is significantly slower than the slowest relaxation time in each one of the clusters, i.e. $\max_i\{\tau_{\text{R}}(\mathcal{C}_i)\} < \min_i\{\tau_{\text{exit}}(\mathcal{C}_i)\}$. Otherwise, the leading eigenfunctions capture only the relaxation process inside the dominant cluster (or clusters), and do not differentiate the remaining clusters, and therefore the clustering fails. Nadler and Galun [26] show that spectral clustering cannot successfully cluster datasets that contain structures at different scales of size and density, as in our setting. Thus, even if using all patches in the image, the dominant eigenvectors could fail to capture the anomaly in the case of a multi-class background.

This limitation is even more pronounced when using a subset of samples from the data. In a case where there are no anomalies in $\overline{\Gamma}$ and it consists only of samples from a single N-dimensional cluster (the background), then the eigenvectors capture only the relaxation process within this cluster [28]. This also occurs when the data contains multiple background clusters, such that the sample set is highly unbalanced in terms of the sample density of each cluster. In these cases, the diffusion map fails to capture the difference between the

anomaly and the background. The OOSE of the diffusion map to the anomaly points will not succeed in assigning them new coordinates that distinguish them from the background. Thus, anomaly detection when the samples from the anomaly are not included in the calculation of the initial diffusion map requires extrapolation of the diffusion coordinates and not interpolation. However, it is not clear how to perform extrapolation on the low-dimensional manifold, if at all possible. This is a "chicken and egg" problem in which it is necessary to sample the anomaly for the purpose of detecting it.

To overcome this limitation of OOSE, we propose an iterative approach, inspired by [4], which drives the sampling process and ensures the inclusion of samples from the anomaly region in $\overline{\Gamma}$. By constructing a sample set which includes a more balanced set of both anomaly and background points, we improve the detection compared to an unbalanced set which is composed mostly, if not entirely, of background points. In our approach, initially random samples are used in $\overline{\Gamma}$ and then more and more suspicious points are sampled from iteration to iteration. This is equivalent to reducing the local density of the large dominant background clusters while increasing the density of the anomaly cluster in the sample set. The diffusion map is then dominated by the anomaly and not the background.

3.2. Iterative Sampling

Given an image, we extract all overlapping patches to obtain Γ, and initialize $\overline{\Gamma}$ with a random subset of m patches from Γ. We then calculate an affinity kernel, calculate the diffusion map for $\overline{\Gamma}$ and extend the embedding to all the patches using an OOSE method. An anomaly score S^{CAS} (presented in Sec. 3.4) is calculated for all patches based on their diffusion embedding. We mark suspicious patches by applying a threshold τ to the anomaly score, where we set τ to be the 95th percentile of the anomaly score. This is the output of the initial iteration. In the next iteration, we set $\overline{\Gamma}$ to include the suspicious patches and the rest of the patches are chosen at random so that $|\overline{\Gamma}| = m$. The process of sampling, dimensionality reduction and anomaly detection continues from iteration to iteration, with each iteration providing prior information on which samples of the dataset will be used in $\overline{\Gamma}$ to construct the diffusion map. At the final iteration, we use a hard threshold on the detection score S^{CAS} and then smooth the resulting image.

Our approach is based on the assumption that anomaly patches will be included among the suspicious patches, and therefore, be represented in the diffusion map of the next iteration. Even if they are not included, it is likely they are more similar to other irregular patches in the background that will be included in $\overline{\Gamma}$. Thus, when the diffusion map is extended from $\overline{\Gamma}$ to the anomaly patches, they will be assigned a representation similar to that of the irregular patches and therefore, will be marked as suspicious for the next iteration.

3.3. Affinity kernel and Out-of-sample Extension

We propose to replace the regular diffusion maps construction described in Sec. 2 with a combined embedding and OOSE approach previously proposed in supervised settings [29–31]. We treat the set of sampled patches $\overline{\Gamma}$ as a reference set and calculate a non-square affinity matrix between Γ and $\overline{\Gamma}$:

$$\mathbf{A}[i,j] = \exp\{-\|\mathbf{p}_i - \overline{\mathbf{p}}_j\|^2/\sigma_i\sigma_j\}, \tag{5}$$

such that $\mathbf{A}$ is an $M \times m$ affinity matrix where $M > m$.

We then define the symmetric matrix $\mathbf{W} = \mathbf{A}^T\mathbf{A}$ which is an $m \times m$ matrix between the sample patches in $\overline{\Gamma}$:

$$\mathbf{W}[i,j] = \sum_{l=1}^{M} \mathbf{A}[l,i]\mathbf{A}[l,j]. \tag{6}$$

The advantage of this construction compared to using the affinity kernel in (1) is that it takes into account not only the distances between the sample patches, but between the sample patches and all patches in the image. The matrix $\mathbf{W}$ can be interpreted as an affinity between any two sample patches via all the patches in the image [30,31]. Thus, two sample patches are similar if they "view" the rest of the image patches the same, i.e. their distances to all other patches in the image are similar. This implies that they belong to the same local "patch neighborhood". Note that by distance here we refer to the high-dimensional difference of the patch intensity values and not the spatial distance within the image. Thus, this kernel is better at suppressing noisy connections between two patches. This is important when the number of samples is limited as in our approach. In addition, this approach provides an elegant and efficient OOSE as follows.

An eigen-decomposition of $\mathbf{W}$ yields the eigenvectors $\{\phi_l\}_l$ and eigenvalues λ_l. The eigenvectors $\{\phi_l\}_l$ are also the singular right vectors of $\mathbf{A}$ and can be used to calculate the singular left eigenvectors $\{\psi_l\}_l, \psi_l \in \mathbb{R}^M$ of $\mathbf{A}$ by simple matrix multiplication [29]

$$\psi_l = \frac{1}{\sqrt{\lambda_l}}\mathbf{A}\phi_l. \tag{7}$$

Thus, an eigen-decomposition of $\mathbf{W}$ provides an efficient manner in which to calculate the singular left eigenvectors $\{\psi_l\}_l$ of $\mathbf{A}$, and obtain the embedding of all patches in Γ following (4).

3.4. Detection Score

We use the anomaly detection score proposed in [32], relying on the diffusion distance between pairs of patches. Let $\{\mathbf{q}_{i,k}\}_{k=1}^{K}$ be the K most similar patches of a patch $\mathbf{p}_i$ under the diffusion distance (3). Then the anomaly detection score is given by:

$$S^{\text{CAS}}(i) = 1 - \exp\left\{-\frac{1}{K}\sum_{k=1}^{K}\frac{d_{\text{DM}}(\mathbf{p}_i,\mathbf{q}_k)/2\sigma_K}{1+3d_{\text{position}}(\mathbf{p}_i,\mathbf{q}_k)}\right\}, \tag{8}$$

where $d_{\text{position}}(\mathbf{p}_i,\mathbf{q}_k)$ is the Euclidean spatial distance between the center pixels of patches $\mathbf{p}_i$ and $\mathbf{q}_k$. The parameter σ_K is a normalizing factor given by the standard deviation of the diffusion distances between all patches and their K^{th} nearest neighbor: $\sigma_K = \underset{i\in\Gamma}{\text{std}}\{d_{\text{DM}}(\mathbf{p}_i,\mathbf{q}_K)\}$.

This score is inspired by the saliency map in [33], quantifying the observation that background patches are similar to both near and far patches, whereas salient patches are grouped together and similar only to nearby patches. We measure the similarity between patches using the diffusion distance, which is preferable to the Euclidean distance between the patches intensities since it is robust to noise. Also, the diffusion map better separates the anomaly from the background, compared to using image patches [32]. Normalizing this distance by the spatial distance d_{position} penalizes background patches by assigning them a low distinctness value. To ensure that most values are spread out in the range $[0,1]$ and therefore comparable to d_{position}, we normalize the diffusion distance by σ_K.

4. EXPERIMENTAL RESULTS

We apply the proposed algorithm to sea-mine detection in real side-scan sonar images, achieving a high detection rate with a low rate of false-alarms. We treat the sea-mines in the images as anomalies and the reflections from the seabed are considered normal background clutter. Automatic detection of sea mines in side-scan sonar imagery is a challenging task due to the high variability in the appearance of the target and sea-bed reverberations (background clutter). Objects in side-scan sonar appear as a strong bright region (highlight) alongside a dark region (shadow), which is due to the object blocking the sonar waves from reaching the seabed. Most algorithms for detection of sea-mines in side-scan sonar make use of a training set, based on real images and/or synthetic ones [6, 34, 35]. Our diffusion-based approach does not require a training set and makes no assumptions regarding the appearance of the mine or its shadow in the image.

We evaluated our approach on a set of 48 side-scan sonar images of size 200×200 pixels. We use patches of size 8×8 and a sampling density of 20% of the patches so that $m \approx 8000$. Note that the size of the images enables denser sampling of the image, however we intentionally use a small percentage of the patches in the image to demonstrate that this framework is applicable also for larger-sized images. The patches are embedded to a diffusion maps space of dimensionality $\ell = 6$.

Detections are found by applying a threshold to the final anomaly score image, resulting in a binary image. A detection is a connected component (CC) in the binary image, where a CC containing the sea-mine is a true positive (TP) and any other CCs are false alarms (FA). We count all detections on the sea-mine as a single TP for a given image, whereas there can be more than one FA in an image. The size of the CC can be used to reject noisy detections by discarding small CCs. We compare two thresholds on the area of the CC: 5 pixels and 10 pixels. Using a larger threshold on the size rejects more FAs, but can also result in a decreased amount of TPs, for small sized anomalies. We compare the percentage of TPs for each method for a given FA rate. Results are given in Fig. 1.

We compare our proposed iterative diffusion detection (IDD) algorithm with three other methods:

- IterStat: An iterative local statistical model-based method as proposed in [4].
- CAS: The saliency method proposed in [33] where detection is performed by applying a threshold to the saliency map.
- MS-CAS: The multiscale method we proposed in [32].

The IDD method is based on the joint diffusion and OOSE construction presented in Sec. 3.3. We compare this method to the "traditional" approach of calculating the diffusion map using only the sample set as in Sec. 2 and performing OOSE, where we employ the Laplacian pyramids approach in [21] (denoted IDD-LP).

For both IDD (blue plot) and IDD-LP (red plot), the second iteration results in a significant increase in performance (comparing the dashed plot for the first iteration to the solid plot for the second iteration). Comparing IDD and IDD-LP, calculating the affinity matrix for the sample set by integrating the affinity to all patches in the image yields a gain of about 4% in the second iteration. Both methods are comparable to the multiscale MS-CAS approach we proposed in previous work (yellow plot) with an advantage of 6% for IDD. Another advantage of our new approach is that MS-CAS is based on a multiscale representation of the image and is thus limited to anomaly detection in images. The IDD method can be adapted to any data provided an anomaly score in the diffusion maps space.

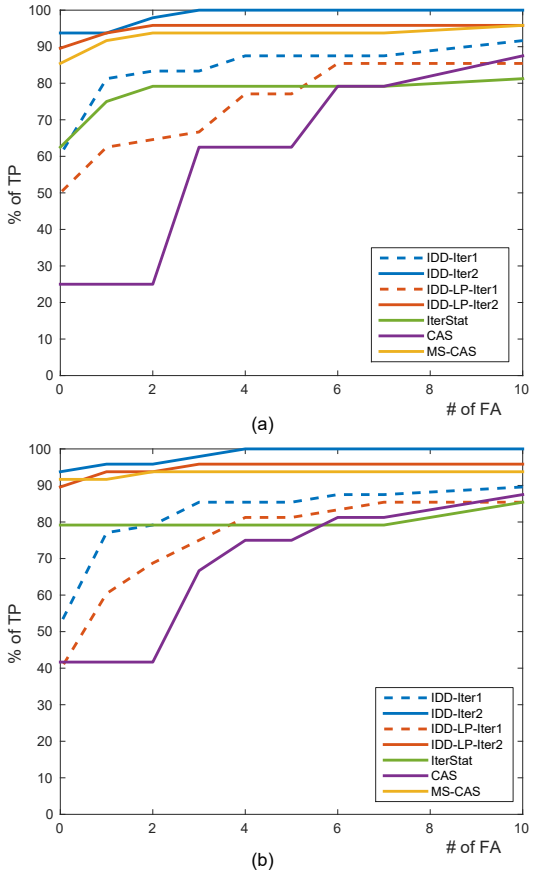

Fig. 1. Percentage of true positives for given number of false alarms for detections of size greater that (a) 5 pixels and (b) 10 pixels.

Comparing the diffusion-based iterative method to the statistical method (green plot), our data-driven approach has significantly higher performance and is not sensitive to parameter selection of the statistical model. Our method also outperforms the CAS method [33]. The reason for this is that the saliency-based method uses a feature space which is suitable for natural images. The images we tested are side-scan sonar and are very noisy. Thus, some of the noise patterns in the background are given a high saliency score. Using the diffusion map as a feature space suppresses the noise, due to the robustness of the diffusion distance, and the noisy patterns are clustered with the background and not detected as anomalies.

5. CONCLUSIONS

We presented a new iterative anomaly detection algorithm using diffusion maps. Based on a spectral clustering analysis, we proposed a method that iteratively improves the sample set used to construct the diffusion map by including suspicious samples detected in the previous iteration. This leads to a representation that better separates the anomaly from the background. In addition, we obtained an improved representation by using an affinity kernel that incorporates all the data and not just the limited sample set. Our algorithm achieved successful performance in the challenging task of automatic target detection in side-scan sonar images and achieved superior results compared to competing methods.

6. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, July 2009.

[2] G. G. Hazel, "Multivariate gaussian MRF for multispectral scene segmentation and anomaly detection." *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1199–1211, 2000.

[3] D. W. J. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker, "Anomaly detection from hyperspectral imagery," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 58–69, Jan 2002.

[4] A. Goldman and I. Cohen, "Anomaly detection based on an iterative local statistics approach," *Signal Process.*, vol. 84, no. 7, pp. 1225–1229, Jul. 2004.

[5] D.-M. Tsai and C.-H. Yang, "A quantile-quantile plot based pattern matching for defect detection," *Pattern Recogn. Lett.*, vol. 26, no. 13, pp. 1948–1962, Oct. 2005.

[6] A. Noiboar and I. Cohen, "Anomaly detection based on wavelet domain GARCH random field modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1361–1373, 2007.

[7] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vision*, vol. 74, no. 1, pp. 17–31, 2007.

[8] E. Madar, D. Malah, and M. Barzohar, "Non-gaussian background modeling for anomaly detection in hyperspecral images," in *Proc. European Signal Process. Conf.*, 2011.

[9] S. Mousazadeh and I. Cohen, "Two dimensional noncausal AR-ARCH model: Stationary conditions, parameter estimation and its application to anomaly detection," *Signal Process.*, vol. 98, no. 0, pp. 322 – 336, 2014.

[10] D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg, "Scale-invariant anomaly detection with multiscale group-sparse models," in *Proc. ICIP*, 2016, pp. 3892–3896.

[11] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, July 2006.

[12] Z. Farbman, R. Fattal, and D. Lischinski, "Diffusion maps for edge-aware image editing," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 145:1–145:10, Dec. 2010.

[13] J. He, L. Zhang, Q. Wang, and Z. Li, "Using diffusion geometric coordinates for hyperspectral imagery representation," *IEEE Geosci. Remote Sens. Letters*, vol. 6, no. 4, pp. 767–771, Oct. 2009.

[14] G. Mishne and I. Cohen, "Multiscale anomaly detection using diffusion maps," *IEEE J. Sel. Topics Signal Process.*, vol. 7, pp. 111 – 123, Feb. 2013.

[15] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[16] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[18] D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci. (PNAS)*, vol. 100, pp. 5591–5596, 2003.

[19] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *NIPS 17*, 2005, pp. 1601–1608.

[20] R. R. Coifman and S. Lafon, "Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 31–52, 2006.

[21] N. Rabin and R. R. Coifman, "Heterogeneous datasets representation and learning using diffusion maps and Laplacian pyramids," in *Proc. SDM12*, 2012.

[22] G. Mishne, U. Shaham, A. Cloninger, and I. Cohen, "Diffusion Nets," *ArXiv e-prints*, Jun. 2015.

[23] Y. Weiss, "Segmentation using eigenvectors: a unifying view," in *Proc. ICCV 1999*, vol. 2, 1999, pp. 975–982.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, 2002.

[25] B. Nadler, S. Lafon, and R. Coifman, "Diffusion maps - a probabilistic interpretation for spectral embedding and clustering algorithms," in *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer, 2007, pp. 242–264.

[26] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering," in *Advances in Neural Information Processing Systems (NIPS) 19*, 2006, pp. 1017–1024.

[27] A. Singer, Y. Shkolnisky, and B. Nadler, "Diffusion interpretation of nonlocal neighborhood filters for signal denoising," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 118–139, Jan. 2009.

[28] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps - a probabilistic interpretation for spectral embedding and clustering algorithms," in *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer, 2007.

[29] D. Kushnir, A. Haddad, and R. R. Coifman, "Anisotropic diffusion on sub-manifolds with application to earth structure classification," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 2, pp. 280 – 294, 2012.

[30] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 1, pp. 130–142, Apr. 2012.

[31] A. Haddad, D. Kushnir, and R. R. Coifman, "Texture separation via a reference set," *Appl. Comput. Harmon. Anal.*, vol. 36, no. 2, pp. 335 – 347, Mar. 2014.

[32] G. Mishne and I. Cohen, "Multiscale anomaly detection using diffusion maps and saliency score," in *Proc. ICASSP-2014*, 2014.

[33] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. CVPR*, 2010, pp. 2376–2383.

[34] S. Reed, Y. Petillot, and J. Bell, "Automated approach to classification of mine-like objects in sidescan sonar using highlight and shadow information," *IEE Proc. Radar Sonar Navig.*, vol. 151, no. 1, pp. 48–56, Feb. 2004.

[35] Y. Petillot, Y. Pailhas, and J. Sawas, "Target recognition in synthetic aperture and high resolution side-scan sonar," in *Proc. European Conference on Underwater Acoustics*, 2010, pp. 99–106.