ROC-based Performance Analysis and Interpretation of Image-based Damage Diagnostic Tools for Underwater Inspections

M. O'Byrne & V. Pakrashi*
Dynamical Systems and Risk Laboratory, School of Mechanical and Materials Engineering, University College Dublin, Ireland; Marine Renewable Energy Ireland (MaREI), University College Dublin, Ireland

F. Schoefs
Université Bretagne-Loire, Université de Nantes, Research Institute of Civil Engineering and Mechanics (GeM)/Sea and Littoral Research Institute (IUML), CNRS UMR 6183/FR 3473, Nantes; IXEAD/CAPACITES Society, Nantes, France

B. Ghosh
Department of Civil, Structural and Environmental Engineering, Trinity College Dublin, Ireland

ABSTRACT: It is of practical importance for inspectors to have knowledge of the efficiency of Non-Destructive Testing (NDT) tools when applied commercially. It has become common practice to model the performance of NDT tools in a probabilistic manner in terms of Probability of Detection (PoD), Probability of False Alarm (PFA) and eventually by Receiver Operating Characteristic (ROC) Curves. Traditionally, these quantities are estimated from training data, however, there are often doubts about the validity of these estimates when the sample size is small. In the case of underwater inspections, the scarcity of good quality training data means that this scenario arises more often than not. Comprehensive studies around the on-site performance of image-based damage diagnostic tools have only recently been made possible through the availability of online resources such as the Underwater Lighting and Turbidity Image Repository (ULTIR), which contains photographs of various damages forms captured under controlled visibility conditions. This paper shows how meaningful information can be extracted from this repository and used to construct ROC curves that can be related to the on-site performance of image-based NDT methods for detecting various damage forms and under a range of environmental conditions. The ability to draw connections between image-based techniques applied in real underwater inspections with ROC curves that can be constructed on-demand provide the engineer/inspector with a clear and systematic route for assessing the reliability of data obtained from image-based methods. As a case study, the general approach has been applied to characterise the performance of image-based techniques for identifying instances of corrosion and cracks on marine structures. A discussion around how the results can be used for further analysis is provided. This includes looking at how the results can be fed into in the decision chain and can be used for risk analysis, intervention and work scheduling, and eventually understanding the value of information.

1 INTRODUCTION

The quality of subsea inspections largely depends on the ability of inspectors to detect and objectively record details of defects. Various NDT tools are often employed to help inspectors to this end, however, choosing the right NDT tool for a given situation is not always straightforward. The type of damage present and the on-site operating conditions are crucial factors that dictate which tool should be adopted. In the highly corrosive marine environment, common damage types include chloride-induced corrosion and cracks that form on concrete structures due to volume expansion of corroding reinforced steel. These damage forms can usually be detected using vision-based systems, which are often the most convenient option, however, the efficacy of vision-based systems in practice is heavily reliant on the underwater environment. The reduced visibility conditions diminish the ability of the camera, and subsequent image-processing algorithms, to effectively identify instances of damage. It is therefore important that inspectors can develop an understanding of the relationship between visibility conditions and the performance of image-processing techniques so that they can rationalise the use of image-processing methods as part of an underwater inspection campaign.

While many image-processing methods have been devised for structural health monitoring applications over the years, comprehensive studies around their on-site performance levels have only recently been made possible through the development of the Underwater Lighting and Turbidity Image Repository (ULTIR) (O'Byrne et al., 2017). ULTIR is a platform

that allows image-based damage detection techniques to be easily investigated under a host of realistic operating conditions, and through studies such as this one, provide meaningful information to inspectors regarding the expected detection accuracy of techniques when applied in the field.

This paper demonstrates how inspectors can gain an insight into this relationship and develop a deeper intuition around on-site conditions by consulting ULTIR. Two case studies are presented that look at how the performance of crack and corrosion detection methods change as visibility conditions vary, and how the information gleaned from ULTIR can be used to forecast the expected change in performance.

2 BACKGROUND AND METHODOLOGY

This section gives a brief overview of the ULTIR database and describes the methodology for evaluating and ranking image-processing techniques.

2.1 Description of the repository

ULTIR is a large database containing hundreds of photographs of various damage instances that were captured under known turbidity and lighting levels. Ground-truth data is also provided which shows the true locations of damage in each image, thereby enabling the performance of applied damage detection algorithms to be evaluated. It is freely-available online at www.ultir.net (O'Byrne et al., 2018).

The repository is partitioned into three categories that relate to the assessment of crack detection techniques, general surface damage detection techniques (such as corrosion), and techniques for recovering 3D shape information. This paper focuses on the first two of these categories.

2.2 Influence of turbidity and lighting

Image quality is assumed to be chiefly affected by luminosity, sharpness (focus accuracy), contrast and noise. These quality factors are directly related to the on-site operating conditions, for which lighting and turbidity are the most significant (Mahiddine et al., 2012). Turbidity is defined as the cloudiness in a liquid caused by the presence of suspended solids that scatter and absorb light and therefore reduce visibility. Turbidity can be caused by organic particles, such as decomposed plant and animal matter; or by inorganic particles such as silt and clay

In this paper, two levels of turbidity are considered: clear water, or 0 NTU (Nephelometric Turbidity Units), and 12 NTU. To put these values in context, water that is visibly cloudy has a turbidity of 6 NTU, while water that is murky has a turbidity of 25 NTU. A high-point of 12 NTU was used in this study as it becomes increasingly difficult to interpret and extract useful information from images beyond this point. In waters above 12 NTU, image-based methods become an increasingly infeasible option as a quantitative inspection tool. Additionally, the turbidity of many rivers and water bodies' lies within the 0 – 12 NTU range, and thus, focusing on these limits is of high practical relevance.

The on-site lighting is also crucial for achieving good visibility. Ambient light may be sufficient for near-surface inspections; however, it is unlikely to be sufficient at greater depths at which point artificial light will be needed. Two light levels were used in this study: 100 lux and 10000 lux. To put this in context, the approximate illuminance on a very dark overcast day is 100 lux, a moderately overcast day is 1000 lux, and a bright day is 10,000 – 25,000 lux.

2.3 Performance measures

The performance of image-processing methods was evaluated and ranked using Receiver Operating Characteristic (ROC) curves. ROC curves offer a convenient way of characterising the performance of NDT methods under various environmental conditions (Rouhan and Schoefs, 2003) and have been expanded to image detection (Pakrashi et al., 2008). The Detection Rate (DR) along with the accompanying Misclassification Rate (MCR) - which are similar to the Probability of Detection (PoD) and Probability of False Alarm (PFA) in the field of probability space and decision theory - are determined by comparing the damaged regions, as identified by a given image-processing method, with a visually segmented image that acts as the control as it is assumed it shows the true extent of damage.

The DR and MCR are represented as a fraction between 0 and 1. Each (MCR, DR) pair form a coordinate in the ROC space that corresponds to a particular decision threshold. The DR and MCR are defined as:

$$DR \approx \frac{Card(E)}{n_c} \text{ with } \quad E = \left\{ g \in \Im ; \gamma_g = 1 \right\} \qquad (1)$$

$$MCR \approx \frac{Card(F)}{n} \text{ with } \quad F = \left\{ g \in \Im ; \gamma_g = -1 \right\} \qquad (2)$$

where Card(.) indicates the cardinal of a particular set, $\Im = \{1, \ldots, n\}$, n is the total number of pixels in the image, n_c denotes the number of damaged pixels and γ_g is an instance label vector, where $\gamma_g = 1$

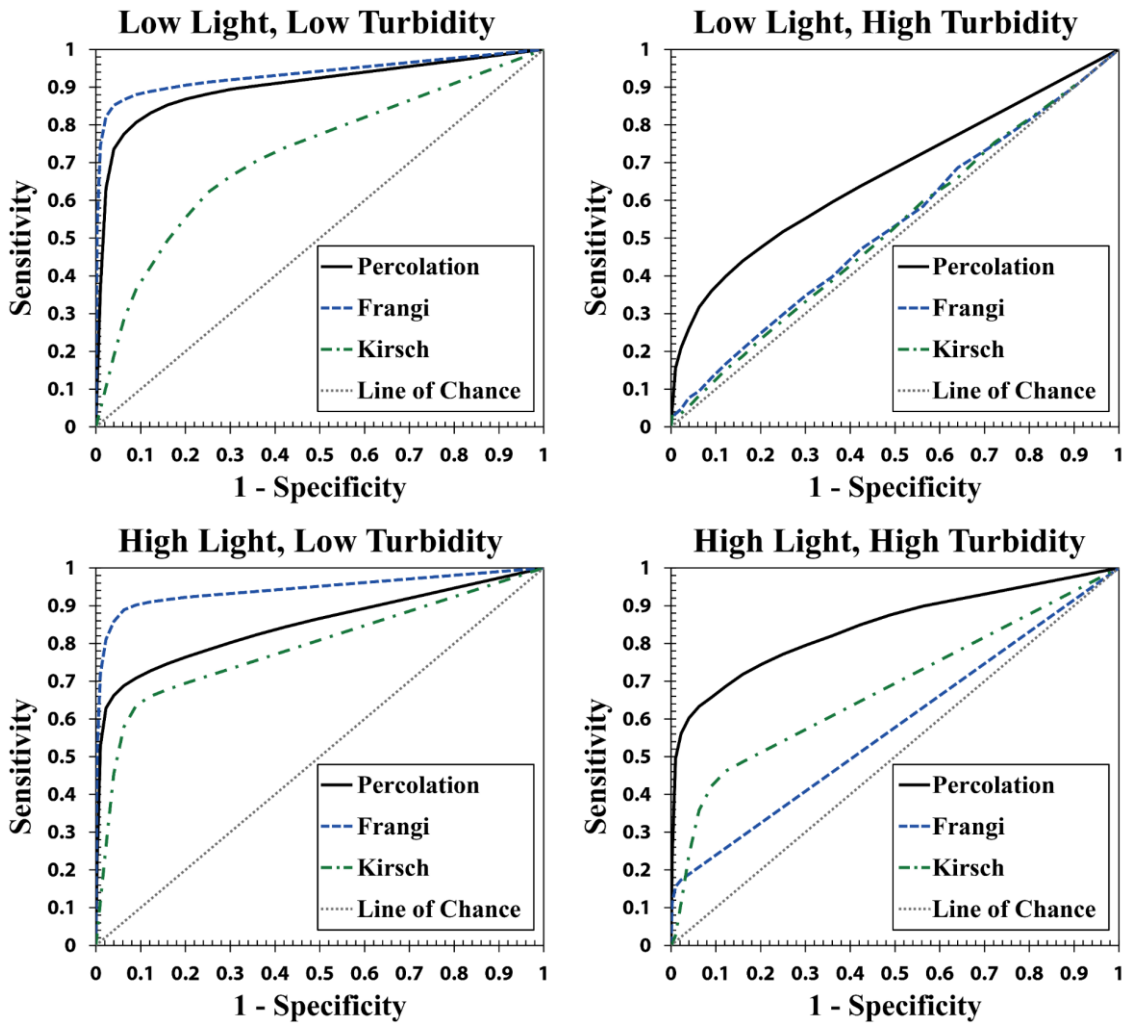

Figure 1. ROC curves of three crack detection techniques under four operating conditions. The techniques are: A percolation-based method (O'Byrne et al., 2014a), Eigenvalue analysis of the Hessian (Frangi et al., 1998), and Kirsch templates (Kirsch, 1971).

corresponds to correctly identified non-damaged pixels, i.e. true positives, and $\gamma_g = -1$ corresponds to incorrectly detected pixels and undetected damaged pixels, i.e. false negatives and false positives. *F* gathers incorrectly detected and undetected damaged pixels while *E* gathers correctly detected pixels.

The α-δ method is employed to find the optimum threshold value that maximises detection. This method provides a measure of how well a parameter can distinguish between two diagnostic groups (i.e. damaged region/non-damaged region) (Baroth et al., 2011, Schoefs et al., 2012). It relies on calculating the angle, α, and the Euclidean distance, δ, between the best performance point, defined as an ideal NDT technique with 100% detection rate and 0% misclassification rate and the considered point to give a measure of the performance of the considered point. As this paper does not deal with risk analysis where the shape of the ROC is a key factor, only the delta, δ, parameter is required as a measure of performance. A low value for δ is indicative of a good performance. Therefore, the closer the ROC curve is to the upper left corner of the plot, the higher the overall accuracy.

3 DATA ANALYSIS

This section considers two case studies that showcase the value of ULTIR for inspectors. The first case study considers underwater crack detection techniques while the second case study deals with corrosion detection techniques. For both case studies, three image processing techniques are applied to controlled imagery in ULTIR and the performances are characterised through ROC analysis for four different visibility conditions. This involves constructing ROC curves for each technique under the following conditions: 1) low light and low turbidity, 2) low light and high turbidity, 3) high light and low turbidity, and 4) high light and high turbidity. The ROC curves for three crack detection techniques are presented in Figure 1, while ROC curves for three corrosion detection algorithms are shown in Figure 2.

By analysing the ROC curves, the performance levels of techniques are found for various on-site operating conditions. The techniques are then applied to new images and the actual performance levels are compared against the expected performance levels as predicted by analysing the ROC curves.

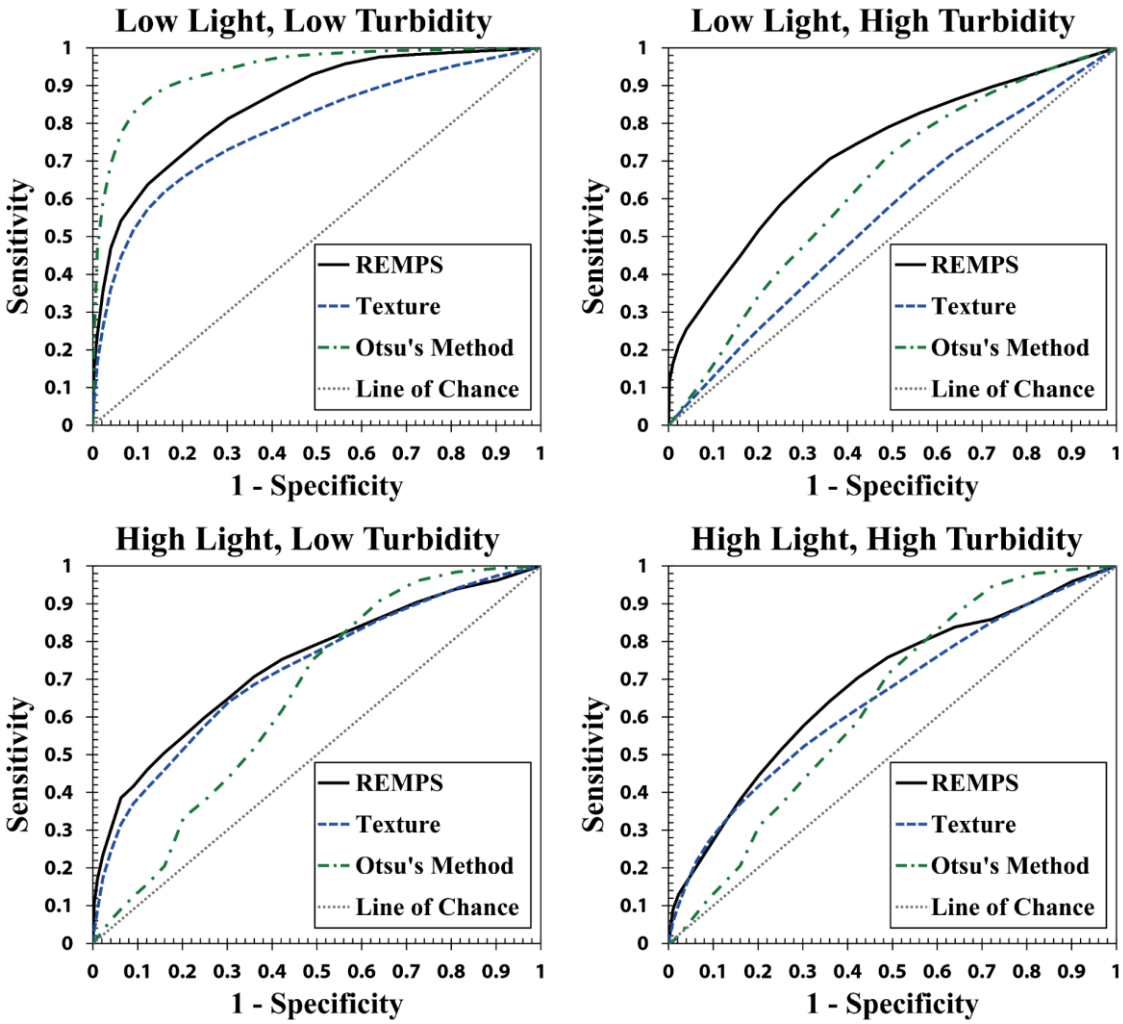

Figure 2. ROC curves of three damage detection techniques under four operating conditions. The techniques are: REMPS (O'Byrne et al., 2014b), Texture analysis (O'Byrne et al., 2014), and Otsu's Method (Otsu, 1979).

3.1 Case Study I: Analysing and ranking crack detection methods

Cracks provide an indication of structural degradation and are an important factor when diagnosing the condition of concrete and steel structures. Crack assessment has been well studied in the past, and numerous image based crack detection algorithms have been devised which are capable of identifying crack-like features. In this case study, the investigated methods are a percolation-based method (O'Byrne et al., 2014a), eigenvalue analysis of the Hessian (Frangi et al. 1998), and Kirsch templates (Kirsch, 1971). These techniques follow different methodologies, and naturally, they will differ in terms of how well they can tolerate deteriorating lighting and turbidity conditions.

These techniques have been applied to photographs of a 1mm crack captured under varying lighting and turbidity levels, as shown in Figure 3.

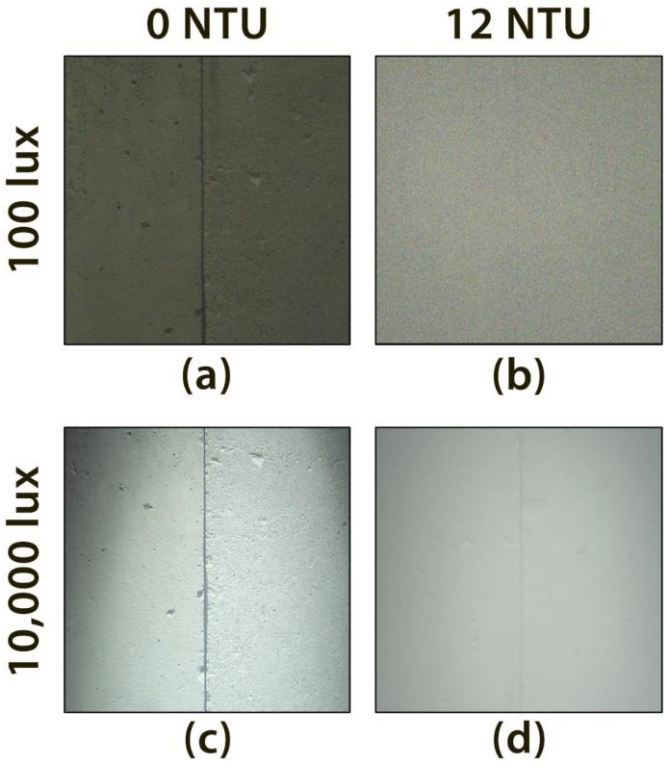

Figure 3. A crack that was photographed under (a) low light and low turbidity, (b) low light and high turbidity, (c) high light and low turbidity, and (d) high light and high turbidity.

The images in Figure 3 are small regions, or sub-images, taken from the 1 mm controlled crack data set in ULTIR. These sub-images represent only a very small fraction of the total images in ULTIR that were used when producing the ROC curves in Figure 1, and therefore, they do not contribute notably to the overall shape of the ROC curves. As such, it is remains valid to infer information from the ROC curves and use this information to conjecture the about the expected performance of methods when applied to the images in Figure 3.

The ground truth data for these images, which shows the location of the crack, was also extracted from ULTIR. Using this ground truth data, the performance of the considered techniques was established. The actual performances – expressed in terms of δ – are summarised in Table 1. The performances are ranked from best (lowest δ) to worst (highest δ value). The expected performances are also presented; these are determined by studying the ROC curves in Figure 1 and finding the δ that corresponds to the optimum decision threshold. The results from the best performing techniques for each set of lighting and turbidity levels are shown in Figure 4. The results from the best performing techniques for each set of lighting and turbidity levels are shown in Figure 4.

Table 1. Performance of the Percolation, Frangi, and Kirsch methods when applied to images in Figure 4. The expected performance value is derived from the ROC curves in Figure 1.

Ref. No.	Method	Operating Condition	Actual δ [rank]	Expected δ [rank]
1	Percolation	Low light, Low turbidity	0.04 [1]	0.22 [3]
2	Percolation	Low light, High turbidity	0.39 [9]	0.57 [9]
3	Percolation	High light, Low turbidity	0.05 [3]	0.32 [4]
4	Percolation	High light, High turbidity	0.47 [10]	0.36 [6]
5	Frangi	Low light, Low turbidity	0.05 [2]	0.17 [2]
6	Frangi	Low light, High turbidity	0.48 [11]	0.69 [11]
7	Frangi	High light, Low turbidity	0.05 [4]	0.13 [1]
8	Frangi	High light, High turbidity	0.39 [8]	0.67 [10]
9	Kirsch	Low light, Low turbidity	0.30 [7]	0.47 [7]
10	Kirsch	Low light, High turbidity	0.89 [12]	0.69 [12]
11	Kirsch	High light, Low turbidity	0.10 [5]	0.36 [5]
12	Kirsch	High light, High turbidity	0.18 [6]	0.55 [8]

It may be noted from Table 1 that while the actual and expected δ values differ quite markedly in terms of magnitude, the relative ranking of methods under various environmental situations is similar for both the actual results and the expected results. For instance, the expected performance of the Kirsch templates method in low light and high turbidity is ranked as the worst scenario for good detection accuracy. This is reflected by the worst performance levels (12th out of 12) in the actual results.

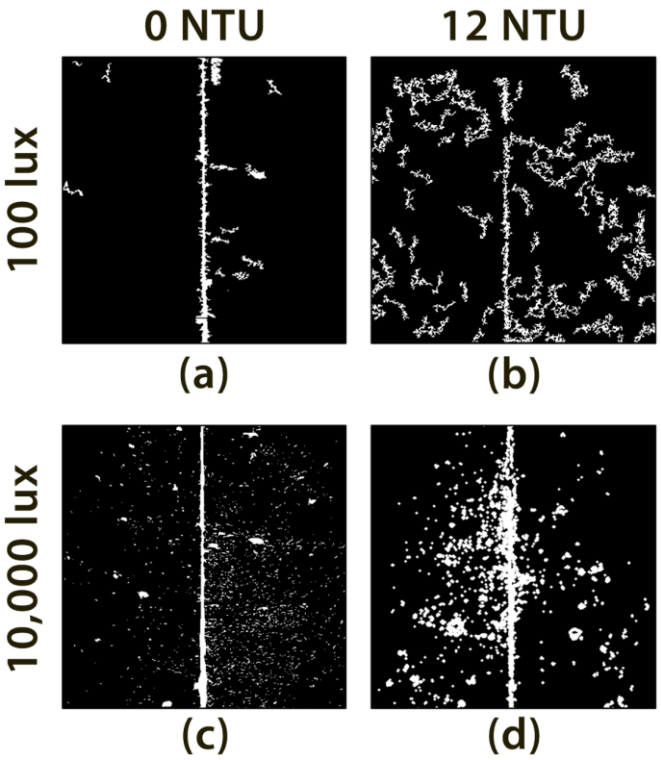

Figure 4. Best performing techniques for each situation (a) percolation, (b) percolation, (c) Frangi, and (d) Kirsch.

3.2 General Observations

It may be observed from the ROC curves that the Frangi technique performs well in low turbidity conditions, however, the performance drastically diminishes in high turbidity conditions. This indicates that this method is highly sensitive to noise/turbidity. On the other hand, the percolation-based method demonstrates consistently good performance across both turbidity levels. The Kirsch template method produces moderately good results; however, it does not perform as well or as consistently as the percolation-based method. Findings of this nature are useful for inspectors as it allows them to choose a technique appropriate to their needs and one that is sufficiently robust to the onsite operating conditions.

3.3 *Case Study II: Analysing and ranking corrosion detection methods*

The corrosion detection methodology should identify and accurately define all corroded regions in an image whilst minimising the inclusion of extraneous regions. In reality, perfect damage detection is difficult to achieve given the inherent chromatic and luminous complexities encountered in natural scenes. Image-processing based damage detection techniques include colour intensity based methods and texture analysis based methods. Naturally, the techniques in each group are suited to different applications. The effectiveness of colour based segmentation algorithms and texture based segmentation algorithms will vary according to the surface and damage type under consideration as certain damages are more separable from the undamaged surface based on either their colour or texture attributes. This section assesses the performance of two colour based methods, REMPS (O'Byrne et al., 2014b) and Otsu's thresholding (Otsu, 1979); along with a texture analysis based technique (O'Byrne et al., 2013) previously proposed in the domain of NDT.

These techniques have been applied to photographs of a corrosion stain captured under a) low light (100 lux) and low turbidity (0 NTU), b) low light (100 lux) and high turbidity (12 NTU), c) high light (100 lux) and low turbidity (12 NTU), and d) high light (10,000 lux) and high turbidity (12 NTU), as shown in Figure 5.

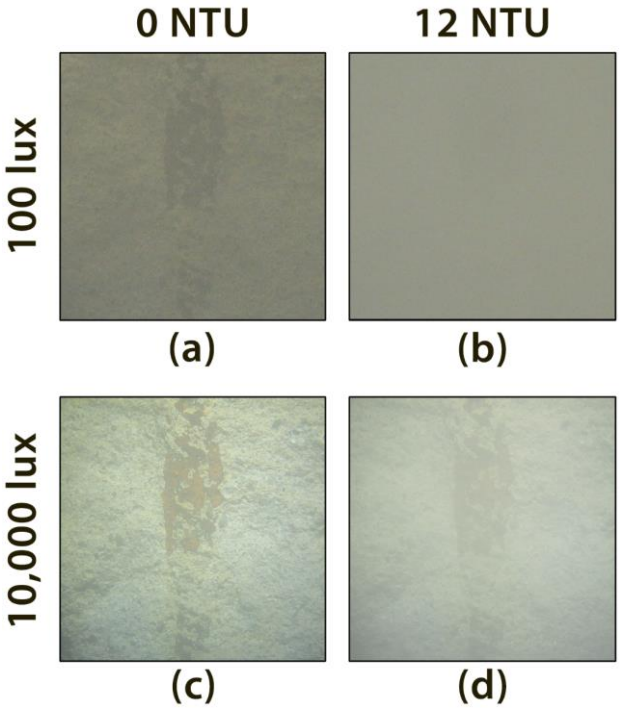

Figure 5. A corrosion stain that was photographed under (a) low light and low turbidity, (b) low light and high turbidity, (c) high light and low turbidity, and (d) high light and high turbidity.

As in the case of the first case study, the three damage detection algorithms are applied to the images in Figure 5 and the performances are ranked from best (lowest δ) to worst (high δ value). The expected performances are obtained from analysing the ROC curves in Figure 2 and finding the δ that corresponds to the optimum decision threshold. These results are tabulated in Table 2, and the outputs from the best performing techniques for each set of lighting and turbidity levels are shown in Figure 6.

Table 2. Performance of the REMPS, Texture Analysis and Otsu's method when applied to images in Figure 5. The expected performances are found from the ROC curves in Figure 2.

Ref. No.	Method	Operating Condition	Actual δ [rank]	Expected δ [rank]
1	REMPS	Low light, Low turbidity	0.32 [2]	0.36 [2]
2	REMPS	Low light, High turbidity	0.62 [9]	0.49 [5]
3	REMPS	High light, Low turbidity	0.57 [7]	0.48 [4]
4	REMPS	High light, High turbidity	0.59 [8]	0.54 [7]
5	Texture	Low light, Low turbidity	0.50 [5]	0.43 [3]
6	Texture	Low light, High turbidity	0.70 [11]	0.67 [12]
7	Texture	High light, Low turbidity	0.38 [3]	0.49 [6]
8	Texture	High light, High turbidity	0.43 [4]	0.58 [8]
9	Otsu	Low light, Low turbidity	0.31 [1]	0.19 [1]
10	Otsu	Low light, High turbidity	0.63 [10]	0.60 [10]
11	Otsu	High light, Low turbidity	0.57 [6]	0.60 [9]
12	Otsu	High light, High turbidity	0.71 [12]	0.64 [11]

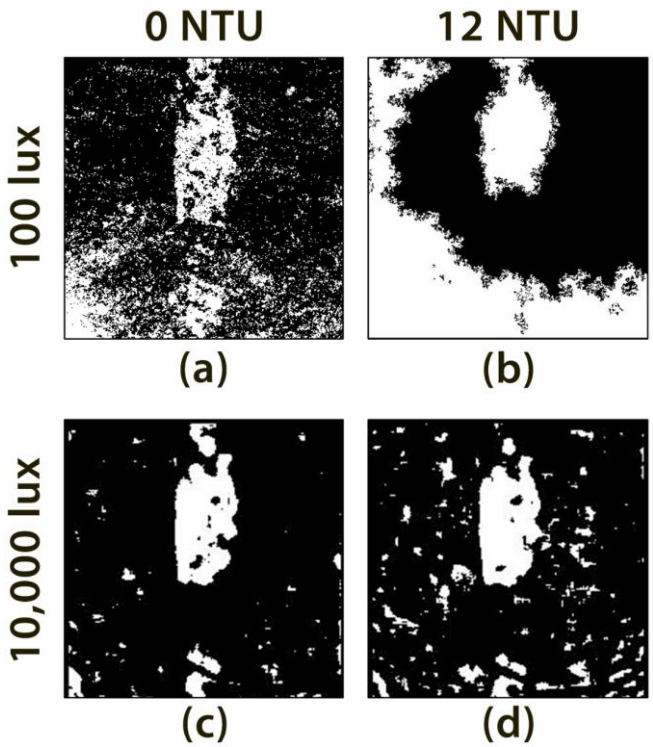

Figure 6. Best performing techniques for each situation (a) Otsu's method, (b) REMPS, (c) Texture analysis, and (d) Texture analysis.

It may be seen from Table 2 that the actual and expected δ values agree quite closely with one another. This is further conveyed in Figure 7, which plots the expected performances alongside the actual performances for each scenario (a description of the scenarios is provided with reference to Table 2).

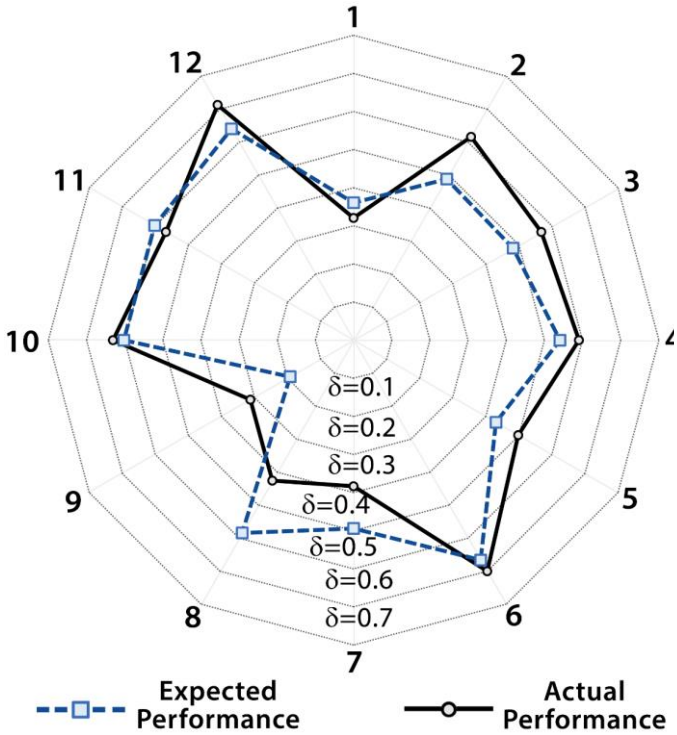

Figure 7. The expected and actual performances for each situation (which are described in Table 2).

3.4 General Observations

It may be noted from these results that REMPS was quite robust - it displayed less sensitivity to the input conditions as evidenced by the similar ROC curves across all lighting and turbidity levels in Figure 2. While Otsu's method demonstrated success in clear conditions (no turbidity), the performance levels sharply declined as the visibility conditions deteriorated.

The texture analysis method outperformed the colour based methods in the high lighting situations. While the high light and the bright surface created luminous complexities that misled the colour based methods, the strong light source illuminated and brought out some of the textural properties of the surface which benefitted the texture analysis technique.

Overall, it is hard to accurately predict the success of image-processing based methods as there are a host of factors that contribute towards successful detection. Nevertheless, turbidity and lighting are major factors and accounting for these parameters represents an important step. This case study shows how inspectors can get a sense of the performance of image-processing based damage detection methods under realistic underwater operating conditions prior to carrying out an inspection. This can add value to any underwater inspection campaign in which image processing based methods are being considered as an NDT tool. The work is relevant for traditional engineering sectors of SHM like bridges (Pakrashi et al., 2013) as well as for bourgeoning sectors like renewable energy (Jaksic et al., 2015[a,b]), leading to an optimised maintenance and management of these infrastructure systems (Weninger-Vycudil et al., 2015).

4 CONCLUSION

Image processing based methods are increasingly being recognised as a valuable tool for inspecting the submerged part of offshore and marine structures. They provide a source of quantitative information that naturally supplements the largely qualitative information obtained from traditional visual methods. As with all NDT techniques, it is of great practical importance for inspectors to know the effectiveness of these techniques when applied in the field. For image-processing methods applied in an underwater setting, this principally means investigating how the visibility conditions affect detection accuracy. Underwater visibility is chiefly governed by the lighting and turbidity levels. This study looks at the

influence of these parameters by drawing on a large image database, known as ULTIR (Underwater Lighting and Turbidity Image Repository), which consists of submerged specimens that have been photographed under controlled lighting and turbidity levels. The specimens feature various forms of damage - the true extent of which is known - and it is therefore possible to measure the performance of all applied image-processing methods. Receiver Operating Characteristics (ROC) curves are employed to rank the techniques under varying operating conditions.

5 REFERENCES

Baroth, J., Breysse, D., & Schoefs, F. (2011). Construction Reliabilit. Hoboken: NJ Wiley.

Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). "Multiscale Vessel Enhancement Filtering." Proc., Medical Image Computing and Computer-Assisted Intervention - MICCAI'98. First International Conference. Proceedings, 11-13 Oct. 1998, Springer-Verlag, Berlin, Germany, 130-137.

Jaksic, V., Wright, C.S., Murphy, J., Afeef, C., Ali, S.F., Mandic, D.P. & Pakrashi, V. (2015). Dynamic Response Mitigation of Floating Wind Turbine Platforms using Tuned Liquid Column Dampers. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373 (2035):20140079:1-20140079:9

Jaksic, V., O'Shea, R., Cahill, P., Murphy, J., Mandic, D.P., & Pakrashi V. (2015). Dynamic Response Signatures of a Scaled Model Platform for Floating Wind Turbines in an Ocean Wave Basin. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373 (2035):20140078:1-20140078:18

Kirsch, R. A. (1971). "Computer Determination of the Constituent Structure of Biological Images." Computers and biomedical research, an international journal, 4(3), 315-328.

Mahiddine, A., Seinturier, J., Boi, D. P. J., Drap, P., Merad, D., and Luc, L. (2012). "Underwater Image Preprocessing for Automated Photogrammetry in High Turbidity Water: An Application on the Arles-Rhone XIII Roman Wreck in the Rhodano River, France." Proc., Virtual Systems and Multimedia (VSMM), 2012 18th International Conference on, Milan, Italy, 2-5 Sept. 2012, 189-194.

O'Byrne, M., Schoefs, F., Ghosh, B., & Pakrashi, V. (2013). Texture analysis based damage detection of ageing infrastructural elements. Computer-Aided Civil and Infrastructure Engineering, 28, 162-177.

O'Byrne, M., Ghosh, B., Pakrashi, V., & Schoefs, F. (2014a). Effects of Turbidity and Lighting on an Image Processing based Crack Detection Technique. Civil Engineering Research Ireland (CERI) Conference, 28-29 August 2014, Belfast, Northern Ireland.

O'Byrne, M., Schoefs, F., Pakrashi, V., & Ghosh, B. (2014b). Regionally enhanced multi-phase segmentation technique for damaged surfaces. Computer-Aided Civil and Infrastructure Engineering, 29, 644-658.

O'Byrne, M., Schoefs, F., Pakrashi, V., & Ghosh, B. (2018). An underwater lighting and turbidity image repository for analysing the performance of image based non-destructive techniques. Structure and Infrastructure Engineering, 14(1), 104-123.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, SMC-9, 62-6.

Pakrashi, V., Harkin, J., Kelly, J., Farrell, A., & Nanukuttan, S. (2013). Monitoring and Repair of an Impact Damaged Prestressed Concrete Bridge, Proceedings of the Institute of Civil Engineers, Journal of Bridge Engineering, 166(1), 16-29.

Pakrashi, V., Schoefs, F., Memet, J. B., & O'Connor, A. (2008). ROC dependent event isolation method for image processing based assessment of corroded harbour structures. Structure and Infrastructure Engineering, 6, 365-378.

Rouhan, A., and Schoefs, F. (2003). "Probabilistic Modeling of Inspection Results for Offshore Structures." Structural Safety, 25(4), 379-399.

Schoefs, F., Boéro, J., Clément, A., & Capra, B. (2012a). The αδ method for modelling expert Judgment and combination of NDT tools in RBI context: application to Marine Structures, Structure and Infrastructure Engineering: Maintenance, Management, Life-Cycle Design and performance (NSIE). SMonitoring, Modeling and Assessment of Structural Deterioration in Marine Environments, 8, 531-543.

Weninger-Vycudil, A., Hanley, C., Deix, S., O'Connor, A. & Pakrashi V. (2015). Cross-Asset Management for Road Infrastructure Networks. Proceedings of the Institution of Engineers – Transport, 168(5), 442-456