MULTISCALE ANOMALY DETECTION USING DIFFUSION MAPS AND SALIENCY SCORE

Gal Mishne and Israel Cohen

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

ABSTRACT

Recently, we presented a multiscale approach to anomaly detection in images, combining diffusion maps for dimensionality reduction and a nearest-neighbor-based anomaly score in the reduced dimension. When applying diffusion maps to images, usually a process of sampling and out-of-sample extension is used, which has limitations in regards to anomaly detection. To overcome the limitations, a multiscale approach was proposed, which drives the sampling process to ensure separability of the anomaly from the background clutter. In this paper, we propose a new anomaly score used in the diffusion map space, which shows increased performance. We show that this algorithm enables improved detection when tested on side-scan sonar images of sea-mines and compare it with competing algorithms.

***Index Terms*—** anomaly detection, diffusion maps, dimensionality reduction, multiscale representation, automated mine detection

1. INTRODUCTION

Anomaly detection is important in many applications in image processing, such as automatic target recognition 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 essential in military applications and automation of quality assurance processes, as the user will be shown only suspicious objects, thus saving valuable time.

Anomaly detection in images is challenging due to the large size of the data set, the presence of noise and high dimensionality of the data [8]. Also, it is usually difficult to obtain labeled data for anomaly detection and the data sets tend to be unbalanced due to the sparseness of anomalies as compared to normal data. This makes unsupervised methods preferable to supervised ones. There are many approaches to anomaly detection in images based on statistical models [2–5], machine learning, saliency based methods [6], sparse representations [1], and more. Such approaches usually require modeling the data or using training data to learn the background or train a dictionary.

Recently, we presented a multiscale algorithm for anomaly detection based on dimensionality reduction using diffusion maps [9]. This approach is data-driven, with the separation of the anomaly from the background arising from the intrinsic geometry of the image, revealed through the use of diffusion maps. The algorithm is unsupervised and requires no prior knowledge regarding the appearance of the anomaly or the background, and no use is made of training data or reference images. The anomaly score is based on local density approximation of the pixels in the reduced dimensionality. It is calculated using the noise-robust diffusion distance, which enables the algorithm to handle very noisy images. Sampling and out-of-sample extension are common practice in applying diffusion maps to images due to the large size of the data set [10, 11]. We show in [9] that this process can limit the success of the dimensionality reduction in revealing the presence of anomalies in the data. To overcome these limitations, a multiscale approach is used, which drives the sampling process to ensure separability of the anomaly from the background clutter.

In this paper, we propose a new anomaly detection score inspired by the saliency map proposed by Goferman, Zelnik-Manor and Tal [12]. The problems of anomaly detection and saliency in images are closely related, where an anomaly can be viewed as a salient object in the image. However, while saliency is typically important in natural images, anomaly detection is usually performed in images which are not natural: multispectral, sonar, microscopy, medical, etc. These images tend to be noisy. Therefore, a feature space which is robust to noise is more appropriate in this context. We adapt the dissimilarity measure used in [12] to the diffusion map feature space used in our algorithm. This new anomaly score has improved performance, requires less parameters and is better at suppressing background regions which are similar to each other yet spatially distant.

The paper is organized as follows. Sec. 2 reviews the diffusion map framework for dimensionality reduction and Sec. 3 describes out-of-sample extension methods and their limitations in anomaly detection. In Sec. 4, the proposed multiscale algorithm is presented. Finally, Sec. 5 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

Real world data is usually represented with features of high dimensionality, yet can be shown to lie on low-dimensional manifolds. Finding a low-dimensional representation of the data is necessary to efficiently handle it and the representation usually reveals meaningful structures within the data. In recent years, a large number of nonlinear techniques for dimensionality reduction have been proposed [13–16]. Some of these methods are spectral methods, based on the eigenvectors of adjacency matrices of graphs on the data [15, 16]. These methods take into account the geometry of the data set and the representation they yield preserves local neighborhood information. Diffusion maps [16] is one such technique, based on the construction of the graph Laplacian of the data set.

Let $\Gamma = \{x_1, .., x_n\}$ be a high-dimensional set of n data points. A weighted graph is constructed with the data points as nodes and the

This research was supported by Robert H. Hillman Foundation for Global Security - Collaboration Technion and University Northeastern, and by the Israel Science Foundation (grant no. 1130/11).

weights of the edges connecting two node is a measure of the similarity between the two data points. The affinity matrix $\mathbf{W} = w(x_i, x_j)$, $x_i, x_j \in \Gamma$ is required to be symmetric and non-negative. A common choice is an RBF kernel $w(x_i, x_j) = \exp\left\{-\|x_i - x_j\|^2/\sigma^2\right\}$, where $\sigma > 0$ is a scale parameter. Then, a random walk is created on the data set by normalizing the kernel:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}, \tag{1}$$

where $\mathbf{D}(i,i) = \sum_{j\in\Gamma} w(x_i, x_j)$. The row-stochastic matrix $\mathbf{P}$ satisfies $p(x_i, x_j) \geq 0$ and $\sum_{j\in\Gamma} p(x_i, x_j) = 1$ and can be viewed as the transition matrix of a Markov chain on the data set Γ. The spectral decomposition of $\mathbf{P}$ yields that t steps of the Markov chain can be presented as

$$p_t(x_i, x_j) = \sum_{l\geq 0} \lambda_l^t \psi_l(x_i)\phi_l(x_j), \tag{2}$$

where ϕ_l and ψ_l are the biorthogonal left and right eigenvectors, respectively, and $|\lambda_0| \geq |\lambda_1| \geq ... \geq 0$ are the sequence of eigenvalues.

A diffusion distance $d_{DM}(x_i, x_j; t)$ between two points $x_i, x_j \in \Gamma$ is defined by

$$d_{DM}(x_i, x_j; t) = \sum_{x_k\in\Gamma} \frac{\left(p_t(x_i, x_k)) - p_t(x_j, x_k)\right)^2}{\phi_0(x_k)}. \tag{3}$$

This measures the similarity of two points according to the evolution of their probability distributions in the Markov chain. This metric is robust to noise, since the distance between two points depends on all possible paths of length t between the points. Using the spectral decomposition given in (2), the diffusion distance can be calculated using the eigendecomposition of $\mathbf{P}$ by

$$d_{DM}(x_i, x_j; t) = \sum_{l\geq 1} \lambda_l^{2t}(\psi_l(x_i) - \psi_l(x_j))^2. \tag{4}$$

Due to the spectrum decay, the diffusion distance can be approximated using only the first ℓ eigenvectors. Thus, the computational complexity of the diffusion distance is low given the eigendecomposition of $\mathbf{P}$.

Equation (4) implies that a mapping can be defined between the original space and the eigenvectors ψ_l, defining a new set of coordinates for the dataset Γ, such that the diffusion distance is equal to the Euclidean distance in this new embedding. Retaining only the first ℓ eigenvectors, the diffusion map is defined by

$$\Psi_t : x_i \to \left(\lambda_1^t\psi_1(x_i), \lambda_2^t\psi_2(x_i), ..., \lambda_\ell^t\psi_\ell(x_i)\right)^T. \tag{5}$$

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.

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. 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 is beneficial, such as the one proposed by Zelnik-Manor and Perona [17]. The scale σ is calculated for each point x_i based on the local statistics of its neighborhood:

$$\sigma_i = \|x_i - x_K\|^2 \tag{6}$$

where x_K is the K-th nearest neighbor. The similarity kernel is then calculated as $w(x_i, x_j) = \exp\left\{-\|x_i - x_j\|^2/\sigma_i\sigma_j\right\}$.

3. FUNCTION EXTENSION

The size of the data set for images is very large. Therefore, it can be computationally inefficient to construct a diffusion map using all the pixels in the image, especially for high-resolution images. Instead, it is a common approach [10, 11] to construct the diffusion map for an image using a subset of random samples, $\Gamma \subseteq \overline{\Gamma}$, and then the diffusion map coordinates Ψ are extended to the set of all patches in the image $\overline{\Gamma}$ 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. Different methods have been proposed to approximate the Nyström extension method [18] or improve upon it, such as the Geometric Harmonics method [19]. Recently, a new algorithm was presented for out-of-sample function extension using the multiscale Laplacian pyramid [20]. At each iteration, the Laplacian pyramid algorithm constructs a coarse approximation of a function f for a given scale. 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 increasingly finer scales. A smooth function can be extended using a coarse scale, i.e. will not require many levels of the pyramid. An oscillating function on the other hand will require finer levels of the pyramid to enable an accurate extension. For more details, see [20]. We perform this extension method for each diffusion coordinate $f = \Psi_l, l \in \{1, ..., \ell\}$ separately. The number of levels in the pyramid extension can differ between coordinates, dependent on their smoothness over Γ.

As discussed in [9], out-of-sample extension methods can cause anomaly detection to fail, depending on the set of random samples $\Gamma \subseteq \overline{\Gamma}$ used to construct the diffusion map. 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 [21]. In such a case, the diffusion map will not capture the difference between the anomaly and the background, and the out-of-sample extension of the diffusion map to the pixels in the anomaly region will not succeed in assigning them new coordinates which separate them from the background. 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. To overcome this limitation of the out-of-sample extension, we propose a multiscale method which drives the sampling process and ensures the inclusion of samples from the anomaly region in Γ.

4. MULTISCALE ANOMALY DETECTION

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 for the construction

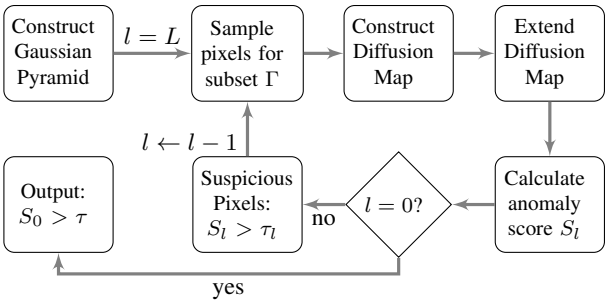

Fig. 1. Flowchart of the multiscale algorithm.

of the diffusion map. Thus, detecting an anomaly at a lower resolution is less likely to fail due to sampling. Since our method performs anomaly detection at different resolutions of the image, even if the anomaly is missed on a coarse level due to its size, 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 [22] representation of the image, yielding $\{G_l\}_{l=0}^{L}$, where G_0 is the original image and G_L is the coarsest resolution. 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, depending on memory constraints. Using the diffusion coordinates, an anomaly score S_L is calculated for all pixels and then a threshold τ_l on the anomaly score determines suspicious pixels. Proceeding to the image G_{L-1}, pixels which correspond to the suspicious pixels found in G_L are included in Γ_{L-1}. The rest of the pixels in Γ_{L-1} are sampled randomly from the image. The threshold τ_l used at the output of each level is chosen to be the 95^{th} 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.

The process of sampling, dimensionality reduction and anomaly detection continues from level to level, with each previous level providing prior information on which samples of the data set will be used in Γ_l to construct the diffusion map. 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 S_0 and then smooth the resulting image. Anomalies have a high score, close to 1. Figure 1 presents a flowchart of the algorithm. This approach greatly increases the detection rate of the diffusion-based anomaly detector, compared with a single scale approach.

At each level, the affinity matrix is calculated for the subset Γ_l. 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 $w(x_i, x_j) = 0$, resulting in a sparse matrix as in [15]. We set $k = 16$.

4.1. Saliency-based Anomaly Score

In our previous paper [9], we used an anomaly score based on a nearest-neighbor approach. The assumption was that the anomaly is in a low density neighborhood in the diffusion map space whereas background pixels belong to dense local neighborhoods in the diffusion map space. Taking advantage of the spatial nature of the data set, the pixel grid, we calculated an affinity measure between each pixel and the pixels in the window surrounding it, using the diffusion distance. This affinity was used to calculate an approximation of the local density of each pixel in the diffusion map space, given its spatial neighborhood. Anomaly pixels were dissimilar to their spatial neighborhoods, resulting in low local density, whereas background pixels were similar to their spatial neighborhoods, resulting in high density. Since we assume the anomaly to be non-pointwise, a mask was used in order to remove the inner region of the window surrounding the pixel.

In this paper we propose using a different measure to calculate the anomaly score. In [12], Goferman et al. defined a local-global dissimilarity measure between two patches in an image:

$$d(p_i, p_j) = \frac{d_{\text{color}}(p_i, p_j)}{1 + c \cdot d_{\text{position}}(p_i, p_j)} \tag{7}$$

with $c = 3$. The distance d_{color} is the Euclidean distance between vectorized patches in CIE *L*a*b* color space normalized to the range $[0, 1]$. The distance d_{position} is the Euclidean distance between the image positions of patches p_i and p_j, normalized by the larger image dimension. This dissimilarity measure is proportional to the difference in appearance and inverse proportional to the positional distance. It realizes the authors' observations that background pixels are similar to both near and far pixels, whereas salient patches are grouped together, therefore similar only to nearby patches. In addition, in order to evaluate the distinctness of a patch it is sufficient to consider its K most similar patches $\{q_k\}_{k=1}^{K}$ ($K = 64$), and not calculate its dissimilarity to all image patches. The saliency value of a pixel was given by

$$S(i) = 1 - \exp\left\{-\frac{1}{K}\sum_{k=1}^{K} d(p_i, q_k)\right\} \tag{8}$$

These observations holds for anomaly detection as well. However, instead of using d_{color}, we propose using the diffusion distance between patches. The diffusion distance is preferable to using the color distance between patches as it is robust to noise. Also, the embedding it yields better separates the anomaly from the background, compared to using image patches. This requires normalizing the diffusion distance such that most values are spread out in the range $[0, 1]$ and therefore comparable to d_{position}. Since we already calculate the K most similar data points for each point in the image in terms of the diffusion distance, we can use the statistics on these distances for the normalization. Empirically, dividing the distances by the standard deviation of the distances to the K^{th} neighbor gave us values in the desired range: $\sigma_K = \underset{i \in \Gamma}{\text{std}}\{d_{\text{DM}}(p_i, q_K)\}$. Finally, the anomaly score is given by

$$S(i)_{DM} = 1 - \exp\left\{-\frac{1}{K}\sum_{k=1}^{K} \frac{d_{\text{DM}}(p_i, p_j)/2\sigma_K}{1 + c \cdot d_{\text{position}}(p_i, p_j)}\right\}. \tag{9}$$

The advantages of using this measure over the one given in [9] are:

1. This grade better suppresses background regions which have similar diffusion coordinates, yet are spatially distant from one another in the image. Using the previous score, a tested pixel was compared to the pixels within a limited spatial

region surrounding it. So if there were background regions which are similar to one other, yet different from their close spatial neighborhood, they could receive a high anomaly score. Now since each pixel is compared to its most similar neighbors in the diffusion space regardless of spatial proximity, these regions receive a low anomaly score. Anomalous pixels, unlike these regions, lie close together yet are different from all other regions, therefore receive a high score.

2. No prior knowledge is required regarding the anomaly size, while in our previous method the size of the anomaly was used to mask the close spatial neighbors of a tested pixel, since they were not used in calculating the anomaly score.
3. The previous score requires more fine-tuning of different parameters (the size of the spatial neighborhood, the size of the mask, the scale used in the diffusion-based affinity measure) and special treatment of the boundary pixels of the image.

In Sec. 5 the previous score is compared to the new score.

5. 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. 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.

Algorithms proposed for detection of mines in side-scan sonar include MRF models for modeling the background [23, 24], a 2-D multiscale GMRF with matched subspace detector (MSD) [4], a multidimensional GARCH model with MSD [3], non-linear matched filters [25], etc. 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 [24, 26]. 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 or its shadow in the image.

We evaluated our algorithm on a set of 27 side-scan sonar images of size 200x200 pixels. The parameters of the multiscale detector are given in Table 1, for a Gaussian pyramid of $L = 3$ levels . Note that the size of the images 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.

Detections are found by applying a threshold to the anomaly score image, resulting in a binary image. A detection is a connected component (CC) in the binary image. A CC containing the sea-mine is a true positive (TP) and any other CCs are false alarms (FA). The size of the CC can be used to reject noisy detections, where small CCs are discarded. We compare two thresholds on the area of the CC: 10 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 percentage of TPs for each method for a given FA rate. Results are given in Table 2.

We compare our proposed algorithm (MS-CAS) with three other methods:

Table 1. Parameters Used in Multiscale Detector

Pyramid Level	Patch size	Embedding Dimension	Percentage of pixels in subset
0	8x8	6	0.10
1	4x4	6	0.33
2	2x2	3	0.5

Table 2. Percentage of True Positives for Given Number of False Alarms

	size=10			size=20		
# of FA	8	4	0	8	4	0
MS	93%	93%	89%	93%	89%	89%
MS-CAS	100%	96%	93%	100%	96%	96%
SS-CAS	89%	89%	44%	89%	89%	37%
CAS	81%	67%	41%	78%	63%	30%

- MS: The multiscale method we proposed in [9], using the local density based score.
- SS-CAS: A single-scale method using the full-size image. The parameters are the same as those in Table 1 for pyramid level 0, however 20% of the image is randomly sampled to construct the diffusion map. The anomaly score is the one described in Sec. 4.1.
- CAS: The method proposed in [12]. Anomaly detection is performed by applying a threshold to the saliency map.

Our new multiscale approach MS-CAS has the highest TP rate. The performance of our new algorithm is better than our previous method MS, with an improvement of 3%-17% for a given number of FAs. Also, as in our previous work, using a multiscale approach results in increased performance compared to using a single scale. The poor performance of the single-scale detector SS-CAS for low FA rate reveals the limitations of sampling the image compared with the multiscale detectors which have a significantly better detection rate. This is due to the propagation of information from level to level. Our method also outperforms [12]. The reason for this is that the Context-Aware Saliency method uses a feature space which is suitable for natural images. The images we tested are side-scan sonar and are very noisy. 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. Our code is available at [27].

6. CONCLUSION

We have presented an anomaly detection algorithm using diffusion maps, which is a state-of-the-art method for manifold learning and dimensionality reduction. To improve the detection process and ensure that the normal pixels and the anomaly regions are separable in the lower dimensional embedding, we implemented a multiscale framework. This multiscale approach overcomes the possible limitations in using diffusion maps with out-of-sample extension algorithms. A new anomaly score is proposed, combining a saliency measure with the noise-robust diffusion distance. Our algorithm performed successfully in the challenging task of automatic target detection in side-scan sonar images. Our method achieved superior results when compared to competing methods, both single scale and multiscale.

7. REFERENCES

[1] Y. Chen, N. Nasrabadi, and T. Tran, "Sparse representation for target detection in hyperspectral imagery," *IEEE Journal of Selected Topics in Signal Process.*, vol. 5, no. 3, pp. 629–640, 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, 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, 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. of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2001, pp. 3–10.

[6] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007.

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

[8] 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.

[9] G. Mishne and I. Cohen, "Multiscale anomaly detection using diffusion maps," *IEEE Journal of Selected Topics in Signal Process.*, vol. 7, pp. 111 – 123, Feb. 2013.

[10] 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.

[11] 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.

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

[13] 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.

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

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

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

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

[18] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 214–225, Jan. 2004.

[19] 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.

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

[21] 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.

[22] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Communications*, vol. 31, no. 4, pp. 532–540, 1983.

[23] S. Reed, Y. Petillot, and J. Bell, "An automatic approach to the detection and extraction of mine features in sidescan sonar," *IEEE Journal Of Oceanic Engineering*, vol. 28, no. 1, pp. 90–105, Jan. 2003.

[24] ——, "Automated approach to classification of mine-like objects in sidescan sonar using highlight and shadow information," *IEE Proc. Radar, Sonar and Navigation*, vol. 151, no. 1, pp. 48–56, Feb. 2004.

[25] G. Dobeck, "Algorithm fusion for automated sea mine detection and classification," in *Proc. MTS/IEEE Oceans Conf. and Exhibition*, vol. 1. Marine Technol. Soc, 2001, pp. 130–134.

[26] 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.

[27] [Online]. Available: http://webee.technion.ac.il/Sites/People/IsraelCohen/