Multiscale Anomaly Detection Using Diffusion Maps

Gal Mishne and Israel Cohen, *Senior Member, IEEE*

***Abstract*—We propose a multiscale approach to anomaly detection in images, combining spectral dimensionality reduction and a nearest-neighbor-based anomaly score. We use diffusion maps to embed the data in a low dimensional representation, which separates the anomaly from the background. The diffusion distance between points is then used to estimate the local density of each pixel in the new embedding. The diffusion map is constructed based on a subset of samples from the image and then extended to all other pixels. Due to the interpolative nature of extension methods, this may limit the ability of the diffusion map to reveal the presence of the anomaly in the data. To overcome this limitation, we propose a multiscale approach based on Gaussian pyramid representation, which drives the sampling process to ensure separability of the anomaly from the background clutter. The algorithm is successfully tested on side-scan sonar images of sea-mines.**

***Index Terms*—Anomaly detection, automated mine detection, diffusion maps, multiscale representation, nonlinear dimensionality reduction, similarity measure.**

I. Introduction

ANOMALY detection is important in many applications in image processing, such as target detection in hyperspectral [1], [2] or sonar images [3], [4], mammographic image analysis [5] and defect detection, for example in wafer or fabric inspection [6], [7]. A robust solution to this problem is important in military applications and automation of quality assurance processes, as the user will be shown only suspicious objects.

Anomaly detection in images is challenging due to several factors:

- Large size of the data set: images have between tens of thousands of pixels and up to millions of pixels.
- Noisy features which may be falsely detected as anomalies.
- Lack of training data: it is usually very hard to attain labeled data for anomaly detection. In addition, the data sets are unbalanced due to the nature of anomalies: there are many examples of normal data, but few of the anomalies. This makes unsupervised methods more desirable than supervised ones.
- High dimensionality of the data: images are usually represented using high-dimensional features such as the patch surrounding each pixel, histogram of gradients, etc.
- Multiple classes of normal data points: in many images the normal datapoints do not belong to a single cluster.

There are many approaches to anomaly detection in images based on statistical models, machine learning, saliency based methods, sparse representations, and more.

Statistical approaches model the data based on its statistical properties and use this information to estimate whether a test sample comes from the distribution describing the normal datapoints [2]–[5]. The problem with statistical approaches is that the choice of the distribution to model the image is not obvious. In cases where the background is multi-class, estimation of the parameters of the statistical model becomes complex. Also, a statistical model which works well for certain images will not necessarily be easily adapted to a new application.

Anomaly detection methods based on machine learning require training data, which is not always available, and they may not be able to detect new types of anomalies they were not trained on. The assumption in anomaly detection using sparse representation is that an anomaly cannot be reconstructed in a sparse manner using a dictionary learned from normal images. In such an approach, it is necessary to learn a dictionary to model the normal regions in the image, which requires training data to model the background.

Chen, Nasrabadi and Tran [1] propose training an additional dictionary to model the anomalies using training samples. In [6], the algorithm proposed by Boiman and Irani is based on the assumption that anomaly patches in an image cannot be composed combining normal patches from the image or from a reference image. The data (image or video) is divided into ensembles of many small patches at multiple scales, along with their relative spatial layout. Image regions that cannot be composed from ensembles of other patches are detected as anomalies. This algorithm presents impressive results, but it has high computational complexity in regards to both memory requirements and run-time. Zontak and Cohen [7] propose an algorithm for wafer defect detection based on anisotropic kernels. Patches from a test image are reconstructed using patches taken from a reference image, and patches which cannot be reconstructed from the reference patches are anomalous. This algorithm requires a reference image or an image with a periodic pattern.

The features used to describe images are typically high-dimensional, but can be shown to lie on a low-dimensional manifold. Dimensionality reduction techniques find a new, lower-dimensional representation for the data, which reveals meaningful structures. This is useful in anomaly detection because such techniques can find a representation which separates the anomaly from the background. The detection itself will then be easier in the reduced dimensionality. In addition, such approaches are data-driven and do not depend on a model for the data. For example, Madar, Malah and Barzohar [8] perform dimensionality reduction using the normalized eigenvectors of the

Manuscript received July 31, 2012; revised October 28, 2012 and November 25, 2012; accepted November 27, 2012. Date of publication December 10, 2012; date of current version January 22, 2013. This work was supported by the Japan Technion Society Research Fund and the Israel Science Foundation (Grant 1130/11). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ery Arias-Castro.

The authors are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel (e-mail: galga@techunix.technion.ac.il; icohen@ee.technion.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2012.2232279

Normalized Laplacian Matrix, constructed on a hyperspectral image. In the lower dimensionality, spectral clustering is employed to model different types of background terrain. These clusters are then used in a combined local-global statistical approach to model the background and detect anomalies. Tsai and Yang [9] introduce a method for defect detection using dimensionality reduction, in cases where a clean reference image is available. Dimensionality reduction is performed on the images using a 1-D vector of quantiles, and the quantile of the input image is compared to that of the template using a quantile-quantile (Q-Q) plot. Abnormalities are detected in the Q-Q plot using Chi square distribution.

We propose using diffusion maps [10] for dimensionality reduction. Diffusion maps is a spectral dimensionality reduction method based on the construction of the graph Laplacian on the data. It has been used successfully in various applications [11]–[16]. The computational burden of the diffusion maps approach may be significant as it requires the computation of an affinity matrix on the data. This requires calculations of the distance between each pair of samples in the data set. The burden can be reduced by sampling a subset of data points for which the diffusion map is calculated and then extending it to all points using an out-of-sample extension method [17], [18]. Sampling and extension is common practice in applying diffusion maps to images due to the large size of the data set [15], [19].

The computational complexity of constructing the affinity matrix can also be reduced by calculating a sparse affinity matrix, using a k-nearest-neighbor search. Thus, instead of calculating the kernel between each sample and all the rest of the samples, the kernel is calculated only between each sample and its nearest neighbors. This results in a sparse matrix and complexity is further reduced by efficient spectral decomposition algorithms adapted for sparse matrices. When using *exact* nearest-neighbor search, it can still be necessary to employ sampling and out-of-sample extension to reduce run-time, dependent on the size of the data set. However, fast algorithms for *approximate* nearest neighbor (ANN) search in which a degree of error is allowed in the query result can enable calculating the matrix for all data-points. This removes the need for sampling and extension. In such methods, the exact k-nearest-neighbors are not necessarily obtained, but k neighbors that are not too distant from the exact ones. These approximate queries can greatly reduce the search time [20]–[22]. For example, the computational complexity of the recently proposed randomized approximate nearest neighbors algorithm (RANN) search method proposed by Jones *et al.* [22] scales nearly linearly with the number of patches. This is useful when the dimensionality of the image features is not too high, since the performance of ANN algorithms deteriorates as the dimension increases. In practice, the performance depends on the intrinsic dimension of the data, which often turns out to be much smaller than the extrinsic dimension, as we assume in our setting. Since often the intrinsic dimension of the data is not known in advance, it is difficult to predict how well an ANN algorithm will do in a specific application.

Rabin and Averbuch recently proposed using diffusion maps for anomaly detection in a different application than image processing: a sensor data fusion framework [23], [24]. Using a hierarchical framework, diffusion maps are applied to the nodes at every level, first fusing groups of sensors together, and then fusing the groups together. The score function used is also a nearest-neighbors based approach, determined by the sum of the diffusion distances between each instance and its nearest neighbors. The anomalies in this application are contextual anomalies: the sensor measurements are not necessarily anomalies by themselves, but their co-occurrence in a particular form makes them anomalies [25]. In [24], the assumption is that the anomaly is within normal levels for each of the individual sensors and only becomes distinct through the fusion of the sensors. At the bottom level of their framework, i.e. the measurements, anomalies have values similar to the normal instances. This assumption usually does not hold in image anomaly detection where the data points are features of image patches or the image patches themselves.

A disadvantage of using spectral dimensionality reduction methods is that they are only useful if the normal and anomalous instances are separable in the lower dimensional embedding of the data [25]. This issue manifests itself in our approach due to the process of sampling and out-of-sample extension. We show how this process can limit the success of the dimensionality reduction in revealing the presence of anomalies in the data and propose an algorithm for overcoming these limitations. We propose a multiscale approach which drives the sampling process to ensure separability of the anomaly from the background clutter. This approach enables to effectively apply diffusion maps to the problem of anomaly detection. We demonstrate on real images that this approach greatly improves the anomaly detection, compared to methods which are single-scale.

The main advantage of using diffusion maps in our framework is that it induces a distance measure over the data set which is robust to noise and preserves local neighborhoods. This enables nearest-neighbor anomaly detection in the reduced dimensionality. Our assumption is that anomalies lie in low-density neighborhoods, whereas normal pixels lie in dense neighborhoods. Based on the local density of the pixel on the lower-dimensional manifold, we compute an anomaly score for every pixel. This score conveys the degree to which the pixel is considered an anomaly. Depending on the application, the score can be thresholded to produce a binary map of anomalies, or the pixels with top-ranking can be outputted to be inspected by the user. The successful performance of our algorithm is demonstrated for real images of side-scan sonar where the anomalies are sea-mines.

Our approach is unsupervised and no prior knowledge is required regarding the appearance of the anomaly or the background. No assumptions are made on the statistical model of the background pixels or if the background can be clustered into several different classes. We do not use training data or a reference image. Our approach is data-driven, and can be used in different applications. The user needs to provide a feature space for the data set and a distance measure which can be used to compare the local similarity of data points. In addition, the size of meaningful anomaly regions in the image can also serve as input, but it is not necessary.

The paper is organized as follows. Section II reviews the diffusion map framework for dimensionality reduction and Section III describes out-of-sample extension methods and

their limitations in anomaly detection. In Section IV, the proposed multiscale algorithm is presented. Finally, Section V demonstrates the application of the proposed algorithm to automatic target detection in real images.

II. Diffusion Maps

Real world data typically has high dimensionality. However, these high dimensional data sets can be shown to lie on low-dimensional manifolds. Finding a low-dimensional representation of the data is necessary to efficiently handle it and usually reveals meaningful structures within the data. This embedding of high-dimensional data into a low-dimensional manifold is done by dimensionality reduction methods. In recent years, a large number of nonlinear techniques for dimensionality reduction have been proposed [10], [16], [26]–[28]. Several of these methods are spectral methods, based on the eigenvectors of adjacency matrices of graphs on the data [10], [16], [28]. These methods take into account the geometry of the data set and the representation they provide preserves local neighborhood information. Diffusion maps [10] is one such technique, based on the construction of the graph Laplacian of the data set. It has been used successfully in various applications such as spectral clustering [11], signal denoising [12], speech enhancement [13], [14], hyperspectral image representation [15] and word recognition based on lip-reading [16].

Let $\Gamma = \{x_1, \ldots, x_n\}$ be a high-dimensional set of n data points. A weighted graph is constructed with the data points as nodes and the weights of the edges connecting two node is a measure of the similarity between the two data points. The weight function $w(x, y)$, $x, y \in \Gamma$ is required to be symmetric and pointwise nonnegative. The choice of the weight function should be determined by the application, since it conveys the local geometry of the data set. A popular choice is to weight the edge between the data points x_i and x_j using a Gaussian kernel:

$$w(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/\sigma^2\right), \tag{1}$$

where $\sigma > 0$ is a scale parameter.

Then, a random walk is created on the data set by normalizing the kernel in an asymmetric manner:

$$p(x, y) = \frac{w(x, y)}{d(x)}, \tag{2}$$

where $d(x) = \sum_{y\in\Gamma} w(x, y)$. The function p satisfies $p(x, y) \geq 0$ and $\sum_{y\in\Gamma} p(x, y) = 1$. Therefore, it can be interpreted as the probability for a random walker to jump from x to y in a single time step. The matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ with $p(\cdot, \cdot)$ as its entries is the transition matrix of this Markov chain on the data set Γ. Taking powers of the matrix is akin to running the Markov chain forward. The kernel $p_t(\cdot, \cdot)$ describes the probability of transition between two points in t steps.

It can be shown that $\mathbf{P}$ has a complete sequence of biorthogonal left and right eigenvectors, ϕ_j and ψ_j respectively, with a sequence of positive eigenvalues: $|\lambda_0| \geq |\lambda_1| \geq \ldots$. The spectral decomposition of $\mathbf{P}$, yields that t steps of the Markov chain can be presented as

$$p_t(x, y) = \sum_{l\geq 0} \lambda_l^t \psi_l(x)\phi_l(y). \tag{3}$$

Because of the fast decay of the spectrum, only a few terms are required to achieve sufficient accuracy in the sum. A mapping can be defined between the original space and the first ℓ eigenvectors. The diffusion map is defined by

$$\Psi_t : x \to \left(\lambda_1^t \psi_1(x), \lambda_2^t \psi_2(x), \ldots, \lambda_\ell^t \psi_\ell(x)\right)^T. \tag{4}$$

Note that ψ_0 is not used in the embedding because it is a constant vector. The mapping Ψ_t embeds the data set Γ into the Euclidean space $\mathbb{R}^\ell$. The spectrum decay of the eigenvalues is the reason why dimensionality reduction can be achieved. The dimension of the new representation depends only on the random walk and is independent of the length of the feature vector used in the original representation of the data.

A diffusion distance $D_t^2(x, z)$ between two points x, z in the data set Γ is defined by

$$D_t^2(x, z) = \sum_{y\in\Gamma} \frac{\left(p_t(x, y)) - p_t(z, y)\right)^2}{\phi_0(y)}. \tag{5}$$

This measures the similarity of two points according to the evolution of their probability distributions in the Markov chain. The diffusion distance between two points is small if there is a large number of short paths connecting them in the graph. This metric is robust to noise, since the distance between two points depends on all possible paths of length t between the points, within the dataset. As opposed to the original distance between two points which is independent of all other points in the dataset, the diffusion distance depends on the location of the other points in the dataset. Using the spectral decomposition given in (3), the diffusion distance in (5) can also be calculated using the eigenvectors by

$$D_t^2(x, z) = \sum_{j\geq 1} \lambda_j^{2t} \left(\psi_j(x) - \psi_j(z)\right)^2. \tag{6}$$

Taking into account the spectrum decay, the diffusion distance can be calculated up to a certain accuracy using only the first ℓ eigenvectors. Thus, the computational complexity of the diffusion distance is low given the eigen-decomposition of $\mathbf{P}$. It was shown [11] that the diffusion distance is equal to the Euclidean distance in the diffusion map space using all eigenvectors in the decomposition:

$$D_t^2(x, z) = \sum_{j\geq 1} \lambda_j^{2t} \left(\psi_j(x) - \psi_j(z)\right)^2 = \|\Psi_t(x) - \Psi_t(z)\|^2. \tag{7}$$

In Section IV, we use this property of the diffusion distance to define a measure of affinity in the diffusion coordinates.

Spectral embedding methods are commonly used in clustering applications [11], [28]–[30]. Most methods suggest to use the first non-trivial eigenvectors (the first eigenvector corresponding to $\lambda_0 = 1$ is constant) to find clusters in the

dataset. This clustering property of the diffusion map is useful for anomaly detection. We expect the background pixels in the image to be clustered together and the anomaly to be distant from this cluster in the new embedding.

A. Setting the Scale Parameter σ

The scale parameter σ is of great significance in constructing the weighted graph. Setting σ to be too small results in a disconnected graph, where many points are connected only to themselves (local neighborhoods of size 1). However, setting σ to be too large results in all the points in 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. Possibilities of setting the scale parameter are using the median distance between points in the dataset or the standard deviation of the distances. These are global parameters.

Zelnik-Manor *et al.* [29] suggest calculating a location dependent σ for each data point instead of selecting a single scaling parameter. Then, the affinity between a pair of points can be written as

$$w(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma_i \sigma_j}\right), \tag{8}$$

where σ_i and σ_j are the local scale parameters for x_i and x_j, respectively. The selection of the local scale σ_i is determined by the local statistics of the neighborhood of point x_i. For example, the scale can be set as

$$\sigma_i = \|x_i - x_K\|^2 \tag{9}$$

where x_K is the K-th nearest neighbor. We adopt this approach in our algorithm, using $K = 7$. This approach is local, since the distance between two points is scaled according to the local statistics of the neighborhoods surrounding the two points. This is desirable since we expect the anomaly to be in a low density neighborhood in contrast with the background, which we expect to be in a dense neighborhood. Setting a single global scale would not be able to address the differences in density of the points.

III. Function Extension

When the data set is very large, it is impractical to compute a diffusion map for the entire dataset $\overline{\Gamma}$. Instead, a diffusion map is constructed for part of the samples $\Gamma \subseteq \overline{\Gamma}$ and then the embedding is extended to all points in $\overline{\Gamma}$ using an out-of-sample extension method.

The Nyström extension method is a common method for the extension of functions from a given training set to new samples. Recently, methods have been proposed to approximate the Nyström extension method [31] or improve upon it, such as the Geometric Harmonics method [17]. In [17], the authors state that low-complexity functions can be easily extended very far from the training set as their behavior is smooth and the extended values are easy to predict. A function with many variations on Γ should have a limited range of extension, as its values off the training set are more difficult to predict.

A. Laplacian Pyramid Extension

Recently, a new algorithm was presented for out-of-sample function extension using the multiscale Laplacian pyramid [18]. At each iteration, the Laplacian pyramid algorithm constructs a coarse approximation of a function f for a given scale l. Then, the difference between f and the coarse approximation is used as input for the next iteration. The difference is approximated at each level using a Gaussian kernel with finer and finer scales.

On the lowest level, the Gaussian kernel is defined on Γ by

$$W_0 \triangleq w_0(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/\epsilon_0\right), \tag{10}$$

with ϵ_0 set to be a large scale. A smoothing operator is obtained by normalizing W_0:

$$K_0 = k_0(x_i, x_j) = q_0^{-1} w_0(x_i, x_j), \tag{11}$$

where $q_0(x_i) = \sum_j w_0(x_i, x_j)$. On the next levels, the Gaussian kernel is computed by

$$W_l = w_l(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/\frac{\epsilon_0}{2^l}\right), \tag{12}$$

and the smoothing operator is

$$K_l = k_l(x_i, x_j) = q_l^{-1} w_l(x_i, x_j). \tag{13}$$

The Laplacian Pyramid representation of a function f on Γ is defined iteratively by:

$$s_0(x_k) = \sum_{i=1}^{n} k_0(x_i, x_k) f(x_i), \quad l = 0 \tag{14}$$

$$s_l(x_k) = \sum_{i=1}^{n} k_l(x_i, x_k) d_l(x_i), \quad l \geq 1 \tag{15}$$

with the difference defined by

$$d_l(x_k) = f - \sum_{m=0}^{l-1} s_m, \quad l \geq 1. \tag{16}$$

The Laplacian pyramid is iterated on finer and finer scales until the difference $\|f - \sum_k s_k\|$ is below a given error threshold.

The function f is extended to a new data point $\overline{x}_k \in \overline{\Gamma}$ by the sum $f(\overline{x}_k) = \sum_i s_i(\overline{x}_k)$, where

$$s_0(\overline{x}_k) = \sum_{i=1}^{n} k_0(x_i, \overline{x}_k) f(x_i), \quad l = 0 \tag{17}$$

$$s_l(\overline{x}_k) = \sum_{i=1}^{n} k_l(x_i, \overline{x}_k) d_l(x_i), \quad l \geq 1. \tag{18}$$

We perform this extension method for each diffusion coordinate $f = \Psi_j$ separately. The number of levels in the pyramid extension can differ between the coordinates, dependent on their smoothness over Γ. A smooth function can be extended using coarse scale, i.e. will not require many levels of the pyramid. An oscillating function on the other hand will require finer and finer levels of the pyramid to enable an accurate extension.

B. Limitations of Out-of-sample Extension for Anomaly Detection

The popular methods for out-of-sample extension are based on interpolation. They are all a variety of calculating the value for a new sample by weighted sum of the values of the test data points in Γ, with the weights dependent on the Euclidean distance between the data points. This is a limitation of extension methods when applied to anomaly detection. In a case where there are no anomalies in Γ and it consists only of examples from a single n-dimensional cluster (the background), then the eigenvectors capture only the relaxation process within this cluster [30]. If the anomaly is not at least partially represented in the subset Γ, the values of the diffusion map will not capture the nature of the anomaly. Extension of the diffusion map to anomaly data points will give these points diffusion coordinates which are not meaningful in separating them from the background. All anomalies or data points which are far removed from the test set, will not be extended to appropriate coordinates representing their distance from the test set. Anomaly detection when the anomaly is not included in 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.

The size of the data set for images is very large. Even for a small image of 100×100 pixels there are 10,000 data points. Therefore, it can be inefficient to construct a diffusion map using all the pixels in the image, especially for high-resolution images. Instead, it is a common approach to construct the diffusion map for an image using a subset of random samples [15], [19]. The subset is embedded in a lower dimensional representation using the first several eigenvectors and then the diffusion map coordinates are extended to all patches in the image using an extension method. If the set of random samples does not include the anomaly, the diffusion map will not capture the difference between the anomaly and the background. Therefore, the out-of-sample extension of the diffusion map to the pixels in the anomaly region will not succeed in separating them from the background. These pixels will be assigned diffusion coordinates which represent the background and the anomaly detection will fail.

IV. Multiscale Diffusion based Anomaly Detection

We propose a multiscale approach combining spectral-based dimensionality reduction and nearest-neighbor-based anomaly detection. Diffusion maps are used to find a lower dimensional representation of the image. Due to the successful use of diffusion maps for spectral clustering, our assumption is that the anomaly regions will be well separated from background regions in the new embedding. In the embedding, background pixels will have similar diffusion coordinates, lying in a dense neighborhood, whereas the anomalies are separated from the background and lie in a low density neighborhood. This enables using a nearest-neighbors based approach in the lower dimensional embedding to determine which pixels are anomalies and which are normal. This approach is based on the assumption that normal data points appear in dense neighborhoods, whereas anomalies lie in neighborhoods with low density [25]. One challenge of such an approach is the computational complexity of computing the distance of each test instance with all other instances, in order to compute its nearest neighbors. Calculating the distances using the low-dimensional diffusion representation, greatly reduces the complexity of the distance computation. Also, as noted in Section II, calculating the distance between points in their diffusion coordinates, i.e., the diffusion distance, has been shown to be robust to noise. These steps are performed in a multiscale framework to overcome limitations of under-sampling the image and out-of-sample extension to the entire image.

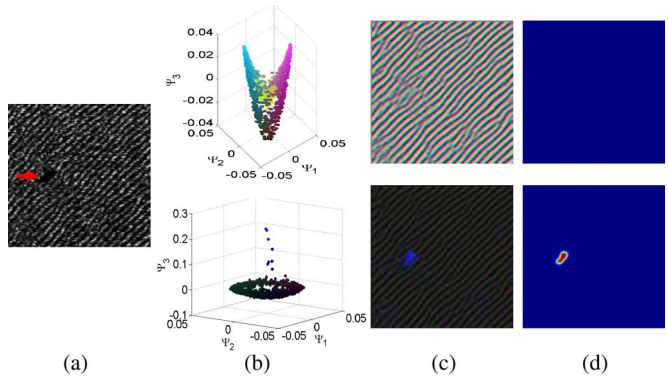


Fig. 1. Demonstration of the affect of random sampling on the diffusion map and the detection results. Results are shown for two different sampling distributions in the top and bottom row. (a) Side-scan sonar image of a sea-mine, visible as the dark shadow and indicated by a red arrow. In (b),(c) the first three coordinates in the diffusion are associated with RGB color in order to display the connection between the location of the pixel in the image, and its diffusion coordinates. (b) First three diffusion map coordinates. (c) Image pixels colored according to the RGB color associated with the first three coordinates of the diffusion map given in (b). (d) Anomaly score.

In Section IV-A we present three anomaly detection methods based on diffusion maps, using a single resolution of the image. We describe the disadvantages of these methods in terms of performance and computational complexity. In Section IV-B, we propose a multiscale anomaly detection method which overcomes the limitations of applying diffusion maps to images. In Section IV-C we describe the implementation details of our algorithm. We compare the performance of our multiscale method with each of the single-scale methods in Section V.

A. Single-Scale Anomaly Detection

One may consider three simple methods for applying diffusion maps to anomaly detection in images, while avoiding the limitations of under-sampling. The first is to apply the process of constructing a diffusion map and detecting anomalies in the low-dimensional embedding several times, for different subsets of random samples. The results can be fused together to detect the anomalies. This method avoids the problem of being too dependent on the random samples. However, it is computationally intensive and the number of times this would have to be performed until the anomaly was detected is unknown, due to the randomness of the sampling. Therefore, this method may result in a miss-detection. An example is displayed in Fig. 1. Fig. 1(a) presents a side-scan sonar image of a sea-mine on a periodic background. The sea-mine is indicated by the red arrow. Two subsets of random samples are used for the image, yielding very different detection results. In the top row there is a miss-detection and in the bottom row there is a positive detection. Note

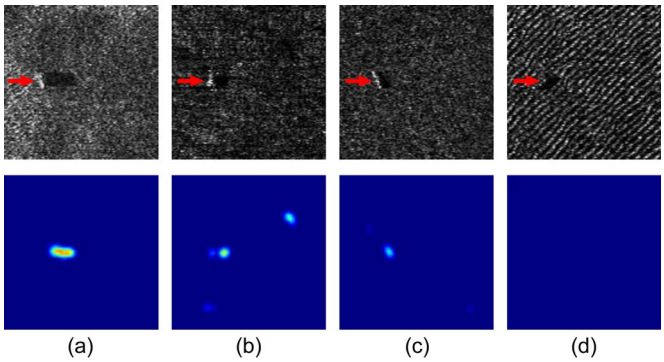


Fig. 2. Top row: original side-scan sonar images, the sea-mines are indicated by red (white in print) arrows. Bottom row: Anomaly score for detection based on coarse resolution of the images. The images were down-sampled by a factor of 2, and a third of the pixels were sampled in the construction of the diffusion map. In (a) the detection is successful. However, this method may result in false alarms (b), low anomaly score (c) or a miss-detection (d).

both subsets have the same number of samples. The diffusion maps for the two sampling schemes are shown in Fig. 1(b). The first three coordinates in the diffusion map (4) are associated with RGB color in order to display the connection between the location of the pixel in the image, and its diffusion coordinates. Each point in the three-dimensional space is assigned RGB values, by applying a simple transform from the diffusion coordinates to RGB values $[0, 255] \times [0, 255] \times [0, 255]$. Then, each pixel in the image is colored (Fig. 1(c)) according to the RGB value assigned to its diffusion coordinates (Fig. 1(b)). Note that this coloring is only for display purposes. In the top row, the diffusion map (Fig. 1(b)) captures the periodic nature of the data, but the anomaly is not sampled sufficiently and is not distinct in the diffusion coordinates. When the diffusion map is extended to the entire image shown in Fig. 1(c), the pixels of the anomaly are given coordinates representing the background, and the anomaly is not visible. Calculating the anomaly score, Fig. 1(d), yields there are no anomalies in the image. In the example on the bottom row, a different subset of random samples is used. In this case, the diffusion map Fig. 1(b) captures both the anomaly and the periodic nature of the background, and separates the anomaly from the background. The anomaly score in Fig. 1(d) displays the existence of an anomaly in the image. These examples demonstrate that the success of the diffusion map in capturing the nature of the anomaly is dependent on the pixels included in Γ. For this image, in average only one out of every five random subsets yielded a detection of the anomaly, when the size of the subset was 15% of the pixels.

A second approach is to perform the detection on a coarser resolution of the image. The advantage of using a coarse resolution is that a higher percentage of samples can be used since the image is down-sampled, and it is more likely that the anomaly will be properly sampled. A disadvantage of this approach is that the chosen scale may limit the ability to detect small anomalies. Also, since the fine details are blurred, the anomaly may be less distinctive from the background. This will require lowering the detection threshold which will result in more false-alarms. An example of anomaly detection on a coarse scale is shown in Fig. 2. The original side-scan sonar images are presented in the top row and the anomaly score for each image is displayed on the bottom. In Fig. 2(a) the detection is successful. In Fig. 2(b), the anomaly is detected as well as other regions in the background. Successful detection of the anomaly in this case, would detect false alarms as well. In Fig. 2(c), the anomaly received a low score. In order to keep the detection rate high, a low threshold would be necessary, which could cause false alarms in other images. The anomaly in Fig. 2(d) is not detected at all.

A third possibility is to divide the image into several sub-images, and perform anomaly detection on each sub-image separately. For each sub-image, a high percentage of samples can be used to avoid sub-sampling. This method is computationally intensive since it requires the calculation of a diffusion map for every sub-image. In addition, it can cause a higher false alarm rate. The reason for this is that regions which are unique in their immediate surroundings, yet similar to other regions in the image, will be treated in separate sub-images and can be detected as anomalies. Also, the anomaly itself might be split between sub-images, making it smaller in each sub-image and reducing the detection rate. To avoid this, the image will have to be divided into overlapping sub-images, raising the computation complexity even more. Finally, even if the sub-image itself is rather homogeneous, the nature of the diffusion maps is that the embedding for such a sub-image will include the inner-cluster variations, and cause possible false alarms.

B. Multiscale Anomaly Detection

Our method aims at reducing the computational complexity while improving the detection rate. To overcome the limitations of random sampling, we propose a multiscale approach. Assume that the anomalies in the image are larger than a single pixel. Therefore, they can be detected at several resolutions of the image. At a lower resolution, it is computationally possible to sample a larger percentage of the image. Thus, detecting an anomaly at a lower resolution is less likely to fail due to sampling. We propose to take advantage of the anomaly detection at different scales to overcome the limitations of random sampling. Since our method performs anomaly detection at different resolutions of the image, even if the anomaly is missed on a coarse level, for example since it is too small at that level, it can still be detected on the following finer levels. In addition, it is possible to lower the threshold for anomaly detection on the coarser levels, since this will not harm the false alarm rate as a decision is only reached at the full-scale level. Thus we are able to detect anomalies on the higher levels, even at the cost of detecting more false alarms, since these false alarms will be removed at the final level.

Our multiscale approach is based on constructing a Gaussian pyramid [32] representation of the image. Starting with the coarsest scale, a diffusion map is constructed, based on a subset of the data set. Since the image is smaller at this scale, a larger percentage of the image can be sampled for the construction of the diffusion map, perhaps even all pixels. Then, an anomaly score is used to determine which pixels are anomalies at this level. These pixels are used as input to the next level as the pixels at the finer level corresponding to the anomalous pixels at the coarser level are included in Γ. The rest of the pixels in Γ are sampled randomly from the image. This algorithm continues from level to level, with each previous level providing prior information on which samples of the data set are used to

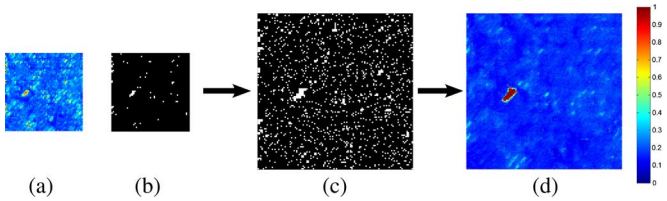


Fig. 3. An example how the anomaly score for a certain level of the pyramid, affects the sampling in the next level. (a) Anomaly score C_l. (b) Suspicious pixels obtained from thresholding C_l. (c) Γ_{l+1} is determined by the suspicious pixels from level l and random pixels. (d) Anomaly score C_{l+1}.

construct the diffusion map. This approach greatly increases the detection rate of the diffusion-based anomaly detector.

Our approach is less computationally intensive than a single-scale detector using an equivalent amount of samples, since on the coarser scales, smaller patches can be used as features, reducing computation time of the calculation of the affinity matrix. Also, the detection process is faster on a coarser scale.

The anomaly score itself is based on a nearest neighbor approach. In the low dimensional embedding, background pixels will have similar diffusion coordinates, lying in a dense neighborhood, whereas the anomalies are separated from the background and lie in a low density neighborhood. The diffusion distances in the low dimensional embedding can be used in a measure of the density of the neighborhood of each pixel, determining which pixels are anomalies and which pixels are normal. Using the diffusion distance in a nearest neighbor approach is both computationally efficient compared to the calculation in the original dimensionality and robust to noise.

C. Implementation

Given an image I, the Gaussian pyramid representation of the image is computed, yielding $\{G_l\}_{l=0}^{L}$, where G_0 is the original image and G_L is the coarsest resolution. At each level l, G_l is calculated by convolving the image from the previous level G_{l-1} with a Gaussian low-pass filter and then down-sampling by a factor of two. Starting with G_L, a subset Γ_L of random pixels is sampled from the image. Since the image at this level is at very low resolution, the subset can include all pixels, if it is feasible given memory constraints. The diffusion map is calculated using this subset, and extended to the remaining pixels. Then, an anomaly score C_L is calculated for all pixels. A threshold τ_l on the anomaly score is used to mark suspicious pixels. We then proceed to the image G_{L-1}. On this level, pixels which correspond to the suspicious pixels found in G_L are included in Γ_{L-1}. The rest of the pixels in the subset are chosen at random.

The threshold τ_l used at the output of each level is chosen to be the 95th percentile of the anomaly score for that level. If the image does not hold an anomaly this will result in random samples with the highest anomaly scores. If the image holds an anomaly, the anomaly will have a high score compared to the rest of the image and it will be sampled more densely in the next level. An example of this process is shown in Fig. 3. Fig. 3(a) shows the calculated anomaly score for level l. Thresholding this score yields a group of suspicious pixels, including both the anomaly and some background pixels. The corresponding pixels on the next level, $l-1$, are included in Γ_{l-1}. The rest of the pixels are randomly sampled. Calculation of the diffusion map and its extension to the image, yields the anomaly score C_{l-1}, in which only the anomaly received a high score, separating it from the background.

The process of sampling, dimensionality reduction and anomaly detection repeats for every level, with the output of each level serving as input to the next level, determining the samples in Γ_l. At the full-scale level G_0, the anomaly score for each pixel determines the existence of anomalies in the image. We use a hard threshold τ on C_0 and then smooth the resulting image. Anomalies have a high score, close to 1. Fig. 4 presents a flowchart of the algorithm.

At each level, the affinity matrix is calculated for the subset Γ_l using (8), with the scaling parameter set as explained in Section II-A. In order to reduce computation time and memory requirements, the matrix is calculated using k nearest neighbors, i.e. patch x_i is connected to patch x_j if x_i is among the k nearest neighbors of x_j or vice-versa. Otherwise $w(x_i, x_j) = 0$, as in [28]. This enables the matrix to be sparse.

The anomaly score for each level is calculated based on a nearest-neighbor approach. This requires calculating the distance to each point's nearest neighbors. Calculating the distances using the low-dimensional diffusion representation, greatly reduces the complexity of the distance computation. To further reduce the complexity, we take advantage of the spatial nature of the original data. We limit ourselves to computing the diffusion distance between each pixel to the pixels in a window surrounding it. Our method is similar to the one presented in [7], [33], where anisotropic kernels were used for defect detection in images of wafers, given a clean reference image. There, anisotropic kernels were used to measure the similarity of a patch in a test image to patches in a window in a reference image. In our approach, instead of calculating the similarity between a patch in a test image and patches in a clean reference image, we compare the test image to itself. In addition, we compare the patches in the embedded diffusion coordinates $\Psi(x_i)$, using the diffusion distance $D^2_{t=1}(x_i, x_j)$ between patches as a similarity measure.

An affinity measure is calculated between each pixel i and the m pixels in the window surrounding the pixel. Similarly to [13], [14], the affinity measure is defined using a Gaussian kernel $\bar{w}$ based on diffusion distances:

$$\overline{w}(i,j) = \exp\left(-\left\|\Psi(x_i) - \Psi(x_j)\right\|^2 / \overline{\sigma}\right),\ 0 < \overline{w}(i,j) \le 1. \tag{19}$$

Unlike the kernel in (1) which relies on the Euclidean distance between grayscale levels of the patch, this kernel relies on diffusion distances.

As opposed to the local scale (9) used in the affinity measure (1), here a single global scale is required for $\overline{\sigma}$. Using a local scale as described in (9), which relies on the distance to the K-th nearest neighbors, would result in each point having approximately K neighbors. Here we do not want to overcome the difference in neighborhood densities between data points. Instead, our purpose is to utilize this difference to detect the anomalies by finding which data points are far removed from their neighbors on the low-dimensional manifold. This scale greatly influences the results as it determines how close pixels are in the

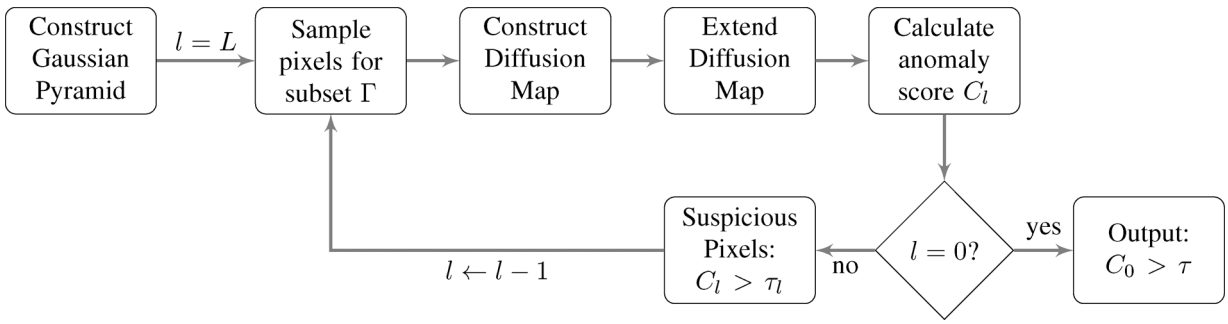


Fig. 4. Flowchart of the multiscale algorithm.

diffusion embedding. Too small a scale will result in all pixels being different and too large a scale will result in the anomaly being considered similar to the background.

We set the scale by the following procedure. We select n_{pair} pairs of pixels in the image and calculate the diffusion distance for each pair: $\|\Psi(x_i) - \Psi(x_j)\|$. Since these are random pixels from the image, most, if not all of them, are background pixels. Thus, these distances represent typical diffusion distances between pixels in the image. The empirical variance of these n_{pair} distances is σ^2_{pair}. We set the scale to be $\overline{\sigma} = r\sigma^2_{\text{pair}}$. The parameter r determines how close we want two normal points to be in the diffusion embedding. This procedure enables a method of automatically setting the scale, with negligible computation time, and gave good empirical results.

To determine whether a pixel is an anomaly, we use the total similarity measure presented in [7], [33]. Our anomaly score of a tested pixel i is defined as

$$C_l(i) = 1 - \frac{1}{m} \sum_{j \in N_i} \overline{w}(i,j). \tag{20}$$

Pixel i is compared to its neighbors $\{j\}$ in the spatial neighborhood denoted N_i, with m being the number of pixel in N_i. The neighborhood N_i is a square window surrounding pixel i of size $2W + 1$ in each dimension. The inner part of the window surrounding the tested pixel is masked, and only the pixels in the outer window are used. Let $2M^{\text{mask}+1}$ be the size of the mask in each dimension. Then the pixel i is compared to all pixels $\{j \in N_i | M^{\text{mask}} \leq d(i,j) \leq W\}$, where $d(i,j)$ is the Manhattan distance. The reason for masking the inner pixels in the window is that we do not want to compare the pixel to its immediate neighbors, since we assume the anomaly is larger than a single pixel. If a pixel belongs to an anomaly, its surrounding pixels are also anomalous and they may all have similar diffusion coordinates, compared to the background pixels. Therefore, if the window is too small, the anomalous pixel will receive a low anomaly score, due to its affinity to its immediate spatial neighbors in the image. To avoid this, the inner pixels are masked and ignored and the window surrounding each pixel is chosen to be large enough in comparison with the expected size of an anomaly.

The sum $\sum_{j \in N_i} \overline{w}(i,j)$ can be seen as a smoothed estimate of the number of close neighbors the data point i has in the window surrounding it, where the notion of closeness is determined by the diffusion distance. Pixels which are anomalous have few close neighbors in the diffusion embedding and therefore a very high anomaly score. Pixels with a low anomaly score are similar to the pixels in the window surrounding them. The size of the window and the masked area should be determined by the application and prior knowledge of the size of possible anomalies.

V. Experimental Results

We demonstrate the proposed algorithm on real sea-mine 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. We compare the multiscale detector with five variations of a single-scale detector, to demonstrate the improvement gained by our multiscale approach.

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) aside a dark region (shadow). The shadow is due to the object blocking the sonar waves from reaching the seabed. Typically, the shadow region is larger than the highlight region in the image.

Research in this field focuses on two aspects of the problem: detection of mine-like-objects (MLO) in the image and classification of these objects as mine or non-mine. Algorithms proposed for detection of the MLOs include MRF models for modeling the background [34], [35], a 2-D multiscale GMRF with matched subspace detector (MSD) [4], a multidimensional GARCH model with MSD [3], non-linear matched filters [36], etc. The detection is sometimes accompanied by extraction of the shadow, for example using snakes [34]. The detection of the shadow increases the ability to correctly classify mines and non-mines.

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 [35], [37]. In [3], a few examples of sea-mines are used for creating the anomaly subspace for the MSD. Our diffusion-based approach does not require a training set and makes no assumptions regarding the appearance of the mine and its shadow in the image. The only prior information used is that the expected size of the sea-mine is approximately 15 pixels by 3 pixels. This information is used in determining the size of the surrounding window and mask for each pixel, as explained in Section IV-C.

TABLE I
PARAMETERS USED IN MULTISCALE DETECTOR

Pyramid Level	Image size	Patch size	Embedding Dimension	Percentage of pixels in subset	Window Size (W)	Mask Size (M^{mask})
0	200x200	8x8	6	0.10	41x41 (20)	9x9 (4)
1	100x100	4x4	6	0.33	21x21 (10)	5x5 (2)
2	50x50	2x2	3	0.5	13x13 (6)	5x5 (2)

We evaluated our algorithm on a set of 28 side-scan sonar images with sea-mines, each image sized 200 $\times$ 200 pixels. For the multiscale detector, we used a Gaussian pyramid of $L = 3$ levels. The parameters used in the multiscale detector are given in Table I. To allow efficient computation times, the affinity matrix was calculated using exact k-nearest-neighbor search with 16 neighbors for each point, resulting in a sparse weight matrix. For the k-nearest neighbors search we use the Matlab function *pdist2*, which uses the exhaustive search method to find the exact k-nearest neighbors. For the spectral decomposition of sparse matrices we use the Matlab function *eigs*. We choose the global scale in (19) to be $\overline{\sigma} = 20\sigma^2_{\text{pair}}$. Note that the size of the images used in our results enables denser sampling of the image than what we used. We intentionally use a small percentage of the pixels in the image to demonstrate that this framework is applicable also for larger images. Using diffusion maps for larger images requires small subsets at the full scale level, due to memory constraints in calculating the affinity matrix.

We compared the performance of our multiscale algorithm (MS) with five single-scale sampling schemes:

1) SS1: 10% of the image was randomly sampled to construct the diffusion map.
2) SS2: 20% of the image was randomly sampled to construct the diffusion map.
3) SS3: The images were blurred with a Gaussian filter and down-sampled to 100 $\times$ 100. 30% of the image was randomly sampled to construct the diffusion map.
4) SS4: The image is divided into 16 overlapping sub-images. A diffusion map is constructed for each sub-image using all the pixels in the sub-image such that no out-of-sample extension is necessary. Anomaly detection is performed on each sub-image separately and for the overlapping pixels, the maximal anomaly score is taken.
5) SS5: The diffusion map is calculated for the entire image at once, without the need for performing sampling and out-of-sample extension. This is done using RANN [22], a recently proposed fast approximate nearest neighbors algorithm. The sparse affinity matrix is calculated for all 8 $\times$ 8 patches using 16 neighbors for each patch. We used 5 iterations of RANN and supercharging; for details about these parameters the reader is referred to [22].

In SS1 the parameters were chosen to be identical to that of the multiscale detector for level $l = 0$, given in Table I. SS2 is intended to demonstrate the effect of using more samples. In addition, the number of samples used in this scheme is equivalent to the total number of samples used in the multiscale detector. We demonstrate that for the same number of samples, the multiscale detector achieves superior results. SS3 has identical parameters to the middle-scale level, $l = 1$, of the multiscale detector given in Table I. This demonstrates the effect of a decimated scale, in which the fine details are blurred, on the detection. In SS4, the dependence on random samples is completely removed. In each sub-image, all pixels are used in the construction of the diffusion map. Instead of using a window surrounding each pixel, the pixels in a sub-image are compared to all other pixels in the sub-image in the calculation of the anomaly score. To avoid border issues, the sub-images are overlapping. In SS5, the dependence on random samples is completely removed. Comparing SS5 with our method demonstrates the effect the multiscale driven sampling has on the final full-scale diffusion map compared to a diffusion map calculated for all points together.

TABLE II
NUMBER OF TRUE POSITIVE FOR GIVEN NUMBER OF FALSE ALARMS

	size=5			size=20		
	FA=7	FA=4	FA=0	FA=7	FA=3	FA=0
MS	100%	89%	89%	100%	86%	82%
SS1	61%	39%	0%	57%	43%	14%
SS2	68%	54%	0%	68%	61%	29%
SS3	61%	61%	29%	61%	57%	29%
SS4	89%	79%	64%	93%	93%	79%
SS5	93%	93%	**	93%	93%	86%

Detections are found by thresholding the anomaly score image resulting in a binary image. A detection is a connected component in the binary image. We considered detection of the sea-mine to be a true positive (TP) and any other detections to be false alarms (FA). The size of the connected component can be used to reject noisy detections. We compare two thresholds on the size of the detection: 5 pixels and 20 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 compared the number of TPs for each method for a given FA rate. Results are given in Table II. Our multiscale approach has the highest TP rate. In SS2, using twice as many samples than in SS1, results in a better detection rate, but at a high computational cost. In addition, the difference in detection for using twice as many samples is not dramatic. Most importantly, it does not overcome the limitations of sub-sampling the image, as the multiscale detector which uses the same number of samples, has a significantly better detection rate. This is due to the propagation of information from level to level. SS3 shows better results than both SS1 and SS2, as it has a lower FA rate. This demonstrates that different scales of the image are useful in detecting the anomalies and combining this information as in our multiscale approach, gives the best results. SS4 demonstrates results which are comparable to that of our multiscale approach. However, as explained in Section IV-A, this method has various false alarms in the background, due to the limited region each diffusion map is calculated for. This results in a higher false alarm

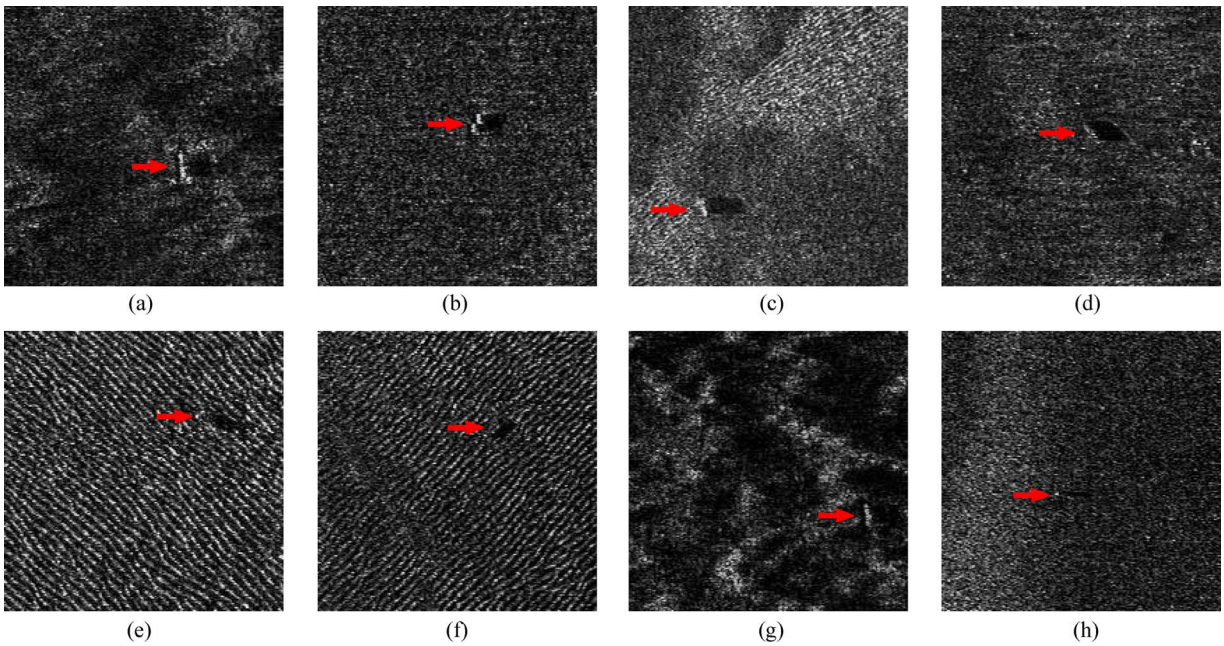


Fig. 5. Side-scan sonar images of sea-mines. The sea-mine locations are marked with a red (white in print) arrow.

rate to ensure positive detection of the anomalies. In addition, such a method faces scalability issues when applied to larger images.

The SS5 approach also gave results which are comparable to that of our multiscale approach. However, for two images, Fig. 5(g) and (h), the SS5 approach was unable to detect the sea-mines at all, even for a detection threshold as low as $\tau = 0.3$. In addition, the MS approach has better results for a small threshold. In fact, for the low threshold on detection size, there was no threshold on the SS5 anomaly score which resulted in zero FAs. For a threshold of $\tau = 1$, the results of SS5 were 3 FAs and 43% detection rate. This because the SS5 results had more small FAs with very high anomaly score values, than the MS algorithm. Therefore it is harder to get a good detection rate with low FAR for a small size anomaly. On the other hand, for the higher threshold on anomaly size, the results of the SS5 approach are slightly better. This is because for a few of the images, the number of anomalous pixels which received a high anomaly score in the MS method was smaller compared to the SS5 method. Therefore, there were more TPs for the SS5 method, for the larger anomaly size and low FAR rate.

Eight of the tested images are shown in Fig. 5. Each image contains one sea-mine on highly cluttered seabed background. The background patterns are diverse. Some appear as noise (Fig. 5(b), (d), and (h)) whereas others contain relatively slow changing backgrounds (Fig. 5(a)). Images with a rapidly changing background (Fig. 5(g) and (c)) or dominant periodical pattern (Fig. 5(e) and (f)) are especially difficult. Also, the size of the mine and its shadow differ from image to image, as well as its orientation. For example, in Fig. 5(a) the mine is quite large, whereas in Fig. 5(h) the mine is very small and its shadow is also thin. In most images, the highlight is rather bright, yet in image Fig. 5(d) its intensity is similar to that of the noise and in Fig. 5(f), the mine highlight is not visible at all, with only its shadow seen in the image.

TABLE III
AVERAGE RUNNING TIMES OF THE ALGORITHM IN SECONDS, COMPARING MULTISCALE APPROACH WITH SINGLE-SCALE EXACT AND APPROXIMATE NN APPROACH

Pyramid Level	Dimensionality Reduction	Anomaly Detection	Total
0	23.70	43.77	67.54
1	5.47	3.77	9.26
2	1.00	0.66	1.67
Total	30.18	48.20	78.47
RANN	10.53	43.70	54.30
Exact NN	190.00	43.79	233.90

Results of the multiscale detector are presented in Fig. 6 and for the single scale detector SS1 in Fig. 7. Positive detection of the sea-mines is achieved in all displayed images using the multiscale detector. The single scale detector on the other hand, does not detect any anomalies in Fig. 7(e)–(h). The single scale detector also suffers from a higher false alarm rate, as can be seen in Fig. 7(a). The multiscale detector has a single false alarm in image Fig. 6(d), on a small shadow in the image. This same false alarm is detected by the single scale detector. The multiscale detector performs very well in detecting both the sea-mine and its highlight in the image, demonstrated for diverse, challenging backgrounds and various sea-mine sizes and orientations.

In Table III, we report the average total running time and the running time for the two parts of the MS algorithm: dimensionality reduction, including sampling the image, constructing a diffusion map and out-of-sample-extension to all image patches, and anomaly detection in the reduced dimensionality. We compare the runtime of the MS algorithm with those of two single-scale schemes: calculating the affinity matrix for the entire image using RANN method (SS5) and using Matlab's exact NN method. In these schemes dimensionality reduction is based only on constructing the diffusion map, without the need

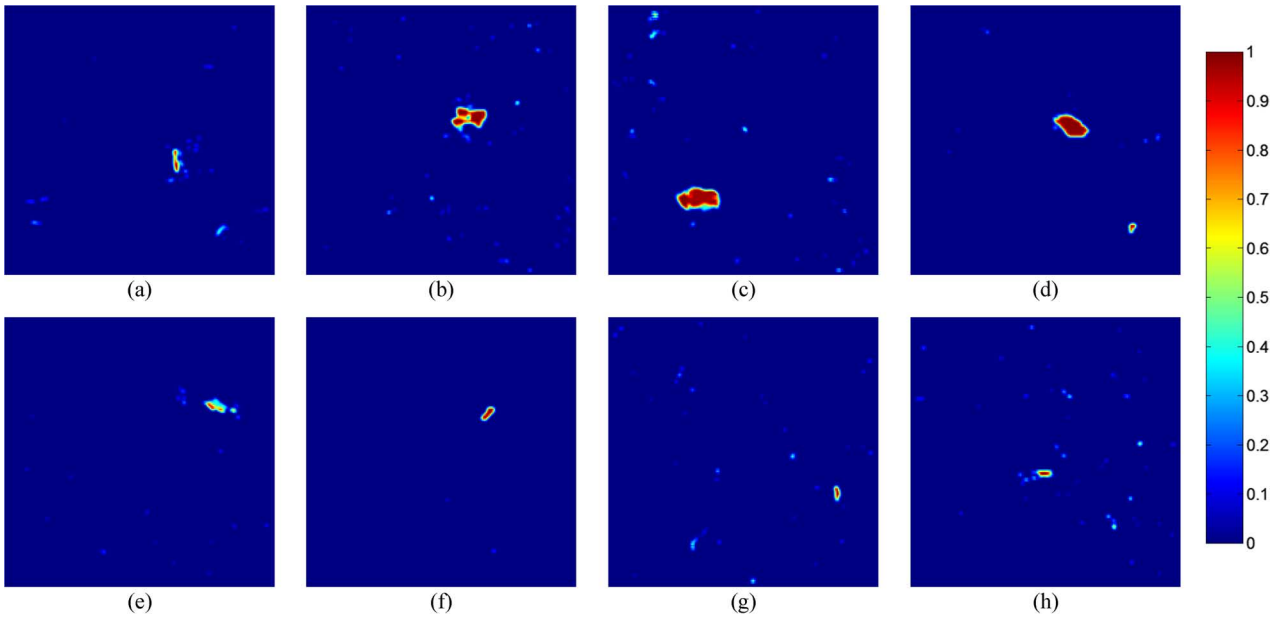


Fig. 6. Results of Anomaly Detection for multiscale detector, corresponding to the images displayed in Fig. 5.

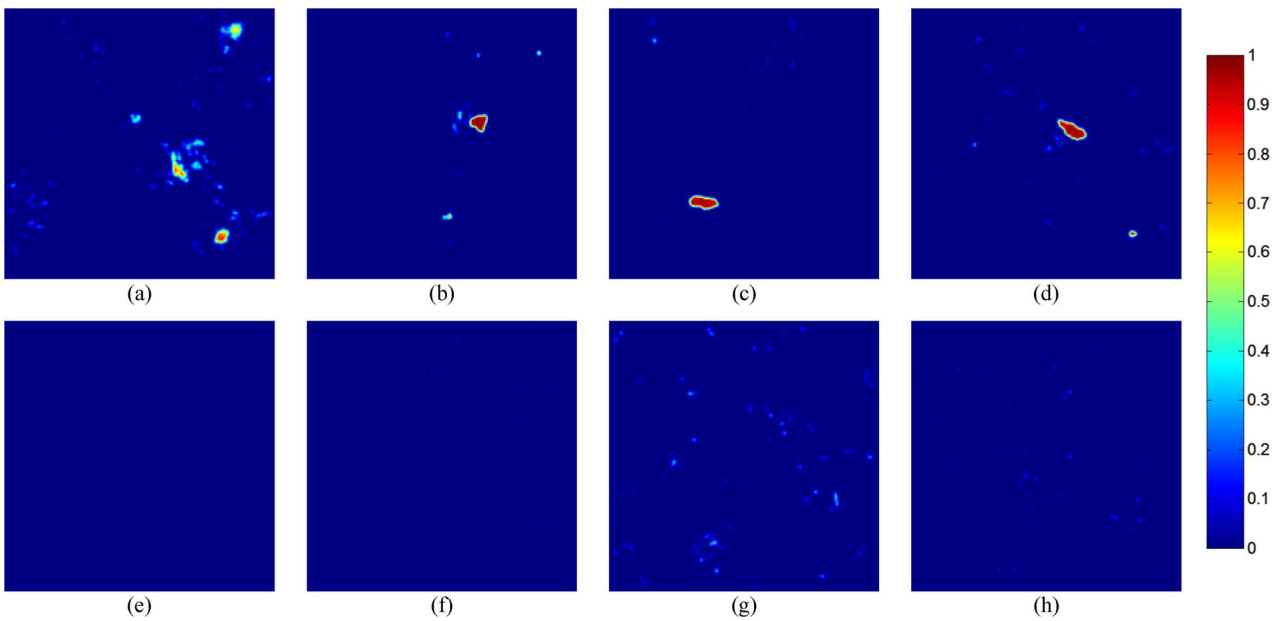


Fig. 7. Results of Anomaly Detection for single scale detector, corresponding to the images displayed in Fig. 5.

for out-of-sample-extension. Results are given in seconds. Our algorithm has been implemented in Matlab and the numerical experiments have been carried out on a Dell laptop computer, with an Intel Core i5 QuadCore CPU 2.67 GHz and 4.0 GB RAM. It should be noted that this a Matlab implementation and it has not been optimized for runtime. The RANN search been implemented in FORTRAN.

The results first enable us to compare between the MS algorithm and SS1, which is equivalent to the runtime of scale $l = 0$ of the MS approach. The MS takes about 15% longer but with greatly improved detection results. Next, comparing RANN with Matlab's exact NN search, the improvement factor in runtime using RANN for dimensionality reduction is around 18. The diffusion maps constructed by the two methods are not identical, as RANN does not always return the true nearest neighbors, so the sparse affinity matrix is different. Overall, the detection statistics for both methods are very similar so we do not report those for the exact NN search in Table II. Based on this comparison, we can assume that using RANN in our multi-scale approach instead of the exact NN search should improve the runtime of our algorithm, without affecting the detection results. This will be verified empirically in future work. In such a framework, RANN will be used initially to construct the affinity matrix for the sampled points in Γ_l, and then a query will be run on each of the points in $\overline{\Gamma}_l$. The improvement in runtime entails a cost in memory on the order of $O(n_l \cdot T))$, with n_l being the number of points in Γ_l and T being the number of iterations used by RANN [22]. For the lower-resolution scales of the pyramid the improvement factor will be modest considering the small size of Γ and the low dimension of the points (small

patches are used as features). However, we expect a meaningful improvement for the higher-resolution scale.

VI. Conclusion

We have introduced an anomaly detection algorithm using diffusion maps representation of the data. Based on the clustering properties of the diffusion map, we proposed to detect anomalies in the reduced dimension based on a nearest-neighbor approach. To improve the detection process and ensure that the normal pixels and the anomaly regions are separable in the lower dimensional embedding of the data, we implemented a multiscale framework to overcome the possible limitations in using diffusion maps with out-of-sample extension.

The successful performance of the algorithm was demonstrated in automatic target detection in side-scan sonar images, which is a challenging task due to the high variability of the target and sea-bottom reverberation. The results show the capability of the proposed model and algorithm to cope with a variety of targets and background clutter patterns. The results also demonstrate the advantage of the multiscale framework over using only a single scale.

Although our algorithm is used in an unsupervised setting, it also has implications for using diffusion maps in a supervised setting, using out-of-sample extension to extend the diffusion map from a training set to a test set. Our results imply that constructing a training set using only background data points will not be successful in a supervised anomaly detection application. The anomalous data points will be assigned coordinates similar to those of the background, and the detection will fail in the lower dimensional embedding.

A possibility for future research is combining the anomaly scores from the different multiscale levels into a single anomaly score. We predict this will improve performance, as the coarser levels have information on the presence of the anomaly, which we currently disregard in our final output. This can also assist in detecting anomalies whose size differ from the expected size. In addition, computational complexity of the algorithm can be reduced by employing the RANN algorithm for computing the affinity matrix for the diffusion map, as explained in Section V.

Acknowledgment

The authors thank Ronen Talmon and Amit Oved for helpful discussions, and Andrei Osipov for providing the RANN code [22] and his insightful explanations on the RANN algorithm. The authors also thank the anonymous reviewers for their constructive comments and useful suggestions.

References

[1] Y. Chen, N. Nasrabadi, and T. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 629–640, Jun. 2011.

[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, May 2000.

[3] 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, May 2007.

[4] A. Goldman and I. Cohen, "Anomaly subspace detection based on a multi-scale Markov random field model," *Signal Process.*, vol. 85, no. 3, pp. 463–479, Mar. 2005.

[5] C. Spence, L. Parra, and P. Sajda, "Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model," in *Proc. IEEE Workshop Math. Meth. Biomed. Image Anal. (MMBIA'01)*, 2001, pp. 3–10.

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

[7] M. Zontak and I. Cohen, "Defect detection in patterned wafers using anisotropic kernels," *Mach. Vis. Applicat.*, vol. 21, no. 2, pp. 129–141, June 2008.

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

[9] 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.

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

[11] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators," in *Neural Information Process. Systems (NIPS) 18*. Cambridge, MA: MIT Press, 2005, pp. 955–962.

[12] 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, January 2009.

[13] R. Talmon, I. Cohen, and S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," in *Proc. 36th IEEE Internat. Conf. Acoust. Speech, Signal Process. (ICASSP-11)*, 2011, pp. 5084–5087.

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

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

[16] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1784–1797, Nov. 2006.

[17] 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.

[18] N. Rabin and R. R. Coifman, "Heterogeneous datasets representation and learning using diffusion maps and Laplacian pyramids," in *Proc. 12th SIAM Int. Conf. Data Mining*, 2012.

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

[20] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, Nov. 1998.

[21] C. Merkwirth, U. Parlitz, and W. Lauterborn, "Fast nearest-neighbor searching for nonlinear signal processing," *Phys. Rev. E*, vol. 62, pp. 2089–2097, Aug. 2000.

[22] P. W. Jones, A. Osipov, and V. Rokhlin, "Randomized approximate nearest neighbors algorithm," *Proc. Nat. Acad. Sci. (PNAS)*, vol. 108, no. 38, pp. 15 679–15 686, Sep. 2011.

[23] "Detection of anomaly trends in dynamically evolving systems," in *Proc. AAAI Fall Symp. Series*, 2010.

[24] N. Rabin and A. Averbuch, Hierarchical sensor fusion with applications for detection of anomaly trends in dynamically evolving systems, submitted for publication.

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

[26] 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.

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

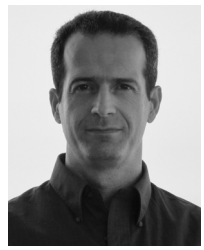
[28] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[29] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Neural Inf. Process. Syst. (NIPS) 17*, 2005, pp. 1601–1608.

[30] 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*. New York: Springer, 2007.
[31] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Jan. 2004.
[32] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.
[33] M. Zontak and I. Cohen, "Defect detection in patterned wafers using multichannel scanning electron microscope," *Signal Process.*, vol. 89, no. 8, pp. 1511–1520, Aug. 2009.
[34] S. Reed, Y. Petillot, and J. Bell, "An automatic approach to the detection and extraction of mine features in sidescan sonar," *IEEE J. Oceanic Eng.*, vol. 28, no. 1, pp. 90–105, Jan. 2003.
[35] 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, Nav.*, vol. 151, no. 1, pp. 48–56, Feb. 2004.
[36] G. Dobeck, "Algorithm fusion for automated sea mine detection and classification," in *Proc. MTS/IEEE Oceans Conf. Exhib.*, 2001, vol. 1, pp. 130–134, Marine Technol. Soc.
[37] Y. Petillot, Y. Pailhas, and J. Sawas, "Target recognition in synthetic aperture and high resolution side-scan sonar," in *Proc. Eur. Conf. Underwater Acoust. (ECUA 10)*, 2010, pp. 99–106.

Gal Mishne received the B.Sc. degree (summa cum laude) in electrical engineering and in physics in 2009 from the Technion—Israel Institute of Technology, Haifa. She is currently pursuing the M.Sc degree in electrical engineering at the Technion.

Her main areas of interests include signal processing, image processing, and computer vision. In 2009 she received the Wilk Family award from the Signal and Image Processing Lab (SIPL) at the Technion for an excellent project.



Israel Cohen (M'01–SM'03) is an Associate Professor of electrical engineering at the Technion—Israel Institute of Technology, Haifa, Israel. He received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001 he joined the Electrical Engineering Department of the Technion. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2008), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), a coeditor of *Speech Processing in Modern Communication: Challenges and Perspectives* (Springer, 20010), and a general co-chair of the 2010 International Workshop on Acoustic Echo and Noise Control (IWAENC).

Dr. Cohen is a recipient of the Alexander Goldberg Prize for Excellence in Research, and the Muriel and David Jacknow award for Excellence in Teaching. He serves as a member of the IEEE Audio and Acoustic Signal Processing Technical Committee (AASP TC) and the IEEE Speech and Language Processing Technical Committee (SLTC). He served as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as Guest Editor of a special issue of the *EURASIP Journal on Advances in Signal Processing* on Advances in Multimicrophone Speech Processing and a special issue of the *Elsevier Speech Communication Journal* on Speech Enhancement.